MASARYKOVA UNIVERZITA
EKONOMICKO-SPRÁVNÍ FAKULTA

Základy ekonometrie

DANIEL NĚMEC

podzim 2009

Obsah

Seznam tabulek

Seznam obrázků

Předmluva

V této fázi se jedná o učební text vycházející obsahem i strukturou z knížky Garyho Koopa *Introduction to Econometrics*[17], doplněný v některých ohledech o zajímavé pasáže z knihy *Principles od Econometrics*, autorů R. Carter Hill. William E. Griffiths a Guay C. Lim [13].

Pokud se jedná o technické zpracování textu, dovolil bych si upozornit na důležitou skutečnost ohledně psaní desetinných čísel, kdy jsem se v rozporu s pravidly českého pravopisu rozhodl pro používání anglického způsobu značení, tedy desetinná čárka je nahrazena desetinnou tečkou.

Veškeré prezentované odhady modelů a většina obrázků je zpracována pomocí nástroje gretl [1], což je volně dostupný software pro ekonometrickou analýzu. Náročnější obrázky jsou zpracovány v programovém prostředí Matlab [2]. Z důvodů kompatibility byl v obou případech zvolen grafický formát `PNG` (Portable Network Graphic), což je i příčinou relativně nižší kvality obrázků.

Kapitola 1

Kouzelný svět ekonometrie

V této kapitole se dozvíme:

- ☞ co je to ekonometrie a proč je dobré ji studovat;
- ☞ co je to ekonomický model a co ekonometrický model;
- ☞ popsat kroky při tvorbě ekonometrického modelu;
- ☞ s jakými typy dat ekonomové obvykle pracují;
- ☞ co můžeme získat z grafické analýzy dat;
- ☞ co nám při analýze dat přinášejí základní popisné statistiky typu rozptyl a střední hodnota;
- ☞ jak pomocí korelačního koeficientu vyjádříme sílu vztahu mezi dvěma proměnnými;
- ☞ že korelace neznamená nutně kauzalitu;
- ☞ jaké počítačové programy nám usnadňují ekonometrickou analýzu;
- ☞ jaká paleta publikací se nám nabízí pro studium ekonometrie.

1.1 Co je ekonometrie a proč ji studovat?

Ekonometrii bychom mohli zjednodušeně popsat jako vědní disciplínu *aplikující statistické nástroje a techniky v oblasti ekonomie*. Mnohem lépe bychom ji ale mohli definovat jako vědní disciplínu, která v sobě propojuje a rozšiřuje zejména poznatky ekonomické teorie, matematické ekonomie, ekonomické statistiky a matematické statistiky. Ekonometrie dává ekonomické teorii empirický rozměr, přičemž obvykle využívá matematickou formulaci ekonomického problému, což je hlavní náplní matematické ekonomie. Pro empirickou analýzu jsou potřebná data, vypovídající o reálném světě, a v jejich získání pomáhá ekonomická statistika. Ekonometrie však není jen pasivní

přejímání statistických dat a ukazatelů. Hodně důležitá je zde jejich aktivní analýza a pečlivý výběr pro účely praktického modelování. Jako příklad si vezměme ekonomický pojem „úroková míra“, který je používán v rámci formulace nějaké hypotézy vycházející z ekonomické teorie. Pro tento pojem lze ale nalézt celou řadu různých statistických ukazatelů a je obvykle jen na nás, který z nich upřednostníme s ohledem na náš cit a zkušenosti. Základem ekonometrických technik a nástrojů, pomocí kterých získáme číselné vyjádření řešení našeho problému, jsou poznatky matematické statistiky. Tyto nástroje jsou obvykle vytvářeny a voleny s ohledem na specifický charakter ekonomických dat a modelů, s nimiž pracujeme.

Ekonomie je věda plná nezodpovězených otázek a problémů, na které ekonometrie dokáže pomoci odpovědět. Příklady otázek z oblasti makroekonomie, financí, podnikové sféry či veřejné sféry mohou být následující:

- Ovlivní změna úrokových sazeb směnný kurz?
- Mají dlouhodobě nezaměstnaní větší problém nalézt zaměstnání než krátkodobě nezaměstnaní?
- Jaký je dopad cen benzínu na rozhodování o tom, jestli lidé jezdí do práce autem nebo veřejnou dopravou?
- Jsou finanční trhy slabě efektivní?
- Jaký je dlouhodobý vztah mezi vývojem cenové hladiny a směnným kurzem?
- Jak predikovat korelaci mezi akciovými indexy dvou zemí?
- Jaký vliv na trestnou činnost bude mít dodatečná finanční injekce pro vyslání více uniformovaných policistů do ulic zkoumaného města?
- Jaký efekt má investice do nějaké reklamní kampaně na prodejnost konkrétního výrobku?
- Co a jak ovlivňuje rozhodnutí zákazníka o nákupu některého z řady vzájemně si konkurujících výrobků?

To všechno jsou příklady toho, „co nevíme“ a „co chceme odhalit“. To, „co víme“ a pozorujeme, jsou ekonomická data. Různé agentury, ať už vládní či soukromé, sbírají fakta o skutečném světě, která se snaží napomoci odkrýt tajemství onoho neznámého. Podíváme-li se do novin, vidíme tam mnoho informací o cenách různých aktiv (úrokové sazby, směnné kurzy, akcie). Z průzkumů vládních či soukromých agentur získáváme poznatky o různorodých aktivitách občanů nějakého státu či oblasti. Takto získaná data pak mohou být využita např. pro porovnání zkušeností dlouhodobě a krátkodobě nezaměstnaných při hledání práce. Ekonomové mohou provádět průzkumy v oblasti dopravy a získané informace mohou být opět využity např. k zodpovězení otázky, jaké faktory ovlivňují volbu toho či onoho dopravního prostředku pro cestovaní za prací, což lze využít třeba v rámci diskuzí nad stanovením cen jízdného v hromadné dopravě, jak to ovlivní počet cestujících a příjmy dopravního podniku nebo jak to změní hustotu provozu ve městě.

Bez odkazu na fakta (tedy data) mohou ekonomické debaty sklouznout do kolotoče neustále se opakujících zavedených názorů a teorií, případně mohou nabývat podob neformálního povídání příběhů a pohádek, kdy se ekonomové snaží podpořit svůj pohled na věc svými oblíbenými vtipy. Ekonometrie nám ukazuje, jak prakticky, efektivně a systematicky používat data k zodpovězení různých ekonomických otázek a problémů. Většinou nám ale nemusí stačit odpověď v podobě *„ano, změna úrokové míry ovlivní směnný kurz“* nebo *„úroková míra ovlivňuje směnný kurz pozitivně“*. V případě řešení rozhodovacích problémů, potřebujeme znát i kvantitativní vyjádření efektu příslušné závislosti, např. pokud by centrální banka potřebovala vědět, o kolik má změnit úrokovou sazbu aby ovlivnila směnný kurz v potřebném rozsahu.

1.2 Ekonomický a ekonometrický model

Podstatou ekonometrie je tedy systematické hledání kvantitativních odpovědí na ekonomické otázky a problémy. Vše se odvíjí od ekonomické teorie, která je spojená s problémem, o který se zajímáme. Samozřejmě ekonometrické nástroje a techniky jsou aplikovatelné i mimo oblast ekonomie, např. v sociologii. V ekonomii vyjadřujeme naše představy o vztahu mezi ekonomickými veličinami v podobě matematických funkcí. Vztah mezi spotřebou a důchodem můžeme formálně zapsat jako

$$\text{spotřeba} = f(\text{důchod}),$$

čímž říkáme, že úroveň spotřeby je nějakou funkcí, $f(\cdot)$ důchodu. Podobně poptávka po nějakém statku, např. Škoda Fabia může být vyjádřena jako

$$Q^d = f(P, P^s, P^c, INC)$$

čímž říkame, že množství poptávaných škodovek, Q^d je funkcí ceny tohoto vozu, P, ceny vozů, které jsou jeho substituty (konkurenty), P^s, cenami komplementárních statků (např. náhradní díly nebo benzín), P^c, a úrovní důchodu, INC. Jako poslední příklad si uveďme nabídku nějaké zemědělské komodity, např. hovězí maso, v podobě

$$W^s = f(P, P^c, P^f),$$

kde Q^s je nabízené množství, P je cena hovězího, P^c je cena substitutu (např. kuřecí maso) a P^f je cena vstupů, což může být např. obilí, které se využívá při produkci hovězího.

Všechny tyto rovnice jsou obecnými ekonomickými modely, které popisují vztah mezi ekonomickými proměnnými. Konkrétní funkční podoba tohoto vztahu vychází z příslušné ekonomické teorie. Ekonomické modely tohoto typu jsou používány v rámci ekonomické analýzy.

Ekonometrický model „převádí“ ekonomický model do podoby, která je s pomocí ekonometrických nástrojů a technik analyzovatelná. Ekonomické vztahy nejsou obvykle přesné. Ekonomická teorie nedokáže predikovat chování jendotlivce nebo firmy, spíše popisuje průměrné či systematické chování velkého počtu firem nebo jednotlivců. Pokud např. studujeme prodej aut, vidíme, že skutečný počet prodaných aut

Škoda Fabia je tvořen součtem této systematické části a nepredikovatelné části, kterou budeme označovat náhodnou složkou, ϵ. *Ekonometrický model* popisující chování prodeje Škoda Fabia tak je

$$Q^d = f(P, P^s, P^c, INC) + \epsilon.$$

Náhodná složka, ϵ, zahrnuje řadu faktorů, které ovlivňují prodeje, ale my jsme je v našem modelu nazahrnuli. Pro úplnou specifikaci ekonometrického modelu musíme vyjádřit i podobu funkčního vztahu mezi ekonomickými proměnnými. V úvodních kurzech ekonomie byla poptávka popisována jako lineární funkce ceny. Tento předpokald můžeme rozšířit i na další proměnné a systematickou část modelu poptávky zapsat jako

$$f(P, P^s, P^c, INC) = \alpha + \beta_1 P + \beta_2 P^s + \beta_3 P^c + \beta_4 INC.$$

Odpovídající ekonometrický model tak má podobu

$$Q^d = \alpha + \beta_1 P + \beta_2 P^s + \beta_3 P^c + \beta_4 INC + \epsilon.$$

V tomto případě se jedná o tzv. lineární regresní model, kterému budeme věnovat následující kapitoly. Dodaná řecká písmenka nám udávají míru vlivu jednotlivých proměnných a označují se jako parametry. Tato funkční podoba reprezentuje hypotézu o vztahu mezi proměnnými. Mnohdy nás může zajímat nalezení právě té podoby, která bude nejlépe odpovídat ekonomické teorii.

1.3 Práce s daty

Podívejme se nyní blíže na typy dat, se kterými se obvykle setkáváme. Data nám reprezentují fakta o reálném světě a jsou nezbytná pro jakoukoli empirickou analýzu. Budeme se zabývat problematikou jejich zpracování a prezentace. Problematice samotného získávání dat je ve stručnosti věnována příloha D.

1.3.1 Typy dat

Časové řady

Hrubý domácí produkt (HDP), ceny akcií, úrokové míry, směnné kurzy, to vše jsou veličiny (*proměnné*), jejichž hodnoty obvykle získáváme v různých časových okamžicích. Data jsou seřazena podle času tak, jak byla získána, a označujeme je jako *časové řady (time series)*. Časové řady se mohou lišit podle toho, s jakou frekvencí je získáváme. Obvykle pracujeme s daty ročními (kdy hodnota příslušné proměnné je zaznamenávána pravidelně každý rok), s daty čtvrtletními (údaj získáváme čtyři krát do roka), s daty měsíčními a s daty denními.

Budeme se držet značení, kdy Y_t odpovídá pozorování proměnné Y (např. směnného kurzu) v čase t. Celá řada dat bude dostupná vždy pro období od $t = 1$ do $t = T$. V tomto případě bude T označovat celkový počet období, která jsou pokryta údaji v naší časové řadě pozorování. Jako příklad můžeme použít měsíční časovou řadu směnného kurzu britské libry vůči americkému dolaru, a to od ledna roku 1947 do října roku

1996. Celkově tak máme období zahrnující 598 měsíců. Čas $t = 1$ odpovídá lednu roku 1947, čas $t = 598$ označuje říjen roku 1996 a celkový počet měsíců je $T = 598$. V našem značení je Y_1 směnný kurz libry vůči dolaru v lednu roku 1947, Y_2 je směnný kurz libry v únoru téhož roku atd. Časové řady mají své chronologické řazení.

Práce s časovými řadami často přináší specifické problémy a otázky, jejichž řešení vyžaduje speciální nástroje. Budou jim proto věnovány samostatné kapitoly.

Průřezová data

Ve spoustě praktických aplikacích pracujeme s daty, která nemají časový rozměr, ale jsou vztažena ke specifickým jednotkám. Těmito jednotkami mohou být firmy, lidé či státy. Například v oblasti financí nás může zajímat analýza teorií týkajících se tvorby portfolia. V rámci takovéhoto výzkumu potřebujeme posbírat data o výnosnosti akcií rozličných podníků. Získaná data tak mají *průřezový (cross-section)* charakter, kdy v jednom čase provedeme průřez skrze jednotky, které nás zajímají (např. získáme údaje o výnosnosti akcií daných firem v konkrétním roce). Oproti časovým řadám nám pramálo záleží na tom, v jakém pořadí jsme daná data získali.

V tomto případě se budeme držet značení, ve kterém Y_i bude označovat pozorování proměnné Y příslušné i-tému jednotlivci či jednotce (anglicky *individuals*). Dostupná pozorování množiny průřezových dat jsou v rozsahu jednotek od $i = 1$ až po $i = N$. Písmenem N se tedy obvykle označuje počet pozorování (např. počet dotazovaných jedinců či zkoumaných firem). Pro ilustraci uvažujme, že potřebujeme získat data o kurzech akcií $N = 100$ firem v určitém časovém okamžiku. V takovémto případě bude Y_1 odpovídat ceně akcie první společnosti, Y_2 ceně akcie druhé společnosti atd.

Je třeba zdůraznit další aspekt rozdělení typu dat. Pokud budeme sbírat data o ceně akcií jednotlivých firem, získáme soubor čísel příslušný jednotlivým firmám (tzn. cena akcie první firmy bude např. 25 dolarů). Tento typ dat označujeme jako *kvantitativní data*.

V řadě případů však získaná data nebudou mít podobu jednoho čísla. Např. v oblasti ekonomie práce můžeme provádět dotazníková šetření mezi zaměstnanci, kdy jedna z otázek bude směřována na členství daného pracovníka v odborové organizaci. V takovémto případě bude odpověď „ano“ nebo „ne“. Tento typ získaných údajů odpovídá *kvalitativním datům*. S tímto typem dat přicházíme do styku v případě řešení ekonomických otázek zahrnujících nějaký druh volby, tedy např. jestli si daný člověk koupil nebo nekoupil daný výrobek (v případě analýzy spotřebitelského chování, či efektů marketingové kampaně), jestli jezdí do práce veřejnou dopravou nebo osobním automobilem (zkoumáme-li otázku, co ovlivňuje volbu dopravního prostředku při cestě do zaměstnání), apod. Ekonometři tento druh kvalitativních odpovědí převádějí do formy numerické. V případě příkladu dotazování zaměstnanců můžeme přiřadit odpovědi „ano“ hodnotu jedna a odpovědi „ne“ hodnotu nula. Pozorování $Y_1 = 1$ pak samozřejmě znamená, že první dotazovaný zaměstnanec je členem odborové organizace, a $Y_2 = 0$ odpovídá situaci, kdy druhý zaměstnanec členem odborové organizace není. Pokud máme proměnnou, která může nabývat pouze hodnot 0 nebo 1, hovoříme o ni jako o *umělé (dummy) proměnné* respektive proměnné *binární*.

Panelová data

Mnohé datové soubory, se kterými pracujeme, mají jak časovou, tak i průřezovou komponentu. Tento typ dat nazýváme *panelová data*. S panelovými daty pracují např. ekonomové zabývající se problematikou ekonomického růstu. Mohou tak pracovat např. s daty hrubého domácího produktu ($Y = HDP$) pokrývajícími 90 zemí v období let 1950–2000. Takový datový soubor bude obsahovat údaje o HDP pro každou zemi v roce 1950 ($N = 90$ pozorování), HDP každé země v roce 1951 (dalších $N = 90$ pozorování), atd. Za celé období T let tak budeme mít TN pozorování proměnné Y. V tomto případě použijeme značení Y_{it}, které bude odpovídat pozorování proměnné Y pro i-tou zemi v čase t. Panelová data jsou využívána i v ekonomii práce. Příkladem mohou být dotazníková šetření vlády (či nějaké vládní instituce), kdy jsou pravidelně respondenti tázáni na otázky týkající se jejich zaměstnnání, příjmu, vzdělání atd. Na základě takovýchto dotazníkových šetření můžeme pracovat např. s proměnnou $Y =$ mzda pro $N = 1000$ jednotlivců (zmaěstnanců) v průběhu pěti let ($T = 5$).

1.3.2 Úprava dat

V řadě aplikací můžeme s úspěchem předpokládat, že máme k dispozici přímo data Y, která nás zajímají. Často však musíme řešit situaci, kdy bereme data z jednoho zdroje a upravujeme si je do podoby, která odpovídá tomu, co zkoumáme. Jako příklad si vezměme situaci, kdy máme k dispozici časově řady proměnných $X =$ zisk společnosti a $W =$ počet akcií a potřebujeme si vytvořit novou proměnnou $Y =$ ziska na akcii. V takovémto případě bude příslušná transformace snadná:

$$Y = \frac{X}{W}.$$

Forma úpravy původních dat samozřejmě souvisí s účelem jejich použití. Nicméně, ukažme si alespoň nejtypičtější úpravu časových řad. V makroekonomických aplikacích nás nemusí zajímat přímo vývoj HDP, ale spíše jeho změny v čase. Představme si ale příklad z oblasti financí. Ani v tomto případě nás nemusí zajímat primárně cena nějakého aktiva, ale spíše výnos, kterého by dosáhl investor jeho nákupem. Tento výnos může záviset na změnách ceny příslušného aktiva v čase (budeme-li abstrahovat od případných dividendových zisků). Předpokládejme, že jsme získali roční data o ceně akcie konkrétní společnosti za období let 1950–1998 (tzn. celkem za 49 let), označené jako Y_t, pro $t = 1, \ldots, 49$. V některých příkladech nás určitě může zajímat přímo tato časová řada. Proměnná tohoto typu je označovaná jako *úrovňová (level)*, a hovoříme tak „o úrovni ceny akcie“. Stejně tak nás ale může zajímat růst ceny akcie. Jednoduchý způsob, jak takovýto růst změřit, je vzít ceny akcie a spočítat jejich procentní změny v každém roce. Procentní změna ceny akcie mezi obdobím $t-1$ a t se spočítá následovně:

$$\%změna = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \times 100 = \left(\frac{Y_t}{Y_{t-1}} - 1 \right) \times 100.$$

Je potřeba zdůraznit, že procentní změna má vždy časový rozměr odpovídající porovnávaným veličinám (např. procentní změna mezi obdobím $t - 1$ a t). S ročními daty tak můžeme získat roční procentní změnu (meziroční tempo růstu), s měsíčními daty

můžeme získat měsíční tempo růstu, atd. Člen $Y_t - Y_{t-1}$ je označován jako dynamika a odpovídá tzv. první diferenci zkoumané veličiny, kterou označujeme obvykle jako ΔY_t. Podíl Y_t/Y_{t-1} je tzv. koeficient růstu a říká nám kolikrát nám příslušná veličina mezi porovnávanými obdobími vzrostla. Odečtením jedničky a vynásobením stem pak získáme odpovídající vyjádření procentní změny.

Není neobvyklé, že při práci s daty využíváme i jejich přirozené logaritmy. S využitím vlastností logaritmu totiž platí, že procentní změna veličiny je přibližně rovna rozdílu logaritmů této veličiny (násobeného stem, pro procentní vyjádření):

$$\%změna \approx [\ln(Y_t) - \ln(Y_{t-1})] \times 100.$$

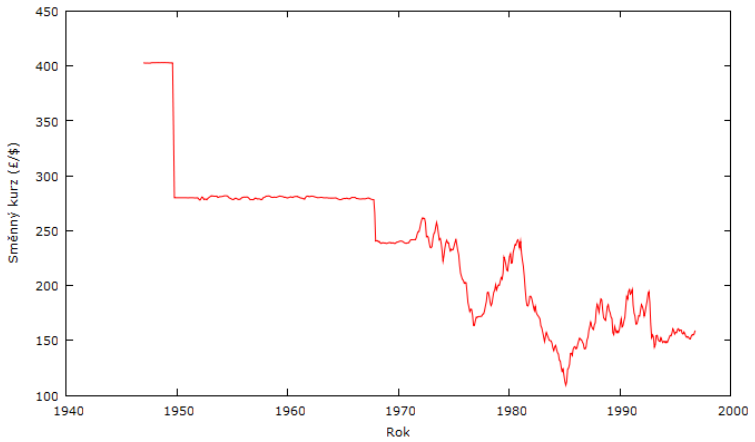
Násobení stovkou obvykle můžeme vypustit, tudíž růst 5% odpovídá hodnotě 0.05. Procentní změnu nějaké veličiny tedy označujeme jako tempo růstu dané veličiny. V makroekonomi hojně uplatníme kromě tempa růstu HDP i inflaci, tedy tempo růstu cenové hladiny (která, jakožto úrovňová veličina, je vyjádřena různými typy indexů - index spotřebitelských cen nebo deflátor).

1.4 Práce s daty – grafické metody

Pracujeme-li s daty, je velmi žádoucí tato data přehledným a výstižným způsobem prezentovat. Asi nikoho nezajímá pročítat jednotlivé položky datové báze, kterou chceme dále využívat v naši práci. To čím se zabývá ekonometrie si je možno představit jako rozvoj metod, kterými jsou v pozorovaných datech rozptýlené informace agregovány do mnohem kompaktnější podoby s vyšší informační hodnotou pro pozorovatele. Různé druhy grafů a tabulek jsou velmi užitečným způsobem, jak naše data prezentovat. Pokud jde o grafy, tak jich existuje celá řada – sloupcové grafy, koláčové grafy atd. V této části si ukážeme několik obvykle používaných typů grafů. Jelikož má většina ekonomických dat podobu časových řad či průřezových dat, zaměříme se stručně na zavedení jednoduchých technik vykreslení právě těchto typů dat.

1.4.1 Spojnicový graf

Jako příklad *spojnicového grafu* si můžeme ukázat časovou řadu měsíčních ukazatelů směnného kurzu britské libry vůči americkému dolaru od ledna roku 1947 do října roku 1996 (data pocházejí z učebnice Koopa [17], v Gretlu odpovídají datovému souboru *exruk.gdt* na záložce *Koop*). Tento graf je vykreslen na obrázku 1.1. Vyobrazená data obsahují 598 pozorování, což je vcelku dost pro prezentaci ve své numerické podobě, aby tím neutrpěla srozumitelnost našeho sdělení zvídavému čtenáři. Pokud se čtenář podívá na příslušný graf této časové řady, snadno z něj vyčte řadu důležitých skutečností. Můžeme zde vidět snahu britské vlády o udržení fixního směnného kurzu až do roku 1971, včetně znatelných devalvací v říjnu roku 1949 a listopadu roku 1967. Stejně tak vidíme např. postupnou depreciaci libry v polovině 70. let, tedy již po přechodu na floating. Díky téměř výlučnému použití pro prezentaci vývoje časových řad můžeme hovořit i přímo o *grafu časové řady*.



Obrázek 1.1: Časová řada vývoje směnného kurzu GBP vzhledem k USD.

1.4.2 Histogram

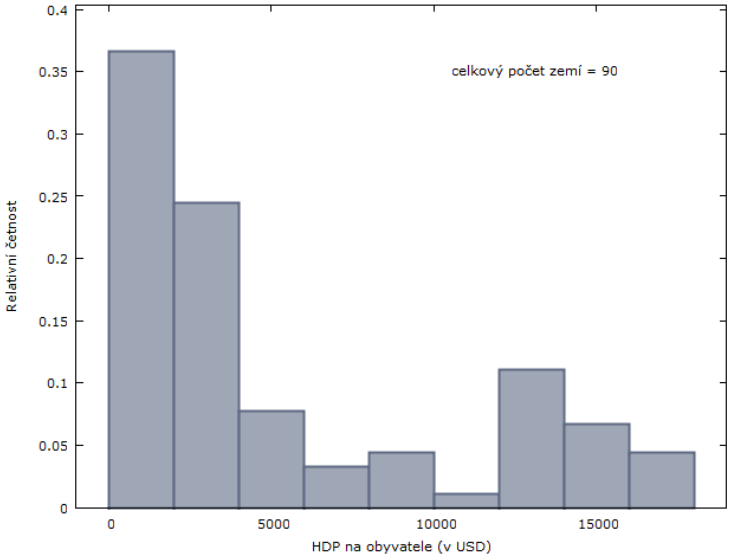
Graf časové řady je velmi informativní, pokud jde o její vývoj v čase. V případě průřezových dat však tento druh zobrazení není zcela vhodný a je třeba shrnout data jiným způsobem.

Tzv. Penn Word Table[1] nám umožňuje získat průřezová data reálného HDP na obyvatele v roce 1992 pro 90 zemí světa (jedná se o data z učebnice Koopa [17], v Gretlu odpovídají datovému souboru *gdppc.gdt* na záložce *Koop*). Reálný HDP na obyvatele byl vyjádřen v amerických dolarech s využitím směnných kurzů dle parity kupní síly. To nám umožňuje přímé porovnávání mezi zeměmi. Jedním ze způsobů souhrnu těchto dat je za použití *histogramu*. Ke konstrukci histogramu je potřeba zvolení *dělících intervalů*(intervalů tříd resp. kategorií), které rozdělí země do jednotlivých skupin (tříd) v závislosti na jejich HDP na osobu. V našem datovém souboru se HDP na osobu pohybuje od $408\$$ pro Čad, až po $17945\$$ v případě USA. Jedna možnost dělících intervalů je $0-2000$, $2001-4000$, $4001-6000$, $6001-8000$, $8001-10000$, $10001-12000$, $12001-14000$, $14001-16000$, $16001-18000$ (vše odpovídá jednotkám amerických dolarů). Každý interval má šířku 2000 (dolarů). Pro každý z intervalů jsme schopni spočítat počet zemí, které mají HDP na osobu v hodnotách odpovídajících rozsahu tohoto intervalu, Např, v našem souboru máme sedm zemí s reálným HDP na osobu mezi $4001\$$ a $6000\$$. Počet zemí nacházejících se v konkrétním intervalu je nazýván jako absolutní četnost zemí odpovídajících danému intervalu. Relativní četností pak nazveme absolutní četnost vydělenou celkovým počtem zemí ve vzorku. Histogram je sloupcový graf, který vykresluje četnosti (absolutní nebo relativní) vzhledem k daným intervalům tříd.

Obrázek 1.2 je histogramem průřezových dat souboru dat HDP na obyvatele s dělením odpovídajícím předchozímu odstavi. Pokud bychom nechtěli specifikovat jak široké mají intervaly být popř. kolik jich má být celkem, tak nemusíme. Každý relevantní

[1] Tyto tabulky, byť ne zcela aktuální, lze získat snadno v rámci Gretlu.

počítačový software toto rozhodnutí udělá za nás. Většina programů dokáže udělat více či méně podrobnější tabulky četností, podobné těm, které jsou obsahem tabulky 1.1.



Obrázek 1.2: Histogram HDP na osobu pro 90 zemí.

Tabulka 1.1: Tabulka četnosti pro HDP na osobu.

Interval (USD)	Četnost	
	Absolutní	Relativní
0-2000	33	36.67 %
2001-4000	22	24.44 %
4001-6000	7	7.78 %
6001-8000	3	3.33 %
8001-10000	4	4.44 %
10001-12000	2	2.22 %
12001-14000	9	10.00 %
14001-16000	6	6.67 %
16001-18000	4	4.44 %

Tabulka četnosti nám ukazuje počet zemí (v absolutním a relativním vyjádření) spadajících do toho či onoho dělícího intervalu. Z tabulky tak vidíme, že 33 zemí má HDP na obyvatele menší než 2000$, kdy tento počet odpovídá téměř 37 % všech pozorovaných zemí. Podobně jen čtyři země (což je asi 4.5 % celkového počtu 90 zemí) mají HDP na obyvatele v rozmězí 16 až 18 tisíc dolarů. Identická informace je obsažena v

histogramu na obrázku 1.2. Grafické zobrazení nám umožňuje okamžitě mít přehled o *rozdělení* HDP na obyvatele mezi jednotlivými zeměmí. Vidíme, že hodně zemí je relativně chudých, nicméně zde je i řada zemí bohatých (19 zemí má HDP na obyvatele vyšší než 12000$). Relativně málo zemí se nachází mezi těmito dvěma skupinami chudých a bohatých zemí (tzn. spadají do intervalu mezi 6001$ a 12000$). Fenomén rozdělení zemí do těchto dvou krajních skupin zemí bohatých a chudých odpovídá rozdělení zemí, které není tzv. *unimodální*, nemá tedy jediný vrchol. Takovouto vlastnost dat snadno vyčteme z pohledu na histogram, nicméně velmi obtížně bychom ji nalézali přímo z pohledu na číselná data pozorování.

1.4.3 Bodový graf

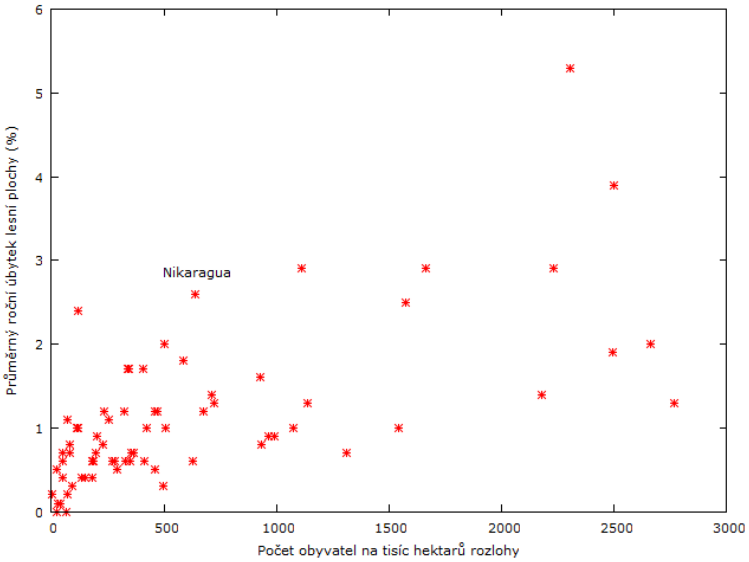
Ekonomy může zajímat vztah či závislost mezi dvěma nebo více proměnnými. Příkladem mohou být otázky typu:

- Jaký je vztah mezi kapitálovou strukturou (tj. rozdělení mezi dluhovým financováním a financováním emise akcií) a výkonností firmy (která může být měřena např. ziskem)?
- Je vyšší úroveň vzdělání a pracovních zkušeností spojena s vyšší úrovní mezd v rámci pracovníků v určitém průmyslovém odvětví?
- Jsou změny v nabídce peněz relevantním indikátorem změn inflace?

Všechny tyto otázky zahrnují analýzu dvou nebo více různých proměnných. Techniky z předchozích částí jsou vhodné pro popis chování jediné veličiny (např. jediné proměnné, kterou může být HDP na hlavu, což je ilustrováno na obrázku 1.2). Nejsou však již vhodné pro analýzu vztahu mezi dvěma veličinami.

Pokud chceme porozumět podstatě vztahů mezi dvěma a více veličinami, málokdy si vystačíme s pouhou grafickou analýzou. V další kapitole se budeme věnovat regresní analýze, což je jeden z nejdůležitějších nástrojů pro práci s více proměnnými. Grafické metody však můžeme použít pro vyjádření jednoduchých vztahů mezi dvěma veličinami. Právě *bodový graf (XY graf)* je toho nejlepším příkladem.

Na obrázku 1.3 vidíme vykreslená data o míře odlesnění (konkrétně o průměrných ročních ztrátách lesní plochy v období let 1981 až 1990, vyjádřené jako procento z celkové rozlohy lesů) pro 70 zemí tropického pásma vzhledem k údajům o hustotě obyvatelstva (což je obvykle počet lidí na tisíc hektarů rozlohy země). Můžeme se s úspěchem domnívat, že země s vyšší husotou obyvatelstva budou odlesňovat své území mnohem rychleji než země s hustotou obyvatelstva nižší. Vysoká hustota obyvatelstva totiž může vyvolávat tlaky na rozšiřování zemědělské půdy, která je nezbytná pro produkci potravin. každý bod obrázku reprezentuje konkrétní zemi. Přečteme-li si pro daný bod údaj na ose y (vertikální ose) zjistíme míru odlesnění v této zemi. Na ose x (horizontální osa) získáme příslušné údaje o hustotě obyvatelstva. Každý bod grafu bychom si mohli popsat jménem příslušné země, což by ale při tak vysokém počtu zemí mohlo být poněkud matoucí. Z tohoto důvodu byla jen pro ukázku označeno pozorování pro Nikaraguu. Tato země má míru odlesnění rovnou $2.6\ \%$ za rok ($Y = 2.6$) a hustotu 640 obyvatel na tisíc hektarů ($X = 640$).



Obrázek 1.3: Bodový graf odlesnění vzhledem k hustotě obyvatelstva.

Bodový graf nám nabízí okamžité, vizuální vyjádření vztahu mezi odlesňováním a hustotou obyvatelstva. Analýza obrázku podporuje myšlenku o existenci vztahu mezi těmito dvěma veličinami. Pokud se podíváme na země s nízkou hustotou obyvatelstva (méně než 500 obvytel na tisíc hektarů), tak téměř všchny mají nízké míry odlesnění (méně než 1% za rok). Pokud se podíváme na země s vysokou hustotou zalidnění (více než 1500 obyvatel na tisíc hektarů), tak téměř všechny mají míry odlesnění vysoké (vyšší než 2 % za rok). To nám napovídá, že zde existuje pozitivní závislost mezi těmito dvěma veličinami, tzn. vyšší hodnoty jedné proměnné mají tendenci odpovídat vyšším hodnotám druhé proměnné a naopak. Podobně jako pozitivní závislost můžeme narazit i na vztah negativní. Pokud bychom nahradili v bodovém grafu hustotu obyvatelstva mírou urbanizace, mohli bychom (teoreticky) sledovat takový vztah, kdy vysoká míra urbanizace je doprovázena nízkými mírami odlesnění, což může být důsledek toho, že koncentrace lidí do měst snižuje tlaky na rozvoj venkovských, zemědělských oblastí, v jejichž okolí se právě s lesy setkáváme. jedná se ale jen o hypotetický předpoklad, neboť i v tomto případě bychom mohli narazit na pozitivní závislost, pokud by bylo odlesňování spojeno s nutností rozšiřování zemědělské půdy, a to kvůli nutnosti uživit rostoucí počet obyvatel (byt' koncentrovaný do měst).

Je třeba opět zdůraznit, že při diskuzi nad vztahem nebo závislosti dvou veličin hovoříme o tendencích, ať už pozitivních či negativních, kterými je tento vztah charakterizován. Pozitivní nebo negativní závislost nemusí platit pro každou zemi. Mohou existovat i výjimky tohoto obecného chování. Na našem bodovém grafu z obrázku 1.3 vidíme jednu zemi, která má vysokou hustotu obvytelstva (1300) a nízkou míru odlesňování (0.7 %). Podobně i nízká hustota obyvatelstva může být u některých zemích doprovázena vysokou mírou odlesnění, což je zjevně případ země s hustotou okolo 150

obyvatel na tisíc hektarů a mírou odlesnění $2.5\ \%$ za rok. Mějme na paměti, že budou existovat výjimky z nalezeného obecnějšího chování, které jsou obvykle označovány *odlehlé hodnoty (outliers)*. Nalezení těchto odlehlých případů pro nás může být zajímavé samo o sobě a zavádí nás obvykle k hlubšímu zkoumání toho, proč se právě jejich chování vymyká obecným tendencím.

1.5 Práce s daty – popisné statistiky a korelace

Grafy a obrázky mají své kouzlo díky okamžitému, vizuálnímu pohledu na věc. V řadě případů je však důležitá i jistá numerická formulace výsledků našich analýz. V následujících kapitolách se detailněji zaměříme na popis běžně užívaných numerických metod sloužících k souhrnnému vyjádření závislosti či vztahů mezi několika proměnnými. Na tomto místě si ale připomeneme ve stručnosti několik popisných statistik, které vypovídají o vlastnostech jedné proměnné či dvou proměnnách. Pro trochu motivace se vraťme ke konceptu rozdělení dat ilustrovaném v části věnované histogramům. Koncept rozdělení pravděpodobnosti je základem matematické statistiky. Příloha B formálně definuje a motivuje jeho principy. Na tomto místě se zaměříme spíše na neformální, intuitivní popis problematiky rozdělení a na způsoby, jakými můžeme chrakterizovat jeho vlastnosti.

V datech o reálném HDP na obyvatele se hodnota této proměnné lišila mezi jednotlivými zeměmi. Tuto variabilitu můžeme vyčíst z pohledu na histogram z obrázku 1.2, který nám rozdělení HDP na osobu mezi jednotlivými zeměmi pěkně ilustruje. Předpokládejme, že bychom chtěli informaci obsaženou v tomto histogramu vyjádřit numericky. Jede ze způsobů je samozřejmě prezentace dat, které vedly k jeho konsturkci, tedy prezentaci relativních či absolutních četností (viz tabulka 1.1). Ovšem i tato tabulka obsahuje poněkud mnoho čísel na to, aby byla snadno intepretovatelná. Místo toho je obvyklé prezentovat dvě jednoduché číselné charakteristiky zvané *střední hodnota* a *směrodatná odchylka*. Střední hodnota (přesněji výběrová střední hodnota) odpovídá aritmetickému průměru. K označení N různých pozorování využíváme značení $Y_1, \ldots, Y_N$. Tuto množinu dat označujeme jako *výběr (sample)*, odtude tedy označení výběrové střední hodnoty, protože je počítána na základě našeho výběru. Matematický vzorec pro její výpočet je snadný

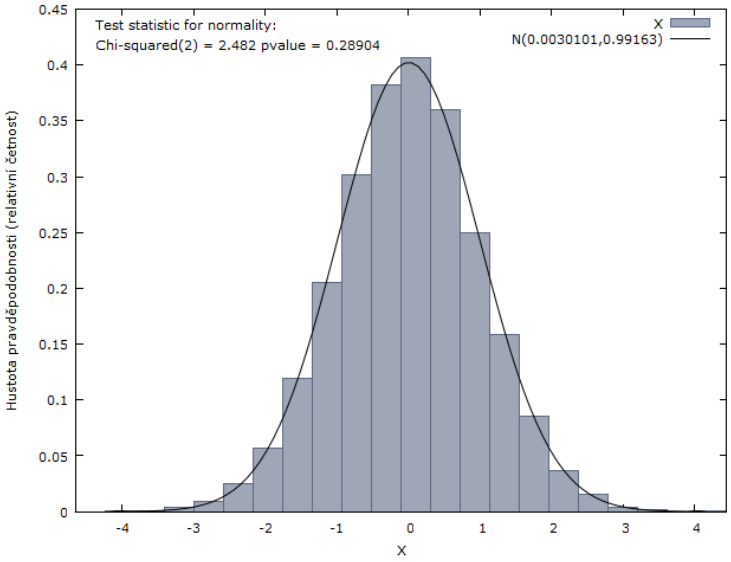
$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N},$$

kde N je označováno jako velikost vzorku (v našem případě počet zemí) a $\sum_{i=1}^{N}$ je operátor sumace (kterým sečteme hodnoty reálného HDP na osobu pro všechny země). Bližší podrobnosti lze nalézt v příloze A. V našem případě je výběrová střední hodnota rovna 5443.8 dolarů. Proměnné s horním pruhem označuje příslušnou střední hodnotu (tzn. $\overline{Y}$ je střední hodnota proměnné Y, $\overline{X}$ je střední hodnota proměnné X, atd.).

Koncept střední hodnoty[2] je spojen se středem rozdělení nějaké proměnné. Na his-

[2]Označení „výběrová“ budeme v textu vynechávat, pokud z kontextu vyplývá, že hovoříme právě o výběrové střední hodnotě jakožto aritmetickém průměru. Podobným konceptem je totiž i tzv. populační střední hodnota, kterou si intuitivně můžeme představit jako aritmetický průměr z celé (nekonečné) populace (což můžou být např. všechny naše země) a obvykle se značí řeckým písmenem μ. ještě na něj ale přijde řeč.

togramu z obrázku 1.2 vidíme, že hodnota 5443.8 dolarů leží právě někde uprostřed. Rozdělení průřezových dat HDP na osobu je poněkud „nestandardní“, neboť má tzv. dva vrcholy, a není tedy unimodální (s jediným vrcholem). Obvykle pracujeme s rozděleními ekonomických veličin, které mají jediný vrchol a tvar zvonu. Obrázek 1.4 je příkladem takovéhoto zvonového rozdělení. Pro tato rozdělení leží střední hodnota uprostřed a pod oním vrcholem (rozdělení je tak symetrické). Nejznámějším příkladem je tzv. *normální rozdělení*, které je právě vykresleno na obrázku 1.4 (plná čára, jsou zde zobrazeny i přislušné parametry). Příslušný histogram je postaven na základě 10000 vygenerovaných náhodných čísel ze *standardizovaného normálního rozdělení*, což je normální rozdělení s nulovou střední hodnotou (tentokrát populační) a jednotkovým rozptylem (opět populačním).



Obrázek 1.4: Histogram pro zvonové rozdělení.

Je zřejmé že střední hodnota či průměr nám nic neříká o variabilitě v datech. Statistiky, které by nám k této chrakteristice mohly říct více jsou minimum a maximum. Pro data o HDP je minimální HDP na osobu 408 dolarů (Čad) a maximum 17945 dolarů pro Spojené státy. Pohledem na vzdálenost mezi minimem a maximem můžem získat jistý pohled na disperzi rozdělení naší proměnné.

Koncept *disperze* je v ekonomii důležitý a je blízký konceptu rozptylu a nerovnosti. V roce 1992 se HDP na osobu v našich datech pohyboval od hodnoty 408 dolarů až po hodnotu 17945 dolarů. Pokud by chudší země v budoucnu rostly rychleji než bohatší (ketré by stagnovaly), potom by disperze v reálném HDP na osobu, řekněme v roce 2012, měla být významně nižší. Mohla by tak nastat hypotetická situace, kdyby nejchudší země měla HDP 12000 dolarů na osobu a nejbohatší by zůstala na svých 17945 dolarech. Pokud by k tomuto došlo, rozdělení HDP na osobu mezi zeměmi by bylo rovnoměrnější (méně disperzní, méně variabilní či rozptýlené). Intuitivně tak jsou

pojmy disperze, variabilita a nerovnost velmi úzce propojeny.

Statistika minima a maxima však není nejlepší charakteristikou variability. Kdyby např. nejchudší Čad nenásledoval náš hypotetický růst ostatních chudých zemí, pak by se určitě disperze mezi zeměmi snížila. Protože ale Čad ani USA nerostly, minimum a maximum zůstává na svých hodnotách a jejich rozdíl nám snížení variability nezachytí. Obvyklejším měřítkem rozptýlenosti v datech, které tento problém nemá, je *směrodatná odchylka (standard deviation)*. Její vzorec je

$$s_Y = \sqrt{\frac{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}{N-1}}.$$

Opět i v tomto případě bychom měli hovořit o výběrové směrodatné odchylce (podobně jako tomu bylo u výběrové střední hodnoty). Druhá mocnina směrodatné odchylky je tzv. *rozptyl (variance)*, (s^2). Jen pro zajímavost si všimněme, že jednotky rozptylu odpovídají druhým mocninám jednotkám příslušné proměnné (tedy např. dolary na druhou). Odmocninou získáme směrodatnou odchylku, která však již je měřena ve stejných jednotkách jako zkoumaná veličina (např. dolary).

Směrodatná odchylka má vcelku jasnou intuici. V našem příkaldu s reálným HDP na hlavu je směrodatná odchylka 5369.5. Mohli bychom ho interpretovat jako prměrné odchýlení od svého průměru, nicméně lepší interpretaci má v komparativním vyjádření. Porovnáme-li standardní odchylky dvou rozdělení, tak rozdělení s menší směrodatnou odchylkou bude méně rozptýlené. Pokud by tedy hypoteticky chudší země z našich dat zažily náhlý ekonomický růst a bohatší by stagnovaly, směrodatná odchylka by měla v průběhu času klesat.

1.5.1 Očekávaná hodnota a rozptyl

V předchozí části jsme hovořili o střední hodnotě a rozptylu. Jak již bylo naznačeno, měly bychom je nazývat výběrovou střední hodnotou a výběrovým rozptylem. Označení „výběrový“ zdůrazňuje skutečnost, že tyto charakteristiky vypočítáváme na základě „výběru“ dat. Tak jsme použitím dat našeho výběru zemí získali hodnoty $\overline{Y} = 5443.8$ a $s_Y = 5369.5$.

Jako další příklad předpokládejme, že jsme získali data o výnosech akcií společnosti za posledních 100 měsíců. Tato data můžeme použít k výpočtu výběrové střední hodnoty a rozptylu. Tato data jsou však počítána na základě historické výkonnosti společnosti. Nás by ale mohla zajímat predikce budoucího vývoje výnosnosti. Již z definice predikce nebudeme vědět přesně, jaká hodnota to ve skutečnsoti bude. Musíme tak rozšířit koncept střední hodnoty a rozptylu na případ, kdy nemáme výběr dat. Investor by se tak mohl zajímat o typický výnos, který může očekávat. Rovněž tak by ho zajímalo riziko spojené s nákupem akcií. Koncept typické resp. *očekávané hodnoty* zní podobně jako koncept střední hodnotě. Koncept rizikovosti je podobný myšlence stojící v pozadí rozptylu. Stručně řečeno, potřebujeme podobný koncept jako jsou výběrová střední hodnota a rozptyl, ale pro případy, kdy nemáme data k jejich výpočtu. Relevatními charakteristikami jsou v tomto případe *populační střední hodnota* a *populační rozptyl*.

Jejich formální definice a vlastnosti jsou obsahem přílohy B. Uveďme si tedy alespoň definici a intuici stojící v pozadí. Předpokládejme příklad, kdy se budeme zabývat výškou každého obyvatele ve Spojených státech (nebo jakékoli jiné zemi). V populaci jako celku existuje nějaká průměrná výška (populační střední hodnota výšky) a určitá variabilita v rámci ní (populační rozptyl). Tyto populační charakteristiky zůstanou neznámé, dokud se nenajde obětavec, který by změřil každou osobu žijící ve Spojených státech. Mohli bychom ale posbírat data o skutečné výšce 100 osob (např. lékař změří výšku 100 svých pacientů). S použitím dat pro 100 osob bychom mohli spočítat výběrový průměr $\overline{Y}$ a výběrový rozptyl s_Y^2. Tyto hodnoty můžeme využít jakožto odhady (či aproximace) toho, co bychom mohli pozorovat v celé zemi (tzn. že výběrový střední hodnota a výběrový rozptyl mohou být využity k odhadu populační střední hodnoty a rozptylu). Striktně odlišovat výběrové a populační charakteristiky je z pohledu statistiky velmi důležité.

Podívejme se tedy, proč např. ve finanční ekonomii je potřeba znát rozdíl mezi populačními a výběrovými statistikami. V našem příkladu z úvodu se potenciální investor zajímal o možné zisky nákupu akcií. Nechť Y označuje výnos v dalším měsíci. Z pohledu investora je Y neznámé. Typický výnos, který by mohl očekávat je měřen jakožto populační střední hodnota a nazývá se očekávanou hodnotou. K označení očekávané hodnoty použijeme $E(Y)$, kde E je operátor očekávání. Očekávaná hodnota $E(Y)$ má své označení pomocí řeckého písmenka μ. „Očekávaná hodnota“ tak již svým jménem označuje to, co můžeme očekávat, že nastane.

Výnos akcie však málokdy bývá ve skutečnosti to, co očekáváme (jen zřídka najdeme Y, které bude přesně rovno $E(Y)$). Akciové trhy jsou vysoce nepredikovatelné, někdy je výnos vyšší než naše očekávání, někdy tomu je naopak. Jinými slovy, existuje zde riziko, spojené s nákupem akcií. Potenciální investor by tak rád znal i míru tohoto rizika. Rozptyl je obvyklý způsob jak toto riziko měřit. Pro rozptyl použijeme značení $var(Y)$, které může být opět nahrazeno řeckým písmenem σ^2. V řadě textů je použito i značení $D(Y)$.

V příloze B nalezneme popis toho, jak získat $E(Y)$ a $var(Y)$ pro dané rozdělení pravděpodobnosti. Pro intuitivní chápání opět využijeme příklad z oblasti financí. Předpokládejme, že se jako investor rozhodujeme, jestli koupíme nebo nekoupíme akcii na základě její výnosnosti pro příští měsíc. Nevíme přesně, jaký tento výnos bude. Domníváme se, ale, že s pravděpodobností $70\ \%\ (0.7)$ budou trhy stabilní, kdy získáme výnos $1\ \%$. Existuje ovšem i možnost, že s 10% pravděpodobností dojde k pádu trhů, kdy výsledkem bude výnos akcie $-10\ \%$. Je zde i 20% pravděpodobnost, že dobré zprávy zvýší náladu na trzích a náš výnos bude $5\ \%$.

V našem případě máme tři možné realizace (dobrou, normální, špatnou), které odpovídají hodnotám 0.05, 0.01 a -0.10 (tedy možné výnosy jsou 5, 1 nebo $-10\ \%$). K označení pravděpodobnosti použijeme symbol Pr. Výraz $\Pr(Y = 0.05) = 0.20$ tak říká, že existuje 20% šance výnosu $5\ \%$. Můžeme tedy definovat očekávaný výnos jako vážený průměr všech možných výsledků, kdy váhy odpovídají pravděpodobnostem je-

jich realizace:

$$\begin{aligned} E(Y) &= \Pr(Y = 0.05)\, 0.05 + \Pr(Y = 0.01)\, 0.01 + \Pr(Y = -0.10)\, (-0.10) \\ &= 0.20 \times 0.05 + 0.70 \times 0.01 + 0.10 \times (-0.10) \\ &= 0.007. \end{aligned}$$

Očekávaná hodnota výnosu akcie v příštím měsíci je $0.7\ \%$ (což je o něco méně než $1\ \%$).

V našem příkladu jsme předpokládali pouze tři možné výsledky. Pokud existuje k možných realizací (označených jako $y_1 = 1, 2, \ldots, y_k$), bude vzorec pro očekávanou hodnotu vypadat následovně:

$$E(Y) = \sum_{i=1}^{k} \Pr(Y = y_i)\, y_i.$$

Vztah pro očekávanou hodnotu spojité proměnné (kdy existuje nekonečně mnoho možných realizací) lze nalézt v příloze B. Intuice je podobná, jen trochu komplikovanější.

Vztah pro rozptyl $var(Y)$ je rovněž obsahem přílohy B. Dostatečná pro nás může být znalost toho, že jej lze spočítat pomocí operátoru očekávané hodnoty:

$$var(Y) = E\left(Y^2\right) - [E(Y)]^2.$$

V našem příkladu jsme si spočítali $E(Y) = 0.007$. K výpočtu rozptylu však potřebujeme umět spočítat $E(Y^2)$. To lze provést jednoduše tak, že místo Y použijeme Y^2. Obecný vztah pro k možných relaizací je

$$E\left(Y^2\right) = \sum_{i=1}^{k} \Pr\left(Y^2 = y_i^2\right) y_i^2.$$

Pro příklad rizika očekávaného výnosu akcie máme možné realizace pro Y^2: $(0.05)^2 = 0.0025$, $(0.01)^2 = 0.0001$ a $(-0.10)^2 = 0.01$. Ze vzorce pro výpočet $E(Y^2)$ dostáváme:

$$\begin{aligned} E(Y^2) =& \Pr\left(Y^2 = 0.0025\right) \times 0.0025 + \Pr\left(Y^2 = 0.0001\right) \times 0.0001 \\ &+ \Pr\left(Y^2 = 0.01\right) \times 0.01 \\ =& 0.20 \times 0.0025 + 0.70 \times 0.0001 + 0.10 \times 0.01 \\ =& 0.00157. \end{aligned}$$

Tyto výsledky využijeme k výpočtu rozptylu:

$$\begin{aligned} var(Y) &= E\left(Y^2\right) - [E(Y)]^2 \\ &= 0.00157 - (0.007)^2 \\ &= 0.001521. \end{aligned}$$

Standardní odchylku získáme použitím odmocniny a činí 0.039. Očekávaný výnos je tedy $0.7\ \%$ s nejistotou, kterou můžeme vyjádřit odpovídající odchylkou $\pm 3.9\ \%$.

Abychom si shrnuli naše poznatky. Měli bychom vědět intuitivní význam použití výběrové střední hodnoty a rozptylu, $\overline{Y}$ a s^2, kterým získáváme pohled na průměrnou hodnotu našeho datového vzorku a jeho rozptýlení (právě kolem této střední hodnoty). Teoretickými, protějšky těchto statistik jsou populační střední hodnota, $E(Y)$ a populační rozptyl $var(Y)$, za kterými stojí podobná intuice, ovšem vztažená k charakteristikám celé populace, které obvykle nejsme schopni rozumě získat a snažíme se je tak vhodně odhadnout (příklad s výnosností příští měsíc byl z tohoto pohledu odlišný, neboť jsme si zde jednoznačně stanovili diskrétní pravděpodobnostní rozdělení veličiny výnosů akcie).

1.5.2 Korelace

Střední hodnota a rozptyl jsou vlastnosti vztažené k jediné proměnné. Nás však může zajímat prozkoumat vztah mezi dvěma (nebo více) proměnnými, což je jednou z náplní ekonometrie, která v tomto ohledu disponuje řadou různých přístupů. Prvním krokem k dalším analýzám je zavedení pojmu *korelace*. Korelace (korelační koeficient) je důležitý způsob číselného vyjádření závislosti mezi dvěma proměnnými. Dříve než se dostaneme k ilustrativním příkladům, podíváme se na trochu teorie.

Nechť X a Y jsou dvě proměnné (např. hustota obyvatelstva a míra odlesnění) a předpokládejme, že máme k dispozici data pro $i = 1, \ldots, N$ různých jednotek (např. zemí). Korelace mezi X a Y označíme písmenkem r a její matematické vyjádření je

$$r = \frac{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sqrt{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}\sqrt{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}}.$$

Proměnné, ke kterým se koeficient korelace váže obvykle vyplývají z kontextu. Nicméně někdy je dobré použít dolní index pro ujasnění, že r_{XY} je korelace mezi proměnnými X a Y, r_{XY} je koeficient korelce mezi proměnnými X a Y atd.

Po vypočtení korelačního koeficient obdržíme nějaké číslo (např. $r = 0.55$). Měli bychom tudíž vědět, jak toto číslo interpretovat. Než se dostaneme k intuitivnímu popisu chápání korelace, podíváme se na její vlastnosti.

Vlastnosti korelace

1. Koeficient r leží mezi -1 a 1.
2. Kladné hodnoty r ukazují pozitivní korelaci mezi veličinami, negativní hodnoty pak korelaci negativní. Pokud je $r = 0$, znamená to, že obě veličiny jsou nekorelované.
3. Vyšší kladné hodnoty r ukazují na silnější pozitivní korelaci. Hodnota $r = 1$ hovoří o perfetktní korelaci. Větší (v absolutní hodnotě) záporné hodnoty r naznačují silnější negativní korelaci, kdy $r = -1$ hovoří o perfektní negativní korelaci.
4. Korelace mezi Y a X je stejná jako korelace mezi X a Y.
5. Korelace mezi jakoukoli proměnnou a sebe samou je vždy rovna jedné.

Ekonometrové používají pojem „korelace“ stejně jako každý jiný, tedy jako míru síly vztahu či závislost mezi dvěma proměnnými. Příklad 1.1 pokračuje v práci s daty o hustotě obyvatel a odlesněním a ukazuje, jak je možno interpretovat získané výsledky korelačního koeficientu.

Příklad 1.1. *Korelace mezi odlesňováním a hustotou obyvatelstva*
Předpokládejme, že nás zajímá analýza vzájemné závislosti vývoje odlesňování krajiny a hustoty obyvatelstva. Příslušná data jsou totožná s těmi, na jejichž základě byl vytvořen obrázek 1.3. Jedná se tedy o průřezová data 70 zemí tropického pásma (data z učebnice Koopa [17], v Gretlu odpovídají datovému souboru *forest.gdt* na záložce *Koop*). Korelace mezi odlesněním (Y) a hustotou obyvatelstva (X) je 0.66. Protože se jedná o hodnotu větší než nula, můžeme vyslovit následující tvrzení:

1. Existuje pozitivní vztah mezi odlesněním a hustotou obyvatelstva.
2. Země s vyšší hustotou zalidnění mají tendenci mít vyšší míry odlesnění. Země s nízkou hustotou obyvatelstva mají tendenci vykazovat nízké míry odlesnění. na tomto místě je třeba zdůraznit význam spojení *„mají tendenci“*. Pozitivní korelace neznamená, že každá země s vysokou hustotou obyvatelstva bude mít nutně vysoké míry odlesnění, jedná se spíše o obecnou tendenci. Je samozřejmě možné, že několik jednotlivých zemí nebude odpovídat tomuto chování.
3. Míry odlesnění se liší mezi jednotlivými zeměmi, stejně jako hustoty obyvatel. Některé země mají míry ztrát lesní plochy vysoké, jiné země je mají zase nízké. Jednotlivé nadprůměrné (podprůměrné) hodnoty mají tendenci odpovídat nadprůměrným (podprůměrným) hodnotám pozorovaným v hustotě obyvatelstva.

Předchozí tvrzení se týkala situace, kdy byl koeficient r pozitivní. Pokud by r bylo negativní, platila by opačná tvrzení. Vysoké (nadprůměrné) hodnoty X by byly doprovázeny nízkými (podprůměrnými) hodnotami Y apod. Není úplně snadné mít nějakou intuici pokud jde o přesný smysl získaného korelačního koeficientu, ve smyslu, jak se korelace 0.66 liší od korelace 0.26. V oblasti sociologie a sociálních věd je populární tabulka[3], která sice ukazuje možnou interpretaci korelační závislosti, ale přece jen je to jen subjektivní kategorizace autora a je k ní tedy třeba přistupovat s rezervou.

Chápání významu korelačního koeficientu napomáhají obrázky 1.5 až 1.7 a ještě se k nim vrátíme při diskuzi nad problematikou regrese. Měřítkem variability stupně míry odlesňování mezi zeměmi může být směrodatná odchylka. Skutěčnost, že míra odlesnění a hustota obyvatelstva jsou pozitvně korelované znamená mimo jiné to, že jsou spolu svázány i variability v příslušných proměnných. Jak bylo zmíněno, pozitivní korelace znamená tendenci, kdy nadprůměrné hodnoty jedné veličiny jsou doprovázeny nadprůměrnými hodnotami veličiny druhé (a to stejné platí i pro veličiny podprůměrné). Čím silnější je korelace a tím i závislost mezi oběma veličinami, tím více spolu

[3] De Vaus, David: Analyzing Social Science Data (50 Key Problems In Data Analysis). Sage Publications Ltd (United Kingdom), 2002.

Tabulka 1.2: Možná intepretace velikosti koeficientu korelace.

Hodnota korelace (v abs. hodnotě)	Interpretace
0.01 – 0.09	triviální, žádná
0.10 – 0.29	nízká až střední
0.30 – 0.49	střední až podstatná
0.50 – 0.69	podstatná až velmi silná
0.70 – 0.89	velmi silná
0.90 – 1.00	téměř perfektní

bude variabilita veličin provázána. Druhá mocnina korelačního koeficientu (r^2) měří v našem příkladu podíl variability odlesnění mezi zeměmi která je svázána respektive vysvětlena variabilitou v hustotě obyvatelstva. Jinými slovy, korelace je číselné vyjádření stupně toho, jak si vzájemně odpovídá chování veličiny X a Y. V příkladu odlesňování a hustoty obyvatelstva je druhá mocnina korelace rovna $0.66^2 = 0.44$, a my můžeme říct, že variabilitita odlesňování mezi zeměmi může být vysvětlena z 44% variabilitou hustoty obyvatelstva mezi zeměmi.

Příklad 1.2, s cenami domů, nám umožňuje zabývat se problematikou *kauzality* mezi veličinami. Zajímá nás, jestli jedna veličina kauzálně působí na jinou. Nebudeme se pokoušet o formální vyjádření pojmu kauzality, neboť je pro nás zcela dostačující to, jak ji chápeme v každodenním životě. V příkladě s cenami domů je vcelku rozumné chápat pozitivní korelaci mezi cenou domů a jeho rozlohou jako kauzální závislost. To znamená, že rozloha domu je proměnná která přímo ovlivňuje ceny domů (kauzálně na ně působí). Ovšem cena domů kauzálně neovlivňuje jejich rozlohu. Jinými slovy, směr kauzality je od rozlohy k ceně, nikoli naopak.

Jiný způsob uvažování na touto problematikou je zamyslet se nad tím, co se asi může stát, když k domu přikoupíme pozemek a zvýšíme tak jeho celkovou rozlohu. Tato aktivita by asi měla zvýšit cenu příslušného domu (zvýšená rozloha kauzálně působí na cenu domu). Pokud si však položíme opačnou otázku, tedy jestli se zvýšením ceny domu zvýší i jeho rozloha, tak tento druh kauzality asi těžko nalezneme (cena domu kauzálně nepůsobí na jeho rozlohu). Pokud by ceny domu z nějakého důvodu náhle vzrostly (např. v důsledku ekonomického růstu oblasti), tak to nezanmená, že domy náhle zvýší svou rozlohu.

Analogická úvaha by byla i v případě, kdybychom „rozlohu domu“ nahradili proměnnou „počet ložnic“. Opět bude rozumné předpokládat, že pozitivní korelace mezi Y = cena domu a Z = počet ložnic. je způsobena kauzálním vlivem Z na Y než kauzalitou opačného směru. Podíváme-li se na korelaci mezi X = rozloha domu a Z = počet ložnic, potom slabou pozitivní korelaci je jen těžké interpretovat v kauzálním slova smyslu. Existuje tendence, že domy s vyššám počtem ložnic mají větší rozlohu, nicméně tato tendence neimplikuje že počet ložnic kauzálně působí na rozlohu domu.

Důležitou věcí při empirické práci je dokázat interpretovat dosažené výsledky. Příklad cen domů tuto věc pěkně ilustruje. Nestačí jen prezentovat, jaký nám vyšel koe-

Příklad 1.2. *Ceny domů ve Windsoru, Kanada*
Datový soubor *hprice.gdt* (viz Gretl, záložka *Koop*) z učebnice Koopa [17] obsahuje data o $N = 546$ domech prodaných ve Windsoru, v Kanadě, v průběhu léta roku 1987. Obsahuje jednak samotné prodejní ceny domů (v kanadských dolarech) a řadu dalších údajů charakterizujících dané domy. Zaměřme se jen na vztah mezi $Y =$ prodejní cena domu a $X =$ rozloha domu ve čtverečních stopách. Korelace mezi oběma proměnnými je $r_{XY} = 0.54$. Na tomto základě můžeme o cenách domů ve Windsoru vyslovit následující závěry:

1. Domy s větší rozlohou mají tendenci mít vyšší hodnotu než domy s rozlohou malou.
2. Existuje pozitivní souvislost mezi rozlohou a prodejní cenou.
3. Variabilita v rozloze domů vysvětluje 29 % (tzn. $0.54^2 = 0.29$) variability cen domů.

Přidáme-li do našich úvah třetí proměnnou, $Z =$ počet ložnic, můžeme spočítat korelaci mezi cenami domů a počtem ložnic. Výsledkem je korelace $r_{YZ} = 0.37$. Tento výsledek nám říká, vcelku v souladu s našim očekáváním, že domy s větším počtem ložnic mají tendenci mít vyšší hodnotu než domy s menším počtem ložnic. Podobně můžeme spočítat korelační koeficient mezi počtem ložnic a rozlohou domu. Výsledkem je hodnota $r_{XZ} = 0.15$, která nám napovídá, že domy s vyšší rozlohou mají tendenci mít vyšší počet ložnic. Ovšem korelační koeficient je v tomto případě relativně malý a poněkud neočekávaně nám napovídá, že vztah mezi velikostí domu a počtem ložnic je vcelku slabý. Jinými slovy, asi bychom očekávali, že domy s velkou rozlohou bývají velké, měly by tedy mít i více ložnic než domy malé. Korelační koeficient nám však oznamuje, že tendence pro tento jev je z pohledu pozorovaných dat velmi slabá.

ficient korelace (např. $r_{XY} = 0.54$), důležitá je rovněž intepretace. Interpretace samozřejmě vyžaduje jak dobré intuitivní znalosti o tom, co je to korelace, tak i dobrou znalost ekonomického problému, kterým se zabýváme. Podíváme se tedy na otázku proč spolu mohou být veličiny korelovány a co nás může inspirovat při interpretaci těchto výsledků.

Proč jsou veličiny korelované?

V příkladu vztahu odlesňování a hustoty obyvatelstva jsme zjistili pozitivní korelaci a závislost mezi těmito veličinami. Jakou formu však tato závislost vlastně má? Obvykle se snažíme uvažovat v pojmech kauzality nebo vlivu a skutečně je obvyklé, že korelace a kauzalita jsou spolu úzce spojeny. Zjištění, že hustota obyvatelstva a míra odlesňování jsou korelovány může znamenat, že hustota obyvatelstva může přímo kauzálně působit na míru odlěsňování. Podobně, zjištění, že pozitivní korelace mezi úrovní dosaženého vzdělání a mzdou může být interpretována způsobem, že vyšší vzdělání má přímý vliv

na výši mzdy. Příklad 1.3 však ukazuje, že interpretace korelace jako kauzality nemusí být vždy korektní a správná.

Příklad 1.3. *Korelace neznamená kauzalitu*
Je obecně akceptovaným faktem, že kouření způsobuje rakovinu plic. Předpokládejme, že jsme získali data o velkém množství obyvatel. Tato data popisují jednak počet cigaret, které jednotlivé osoby ze vzorku vykouří v průběhu týdne (proměnná X) a zda-li osoba prodělala či má diagnostikovanou rakovinu plic (proměnná Y). Protože předpokládámě kauzalitu, že kouření způsobuje rakovinu, nepochybně získáme hodnotu korelačního koeficientu mezi těmito veličinami větší než nula, tedy $r_{XY} > 0$. To znamená, že mezi kuřáky lidé mají tendenci mít větší míru výskytu rakoviny plic, než nekuřáci. V tomto případě nám pozitivní korelace mezi X a Y indikuje přímou kauzalitu.
Nyní předpokládejme, že budeme mít další údaj týkající se téhož vzorku populace, který bude odpovídat množství alkoholu vypitého v průběhu týdne, proměnná Z. Obvykle můžeme pozorovat skutečnost, že težcí pijáci mají tendenci zároveň kouřit, tedy $r_{XZ} > 0$. Tato korelace neznamená kauzalitu, že kouření vede lidi k alkoholismu. Spíše to reflektuje určité sociální či psychologické faktory, kdy lidé kteří kouří mají tendenci i pít. Korelace mezi dvěma proměnnými tedy neznamená, že jedna proměnná kauzálně působí na druhou. Je totiž dost dobře možné, že zde existuje třetí proměnná, kderá tuto korelaci způsobuje.
Předpokládejme korelaci mezi rakovinou plic a pitím alkoholu. Protože kuřáci mají tendenci vyššího výskytu rakoviny plic a současně kuřáci mají obvykle tendenci i více pít, není moc nereálné předpokládat, že výskyt rakoviny plic bude vyšší právě mezi težkými pijáky (tzn. $r_{YZ} > 0$). Je nutné zdůraznit, že tato pozitivní korelace neimplikuje kauzalitu, že spotřeba alkoholu vede k rakovině plic. Je to kouření cigaret, které vede k rakovině a je jen souhrou psychologických a sociálních faktorů, že kouření a pití alkoholu jdou spolu v řadě případů ruku v ruce. Při interpretaci výsledků je tak velmi důležité dobře zvážit, jestli korelaci můžeme v tom či onom příkladě spojovat i s kauzalitou.

Další důležitý rozdíl je mezi přímou a nepřímou kauzalitou. Nalezli jsme pozitivní korelaci mezi hustotou obyvatelstva (X) a mírou odlesňování (Y), kdy $r_{XY} > 0$. Příčinou této korelace může být skutečnost, že tlaky vyvstávající z vysoké hustoty obyvatelstva v zemědělských oblastech nutí farmáře ke kácení lesů (či pralesů), aby se tak uvolnilo místo pro novou zemědělskou půdu pro pěstování plodiny, které jsou nutné pro uživení obyvatelstva. Je to právě proces zemědělské expanze, který má přímý vliv na odlesnění. Pokud bychom spočítali korelační koeficient mezi mírou odlesnění a zemědělskou expanzí (proměnná Z), zjsitli bychom, že $r_{YZ} > 0$. V tomto případě hustota obyvatelstva nepřímo kauzálně působí na míru odlesnění. Můžeme tedy říct, že X (populační tlaky), způsobují Z (zemědělská expanze), což má kauzální důsledek na Y (míru odlesnění). Tento scénář odpovídá zjištění, kdy $r_{XY} > 0$ a $r_{YZ} > 0$. V příkladu s cenami domů je pravděpodobné, že pozorovaná pozitivní korelace odpovídá přímé kauzalitě. Mít dům s velkou rozlohou považují lidé již sami o sobě za výhodu, tudíž zvýšení rozlohy domu přímo zvyšuje jeho cenu. Není tu žádná třetí proměnné a

tudíž i kauzalita je přímá.

Co si tedy odnést z těchto příkladů? Korelace sama o sobě nemůže automaticky vést naše úvahy směrem ke kauzální závislosti. Na příkladu kouření a rakoviny (viz příklad 1.3) nás zjištění o pozitivní korelaci mezi kouřením a výskytem rakoviny plic spolu s poznatky lékařské vědy, že látky obsažené v cigaretách ovlivňují lidské tělo natolik, že můžeme přijmout závěr, že kouření způsobuje rakovinu- V příkladu o cenách domů (viz příklad 1.2) nám náš cit napovídá, že proměnná „počet ložnic“ může přímo ovlivnit cenu domů. V ekonomi je tak možno koncept korelace propojit s přesvědčivou ekonomickou teorií nebo jinými logickými úvahy, abychom dospěli k závěru o kauzalitě.

Korelace a bodový graf

Trochu intuice o významu korelace můžeme získat z bodových grafů. Vzpomeňme si, že jsme při diskuzi nad obrázkem 1.3 zmiňovali pozitivní nebo negativní vztah mezi proměnnými, a to na základě toho, jestli příslušný graf vykazoval rostoucí nebo klesající tendenci. Pokud jsou dvě proměnné korelovány, bude bodový graf znazorňující vzájemnou závislost těchto proměnných vykazovat právě takovéto chování. Bodový graf hustoty obyvatelstva vzhledem k míře odlesnění vykazuje rostoucí sklon (viz obrázek 1.3). Tento obrázek nám naznačuje, že tyto proměnné by měly být pozitivně korelovány, což se potvrzuje na základě hodnoty korelačního koeficientu $r = 0.66$. Důležité je zdůraznit, že pozitivní korelace znamená rostoucí sklon chování bodového grafu a negativní korelace sklon klesající.

Obrázek 1.5 využívá data o cenách domů ke konstrukci bodového grafu, kde $X =$ rozloha domu a $Y =$ cena domu.Koeficient korelace je $r_{XY} = 0.54$, což je kladné číslo. Tato pozitivní závislost (rostoucí sklon) je patrný z obrázku 1.5. Domy s malou rozlohou (malé hodnoty na horizontální ose x) mají tendenci mít i nízkou cenu (malé hodnoty na vertikální ose y). Podobně platí i vztah velkých hodnot těchto proměnných.

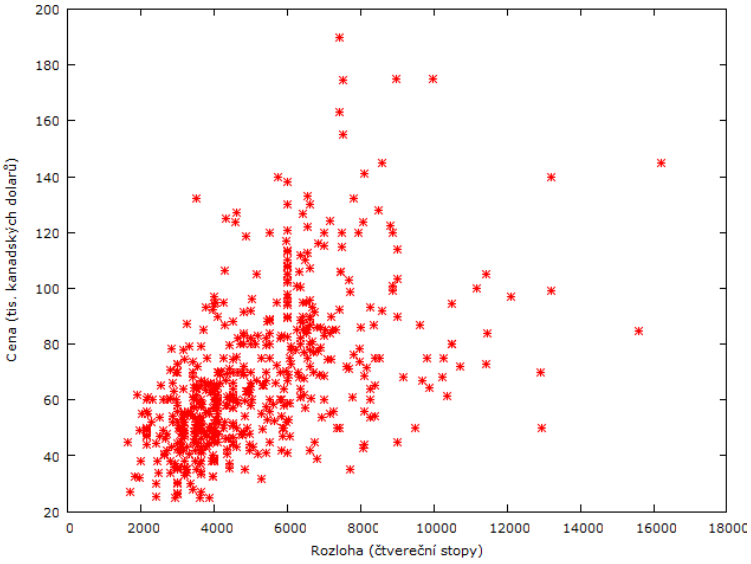
Zatím jsme diskutovali problematiku znaménka korelace. Bodový graf nám však dokáže naznačit to, jak silná tato korelace je. Obrázek 1.6 představuje dvě dokonale korelované veličiny ($r = 1$). Jedná se samozřejmě jen o umělá, simulovaná data. Všechna pozorování leží na jedné přímce.

Obrázek 1.7 představuje dvě nekorelované veličiny ($r = 0$). Všimněme si, že body jsou náhodně roztroušeny po celém grafu. Reálná data obvykle odpovídají situacím mezi těmito extrémními případy.

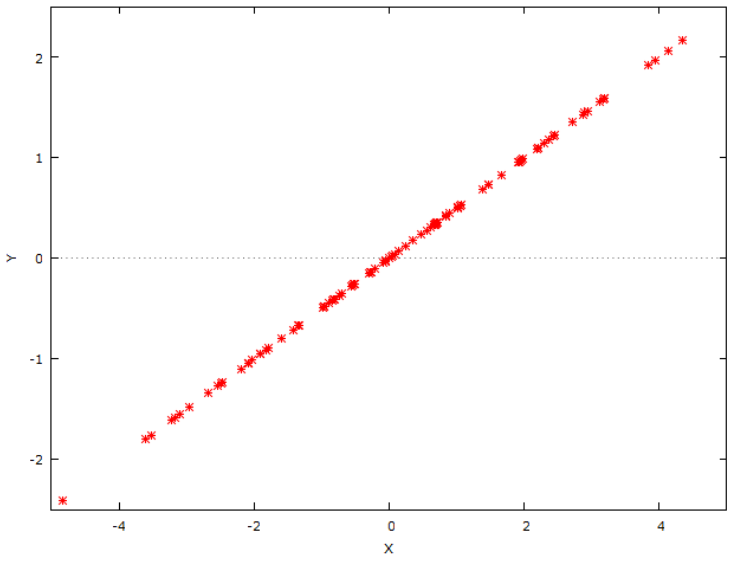
Ačkoli jsme zde prezentali zejména pozitivní korelaci, podobné závěry platí i pro korelaci negativní, kdy chování pozorování zobrazených na bodovém grafu má klesající tendenci. Korelace nám fakticky říká, jak dobře můžeme pozorovanými body proložit přímku. Silně korelované veličiny odpovídají pozorováním ležícím blízko takovéto přímky. Slabě korelované veličiny jsou více rozptýlené.

Korelace mezi několika proměnnými

Korelace je charakteristika vztažená ke dvěma proměnným. Obvykle však pracujeme s více proměnnými. Ceny domů tak závisí nejen na rozloze těchto domů, ale i na počtu

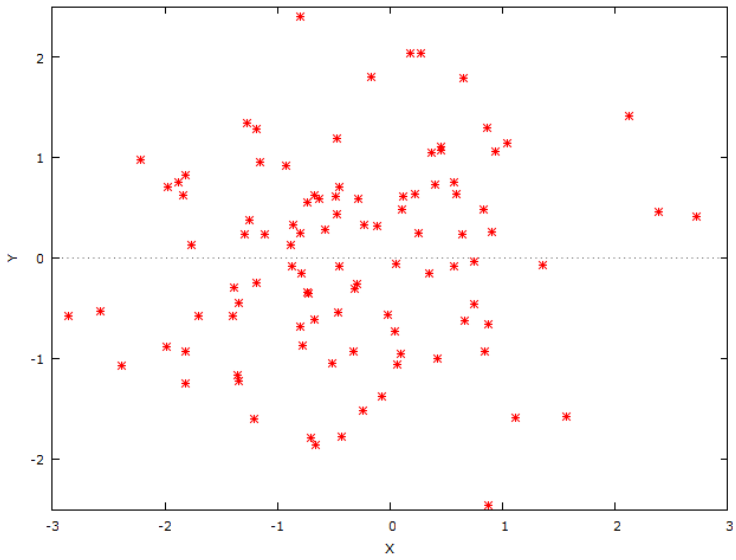


Obrázek 1.5: Bodový graf rozlohy domů vzhledem k jejich ceně.



Obrázek 1.6: Bodový graf dvou dokonale korelovaných proměnných.

ložnic, počtu koupelen apod. Jak uvidíme v dalších kapitolách, regrese je nejvhodnějším nástrojem pro analýzu více než dvou proměnných. Není však neobvyklé, že si i při práci s více proměnnými spočítáme vzájemné korelace těchto proměnných. Máme-li



Obrázek 1.7: Bodový graf nekorelovaných proměnných.

tři proměnné (X, Y, Z), počet možných korelací je tři (r_{XY}, r_{XZ}, r_{YZ}). Nepočítáme tedy korelace jednotlivých veličin se sebe samou, které jsou rovny jedné. Víme také, že $r_{XY} = r_{YX}$. Přidáme-li čtvrtou proměnnou W, počet možných korelací se nám rozšíří na šest. Obecně, pro M různých proměnných máme $\frac{M(M-1)}{2}$ možných korelací. Obvyklý způsob prezentace těchto korelačních koeficientů je za pomocí tzv. korelační matice, či tabulky korelací. Příkladem je tabulka 1.3 pro umělá data popř. tabulka 1.4 s proměnnými o cenách domů (v tomto případě je X cena domu, Y je orzloha domu a Z počet ložnic. Číslo 0.318 odpovídá korelačnímu koeficientu r_{XY} (nachází se na řádku X a sloupci Y). Podobně $r_{XZ} = -0.131$ a $r_{YZ} = 0.097$. Hodnota 1.000 odpovídá zmiňované skutečnosti, že jedna veličina je vždy perfektně korelována se sebe samou. Prostor nad horní diagonálou zůstává obvykle prázdný, protože korelační matice je symetrická, tedy $r_{XY} = r_{YX}$.

Tabulka 1.3: Korelační matice - umělá data.

	X	Y	Z
X	1.000		
Y	0.318	1.000	
Z	-0.131	0.097	1.000

Tabulka 1.4: Korelace cen domů, rozlohy a počtu ložnic.

	X	Y	Z
X	1.000		
Y	0.5358	1.000	
Z	0.3664	0.1519	1.000

1.5.3 Populační korelace a kovariance

Když jsme hovořili o střední hodnotě a rozptylu, rozlišovali jsme, jestli se jedná o populační nebo výběrové statistiky. Totéž rozlišení platí i pro korelace. Pro označení *populační korelace* použijeme značení $corr(X, Y)$, pro označení výběrové korelace setrváme u značení pomocí písmenka r. Formální definic populační korelace lze nalézt v příloze B. Na tomto místě se pokusíme o neformální přístup za použití příkladu ze světa financí. Předpokládejme portfolio skládající se z akcií dvou společností s výnosností X a Y. Očekávaná hodnota portfolia závisí na očekávaných výnose jednotlivých akcií, tedy $E(X)$ a $E(Y)$. Jaká je rizikovost tohoto portfolia? Riziko jedné akcie (či akcií jedné společnosti) lze popsat příslušnou variabilitou výnosů. Ovšem v rámci portfolia akcií je důležitá vzájemná korelace jednotlivých akciových podílů. Při hodnocení rizikovosti portfolia nás tedy zajímá $corr(X, Y)$.

Pro ilustraci předpokládejme investora, který chce po dobu letních měsíců investovat do akcií dvou společnsotí: společnosoti na výrobu deštníků a společnosti vyrábějící zmrzliny. Prodeje těchto společnsotí jsou závislé na počasí. Pokud je horké, slunečné léto, výrobci zmrzliny si přijdou na své (a vlastníci jejich akcií realizují dobré výnosy). Pokud je léto deštivé, prodejnost zmrzlin není z nejlepších (a vlastníci mohou získat jen malé zisky, nebo mohou zaznamenat dokonce ztráty). Zdá se tedy, že investice do výroby zmrzliny je vcelku riziková. Stejně tak je i riziková investice do společnsoti na výrobu deštníků, a to ze zcela opačných důvodů. Slunečné léto moc deštníků neprodá, naopak deštivé počasí zajistí dobré prodeje. Celé portfolio tvořené podíly do těchto dvou společností je méně rizikovější, než investice do podílu jediné z těchto společností. Pokud se jedné společnosti daří dobře, druhá na tom bude asi hůře. Při deštivém létě získají investoři dobré výnosy z části portfolia do akcií deštníkové firmy a horší výnosy z části podílů do akcií zmrzlinové společnosti. Slunečné léto bude znamenat analogicky opačnou situaci pokud jde o výnosy jednotlivých částí portfolia. Celé portfolio tak bude relativně bezpečnou investicí, přinášející adekvátní zisky bez ohledu na počasí.

Předchozí příklad ukazuje, jak důležitá je korelace mezi výnosností jednotlivých akcií při ocenění rizikovosti portfolia. V našem příkladu dvou společností tato korelace byla negativní (když jedna společnsot má dobré výnosy, druhá je bude mít špatné). V praxi samozřejmě mohou být korelace mezi výnosy z akcií dvou společnsotí pozitivně i negativně korelovány. Tento příkald by mohl být dobrou motivací proč je korelace důležitá minimálně v oblasti finanční ekonomie.

Abychom si odvodili vzorec pro populační korelaci, musíme si zavést koncept *kovariance*. Už název nám napovídá, že bude vyjadřovat jakousi společnou varibilitu

dvou veličin. Kovariance je definována následovně:

$$cov(X,Y) = E(XY) - E(X)E(Y).$$

Víme již, jak spočítat střední hodnoty $E(X)$ a $E(Y)$. Analogicky lze spočítat $E(XY)$ jen jako proměnnou vezmeme součin XY. Populační korelace je kovariance normalizovaná tak, že má stejné vlastnosti jako výběrové korelace (stačí jen v přehledu vlastností nahradit r pomocí $corr(X,Y)$). Všimněme si také, že jelikož dochází ve vztahu pro kovarianci k násobení veličin, odpovídají jednotky naměřené kovariance součinu jednotek těchto veličiny (jsou-li jednotky stejné, pak se jedná o druhou mocninu). Normováním se z kovariance stává bezrozměrná veličina, stejně jako výběrový korelační koeficient. Populační korelace tedy odpovídá výrazu

$$corr(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}.$$

Znalost tohoto vztahu určitě není na škodu, i když se s ním v textu málokdy setkáme. Je však důležité mít určitou intuici, co to korelace je a jak závisí na variabilitě a kovarianci příslušných dvou veličin.

Podobně jako u středních hodnota a rozptylů jsou i v tomto případě výběrové stastitiky používány jako odhady svých populačních protějšků. U příkaldu zmrzliny a deštníků by manažér portfolia rád znal $corr(X,Y)$, tedy populační korelaci mezi výnosy akcií příslušných společností. Pravděpodobně by byl schopen shromáždit data o výnosnosti těchto společností za posledních 20 letních období, čímž by snadno spočetl r, tedy výběrovou korelaci. Výběrová korelace může být použita jako dobrý odhad populační korelace, $corr(X,Y)$.

1.6 Ekonometrie a počítače

Současná ekonometrie se neobejde bez využití výpočetní techniky. Přehled progrmů využitelných pro ekonometrickou analýzu nabízí tabulka 1.5. Nejedná se určitě o vyčerpávající přehled použitelných programů, nicméně rozhodně se jedná o vhodné reprezentanty.

Software je rozdělen do dvou skupin. První skupinu tvoří volně dostupný software (GNU General Public Licence), druhou pak komeční software, kdy potřebnou licenci je třeba zakoupit. Jednotlivé programy se od sebe odlišují (kromě ceny) obvykle grafickým rozhraním, možnostmi vlastního programovacího jazyka (jazyka pro tvorbu skriptů, díky kterému si uživatel může naprogramovat vlastní funkce či procedury) a v množství funkcí, které nabízejí. Některé jsou zaměřeny spíše na oblast čistě statistické analýzy (SPSS), některé do oblasti časových řad (JMulTi), pro některé je specialitou práce s tzv. modely kvalitativních a omezených vysvětlovaných proměnných (LIMDEP) a jiné mají své zaměření primárně do oblasti finanční ekonometrie (R/Rmetrics). Stránky příslušného software nám nejlépe napoví, k čemu je daný software nejvíce vhodný, nebo jestli se jedná o univerzální ekonometrický nástroj.

Velmi mocnými nástroji, i když ne pro každého uživatele přijatelnými, jsou programy Octave, R/Rmetrics a Matlab. Jedná se v podstatě systémy s vlastním programovacím jazykem (Octave a Matlab ho mají totožný) navržené pro složité numerické

Tabulka 1.5: Přehled ekonometrického software.

Typ - Název	Výrobce	Poznámka
Open source		
gretl	„By econometricians, for econometricians“	
JMulTi	Benkwitz, Krätzig	časové řady
Octave	University of Wisconsin	
R/Rmetrics	Free Software Foundation, Inc.	finanční
Komerční		
EViews	QMS Software, Inc.	
GAUSS	Aptech Systems, Inc.	
LIMDEP	Econometric Software, Inc.	
Matlab	MathWorks, Inc.	
RATS	Estima	
SAS	SAS Institute	
SHAZAM	Northwest Econometrics, Ltd.	
SPSS	SPSS, Inc.	statistický
Stata	StataCorp	

výpočty. Disponují celou řadou implementovaných či externě stáhnutelných toolboxů či pluginů (což jsou balíčky již hotových a naprogramovaných funkcí pro různé účely použití). Matlab tak má k dispozici implementovaný statistický toolbox (funkce pro základní regresní techniky a statistiku), ekonometrický toolbox (pro práci s časovými řadami) a optimalizační toolbox (nástroje pro optimalizaci). Zcela volně je pro Maltab dostupný tzv. Econometrics Toolbox, z části naprogramován a spravován Jamesem P. LeSagem [19], který obsahuje velké množství funkcí a nástrojů pro odhad široké škály ekonometrických modelů. Nevýhodou zde je uživatelské rozhraní, které není tzv. GUI (Graphical User Interface), což znamená, že veškeré odhady modelů musíme obvykle zapsat pomocí příkazů. Výhodou tohoto typu programů je však obrovská flexibilita, kdy je možné naprogramovat si funkce pro řešení i těch nejsložitějších problémů. Jako příklad můžeme uvést nástroje pro makroekonomické modelování, Juillardův Dynare [14] či IRIS Jaromíra Beneše [4], které běží jako toolboxy pod Matlabem (v případě Dynare i pod Octave).

Pro běžného uživatele přívětivější jsou programy s grafickým uživatelským rozhraním, kde se obvykle velmi snadno „proklikáme“ k provedení požadovaných odhadů. Příkladem je volně dostupny gretl nebo komerční EViews a Stata. Všechny tyto programy nabízí v případě potřeby i využití vlastních programovacích jazyků, kterými lze jednak provést všechny operace prováděné „klikáním“, jedna si tak lze naprogramvovat vlastní spustitelné skripty a vytvářet si tak vlastní funkce či procedury, které nejsou automaticky v programech obsaženy a my je z nějakého důvodu potřebujeme. Ovládání je obvykle intuitivní a člověk se do něj velmi snadno dostane, nejlépe pokud si chvíli na jednoduchých příkladech sám zkouší, co všechno program dokáže (samozřejmě pře-

číst si nápovědu či dokumentaci taky není od věci). Nutno podotknout, že výstup všech těchto zmiňovaných programů (tedy výsledky nějakého odhadu) jsou velmi podobné a člověk se tak v nich velmi rychle zorientuje. Komerční software nabízí trochu větší možnosti pro grafickou prezentaci výstupu a správu dat. Není moc problémů, které by gretl nebyl automaticky schopen řešit jako jeho komerční protějšky.

Je třeba zdůraznit ale jednu věc. Nezáleží na tom, s jakým programem pracujeme. Důležité je řešený problém chápat, vědět jaký nástroj či nástroje na něj chceme implementovat a uměte interpretovat dosažené výsledky. V čem daný problém technicky spracujeme je spíše otázka zvyku.

1.7 Z čeho studovat?

Podobně jako je celá řada dostupného software, je také celá řada knih zaměřených pojednávajících o ekonometrii. Přehled některých z nich nabízí tabulka 1.6. Existuje několik českých knih věnovaných ekonometrii, nicméně přes mnohdy velmi reprezentativní rozsah zpracovaných témat nedosahují kvalit zahraničních učebnic, a to z hlediska vhodnosti pro samostudium. Obvykle totiž vyžadují předchozí znalosti dané problematiky a představují tak velmi dobré referenční příručky. Tabulka opět neobsahuje všechny dostupné knihy, nicméně jedná se o reprezentativní vzorky. Knihy se mezi sebou liší úrovní obtížnosti, která odpovídá spíše hloubce zpracování jednotlivých témat, a stylem zpracování (některé jsou více povídavé, jiné mají spíše formální pojetí výkladu). Většina z nich pojednává o zcela identických problémech a tématech, nicméně každá ma jíný styl výkladu, kdy záleží na každém z nás, co mu sedne nejlépe. Ať sáhneme po jakékoli knize z oblasti „mírně“ nebo „středně“ pokročilé, nemůžeme udělat chybu (samozřejmě některé publikace jsou koncipovány výhradně na ekonometrii časových řad, panelových dat či jinou, úžeji zaměřenou problematiku). „Pokročilé“ knihy se zaměřují více na ekonometrickou teorii, ale nechybí tam samozřejmě i ukázky praktických aplikací, kterých je ale nesrovnatelně více v knížkách s „nižší“ úrovní náročnosti.

Tabulka 1.6: Přehled ekonometrických učebnic.

Úroveň – Autoři	Název	Poznámka
Základní		
Koop [18]	Analysis of Economic Data	
Mírně pokročilá		
Koop [17]	Introduction to Econometrics	širší rozsah témat
Hill, Griffiths, Lim [13]	Principless of Econometrics	
Stock, Watson [20]	Introduction to Econometrics	netypický přístup
Stock, Watson [21]	Introduction to Econometrics	netypický přístup, zkráceno
Wooldridge [23]	Introductory Econometrics (A Modern Approach)	
Středně pokročilá		
Brooks [5]	Introductory Econometrics for Finance	
Cipra [6]	Finanční ekonometrie	spíše přehledová, česká
Dougherty [8]	Introduction to Econometrics	
Enders [9]	Applied Econometric Time Series	časové řady
Gujarati, Porter [11]	Basic Econometrics	
Kennedy [15]	A Guide to Econometrics	praktická příručka
Verbeek [22]	A Guide to Modern Econometrics	
Pokročilá		
Baltagi [3]	Econometric Analysis of Panel Data	panelová data
Hayashi [12]	Econometrics	
Greene [10]	Econometric Analysis	„klasika“
Davidson, Russel, MacKinnon [7]	Econometric Theory and Methods	

1.8 Shrnutí

Na základě této kapitoly tedy již víme, že:

- ekonometrie je vědní disciplína, která rozvíjí statistické metody k odhadu ekonomických vztahů, testování ekonomických hypotéz a teorií a k hodnocení vládní či obchodní politiky;
- ekonomických model je východiskem pro sestavení ekonometrického modelu;
- ekonometrie pracuje s reálnými daty, které nesou informaci o zkoumaných problémech reálného světa;
- ekonomická data získáváme v mnoha podobách, kdy obvyklými typy jsou časové řady, průřezová data a panelová data;
- zdrojů ekonomických dat je velká spousta, nejužitečnějším u nich je internet a díky němu přístupné elektronické databáze statistických úřadů, centrálních bank či mezinárodních institucí;
- informace obsažené v datech lze získat pomocí jednoduchých grafických technik, mezi které patří histogramy či bodové grafy;
- k vyjádření informace obsažené v datech slouží rozličné numerické statistiky, kdy nejvýznamnější z nich jsou výběrový průměr (vyjadřující hodnotu, kolem které jsou data rozprostřena) a směrodatná odchylka (vyjadřující míru rozptýlení dat kolem svého průměru);
- pokud může mít veličina Y různé relaizace, potom je očekávaná (střední) hodnota, $E(Y)$, měřítkem typické nebo očekávané realizace a rozptyl, $var(Y)$ je měřítkem rozptýlení (nejistoty) spojené s možnými realizacemi;
- korelace je numerickým měřitkem vztahu (závislosti) mezi dvěma veličinami;
- korelace může být vyjádřena graficky skrze bodový graf;
- znaménko korelačního koeficientu souvisí se sklonem křivky, která nejlépe vyrovnává pozorování bodového grafu;
- velikost korelace odpovídá tomu, jak hodně rozptýlená jsou data kolem přímky nejlepšího vyrovnání;
- existuje řada důvodů, proč mohou být dvě veličiny korelovány, nicméně korelace neznamená automaticky i kauzalitu mezi dvěma proměnnými;
- $corr(X, Y)$ je populační korelace a jedná se o užitečný koncept využívaný v řadě problémů světa ekonomie a financí (např. portfolio managementu);
- existuje řada publikací věnovaných problematice ekonometrické analýzy, které se liší svým rozsahema náročností témat;

- ✌ máme k dispozici širokou paletu počítačových programů, které nám usnadňuji používání ekonometrických nástrojů a technik.

Měli bychom tak již znát a umět vysvětlit obsah následujících klíčových pojmů:

- ❀ Ekonomický model
- ❀ Ekonometrický model
- ❀ Průřezová data
- ❀ Časové řady
- ❀ Panelová data
- ❀ Střední hodnota
- ❀ Rozptyl
- ❀ Kovariance
- ❀ Korelace a korelační koeficient
- ❀ Korelační matice

Kapitola 2

Netechnický úvod do regrese

V této kapitole se dozvíme:

- ☞ co je to jednoduchá regrese a jak ji graficky interpretovat v rámci bodového grafu;
- ☞ jak interpretovat parametry jednoduchého regresního modelu;
- ☞ co je to metoda nejmenších čtverců;
- ☞ jak interpretovat koeficienty jednoduchého regresního modelu;
- ☞ jaké měřítko můžeme použít k vyjádření kvality našeho regresního modelu;
- ☞ co jsou to intervaly spolehlivosti regresních koeficientů a jak můžeme testovat vhodnost zahrnutí té či oné vysvětlující proměnné;
- ☞ co je to model vícenásobné regrese a jak se liší interpretace jeho koeficientů oproti koeficientům jednoduchého regresního modelu;
- ☞ jaké problémy přináší opomenutí důležité vysvětlující proměnné;
- ☞ jaké nepříjemnosti jsou spojené v situaci vysoké korelace vysvětlujících proměnných;
- ☞ co jsou to umělé proměnné a jaké možnosti se nám nabízí pro jejich využití v regresních modelech.

2.1 Úvod

Základním stavebním kamenem ekonometrie je model. V této kapitole se budeme věnovat nejobvykleji používanému modelu v ekonometrii: lineárnímu regresnímu modelu. Tento model je určitě zajímavý sám o sobě. Řada komplikovanějších modelů pak bývá obvykle interpretována jako rozšíření regresního modelu. Pochopení podstaty regresního modelu je tedy klíčové. V této kapitole se budeme věnovat „netechnickému“

vysvětlení problematiky regrese, čímž je myšlen přístup s minimem formální matematiky a více zaměřený na intuitivní chápání problému. V dalších kapitolách na takto získané poznatky navážeme v diskuzi nad stejnou problematikou, ale již v mnohem rogoróznějším duchu s využitím formálních odvození. Důležité nicméně je, že již na základě znalostí této kapitoly budeme schopno začít s praktickou empirickou prací s daty s využitím počítačů. Konkrétně půjde o práci s průřezovými daty, které nebudou zatíženy problémy diskutovanými v kapitole 5. Navíc, intuitivním pochopením práce s regresním model jsme schopni přejít na formálnější chápání ekonometrické teorie dalších kapitol v již dobře známém kontextu.

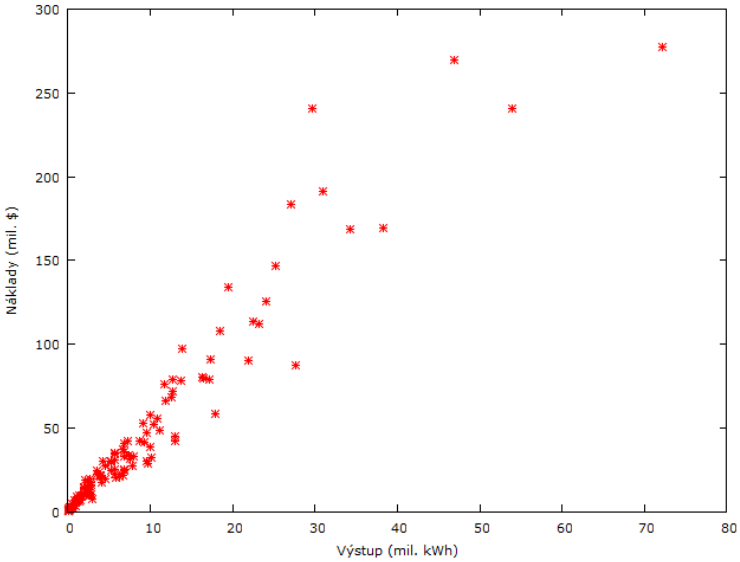
2.2 Regresní model jedné vysvětlující proměnné

Regrese je nejdůležitějším nástrojem využívaným v aplikované ekonomii pro analýzu a pochopení vztahu mezi dvěma a více proměnnými. V této kapitole se zaměříme na problematiku jednoduchého regresního modelu, tedy modelu s jedinou vysvětlující proměnnou. Přestože ve většině ekonomických aplikací se setkáváme s případy více proměnných, bude nám jednoduchý regresní model sloužit jako základ, na kterém si můžeme vybudovat potřebné teoretické koncepty s využitím grafické ilustrace. Koncept vícenásobné regrese (regrese, kdy máme analyzovat více než dvě proměnné) je v podstatě zcela analogickým a jednoduchým rozšířením regrese jednoduché.

2.2.1 Regrese jako přímka nejlepšího vyrovnání

Abychom si ilustrovali potřebné teoretické koncepty na praktické bázi, budeme předpokládat jednoduchý příklad z oblasti mikroekonomie. Naše data (*electric.gdt*, viz Gretl, záložka *Koop*) z učebnice Koopa [17]) obsahují údaje o nákladech produkce (vyjádřené v milionech dolarů) pro 123 společností vyrábějících elektřinu ve Spojených státech v roce 1970. Může nás zajímat jejich nákladová funkce, respektive faktory, které ovlivňují náklady. Náklady takové společnosti zabývající se výrobou elektřiny budou záviset asi na celé řadě faktorů. Jeden z těch nejdůležitějších bude ale výstup dané společnosti. Můžeme očekávat, že společnosti vyrábějící více elektřiny budou mít vyšší náklady, např. z toho důvodu, že k její výrobě potřebují více paliva. Kromě nákladů jednotlivých společností tak naše data obsahují údaje o jejich výstupu (měřeném v kilowatthodinách produkované elektřiny). Obrázek 2.1 je bodovým grafem těchto dvou proměnných, tedy výstupu a nákladů. Každý bod na obrázku znázorňuje konkrétní společnost, s daným výstupem (nalezneme na vodorovné ose x) a s odpovídajícími náklady (jeho hodnotu najdeme na horizontální ose y). Máme tedy stanoven ekonomický problém, tzn. chceme zkoumat jak výstup ovlivňuje náklady v odvětví výroby elektřiny, a příslušná data jsou právě to, na čem budeme náš výzkum stavět.

Už samotný obrázek podobný tomu našemu obrázku 2.1 nám řekne hodně o tom, jaký vztah mezi danými proměnnými je. Obrázka a grafy však poskytují jen neformální nástin tohoto vztahu a je tak žádoucí spočítat a vyjádřit tento vztah číselně. To je jedna z věcí, kterou nám regresní model umožní. Na obrázku 2.1 vidíme, že společnosti s vyšším výstupem mají tendenci mít i vyšší náklady. Správného mikroekonoma ale může zajímat, jaké jsou mezní náklady v tomto odvětví průmyslu, tedy jak se zvednou ná-



Obrázek 2.1: Bodový graf výstupu vzhledem k nákladům.

klady, změní-li se výstup o jednotku. Takováto otázka vyžaduje přesnou odpověď v podobě číselného vyjádření. A právě regrese nám takovou odpověď dá.

Lineární regresní model je formulován na základě předpokladu, že existuje lineární vztah mezi dvěma proměnnými, X a Y (v našem případě náklady a výstupem). Vyjádření lineárního vztahu (tedy fakticky rovnice přímky) lze zapsat v podobě

$$Y = \alpha + \beta X,$$

kde α je úrovňová konstanta (anglicky *intercept*), tedy hodnota, ve které protíná příslušná přímka ypsilonovou osu, a β je sklon této přímky (*slope*). Pokud bychom snad měli problémy s konceptem rovnice přímky, je dobré podívat se do přílohy A, pojednávající o základech matematiky. Tato přímka je označována jako *regresní přímka*. Pokud bychom znali skutečné hodnoty α a β, potom bychom znali i závislost mezi Y a X. Ve skutečnosti však jsou pro nás koeficienty α a β neznámé. Navíc, i kdyby byl náš model lineární závislosti proměnných Y a X skutečně správný a přesně odpovídající chování vysvětlované veličiny, nikdy bychom ve skutečnosti nepozorovali data, která by ležela přesně na této přímce. Tím není myšleno nic jiného, než, že se v realitě setkáváme s faktory jako chyby měření, díky kterým jednotlivá pozorování budou ležet v blízkosti pomyslné přímky, ale ne přesně na ní.

Předpokládejme například, že náklady produkce (Y) budou záviset na výstupu X, a to v následující podobě: $Y = 2 + 5X$ (tzn. $\alpha = 2$ a $\beta = 5$). Pokud je $X = 1$ (tedy výstup je jeden milión kilowatthodin), pak nám model říká, že náklady firmy by měly být $Y = 2 + 5 \times 1 = 7$ (tedy náklady by měly být sedm miliónů dolarů). Ovšem ne každá firma bude mít při výstupu jedné kilowatthodiny náklady přesně sedm miliónů dolarů. Některé firmy budou efektivnější než druhé a budou schopny produkovat stejný

objem výstupu při nižších nákladech. Regresní model stejně tak může s velkou pravděpodobností opomenout některé důležité proměnné, které mohou ovlivňovat náklady. Náklady produkce nezávisejí jen na vyprodukovaném výstupu, ale také na cenách použitých vstupů. Pokud firma A dokáže najmout zaměstnance za nižší mzdu než firma B, potom by firma A měla mít nižší náklady než firma B i v případě produkce stejného množství elektřiny (a stejné efektivitě). Z tohoto důvodu, ačkoli $Y = 2 + 5X$ nabízí přesný popis skutečného vztahu mezi Y a X, nenastane reálný případ, kdyby data ležela přímo na této přímce. Z tohoto důvodu dodáváme do modelu chybový člen (náhodnou složku), ϵ, abychom získali regresní model v podobě:

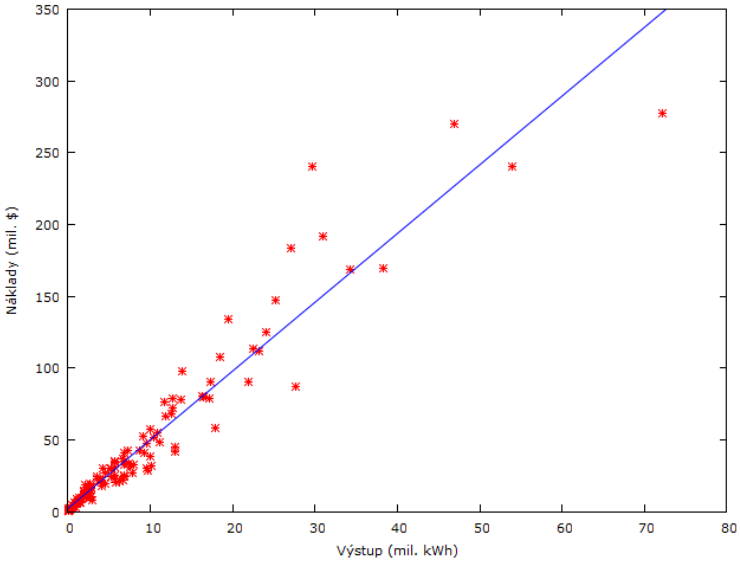
$$Y = \alpha + \beta X + \epsilon.$$

V dalších kapitolách uvidíme, jaké vlastnosti musí nezbytně ϵ splňovat, aby naše ekonometrické výsledky byly korektní a nenapadnutelné. Protože se však zabýváme netechnickým úvodem, můžeme si tento chybový člen představit způsobem, že dovoluje, aby se skutečné nákaldy firem odchylovaly od hodnot, které by předpovídala regresní přímka.

Zaveďme si nyní určitou terminologii a další intuici týkající se regresního modelu. Proměnná na levé straně rovnice, Y, je označována nazývána jako *vysvětlovaná* nebo *závisle proměnná* (*dependent variable*). Proměnná X je nazývána jako *vysvětlující proměnná* (*explanatory variable*) či *nezávisle proměnná* (*independent variable*). Písmenky α a β označujeme *koeficienty* resp. *parametry* modelu. Jak nám názvy napovídají, můžeme s úspěchem předpokládat, že vysvětlující proměnná bude vysvětlovat (ovlivňovat) závisle proměnnou, a koeficient β bude měřit vliv proměnné X na Y.

Na tomto místě bude užitečné udělat odbočku k problematice toho, proč jedna proměnná je vybrána jako závisle proměnná a druhá jako vysvětlující proměnná. V kapitole 1 jsme se otázkou chápání toho, proč mezi sebou proměnné jsou proměnné korelovány. Důležitým závěrem byla skutečnost, že korelace neznamená kauzalitu. Podobný závěr platí i pro regresi. Je zcela nezbytné zdůraznit, že musíme věnovat velkou pozornost intepretaci výsledků regrese v tom smyslu, jstli tento vztah znamená kauzalitu mezi danými proměnnými. Ideální situace je samozřejmě ta, kdy vysvětlující proměnná kauzálně působí na proměnnou vysvětlovanou. Tato situace však nemusí vždy nastat. Kdy je tedy rozumné interpretovat výsledky regrese v podobě kauzální závislosti? V mnoha případech nám to naznačí samotná ekonomická teorie, ze které vycházíme. Mikroekonomickou teorii je možno využít k odvození nákladové funkce. Tato teorie nám říká, že náklady jsou závisle proměnná a produkce a ceny vstupů jsou vysvětlující proměnné. V jiných případech nám otázku kauzality pomůže vyřešit správné pochopení problému. Vezměme si příklad s prodejními cenami domů z kapitoly 1 a jejich příslušnými charakteristikami (např. počet ložnic). V tomto případě nám logika říká, že počet ložnic by měl být vysvětlující proměnnou a cena domů závisle proměnnou. Počet ložnic ovlivní cenu domů (postavíme-li novou ložnici, můžeme tím zvýšit cenu našeho domu). Opačný směr kauzality nedává smysl. Pokud ceny v mnoha zemích v posledních letech rostly (nebo klesaly), lidé mohli sledovat, že hodnota jejich domů roste (popř. klesá). To ale určitě nemělo kauzální dopad na to, že příslušné domy začaly mít větší (resp. menší) počet ložnic.

V řadě případů však nemusí být jasné, která z proměnných kauzálně ovlivňuje druhou. Způsobuje mzdové inflace inflaci cenovou, protože firmy jsou v důsledku nárůstu



Obrázek 2.2: Bodový graf výstupu vzhledem k nákladům s regresní přímkou vyrovnání.

mezd nuceni promítnout tento nárůst do konečných cen svým zákazníkům? Nebo to je cenová inflace, která vede k inflaci mzdové, protože pracovníci požadují v rámci mzdových vyjednávání vyšší mzdy pro zachování jejich reálné hodnoty? Obě možnosti dávají smysl. V regresi zahrnující dvě proměnné, mzdovou a cenovou inflaci, tak je nejasné, která z obou má být závisle proměnná a která vysvětlující. Ja k později uvidíme, tento problém může být ekonometricky testován v rámci tzv. Grangerovských kauzalit (viz kapitola 8). Na tomto místě si jen dobře zapamatujme, že si musíme dobře promyslet situaci, kdy chceme výsledky regrese interpret v podobě existence kauzálního vztahu.

Vrat'mě se nyní zpět k našemu regresnímu modelu. Máme zaveden chybový člen, ϵ, a nevíme jaké hodnoty koeficientů α a β vlastně jsou. Prvním problémem regresní analýzy je to, určit, jak velké tyto koeficienty přibližně budou, chceme tedy *odhadnout* velikost koeficientů α a β. Využitím standardního značení budeme tyto odhad označovat jako $\widehat{\alpha}$ a $\widehat{\beta}$. Skutečné, neznáme regresní koeficienty jsou α a β, a $\widehat{\alpha}$ a $\widehat{\beta}$ jsou jejich odhady. Odhad je jedna z nejdůležitějších aktivit, kterou ekonometr provádí. Budeme si ilustrovat princip odhadu jednoduchým intuitivním způsobem s využitím obrázku. V dalších kapitolách přejdeme do formálnějšího vyjádření.

Podívejme se na obrázek 2.2, který je bodovým grafem výstupu a nákladů 123 společností vyrábejících elektřinu, přičemž je těmito body protažena přímka. Takových přímek bychom mohli vykreslit celou řadu. Proč byla ale vybrána právě tato? Odpověď je jednoduchá, tato přímka nejlépe prokládá (*fit*) data a nejlépe tak vystihuje závislost mezi výstupem a náklady. Takováto odpověď samozřejme automaticky generuje další otázku, týkající se toho, co je to „nejlépší proložení“ (*best-fitting*). Než si tuto druhou otázku zodpovíme, vrat'me se ke značení z kapitoly 1. Používali jsme zna-

čení Y (popř. X) pro označení závisle proměnné (resp. vysvětlující proměnné). Data máme pro každou proměnnou a pro řadu jednotek pozorování (tzn. pro řadu společností vyrábějících elektřinu). Jednotlivé pozorované jednotky jsou v anglických textech označovány jako *individuals*, tedy jednotlivci či jednotky, pod kterými si můžeme představit firmy, země či cokoli jiného. Písmenem N označíme počet pozorování a příslušným dolním indexem označíme jednotlivé pozorování. V takovém případě bude Y_1 označovat závisle proměnnou pro prvního jednotlivce, X_4 vysvětlující proměnnou pro čtvrtého jednotlivce, atd. Indexem i tedy označíme příslušná jendotlivá pozorování, kdy Y_i pro $i = 1, \ldots, N$ bude vyjadřovat všechna naše pozorování závisle proměnné. Regresní přímka pro každou jednotku je vyjádřena jako

$$Y_i = \alpha + \beta X_i + \epsilon_i.$$

Důležité je chápat rozdíl mezi chybovými členy (náhodnými složkami) a *rezidui* (odhady těchto náhodných složek). Náhodná složka je definována jako rozdíl mezi určitou pozorovanou hodnotou a odpovídající hodnotou na skutečné regresní přímce. Mathematicky řečeno, můžeme si předělat regresní model tak , že získáme výraz pro náhodnou složku

$$\epsilon_i = Y_i - \alpha - \beta X_i.$$

Pokud nahradíme α a β jejich odhady, získáme přímku, která je obecně odlišná od skutečné regresní přímky. Odchylky od této odhadnuté regresní přímky se nazývají rezidua a budeme je označovat jako $\widehat{\epsilon}$. Rezidua jsou tedy dána jako

$$\widehat{\epsilon} = Y_i - \widehat{\alpha}\alpha - \widehat{\alpha}\beta X_i.$$

Pro každou jednotku (pozorování) existuje chyba a reziduum. Na obrázku 2.2 jsou rezidua rozdílem vzdáleností mezi skutečným bodem pozorování a regresní přímkou vyrovnání. *Regresní přímka vyrovnání* je dána jako

$$\widehat{Y_i} = \widehat{\alpha} + \widehat{\beta} X_i,$$

kdy $\widehat{Y_i}$ označujeme jako *vyrovnané hodnoty*.

Vraťme se k otázce jak najít dobré odhady koeficientů α a β. Na obrázku 2.2 nenajdeme přímku, která by procházela všemy body. Každá přímka však s sebou přináší rezidua pro jednotlivá pozorování. Regresní přímka, která tato rezidua učiní nejmenšími možnými je právě přímka nejlepšího vyrovnání dat. Obvyklý způsob měření velikosti reziduí (je jich přece jen hodně a my bychom potřebovali jednu jejich souhrnnou charakteristiku) je *součet čtverců reziduí* (*sum of squared residuals – SSR*)[4]. Součet

[4]Značení bývá obvyklé i *sum of squared errors*, tedy *SSE*. Nicméně, v tomto případě je chybou myšlena chyba vyrovnání, tedy reziduum a nikoli náhodná složka, jako chybový člen regresního modelu.

čtverců reziduí je možno zapsat následujícími (ekvivalentními) způsoby:

$$\begin{aligned} SSR &= \sum_{i=1}^{N} \widehat{\epsilon_i}^2 \\ &= \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta} X_i\right)^2 \\ &= \sum_{i=1}^{N} \left(Y_i - \widehat{Y_i}\right)^2 . \end{aligned}$$

Předchozí rovnice využívá operátor sumace. V případě problémů s jeho používáním je možno využít přílohy A.

K získání nejlepší přímky vyrovnání chceme najít takové hodnoty $\widehat{\alpha}$ a $\widehat{\beta}$, které učiní výraz SSR nejmenším možným. V následujících kapitolách uvidíme, že řešení problému „nalezení hodnot $\widehat{\alpha}$ a $\widehat{\beta}$, které minimalizují SSR“ můžeme získat s využitím jednoduchých matematických postupů. Není ani překvapením, že odhady parametrů $\widehat{\alpha}$ a $\widehat{\beta}$ nám poskytne jakýkoli ekonometrický software (viz tabulka 1.5) nebo dokonce i Excel. Tyto odhady se nazývají odhady *metodou nejmenších čtverců*, anglicky *ordinary least squares*, a budeme je tak nazývat OLS odhady (i když existuje i pěkné české označení MNČ). Pro data společností produkujících elektřinu jsou OLS odhady $\widehat{\alpha} = 2.19$ a $\widehat{\beta} = 4.79$. Rovnice přímky z obrázku 2.2 je tedy

$$\widehat{Y_i} = 2.19 + 4.79 X_i .$$

2.2.2 Interpretace výsledků OLS odhadů

v předchozí části jsme se seznámili s tím, jak získat OLS odhad parametrů α a β. Jak ale tyto odhady (a stejně tak i samotné parametry) interpretovat? OLS odhad úrovňové konstanty v regresním modelu můžeme interpretovat jako predikovaná hodnota vysvětlované proměnné při nulové hodnotě všech vysvětlujících proměnných. A skutečně, pokud položíme $X_i = 0$ v rovnic regresní přímky vyrovnání $\widehat{Y_i} = \widehat{\alpha} + \widehat{\beta} X_i$ získáme $\widehat{Y_i} = \widehat{\alpha}$. Úrovňová konstanta nemusí mít často nějakou zajímavou ekonomickou interpretaci. V našem příkladu o společnostech vyrábějících elektřinu udává úrovňová konstanta náklady produkce společnosti s nulovým výstupem. S trochou fantazie bychom asi mohli tento parametr chápat jako úroveň fixních nákladů, (byť za předpokladu, že tato úroveň je pro všechny firmy stejná). Nicméně, z praktického hlediska její odhadnutá hodnota pro nás význam asi mít nebude.

Důležitou ekonomikou interpretaci by měla ale úrovňová konstanta v regresi vycházející z modelu oceňování kapitálových aktiv (Capital Asset Pricing Model – CAPM). Pokud jsou kapitálové trhy dokonale efektivní, měla by být její hodnota nulová. Pro CAPM je tedy klíčový právě odhad úrovňové konstanty a testování toho, jestli je skutečně její hodnota nulová.

Odhad koeficientu sklonu, $\widehat{\beta}$, má pro ekonomy význam obvykle větší. Tento odhad udává sklon (směrnici) přímky nejlepšího vyrovnání bodového grafu jako je na obrázku

2.2. Jinou interpretaci odhadu $\widehat{\beta}$ získáme derivací regresní přímky vyrovnání podle vysvětlující proměnné:

$$\frac{\mathrm{d}\widehat{Y}_i}{\mathrm{d}\widehat{X}_i} = \widehat{\beta}.$$

I kdyby nám princip derivace nebyl úplně jasná, intuice spojená s předchozím výrazem je vcelku jasná. Derivace nám měří to, o kolik se nám změní Y, když změníme X o velmi malou (nepatrnou, tedy marginální) hodnotu. Odhad $\widehat{\beta}$ může být interpretován jako mezní (marginální) efekt X na Y. Je to míra jakou vysvětlující veličina ovlivňuje veličinu vysvětlovanou. Abychom byli přesnější, můžeme $\widehat{\beta}$ interpretovat jako měřítko toho, jak moc má Y tendenci změnit se, když se změní X o jednotku. Definice „jednotky“ v předchozí větě závisí na konkrétní množině dat, ze které vycházíme. V našem případě, který zahrnuje náklady produkce výrobní jednotky produkující elektřinu, jsme získali $\widehat{\beta} = 4.79$. To je míra toho, o kolik mají náklady tendenci změnit se, když se změní výstup o jednu jednotku. Protože jsou nákaldy měřeny v miliónech dolarech a výstup je měřen v miliónech kilowatthodin produkované elektřiny, může být výsledek $\widehat{\beta} = 4.79$ interpretován následovně: jestliže se výstup zvýší o jeden milión kilowatthodin (tzn. jednu jednotku vysvětlující proměnné), budou mít nákaldy tendenci zvýšit se o 4790000 dolarů. Parametr sklonu v našem případě mezním nákladům typické firmy v odvětví, protože mezní náklady jsou definovány způsobem, „o kolik se změní celkové náklady, změní-li se výstup firmy o jednotku.“

2.2.3 Měření kvality vyrovnání regresního modelu

Nyní již víme jak spočítat a interpretovat regresní koeficienty. OLS odhady odpovídají nalezení regresní přímky, která bude „nejépe prokládat“ data, tedy bude minimalizovat součet čtverců reziduí. Nemusí však platit, že z nejlepší vyrovnání z pohledu metody bude i dobrým vyrovnáním jako takovým. Je tedy žádoucí mít měřítko kvality vyrovnání, které nám řekne, jak dobrá je naše přímka nejlepšího vyrovnání. Nejobvyklejším měřítkem je tzv. *koeficient determinace*, R^2.

Rezidua nám měří vzdálenost jednotlivých dat od regresní přímky. Analýza reziduí nám tedy může poskytnout dostatečnou informaci i kvalitě vyrovnání příslušnou regresní přímkou. K dispozici máme celkem N reziduí a pouhým pohledem na ně (či jejich jednotlivé vypsání v tabulce) asi jasnou odpověd na otázku kvality vyrovnání nemusíme získat. Potřebujeme tedy nejlépe jedno číslo, které by nám shrnulo informaci obsaženou v reziduech, A to je právě náš koeficient determinace, R^2.

Rozptyl je měřítkem disperze (rozptýlení) dat (od svého průměru). Rozptyl jakékoli proměnné můžeme odhadnout jako

$$var(Y) = \frac{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}{N-1},$$

kde $\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$ je aritmetický průměr či výběrová střední hodnota přísušné proměnné. Zadefinujme si charakteristiku zvanou *celkový součet čtvercú (Total Sum of*

Squares – TSS):

$$TSS = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2,$$

což je výraz proporcionální rozptylu. Volně řečeno, člen $N-1$ ve jmenovateli výrazu pro rozptyl $var(Y)$ se nám v konečném výrazu pro R^2 vykrátí s jiným členem $N-1$. O celkovém součtu čtvercu tak můžeme uvažovat jako o měřítku variability proměnné Y. Regresní model se snaží vysvětlit variabilitu Y s využitím vysvětlující proměnné X. Lze ukázat, že celkovou variabilitu Y můžeme rozdělit do dvou částí:

$$TSS = RSS + SSR,$$

kde RSS je *regresní součet čtverců (Regression Sum of Squares)*, tedy měřítko míry vysvětlení chování zkoumané veličiny pomocí regresního modelu. Regresní součet čtverců je dán jako

$$RSS = \sum_{i=1}^{N} \left(\widehat{Y}_i - \overline{Y}\right)^2.$$

Využijeme-li náš příklad s produkcí elektřiny, můžeme říct, že celkové náklady produkce (Y) jsou v rámci firem různé, mají tedy svou variabilitu měřenou celkovým součtem čtverců, TSS. Tato variabilita může být rozložena na část, která může být vysvětlena skutečností, že různé firmy produkují různá množství elektřiny (tuto variabilitu vystihneme našim regresním modelem a odpovídá regresnímu součtu čtverců, RSS), a na část kterou nedokážeme vysvětlit a zůstává nám v reziduálních složkách (jejichž variabilita je měřena součtem čtverců reziduí, SSR).

Koeficient determinace vychází z rozkladu celkové variability vysvětlované proměnné a je definován jako

$$R^2 = \frac{RSS}{TSS},$$

což je ekvivalentní výrazu

$$R^2 = 1 - \frac{SSR}{TSS}.$$

Intuitivně řečeno, R^2 vyjadřuje podíl celkové variability závisle proměnné Y, která může být vysvětlena vysvětlující proměnnou X. Výrazy TSS, RSS a SSR jsou čtverce nějakých čísel a jsou tedy všechny nezáporné. Protože TSS odpovídá součtu RSS a SSR, musí platit, že $TSS \geq RSS$ a $TSS \geq SSR$. Na tomto základě pak platí, že $0 \leq R^2 \leq 1$.

Další intuici ohledně koeficientu determinace získáme, pokud si uvědomíme, že malé hodnoty SSR naznačují, že rezidua jsou malá, a tím by měl regresní model dobře vystihnout chování pozorované v datech (data leží v blízkosti regresní přímky $\widehat{\alpha}+\widehat{\beta}X$). Toto tvrzení samozřejmě není zcela korektní, protože závisí samozřejmě na samotném měřítku vysvětlované proměnné. Pokud pozorování jsou v řádu tisíců a rezidua v řadu jednoteke, pak jsme řad vyrovnali dobře, pokud by ale rezidua byla v řádu jednotek a pozorování rovněž v řádu jednotek (či v řádu desetin), pak o dobrém vyrovnání mluvit nemůžeme. Ve vztahu pro koeficient determinace, R^2, toto vše v příslušných podílech zohledněno je (TSS a RSS odpovídají variabilitě kolem společné střední hodnoty a

SSR vystupuje v poměru k celkové variabilitě, TSS). Regresní přímka, která perfektně vyrovná pozorovaná data nebude mít žádnou chybu vyrovnání, tudíž $SSR = 0$ a $R^2 = 1$. Je tedy zřejmé, že hodnoty blízké jedničce implikují dobrou kvalitu vyrovnání, kdy $R^2 = 1$ znamená dokonalé vyrovnání. Nízké hodnoty R^2 odpovídají horší kvalitě vyrovnání. Pokud je $R^2 = 0$, znamená to, že model nevysvětluje žádnou variabilitu chování závisle proměnné a fakticky takový model odpovídá situaci, kdy všchny vyrovnané hodnoty odpovídají aritmetickému průměru dat $RSS = 0$. Tato situace nastane v případě, kdy v modelu vystupuje pouze úrovňová konstanta (a žádná vysvětlující proměnná). Její odhad bude odpovídat aritmetickému průměru, regresní přímka bude rovnoběžná s osou x a je logické, že když všechny vyrovnané hodnoty jsou stejné, nemají žádnou variabilitu. S nulovou variabilitou data generovaných modelem (vyrovnané hodnoty) nemáme šanci vysvětlit nenulovou variabilitu v datech.

Další možností pochopení koeficientu determinace je tedy pomocí RSS. Reziduální součet čtverců nám říká, kolik variability v Y je vysvětleno chováním (variabilitou) vysvětlujících proměnných. Pokud je RSS blízké TSS, znamená to, že vysvětlující proměnné se podílí na vysvětlení téměř celé varibility v datech a kvalita vyrovnání tak bude velmi dobrá. Ze vzahu pro R^2 vyplývá, že koeficient determinace bude blízký hodnotě jedna.

V regresi, kdy Y = náklady produkce a X = výstup pro 123 elektrárenských jednotek produkujících elektřinu je $R^2 = 0.92$. Toto číslo můžeme interpretovat tak, že 92% variability v nákladech mezi společnostmi může být vysvětleno variabilitou v jejich výstupu.

2.2.4 Základní statistické koncepty v regresním modelu

Parametry α a β vyjadřují závislost mezi závisle proměnnou Y a vysvětlující proměnnou X. Ve skutečnosti ale tuto závislost neznáme a musíme k jejímu vyjádření použít odhady $\widehat{\alpha}$ a $\widehat{\beta}$. Vzniká nám tedy otázka, jak přesné tyto odhady jsou. Obvyklý způsob zodpovězení této otázky je za pomocí tzv. *konfidenčních intervalů* či *intervalů spolehlivosti*. S nimi je úzce svázán koncept *testování hypotéz*. Na tomto místě se budeme snažit o intuitivní pochopení této problematiky. Formálnější přístup výkladu využijeme až v kapitole 3.

Metoda nejmenších čtverců nám poskytuje bodové odhady parametrů α a β (např. $\widehat{\beta} = 4.79$ je bodový odhad parametru β v regresi nákladů produkce vzhledem k výstupu pro náš příklad dat společností vyrábějících elektřinu). Bodový odhad můžeme chápat jako náš nejlepší odhad toho, jakou hodnotu má ve skutečnsoti nám neznámý parametr (např. β). Intervaly spolehlivosti nám dávají intervalové odhady rozsahu, v jakém se s vysokou pravděpodobností skutečná hodnota parametru β nachází. Tyto intervaly nám tedy představují míru nejistoty, s jakou musíme počítat při úvahách o skutečné hodnotě neznámého parametru (v našem příkladu to znamená, že „jsme si s velkou pravděpodobností jistí, že β je větší než 4.53 a menší než 5.05.). Můžeme obdržet různé intervaly spolehlivosti odpovídající různé hladině spolehlivosti. V případě 95% intervalu spolehlivost můžeme říct, že *„jsme si na* $95\ \%$ *jisti, že* β *leží v tomto intervalu,“* v případě 90% intervalu spolehlivosti tuto jistotu snížujeme na $90\ \%$ apod. Právě onu míru důvěry či jistoty, kterou máme ke zvolenému intervalu (např.

95 %) nazýváme jako *hladina spolehlivosti (confidence level)*. V následující kapitole si formálně odvodíme vztah pro výpočet intervalu spolehlivosti. Důležité je, že tyto intervaly spolehlivosti je schopen spočítat jakýkoli relevantní ekonometrický program. Předchozí řádky tak snad poskytly dostatečnou intuici pro práci s intervaly psolehlivosti. V příkladu elektrárenských společností je 95% interval spolehlivosti parametru β $[4.53, 5.05]$ (viz tabulka 2.1). To můžeme vyjádřit tak, že „jsme si na 95% jisti, že mezní efekt výstupu na náklady je nejméně 4.53 a nejvýše 5.05." Podobně můžeme interpretovat výsledky pro odhad úrovňové konstanty, α.

Testování hypotéz je další důležitou dovedností v empirické ekonomii. Formálně si vše popíšeme v následující kapitole. Na tomto místě se však zaměříme na praktické detaily toho, jak testování hypotéz provádíme a jak interpretujeme výsledky. Klasické testování hypotéz zahrnuje specifikaci hypotézy, kterou testujeme. Tato hypotéza se nazývá *nulová hypotéza* a označuje se jako H_0. Její testování probíhá oproti *alternativní hypotéze*, označované jako H_1. V jednoduchém regresním modelu obvykle testujeme hypotézu, že $\beta = 0$. Pokud je $\beta = 0$, znamená to, že příslušná vysvětlující proměnná nemá žádnou vysvětlující sílu. Jaký druh otázek nás při analýze ekonomických problémů může zajímat? Příkladem jsou otázky typu: „Zvyšuje úroveň vzdělání jednotlivce jeho potenciál, pokud jde o výši jeho pracovních příjmů?", „Zvyšuje určitý typ reklamní strategie nebo kampaně celkové tržby?", „Sníží nový vládní systém rekvalifikačních programů nezaměstnanost?". Většina otázek je tedy typu „Má vysvětlující proměnná vliv na závisle proměnnou?", resp. „Je $\beta \neq 0$ v regresi Y na X?". Účelem testování hypotéz $\beta = 0$ je právě zodpovězení těchto otázek. Formálně tedy testuje $H_0 : \beta = 0$ oproti $H_1 : \beta \neq 0$.

Testování hypotéz a intervaly spolehlivosti představují rovnocenné pohledy na stejný problém. Chceme-li testovat hypotézu, že $\beta = 0$, stačí se podívat na interval spolehlivosti a zjistit, jestli tento interval obsahuje nulu. Pokud tomu tak není, potom „zamítáme hypotézu, že $\beta = 0$" popř. řekneme, že „X statisticky významně vysvětluje Y" nebo, že „koeficient β je statisticky významný". Pokud interval spolehlivosti nulu obsahuje, potom místo „zamítnutí" hypotézy použijeme spojení, že nulovou hypotézu „příjmeme" nebo ještě lépe, že nulovou hypotézu „nezamítáme". V takovémto případě „příslušná vysvětlující proměnná nemá statisticky významný vliv pro vysvětlení chování závisle proměnné".

Intervaly spolehlivosti odpovídají různým hladinám spolehlivosti (obvyklá volba je 95 %). Podobně i testování hypotéz má svou *hladinu významnosti*[5]. Použijeme-li k testování hypotéz přístup přes interval spolehlivosti, potom hladina významnosti je 100 % mínus hladina spolehlivosti. To znamená, že pokud 95% interval spolehlivosti neobsahuje nulu, potom můžeme říct, že „zamítáme hypotézu, že $\beta = 0$ na 5% hladině významnosti" (tj. 100 % − 95 % = 5 %). Použijeme-li 90% interval spolehlivosti a zjistíme-li, že neobsahuje nulu, potom bychom řekli: „Zamítáme hypotézu, že $\beta = 0$ na hladině významnosti 10 %".

Standardní způsob, jak testovat hypotézy, je začít specifikací nulové hypotézy a zvolit hladinu významnosti. V jednoduchém regresním modelu bude obvykle stanovena $H_0 : \beta = 0$ a 5% hladina významnosti. Následně se spočítá testová statistika a

[5] Obvykle je tato hladina označována řeckým písmenem α, což ale nezaměňujme s úrovňovou konstantou, která má v této kapitole podobné značení.

porovná se s *kritickou hodnotou*, o níž se formálněji zmíníme v kapitole 3. V případě testu statistické významnosti parametru, tedy $\beta = 0$, je testová statistika známá jako *t-statistika (t-statistic, t-ratio, t-stat)*. Spočítá se jako

$$t = \frac{\widehat{\beta}}{s_b},$$

kde s_b je *směrodatná odchylka odhadu parametrů* (respektive standardní chyba odhadu parametru, $\widehat{\beta}$). V tuto chvíli nám stačí vědět, že ji spočítá každý relevantní software. Myšlenka v pozadí je takova, že přijmeme resp. nezamítneme hypotézu H_0, pokud je hodnota testové statistiky konzistentní s hodnotou, která by nám vyšla, pokud by H_0 byla pravdivá. Jestliže je H_0 pravdivá a $\beta = 0$, potom bychom očekávali, že odhad tohoto parametru, $\widehat{\beta}$, bude velmi malý. Pokud je hodnota $\widehat{\beta}$ velká, je to důkaz proti platnosti hypotézy H_0. Formálně řešíme otázka toho, jestli je $\widehat{\beta}$ velké nebo malé vzhledem ke své směrodatné odchylce. Proto je v t-statistice tento výraz obsažen. Otázkou ale zůstává, co budeme myslet „velkou“ nebo „malou“ hodnotou t-statistiky. Ve formálním, statistickém významu to znamená porovnání vzhledem k příslušné kritické hodnotě Studentova t-rozdělení. V praxi se však i bez nutnosti hledání v tabulkách můžeme obejít. Většina ekonometrických počítačových programů k výsledkům testování hypotéz (i jiným statistickým testům) poskytuje tzv. *p-hodnoty* (i tyto p-hodnoty si můžeme spočítat sami, je nutné pouze nalézt kvantil odpovídající přslušné testové statistice pro dané rozdělení). V takovém případě statistické tabulky nepotřebujeme. Příslušná p-hodnota totiž odpovídá hladině významnosti, pro kterou můžeme na zákaldě našich dat zamítnout nulovou hypotézu, H_0. Pokud pracujeme s 5% hladinou významnosti a vypočtená p-hodnota je rovna 0.05, potom můžeme H_0 zamítnout. Pokud je p-hodnota menší než 0.05, tak samozřejmě rovněž můžeme zamítnou naši nulovou hypotézu. To vyplývá ze skutečnosti, že pokud můžeme zamítnout hypotézu na dané hladině významnosti (řekněme 4 %), potom ji můžeme zamítnou i na vyšší haldině významnosti (např. 5 %).

Často se setkáváme s interpretací, že p-hodnota v tomto případě měří pravděpodobnost, že $\beta = 0$. Pokud je p-hodnota menší než 0.05, můžeme se setkat s interpretací, že „existuje menší než 5% pravdepodobnost, že $\beta = 0$, a protože je to hodnota realtivně nízká, zamítáme hypotézu, že $\beta = 0$“. Toto není formálně správné tvrzení, p-hodnota *není* pravděpodobnost, že $\beta = 0$. Pro neformální intuici a motivaci tohoto konceptu (tedy že pro nízké p-hodnoty zamítáme H_0) však je možno přijmout i tuto interpretaci.

Shrňme si naše dosavadní poznatky. Chceme-li testovat $H_0 : \beta = 0$, získáme z počítačového výstupu p-hodnotu testu této hypotézy:

1. Jestliže je p-hodnota menší než 5 % (počítač nám tedy dá hodnotu menší než 0.05), potom je t-statistika „dostatečně velká“ na to abychom učinili závěr, že $\beta \neq 0$ (zamítáme nulovou hypotézu na hladině významnosti 5 %).

2. Jestliže je p-hodnota větší než 5 % (počítač nám tedy dá hodnotu větší než 0.05), potom je t-statistika „dostatečně malá“ na to abychom učinili závěr, že $\beta = 0$ (nezamítáme nulovou hypotézu na hladině významnosti 5 %).

Asi se můžeme zeptat, co když nám počítač vrátí hodnotu přesně 0.05. V tomto případě musíme mít na paměti, že mohlo dojít k zakrouhlení na dvě desetinná místa a je

tak nutné podívat se na nezaokrouhlenou hodnotu a rozhodnout jestli skutečná hodnota je větší nebo menší než 0.05 (tedy naše hladina významnosti. Předchozí testy počítali s hladinou významnosti $5\ \%$, pokud však tuto hodnotu nahradíme hodnotou $1\ \%$ nebo $10\ \%$, budeme příslušnou p-hodnotu porovnávat právě s těmito hladinami významnosti a získané závěry o zamítnutí či nezamítnutí nulové hypotézy budou odpovídat této hladině významnosti.

Počítačový software nám obvykle poskytne minimálně následující informace o parametru β (nebo jakémkoli jiném koeficientu, který chceme odhadnout):

- $\widehat{\beta}$, bodový odhad metodou nejmenších čtverců, což je nejlepší odhad skutečné hodnoty β;
- 95% interval spolehlivosti, který nám dává informaci o intervalu, ve kterém s 95% pravděpodobností leží skutečná hodnota parametru β;
- směrodatnou odchylku (standardní chybu) odhadu parametru ($\widehat{\beta}$), s_b, což je měřítko toho, jak přesný náš odhad je (s_b je důležitá součást vztahů pro konstrukci intervalů spolehlivosti a testové statistiky pro testování hypotézy $H_0 : \beta = 0$);
- testovou t-statistiku pro testování $H_0 : \beta = 0$;
- p-hodnotu pro testování $H_0 : \beta = 0$.

2.2.5 Testování hypotéz zahrnující R^2: F-statistika

Většina ekonometrických počítačových programů obsahuje v rámci prezentace výsledků regrese i výsledky testu hypotézy $H_0 : R^2 = 0$. Koeficient determinace udává kvalitu vyrovnání dat modelem. Pokud $R^2 = 0$, nemá vysvětlující proměnná statisticky významnou sílu k vysvětlení chování závisle proměnné Y. Test hypotéza $R^2 = 0$ lze interpretovat jako test, jestli vůbec naše regrese něco vysvětluje. V případě jednoduché regrese je tento test ekvivalentní testu $\beta = 0$. V případě vícenásobné regrese se však jedná o test hypotézy, zda-li jsou všechny regresní koeficienty společně rovny nule. O tom ale bude řeč později.

Procedura samotného testování je standardní. Je vypočítána testová statistika, kterou musíme porovnat s kritickou hodnotou. Alternativně obvykle máme k dispozici příslušnou p-hodnotu, z které přímo dokážeme rozhodnout o nezamítnutí či zamítnutí nulové hypotézy $H_0 : R^2 = 0$ vzhledem k alternativní hypotéze $H_1 : R^2 \neq 0$.

V tomto případě je odpovídající testová statistika tzv. F-statistika a spočítá se jako

$$F = \frac{(N-2)R^2}{1-R^2}.$$

Z příslušných statistických tabulek můžeme získat kritickou hodnotu pro F-rozdělení (Fisherovo-Snedecorovo rozdělení), díky němuž se příslušná statistika jmenuje tak, jak se jmenuje. Naše F-statistika tvoří ale jenom část celé třídy testových statistik, jejichž kritické hodnoty je nutné hledat z F-rozdělení.

Zcela analogicky k testu statistické významnosti parametru vycházíme z toho, že „dostatečně velké“ hodnoty testové statistiky naznačují, že $R^2 \neq 0$, a „malé“ hodnoty

Tabulka 2.1: Jednoduchý regresní model pro data o spotřebě elektřiny

Proměnná	Koef.	Sm. odch.	t-stat.	p-hodnota	95% int. spol.
Konstanta	2.19	1.88	1.16	0.25	[-1.53;5.91]
Výstup	4.79	0.13	36.36	0.00*	[4.53;5.05]

* Hodnota 0.00 znamená, že se jedná o nulovou hodnotu po zaokrouhlení na dvě desetinná místa. Hodnota daná programem je 5.4×10^{-67}.

hovoří proti naši nulové hypotéze, že $R^2 = 0$. Díky p-hodnotě rozhodneme, jestli je hodnota statistiky „velká“ nebo „malá“ a jestli je tedy koeficient determinace, R^2, statisticky významně různý od nuly, nebo tomu tak není. Test provádíme stejně jako při testování statistické významnosti parametru:

1. Jestliže je p-hodnota menší než $5\ \%$ (počítač nám tedy dá hodnotu menší než 0.05), potom je F-statistika „dostatečně velká“ na to abychom učinili závěr, že $R^2 \neq 0$ (zamítáme nulovou hypotézu na hladině významnosti $5\ \%$).

2. Jestliže je p-hodnota větší než $5\ \%$ (počítač nám tedy dá hodnotu větší než 0.05), potom je F-statistika „dostatečně malá“ na to, abychom učinili závěr, že $R^2 = 0$ (nezamítáme nulovou hypotézu na hladině významnosti $5\ \%$).

Hodnota přesně rovna 0.05 je obvykle důsledkem zaokrouhlení a je nutno podívat se nezaokrouhlenou hodnotu. Analogická tvrzení platí i pro testy na jiných hladinách významnosti, stačí nahradit hodnoty $5\ \%$ (0.05) hodnotami $1\ \%$ (0.01) popř. $10\ \%$ (0.10).

Poznatky této podkapitoly shrnuje příklad 2.1.

2.3 Regresní model více vysvětlujících proměnných

Doposud jsme se bavili o jednoduchém regresním modelu, kdy chování veličiny Y bylo vysvětlováno chováním jediné vysvětlující proměnné X (a úrovňovou konstantou, která ale pro svou „konstantnost“ žádné chování vysvětlit nedokáže). Vícenásobná regrese rozšiřuje regresní analýzu na případ, kdy chování jedné veličiny chceme vysvětlit chováním více než jedné vysvětlující proměnné.

Intuice stojící v pozadí je analogická té, se kterou jsme pracovali v případě jednoduché regrese. V případě jednoduché regresi jsme učinili níže uvedené poznatky.

1. Graficky chápeme regresní techniky jako proložení přímky skrze bodový graf pozorovaných hodnot.

2. Regresní koeficienty vyjadřují marginální vliv vysvětlujících proměnných.

3. Odhady metodou nejmenších čtverců odpovídají parametrům nejlepší regresní přímky vyrovnání pozorovaných dat, kdy nejlepší je ve smyslu minimalizace součtu čtverců reziduí.

4. Koeficient determinace, R^2 je měřítkem kvality vyrovnání dat regresním modelem.

Příklad 2.1. *Náklady produkce v odvětví výroby elektřiny*
Předpokládejme opět data z příkladu věnovanému výrobě elektřiny, kdy $Y =$ náklady produkce a $X =$ objem vyrobené elektřiny pro 123 společností. Tabulka 2.1 obsahuje výsledky regresní analýzy v podobě, která v zásadě odpovídá výstupu většiny softwarových balíčků.
Navíc, $R^2 = 0.92$ a p-hodnota testování hypotézy $H_0 : R^2 = 0$ je 0.00. Sloupec *„Koef.“* ukazuje výsledky OLS odhadů parametrů $\widehat{\alpha}$ a $\widehat{\beta}$, které přísluší „Proměnným“ úrovňové konstanty (*„Konstanta“*) a vysvětlující proměnné objemu produkce elektřiny (*„Výstup“*). Sloupec *„Sm. odch“* obsahuje směrodatné odchylky odhadů parametrů $\widehat{\alpha}$ a $\widehat{\beta}$. Sloupec označený jako *„t-stat“* ukazuje hodnoty t-statistiky příslušné testování hypotézy $H_0 : \alpha = 0$ (v řádku proměnné *„Konstanta“*) a $H_0 : \beta = 0$ (v řádku proměnné *„Výstup“*). Sloupec *„p-hodnota“* ukazuje p-hodnoty příslušných testů. V posledním sloupci jsou uvedeny 95% intervaly spolehlivosti pro parametry α a β.
Při písmené prezentaci výsledků našich odhadů obvykle prezentujeme výsledky našich odhadů podobným zposobem jak tomu je v tabulce 2.1, kdy však nesmíme opomenout diskuzi nad ekonomickou interpretací našich výsledků. Text příslušné zprávy či studie by mohl být následující:

> Tabulka 2.1 ukazuje výsledky OLS odhadů založených na datech o nákladech a produkci společností v odvětví energetiky. Zajímá nás prozkoumat otázku, jak volba výstupu firmy ovlivní její náklady produkce. Z tohoto důvodu zvolíme nákaldy produkce jako závisle proměnnou a výstup jako vysvětlující proměnnou. Tabulka nám ukazuje, že odhad parametru výstupu je 4.79, což nám naznačuje, že společnosti s vyššími objemy výstupu mají tendenci mít vyšší náklady produkce. Konkrétně, zvýšení výstupu o jeden milión kilowatthodin má tendenci zvyšovat náklady o 4790000 dolarů.
>
> Vidíme, že mezní efekt výstupu na náklady (což odpovídá mezním nákladům odvětví) je statisticky významný, neboť odpovídající p-hodnota je velmi malá (menší než 1 %). Podíváme-li se na 95% interval spolehlivost, můžeme říci, že zvýšení výstupu o jeden milión kilowatthodin je spojeno se zvýšením nákladů o 4.53 až 5.05 miliónů dolarů (při dané hladine významnosti). Koeficient determinace R^2 potvrzuje, že velikost výstupu vysvětluje velkou část variability nákladů mezi firmami. Přesněji, 92 % variability v nákladech produkce mezi firmami je vysvětleno úrovní výstupu (a jeho variabilitou mezi těmito firmami). Odpovídající p-hodnota F-statistiky je menší než 1 %, což znamená, že R^2 je statisticky významný na hladině významnosti 1 %.

Ve zprávě by samozřejmě měl být obsažen i popis a charaketeristika využívaných dat, přičemž tabulky s odhady by mohly obsahovat i různé charakteristiky kvality vyrovnání (R^2) a jiné statistické testy, je-li potřeba. Blíže se prezentaci empirických odhadů věnuje příloha C.

5. Další statistická analýza odhadů zahrnuje konstrukci intervalů spolehlivosti a testování hypotéz.

Až na drobné výjimky, o kterých bude ještě řeč, odpovídají výše uvedené body i poznatkům, které získáme analýzou modelu vícenásobné regrese. Odlišnosti se týkají interpretace parametrů a vzniká nám rovněž otázka jak volit vhodné vysvětlující proměnné. Ilustrační příklady budou vycházet z dat o cenách domů v kanadském Windsoru.

Příklad 2.2. *Vysvětlení prodejních cen domů*
Výzkum v oblasti mikroekonomie a marketingu bývá zaměřen na problematiku toho, čím je ovlivněna cena nějakého zboží. Jedním z přístupů je výstavba modelu, ve kterém cena zboží závisí na různých charakteristikách tohoto zboží. Náš datový souboru soubor *hprice.gdt* (viz Gretl, záložka *Koop*) z učebnice Koopa [17] obsahuje data o $N = 546$ domech prodaných ve Windsdoru (Kanada). Naší závislou proměnnou, Y, tedy bude prodejní cena domu v kanadských dolarech. Cena domu může být ovlivněna řadou charakteristik daného domu. Datový soubor obsahuje celkem jedenáct možných vysvěltujících proměnných. Zaměříme se v tuto chvíli na čtyři z nich:

- $X_1 =$ celková rozloha domu (ve čtverečních stopách);
- $X_2 =$ počet ložnic;
- $X_3 =$ počet koupelen;
- $X_4 =$ počet pater (kromě přízemí).

2.3.1 OLS odhad modelu vícenásobné regrese

Model vícenásobné regrese s k vysvětlujícími proměnnými můžeme zapsat v podobě

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i,$$

kde opět použijeme index i pro označení jednotlivých pozorování, kdy $i = 1, \ldots, N$. V rámci jednoduché regrese jsme odhadovali dva parametry či koeficienty, α a β. V modelu vícenásobné regrese odhadujeme úrovňovou konstantu, α, a další parametry $\beta_1, \ldots, \beta_k$. Nalezení odhadů těchto parametrů je analogické postupu v rámci jednoduché regrese. To znamená, že definujeme součet čtverců reziduí:

$$SSR = \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{1i} - \widehat{\beta}_2 X_{2i} - \ldots - \widehat{\beta}_k X_{ki} \right)^2.$$

Odhady metodou nejmenších čtverců získáme nalezením takových hodnot odhadů parametrů, tedy $\widehat{\alpha}$ a $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$, pro které bude součet čtverců reziduí, SSR, minimalizován. Jedná se o jednoduchý matematický problém hledání minima funkce, nicméně

výsledný vztah pro příslušný *estimátor*[6] je bez využití maticového zápisu komplikovaný. Každý ekonometrický software nám je ale bez problémů spočítá. Výsledná rovnice odhadu tohoto modelu již není rovnicí přímky, ale jedná se o rovnici roviny (v případě dvou vysvětlujících proměnných) a hůře představitelných nadrovin (vyšších a vyšších dimenzí) pro případy více než dvou vysvětlujících proměnných. Místo regresní přímky tak budeme požívat spíše pojem *rovnice vyrovnání*, *regresní rovnice* nebo prostě *odhadnutý model*.

2.3.2 Statistické aspekty vícenásobné regrese

V případě modelu vícenásobné regrese využijeme identické statistické koncepty jako v případě regresního modelu jediné vysvětlující proměnné. Koeficient determinace R^2 zůstává měřítkem kvality vyrovnání (souladu modelu s daty) a počítá se stejným způsobem. V tomto případě jej ale interpretujeme jako měřítko vysvětlující síly všech vysvětlujících proměnných dohromady. Podobně i F-statistika stále slouží k testování jeho významnosti (testujeme, zda-li $R^2 = 0$), pouze člen $N - 2$ je nahrazen výrazem $N - k - 1$, což je tzv. *počet stupňů volnosti*. Tomu se ale budeme věnovat v následujících kapitolách. Pohledem na příslušnou p-hodnotu testu snadno příslušnou hypotézu ověříme. Pokud zjistéme, že $R^2 \neq 0$, můžeme říct, že *„vysvětlující proměnné v regresi jsou společně schopny vysvětlit chování závisle proměnné“*. V případě, že zjistíme, že $R^2 = 0$, můžeme analogicky říct, že *„vysvětlující proměnné nejsou významné a neposkytují žádné vysvětlení chování závisle proměnné“*.

Vztahy pro výpočet konfidenčních intervalů regresních koeficientů a pro testování jejich statistické významnosti (jsou-li statisticky významně odlišné od nuly) jsou identické s těmi jako v případě jednoduché regrese. Jednotlivé proměnné potřebné pro jejich konstrukci se počítají trochu složitěji (např. směrodatná odchylka odhadu parametrů, s_b). Pokud však získáme příslušné intervaly spolehlivost, interpretujeme stále stejně, tedy pro 95% interval spolehlivosti platí, že se jedná interval, ve kterém si *„můžeme být na* $95\ \%$ *jisti, že skutečná hodnota leží v tomto rozmezí“*. Stejně jako v případě jednoduché regrese interpretujeme i výsledky testu statiské významnosti parametrů na zákaldě příslušných p-hodnot. Získáme-li p-hodnotu menší než 0.05, můžeme říct, že příslušná vysvětlující proměnné je relevantní pro vysvětlení chování závisle proměnné na hladině významnosti $5\ \%$. Tyto statistiky získáme pro každý z koeficientů, tedy α a $\beta_1, \ldots, \beta_k$.

2.3.3 Interpretace OLS odhadů v modelu vícenásobné regrese

Jak tedy interpretovat výsledky odhadů modelu vícenásobné regrese? Velmi podobně jako v případě modelu jednoduché regrese, i když zde existují důležité odlišnosti. Než se k tomu dostaneme, ujasněme si používané značení. Budeme-li hovořit o obecné vlastnosti, která platí pro všechny parametry, budeme je značit jako β_j (tzn. koeficient pro j-tou vysvětlující proměnnou, kde j celé číslo mezi 1 a k). Budeme-li chtít hovořit o specifickém parametru, přiřadíme mu konkrétní hodnotu indexu j, tedy např. pro β_2 je $j = 2$ a jedná se o koeficient u druhé vysvětlující proměnné.

[6] Estimátor je statistika respektive funkce hodnot pozorovaného výběru dat, kdy pro jejich konkrétní hodnoty pozorování získáme příslušné odhady.

V jednoduché regresi byl parametr sklonu regresní přímky, β, interpretován jako marginální (mezní) vliv vysvětlující proměnné (tzn., že vyjadřoval vliv, jaký má změna X na Y). V modelu vícenásobné regrese lze β_j interpretovat jako mezní vliv, ale v trochu odlišném smyslu. Konkrétně, β_j vyjadřuje marginální vliv vysvěltující proměnné X_j na Y, *pokud ostatní vysvětlující proměnné zůstávají neměnné*. Toto je důležité si uvědomit pro korektní interpretaci výsledků regresní analýzy. Prakticky si interpretaci můžeme ilustrovat na příkladu s cenami domů (příklad 2.3).

Příklad 2.3. *Vysvětlení prodejních cen domů (pokračování příkladu 2.2)*
Tabulka 2.2 obsahuje výsledky regrese závisle proměnné Y (prodejní cena domu) na vysvětlujících proměnných X_1 (celková rozloha domu ve čtverečních stopách), X_2 (počet ložnic), X_3 (počet koupelen) a X_4 (počet pater bez přízemí). V modelu je rovněž uvažována úrovňová konstanta. Koeficient determinace $R^2 = 0.54$ a p-hodnota testu hypotézy $H_0 : R^2 = 0$ je 0.00.
První sloupec tabulky popisuje jednotlivé vysvětlující proměnné. V našem příkladu máme tedy čtyři vysvěltující proměnné a úrovňovou konstantu. Každý řádek obsahuje tutéž informaci, jako tomu bylo v případě jednoduché regrese (viz tabulka 2.1), tedy postupně OLS odhady parametrů, t-statistiku, p-hodnotu testu zda $\beta_j = 0$ a 95% interval spolehlivosti parametru.
Výsledky můžeme prezentovat alternativně v podobě zapsání odhadnutého regresního modelu, kdy pod odhady jednotlivých koeficientů jsou uváděny t-statistiky, směrodatné odchylky odhadu parametrů nebo příslušné p-hodnoty (co je obsahem závorek musí být vždy zdůrazněno, protože všechny tři varianty jsou obvyklé):

$$\widehat{Y} = \underset{(-1.11)}{-4009.55} + \underset{(14.70)}{5.43} X_1 + \underset{(2.33)}{2824.61} X_2 + \underset{(9.86)}{17105.17} X_3 + \underset{(7.57)}{7634.90} X_4.$$

Interpretovat výsledky opět můžeme různými zposoby. Jako příklad předpokládejme koeficient u první vysvětlující proměnné, což je rozloha domu. Odhad tohoto parametru je $\widehat{\beta_1} = 5.43$. Na tomto základě můžeme říct:

1. Dodatečná stopa čtvereční rozlohy domu má tendenci zvýšit cenu domu o dalších 5.43 dolarů, *ceteris paribus*, tedy za jinak nezměněných podmínek (zejména, žeostatní vysvětlující proměnné zůstanou zachovány).

2. Budeme-li předpokládat domy se stejným počtem ložnic, koupelen a pater, potom každá čtvereční stopa rozlohy domu navíc bude mít tendenci zvýšit cenu těchto domů o dodatečných 5.43 dolarů.

Obě dvě interpretace jsou rovnocenné a určitě bychom našli další varianty toho, jak podat obdobnou informaci na zákaldě našich odhadů. Abychom nehovořili stále o „tendenci“ můžeme klidně použít i volnější spojení „v průměru“.

V příkladu 2.3 je dobré rozvést motivaci pro závěrečnou interpretaci výsledků odhadu parametru první vysvětlující proměnné. Nemůžeme totiž tvrdit, že *„domy s větší rozlohou jsou mnohem hodnotnější“*, protože to není náš případ. Některé pěkné domy

Tabulka 2.2: Regresní model pro data o cenách domů.

Proměnná	Koef.	Sm. odch.	t-stat.	p-hodnota	95% int. spol.
Konstanta	-4009.55	3603.11	-1.11	0.27	[-11087.3;3068.25]
Rozloha	5.43	0.37	14.70	0.00	[4.70;6.15]
Počet					
ložnic	2824.61	1214.81	2.33	0.02	[438.30;5210.93]
koupelen	17105.17	1734.43	9.86	0.00	[13698.12;20512.22]
pater	7634.90	1007.97	7.57	0.00	[5654.87;9614.92]

s malou rozlohou mohou mít vyšší cenu než chudší domy s velkou rozlohou. Můžeme však říct, že *při uvažování domů s různou rozlohou, ale srovnatelných, co do jiných charakteristik, jsou domy s vyšší rozlohou v průměru hodnotnější než domy s rozlohou menší*". Důležité je při interpretaci zdůraznit ono „srovnatelné, co do jiných charakteristik". V případě jednoduché regrese byla jediná vysvětlující proměnná a tyto starosti jsme tak nemuseli řešit. V případě jedoduché regrese jsme ohli konstatovat, že *„β měří vliv X na Y"*, v případě vícenásobné regrese můžeme říct, že *„β_j měří vliv X_j na Y, při neměnných ostatních vysvětlujících proměnných"*.

V příkladu 2.3 je ukázána interpretace odhadu parametru první vysvěltující proměnné, $\widehat{\beta}_1$. Analogicky můžeme interpretovat (na základě tabulky 2.2) i ostatní parametry. Protože $\widehat{\beta}_2 = 2824.61$, můžeme říct, že *„při úvahách o srovnatelných domech (např. o rozloze* 5000 *čtverečních stop, se dvěma koupelnami a dvěma podlažími), budou domy se třemi ložnicemi mít tendenci být o* 2824.61 *dolarů dražší než domy se dvěma ložnicemi"*. Na základě $\widehat{\beta_3} = 17105.17$ můžeme říct, že *„domy s extra koupelnou navíc mají o* 17105.17 *dolarů větší cenu, ceteris paribus"*. Odhad $\widehat{\beta_4} = 7634.90$ nám napovídá, že *„pro srovnatelné domy ve všech ostatních aspektech znamená každé patro navíc nárůst hodnoty domu o* 7634.90*"*.

Statistická analýza výsledků odhadů regresních koeficientů je opět podobná té pro jednoduchou regresi. Protože p-hodnoty pro všechny vysvětlující proměnné (s výjimkou úrovňové konstanty) jsou menší než 0.05, můžeme říct, že *„koeficienty β_1, β_2, β_3 a β_4 jsou statisticky významné na* 5% *hladině významnosti"*. Ekvivalentně samozřejmě můžeme říct, že *„zamítáme jednotlivé hypotézy, že tyto čtyři parametry jsou nulové na hladiné významnosti* 5 *%"*. Jako další příklad uvažujme 95% interval spolehlivosti pro parametr β_2, který je $[438.2761; 5210.932]$. Tuto informaci můžeme prezentovat způsobem, že *„náš bodový odhad říká, že marginální vliv počtu ložnic na cenu domů je* 2842.61 *dolarů, což je však jen nejlepší odhad. Ve skutečnosti nám* 95% *interval spolehlivosti naznačuje, že si můžeme být s touto pravděpodobností jisti, že tento mezní vliv se pohybuje mezi* 438.28 *dolary a* 5210.92 *dolary"*.

Test hypotézy o nulovosti koeficientu determinace vrací p-hodnotu menší než 0.05. To znamená, že všechny vysvětlující proměnné (i s úrovňovou konstantou) mají společně statisticky významnou vysvětlující sílu pro vysvětlení ceny domů. Variabilita v rozloze domů, počtu ložnic, počtu koupelen a počtu pater vysvětluje 54% variability v cenách domů.

2.3.4 Jaké vysvětlující proměnné zvolit?

V případě vícenásobné regrese vyvstává otázka, jaké proměnné bychom měli vybrat, tedy čím se při jejich výběru řídit? Na jedné straně zde je snaha zahrnout co možná nejvíce vysvětlujících proměnných, které by napomohli vysvětlit chování závisle proměnné. Na druhé straně, zahrnutí irelevantních proměnných (tzn. statisticky nevýznamných) snižuje statistickou významnost i ostatních vysvětlujících proměnných. Z tohoto pohledu bychom se měli naopak snažit o zařazení co možná nejmenšícho počtu vysvěltujících proměnných. V empirické praxi musíme tyto dva protichůdné požadavky uvést do souladu. K tomu nám pomáhá testování hypotéz. Pokud nám test statistické významnosti parametru říká, že se jedná o nevýznamný parametr, měli bychom tuto proměnnou z naší regrese vypustit.

Začněme tedy s diskuzí nad tím, proč je důležité neopomenout zahrnout důležité vysvětlující proměnné. Pro ilustraci se vraťme k příkladu s cenami domů a odpovídajícími charakteristikami těchto domů. Provedli jsme regresi se čtyřmi vysvětlujícími proměnnými (viz tabulka 2.2). Můžeme tyto výsledky porovnat s výsledky jednoduché regrese ceny domů (Y) na počet ložnic (X). OLS odhady nám dávají následující odhad regresní přímky:

$$\widehat{Y} = 28773.43 + 13269.98X.$$

Protože $\widehat{\beta} = 13269.98$, můžeme říct, že *„mezní vliv počtu ložnic na cenu domů je* 13269.98 *dolarů“* popřípadě, že *„domy s dodatečnou ložnicí mají tendencí stát o* 13269.98 *dolarů více“*. Porovnejme si tyto výsledky s výsledky modelu vícenásobné regrese. V případě jednoduché regrese jsme jednoduše opustili podmínku *ceteris paribus* využívanou v intepretaci výsledku vícenásobné regrese, tedy *„pokud uvažujeme porovnatelné domy (např. o rozloze* 5000 *čtverečních stop, se dvěma koupelnami a jedním patrem)“*.

V případě jednoduché regrese je koeficient u proměnné *„počet ložnic“* v modelu jednoduché regrese výrazně vyšší (v případě modelu vícenásobné regrese byl odhad $\widehat{\beta}_2 = 2824.61$). Čím je to způsobeno? Představme si, že máme kamaráda, který si chce ve Windsdoru postavit přistavit další ložnici, a požádá nás o radu, jak mu tato investice zvýší hodnotu domu. Použili bychom pro zodpovězení této otázky výsledky jednoduché nebo vícenásobné regrese?

Jednoduchý regresní model obsahuje data o cenách domů a počtu ložnic. Můžeme z nich usuzovat, že domy s větším počtem ložnic jsou hodnotnější (dům se třemi ložnicemi je o 13269.98 dolarů dražší než dům se dvěma ložnicemi). To ale neznamená, že přidání ložnice zvýší cenu domu o tuto částku. Existuje zde totiž celá řada dalších faktorů, které ovlivňují cenu domu. Tyto faktory mohou být navzájem korelovány, tedy např. velké domy mohou mít v průměru více ložnic, více koupelen, více pater a velkou celkovou rozlohu. Pro prozkoumání této možnosti není od věci podívat se na korelační matici těchto vysvětlujících proměnných. V případě modelu z tabulky 2.2 máme celkem pět proměnných (jednu závisle proměnnou a čtyři vysvětlující). Korelace nám měří těsnost vztahu mezi dvěma proměnnými. Pro pět proměnných tedy můžeme mít deset možných korelací (cena a počet ložnic, cena počet koupelen, atd.). Příslušnou korelační matici máme obsaženou v tabulce 2.3.

Korelaci mezi odpovídajícími si proměnnými najdeme na průniku příslušného řádku a sloupce (jedná se o symetrickou matici, proto ta prázdná místa v horní části, a

Tabulka 2.3: Korelační matice dat cen domů

	Cena	Rozloha	Počet		
			ložnic	koupelen	pater
Cena	1				
Rozloha	0.54	1			
Počet ložnic	0.37	0.15	1		
Počet koupelen	0.52	0.19	0.37	1	
Počet pater	0.42	0.08	0.41	0.32	1

s jedničkami na hlavní diagonále, neboť každá proměnná je perfektně korelována se sebou samou). Hodnota 0.32 udává korelaci mezi počtem koupelen a počtem pater.

Všechny prvky korelační matice jsou kladné, tudíž každá proměnná je pozitivně korelována s ostatními (např. korelace mezi počtem ložnic a počtem koupelen je 0.37, což znamená, že domy s větším počtem ložnic mají tendenci mít i více koupelen). V takovémto případě nemůže jednoduchá regrese rozlišit vliv jednotlivých proměnných na cenu domu. V případě jednoduché regrese tak zjištění, že domy s větším počtem ložnic stojí více neznamená, že přidání další ložnice povede zvýší cenu domu. Kupující si mohou více cenit koupelen nebo celkové rozlohy domu než ložnic. Předpokládejme, že jsou to právě koupelny, které kupující dokáží ocenit. Domy s větším počtem koupelen mají tendenci mít více ložnic. Jednoduchý regresní model se podívá na cenu domů a počet ložnic a vidí, že ty domy, které mají více ložnic, mají i větší cenu. Nedokáže však objevit skutečnost, že je to právě počet koupelen, který zvyšuje cenu domu. Pokud bychom našemu kamarádovi poradili, že nová ložnice zvýší v průměru cenu jeho domu o 13269.98 dolarů, byla by to zavádející informace. V jednoduchém regresním modelu jsme tak zanedbali důležité vysvětlující proměnno jako rozloha, počet koupelen nebo počet pater. Jednoduchá regrese zkombinuje příspěvek všech těchto faktorů k vysvětlení ceny domů a přiřadí ho té jediné vysvětlující proměnné, které může, tedy počtu ložnic. Tudíž je $\widehat{\beta}$ obrovské.

Oproti tomu, vícenásobná regrese nám dokáže rozlišit příspěvky jednotlivých vysvětlujících proměných k vysvětlení závisle proměnné. Proto hovoříme o vysvětlujících proměnných jako o *kontrolních* či *řídících* proměnných, protože jimi kontrolujeme (řídíme) další možné vlivy, které mohou působit na vysvětlovanou proměnnou. Hodnota $\widehat{\beta}_2 = 2824.61$ se zdá být vhodnějším měřítkem vlivu dodatečné ložnice na cenu domu. Tímto údajem našeho kamaráda určitě nepřivedeme do případného nerozumného budování dalších ložnic. V tomto případě si můžeme být jistější, že příspěvek počtu ložnic k vysvětlení ceny domu je odpovídající a není zkreslen nepřítomností dalších důležitých vysvětlujících proměnných.

Tato problematika se týká statistického problému zvaného *zkreslení při opomenutí důležité vysvětlující proměnné (omitted variable bias)*. Formálně se k němu vrátíme v kapitole 4. Neformálně ale můžeme říct, že pokud opomeneme vysvětlující proměnné, která by v regresi být měly a tyto opomenuté proměnné jsou korelovány s těmi, které v regresi zahrnuté jsou, potom odhady koeficientů těchto proměnných budou chybné.

Předchozí příklad jednoduché regrese se zahrnutím pouze počtu ložnic je toho krásným příkladem a vidíme zde i velikost takovéhoto zkresleno (více než 10000 dolarů).

Do regrese bychom měli zkoušet zahrnout všechny vysvětlující proměnné, které by mohly ovlivňovat závisle proměnnou. Ceny domů závisí na řadě proměnných, které nejsou obsaženy v dostupných datech (např. stav údžby daného domu, příjemnost sousedů, má-li dům dřevěnou podlahu, kvalita zahrady apod.). V praxi je nemožné získat data o všech možných vysvětlujících proměnných (např. příjemnosti sousedů). Vždycky zde bude určité zkreslení při nezahrnutí důležité vysvětlující proměnné. S tím nic nenaděláme, kromě toho, že budeme doufat, že tyto opomenuté proměnné mají minimální vysvětlující sílu a nejsou korelovány s proměnnými, které do našich analýz zahrnujeme.

Tímto jsme se částečně vypořádali s problémem zahrnutí co možná nejvíce vysvětlujících proměnných. Na druhé straně zde máme požadavek mít vysvětlujících proměnných co nejméně. Lze totiž ukázat, že zahrnutí irelevantních proměnných vede ke zkreslení přesnosti odhadů u všech koeficientů (i u těch, které irelevantní nejsou). Tato nepřesnost se projeví v širokých intervalech spolehlivosti a velkých p-hodnotách.

Jak tedy balancovat mezi výhodou zahrnutí hodně proměnných (a vypořádání se s problémem opomenutí důležité proměnné) a náklady spojenými s možností zahrnutí irelevantní proměnné (tzn. snížení přesnosti odhadu)? Obvyklý postup je ten, začít s co možná největším počtem vysvětlujícíh proměnných a postupně vyhazovat ty proměnné, které nejsou statisticky významné a provést regresi s menším počtem proměnných. Výsledné regrese by měly obsahovat pokud možno jen proměnné se statisticky významnými parametry (výjimkou je úrovňová konstanta, která je důležitá pro korektní interpretaci koeficientu determinace, o čemž ještě bude řeč) případně proměnnou (proměnné), u kterých nás z povahy zkoumaného řešeného ekonomického problému zajímá, jestli je příslušný parametr statisticky významný. Kvalitu jednotlivých modelů (z hlediska souladu modelu s daty) pak můžeme porovnat pomocí koeficientu determinace.

2.3.5 Multikolinearita

Multikolinearite je jeden z dalších problémů, kterého se je třeba vyvarovat. Jedná se o problém, který nastává v případě, kdy jsou některé nebo všechny proměnné velmi silně navzájem korelovány. V takovém případě má regresní model problém určit, která z vysvětlujících proměnných ovlivňuje závisle proměnnou. Problém multikolinearity se projevuje skrze nízké hodnoty t-statistik a vysoké p-hodnoty. Intervaly spolehlivosti dotčených parametrů jsou velmi široké a dospíváme obvykle k závěru, že koeficienty jsou nevýznamné a odpovídající proměnné by tak měly být z regrese vyjmuty. V extrémním případě můžeme dojít k závěru, že všechny parametry jsou statisticky nevýznamné a koeficient determinace R^2 je přitom vcelku vysoký a statisticky významný. To intuitivně znamená, že všechny vysvětlující proměnné společně dostatečně vysvětlují chování závisle proměnné, nicméně problém multikolinearity regresi neumožňuje rozhodnout, která z vysvětlujících proměnných chování závisle proměnné vlastně vysvětluje.

Pro řešení problému nenaděláme nic více, než že některou ze silně korelovaných proměnných z regrese vypustíme. V některých případech to nemusí být samozřejmě

žádoucí, např. v příkladu s cenami domů počet ložnic a počet koupelen je korelován a multikolinearita by tak mohla být problém. Nicméně, zdravý rozum velí, že obě proměnné by mohly chování cen domů vysvětlit. Příklad 2.4 ukazuje situaci, kdy problém multikolinearity existuje a jak jej lze vyřešit vypuštěním příslušných vysvětlujících proměnných.

Příklad 2.4. *Vliv úrokových sazeb na směnný kurz*
Předpokládejme, že nás zajímá analýza vlivu politiky stanovení úrokových sazeb na směnný kurz. Jeden ze způsobů by byly volba směnného kurzu (např. libry vzhledem k dolaru) jako závisle proměnné a provedení regrese na úrokovou míru. Nicméně, k dispozici máme celou řadu úrokových měr, které lze jako vysvětlující proměnné využít (záleží na směnném kurzu, který chceme sledovat, ale v úvahu případá obecně např. krátkodobá tříměsíční úroková míra (PRIBOR, LIBOR, EURIBOR), dlouhodobá roční úroková míra, úroková míra pokladničních poukázek, bankovní prime rate apod.). Tyto úrokové míry jsou si, co do dynamiky svého chování, velmi podobné a budou tedy silně korelovány. Pokud bychom do regrese zahrnuli více než jednu z nich, narazíme na problém s multikolinearitou. Protože však tyto úrokové míry popisují v zásadě tentýž fenomén, může nám stačit použít jen jednu z nich, čímž se vyhneme problému multikolinearity, a to bez toho, aniž by tím utrpěla naše snaha vysvětlit chování směnného kurzu.

Je nezbytně důležité uvědomit si, že multikolinearita se týká korelace mezi vysvětlujícími proměnnými, a ne korelace mezi vysvětlující a závisle proměnnou, kde je korelace naopak přímo žádoucí. Aby se jednalo o skutečný problém, musí se hodnoty korelačních koeficientů pohybovat v blízkosti extrémních hodnot 1 nebo -1. Řekněme, že korelace nad 0.9 už by měla zvýšit naši pozornost pokud jde o volbu a vztah mezi vysvětlujícími proměnnými. V případě cen domů se korelace mezi vysvětlujícími proměnnými pohybují v rozmezí hodnot 0.3 až 0.4. Tato mírná korelace k multikolinearitě nevede, neboť všechny koeficienty jsou statisticky významně různé od nuly.

2.3.6 Vícenásobná regrese s umělými proměnnými

V našich příkladech byly používané proměnné kvantitativního charakteru. Náklady elektrárenských společností, produkce elektřiny, ceny domů, jejich rozloha, počet koupelen, ložnic atd., to vše mělo své měrné ednotky, popřípadě se jednalo o počet něčeho (jednotky odpovídaly fakticky „počtu kusů“). Mnohá data, která ekonomové používají mají kvalitativní charakter. v aplikacích ekonomie práce jsou data získávána z dotazníkových šetření a některá z otázek může znít: *„Jste členem odborové organizace?“*. Odpověď může být „ano“ nebo „ne“, tedy jedná se o kvalitativní odpovědi. Stejně tak je možno dotazovat se na pohlaví respondenta, kdy odpověď „muž“ nebo „žena“ je opět kvalitativního ražení. Umělé proměnné představují způsob, jak převést kvalitativní vysvětlující proměnné do podoby kvantitativní vysvětlující proměnné. Po tomto převodu již lze standardně využít model vícenásobné regrese. Formálně je umělá proměnná taková proměnná, která nabývá pouze dvou hodnot, a to 0 nebo 1. Jako ilustrace může sloužit příklad 2.5.

Příklad 2.5. *Vysvětlení prodejních cen domů (pokračování příkladu 2.3)*
Cenu domů jsme se v předchozích částech příkladu snažili vysvětlit pomocí proměnných, které měly kvantitativní charakter (rozloha, počet ložnic atd.). Existují však faktory, které ovlivňují cenu domu, nicméně nemají přímo kvantitativní charakter. Příkladem může být existence příjezdové cesty, vybavení klimatizací, přítomnost místnosti pro zábavu a relaxaci(recreation room), zabudovaný sklep či centrální vytápění. Všechny tyto proměnné jsou kvalitativního typu „ano/ne“, tedy např. „ano“ pro případ, kdy dům má příjezdovou cestu a „ne“, pokud dům tuto cestu nemá.
Abychom tyto proměnné mohli zakomponovat do naši regresní analýzy, musíme je transformovat do podoby umělých proměnných, které nabývají hodnoty 1 („ano“) nebo 0 („ne“). Použijme pro označení těchto umělých (dummy) proměnných písmenko D. Pak můžeme zavést jejich následující specifikaci:

- $D_1 = 1$ pokud má dům příjezdovou cestu (jinak $D_1 = 0$).
- $D_2 = 1$ pokud má dům místnost pro zábavu a relaxaci (jinak $D_2 = 0$).
- $D_3 = 1$ pokud má dům sklep (jinak $D_3 = 0$).
- $D_4 = 1$ pokud má dům centrální vytápění (jinak $D_4 = 0$).
- $D_5 = 1$ pokud má dům klimatizaci (jinak $D_5 = 0$).

Stejně jako u ostatních proměnných, můžeme proměnným dodat další index označující konkrétní pozorování. Pokud tedy má i-tý dům příjezdovou cestu, sklep a centrální vytápění, ovšem nedisponuje klimatizací a místnsotí pro relaxaci, budou námi definované umělé proměnné příslušné tomuto pozorování odpovídat hodnotám $D_{1i=1}$, $D_{2i=0}$, $D_{3i=1}$, $D_{4i=1}$ a $D_{5i=0}$.
Díky takto definovaným umělým proměnným provedeme regresi standardním způsobem a můžeme využít veškerou nám známou teorii a intuici, jako by se jednalo o „standardní“ proměnnou. Umělé proměnné tedy s sebou nepřinášejí žádné nové statistické speciality a stále zde budeme získávat OLS odhady parametrů jim příslušným, jejich konfidenční intervaly apod. Nicméně, interpretace regresních koeficientů příslušných umělým proměnným má svá drobná specifik, kterým se je třeba podrobněji věnovat.

Předpokládejme jednoduchý regresní model s jedinou umělou proměnnou, D:

$$Y_i = \alpha + \beta D_i + \epsilon_i,$$

pro $i = 1, \ldots, N$ pozorování. Odhadem metodou nejmenších čtverců získáme $\widehat{\alpha}$ a $\widehat{\beta}$ a regresní přímku vyrovnání můžeme zapsat jako

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} D_i.$$

Protože D_i nabývá hodnot 0 nebo 1, mohou vyrovnané hodnoty nabývat hodnoty $\widehat{Y}_i =$

$\widehat{\alpha}$ nebo $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}$. Jako ukázka regrese s umělými proměnnými může sloužit příklad s cenami domů, konkrétně 2.6 a 2.7.

Příklad 2.6. *Vysvětlení prodejních cen domů (pokračování příkladu 2.5)*
Provedeme-li regresi ceny domu, Y na umělou proměnnou D, označující, jestli dům má nebo nemá klimatizaci, získáme s využitím dat o cenách domů následující odhad regresní přímky:

$$\widehat{Y}_i = 59884.85 + 25995.74 D_i.$$

Jak tyto výsledky interpretovat? Můžeme samozřejmě říct, že $\widehat{\beta}$ je měřítkem toho, o kolik se nám v průměru změní závisle proměnná, pokud se vysvětlující proměnná změní o jednotku. Ovšem v případě umělé proměnné „změna o jednotku“ odpovídá v našem příkladu změně charakteristiky domu „bez klimatizace“ na „s klimatizací“. Můžeme tedy říct, že domy s klimatizací mají tendenci mít o 25996 dolarů větší hodnotu něž domy bez klimatizace.
Podobně můžeme výsledky interpretovat tak, že v případě domů bez klimatizace je $D_i = 0$ a tedy $\widehat{Y}_i = 59884.85$. Náš model tedy ukazuje, že domy bez klimatizace mají v průměru hodnotu 59885 dolarů. V případě domů s klimatizací je $D_i = 1$ a regresní model nám říká, že $\widehat{Y}_i = 59884.85 + 25995.74 = 85880.59$. Domy s klimatizací tak v průměru stojí 85881 dolarů. To je tedy další atraktivní způsob prezentace našich výsledků.
Poznamenejme, že pokud bychom spočítali průměrnou cenu domů s klimatizací, získali bychom částku 85881 dolarů. Průměrná cena domů bez klimatizace by byla 59885 dolarů. To odpovídá výsledkům námi specifikovaného regresního modelu.
Předchozí model však může být zatížen problémem zkreslení, díky vynechané podstatné vysvětlující proměnné (tzv. *omitted variable bias*). Předchozí výsledky bychom asi těžko mohli použít k tvrzení, že „vybavení domu klimatizací zvýší jeho cenu v průměru z 59885 dolarů na 85881 dolarů“. Klimatizace může stát maximálně několik tisíc dolarů, a tak je předchozí tvrzení velmi chabé. V praxi je třeba vyzkoušet různé specifikace regresního modelu, tedy zkoušet zahrnout různé vysvětlujících proměnných.

V praxi se samozřejmě setkáme s využitím různých typů vysvětlujících proměnných. Nejjednodušší z nich může být případ s jednou umělou vysvětlující proměnnou (D) a jednou standardní, neumělou, vysvětlující proměnnou (X):

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_i.$$

Interpretaci výsledků takovéto regrese ukazuje příklad 2.8.

V příkladu 2.8 fakticky odhadujeme dvě regresní přímky se stejným sklonem, ale s různou úrovňovou konstantou. To je zcela běžné v případě modelu vícenásobné regrese, kde kombinujeme standardní i umělé vysvětlující proměnné. Umělé proměnné nám definují různé kategorie pozorování (např. různé skupiny domů) a tyto kategorie mají v rámci svých regresních rovnic různé úrovňové konstanty. Všechny kategorie

Příklad 2.7. *Vysvětlení prodejních cen domů (pokračování příkladu 2.6)*
Jak tedy interpretovat výsledky při zahrnutí více umělých proměnných? Příslušný regresní model můžeme zapsat jako

$$Y_i = \alpha + \beta_1 D_{1i} + \ldots + \beta_k D_{ki} + \epsilon_i,$$

pro $i = 1, \ldots, N$ pozorování. Odhad metodou nejmenších čtverců lze provést standardním způsobem, stejně jako další statistické analýzy či testy.
Předpokládejme příklad s cenami domů, kde máme dvě umělé proměnné. Konkrétně, $D_1 = 1$ pokud dům má příjezdovou cestu ($D_1 = 0$ pokud tuto cestu nemá) a $D_2 = 1$, pokud má dům místnost pro zábavu a relaxaci ($D_2 = 0$ pokud tuto místnost nemá). Tato klasifikace nám rozdělí domy do čtyř skupin, jak uvidíme v rámci interpretace odhadnutého regresního modelu:

$$\widehat{Y}_i = 47099.08 + 21159.91 D_{1i} + 16023.69 D_{21}.$$

Pro různé kombinace možných hodnot umělých vysvětlujících proměnných získáme vyrovnané hodnoty pro následující čtyři kategorie domů (odhady parametrů jsou zaokrouhleny na celá čísla):

1. Domy s příjezdovou cestou a místností pro zábavu a relaxaci ($D_1 = 1$ a $D_2 = 1$) mají v průměru hodnotu $\widehat{Y}_i = 47099 + 21160 + 16024 = 84283$ dolarů.

2. Domy s příjezdovou cestou ale bez místností pro zábavu a relaxaci ($D_1 = 1$ a $D_2 = 0$) mají v průměru hodnotu $\widehat{Y}_i = 47099 + 21160 = 68259$ dolarů.

3. Domy bez příjezdové cesty ale s místností pro zábavu a relaxaci ($D_1 = 0$ a $D_2 = 0$) mají v průměru hodnotu $\widehat{Y}_i = 47099 + 16024 = 63123$ dolarů.

4. Domy bez příjezdové cesty a bez místností pro zábavu a relaxaci ($D_1 = 0$ a $D_2 = 0$) mají v průměru hodnotu $\widehat{Y}_i = 47099$ dolarů.

však mají stejné parametry sklonu, tedy pro všecha pozorování existují stejný marginální vliv X na Y. Pokud bychom v rámci kategorií chtěli uvažovat i možnost různých marginálních vlivů, potom je třeba vytvořit novou vysvětlující proměnnou jako kombinaci umělé proměnné a příslušné neumělé proměnné. Co tím máme na mysli je nejlépe představitelné na příkladu následujícího regresního modelu:

$$Y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \beta_3 Z_i + \epsilon_i,$$

kde D je umělá proměnná a X je naše neumělá či standardní vysvětlující proměnná. Do rovnice jsme přidali novou proměnnou Z, kterou definujeme jako součin D a X, tedy $Z = DX$.

Jak interpretovat výsledky odhadnutého modelu? Odpověď je snadná, pokud si uvědomíme, že Z_i nabývá hodnoty 0 pro pozorování, kde $D_i = 0$, nebo nabývá hodnoty X_i pro pozorování, kde $D_i = 1$. Stejně jako v příkladu 2.8 budeme uvažovat dvě

Příklad 2.8. *Vysvětlení prodejních cen domů (pokračování příkladu 2.7)*
Pokud provedeme regresi cen domů, Y, na umělou proměnnou označující vybavení klimatizací, D, a proměnnou udávající rozlohu domu, X, získáme odhady: $\widehat{\alpha} = 32693$, $\widehat{\beta_1} = 20175$ a $\widehat{\beta_2} = 5.638$. Umělá proměnná nabývá hodnot 0 nebo 1 a v předchozí části příkladu (2.7) jsme si ukázali, jak lze interpretovat vyrovnané hodnoty pro odpovídajícím způsobem rozdělené kategorie domů. V tomto případě opět získáváme regresní přímky

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 + \widehat{\beta}_2 X_i = 52868 + 5.638 X_i,$$

pro $D_i = 1$ (dům s klimatizací) a

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_2 X_i = 32693 + 5.638 X_i,$$

pro $D_i = 0$ (dům bez klimatizace). Jinými slovy, máme dvě různé regresní přímky lišící se v tom, jestli hovoříme o domu s klimatizací nebo bez ní. V příkladu s jedinou vysvětlující umělou proměnnou (2.6) jsme zkoumali situaci, kdy se průměrná cena lišila jen v závislosti na tom, jestli dům má nebo nemá klimatizaci. V tomto případě ale říkáme, že existují odlišné regresní přímky a do ceny domu nám vstoupil i faktor jeho rozlohy. Rozdíl mezi domy s klimatizací a bez ní není již 26000 dolarů, ale jen něco málo přes 20000 dolarů, což nám říka $\widehat{\beta}_1$.

vyrovnané regresní přímky pro jednotlivce (resp. pozorování), kde $D_i = 0$ a $D_i = 1$. Jedná se o regresní přímky, protože máme stále jen jednu neumělou vysvětlující proměnnou:

- Pokud je $D_i = 0$, potom $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_2 X_i$,
- Pokud je $D_i = 0$, potom $\widehat{Y}_i = \left(\widehat{\alpha} + \widehat{\beta}_1\right) + \left(\widehat{\beta}_2 + \widehat{\beta}_3\right) X_i$.

Jinými slovy, máme dvě odlišné regresní přímky odpovídající kategoriím pozorování či jednotlivců, kdy $D = 0$ a $D = 1$, které mají odlišnou úrovňovou konstantu i odlišný sklon. Důležitou implikací tedy je to, že marginální vliv X na Y se liší v závislosti na tom, jestli je $D_i = 0$ nebo $D_i = 1$. Pěkná ilustrace je obsažena v příkladu 2.9.

2.3.7 Co když je vysvětlovaná proměnná umělá?

Doposud jsme uvažovali případy, kdy umělá proměnná byla na straně vysvětlujících proměnných. Můžeme se setkat i se situací, kdy potřebujeme umělou proměnnou na straně vysvětlované proměnné. Aplikací je celá řada. Může nás zajímat analýza volby dopravního prostředku na cestu do práce, tedy jestli to bude auto nebo hromadná doprava. To znamená, že pracujeme s daty z dotazníkových šetření, kde se ptáme respondentů na jejich zvyklosti ohledně volby dopravního prostředku, kdy potenciální vysvětlující proměnné může být vzdálenost z bydliště do práce nebo doba jízdy, příjem apod. Závisle proměnná bude kvalitativního charakteru, protože každý respondent

Příklad 2.9. *Vysvětlení prodejních cen domů (pokračování příkladu 2.8)*
Pokud provedeme regresi cen domů, Y, na umělou proměnnou označující vybavení klimatizací, D, proměnnou udávající rozlohu domu, X, a novou proměnnou $Z = DX$, získáme odhady: $\widehat{\alpha} = 35684$, $\widehat{\beta_1} = 7613$ a $\widehat{\beta_2} = 5.02$ a $\widehat{\beta_3} = 2.25$. Marginální vliv rozlohy domu je 7.27 dolarů pro domy s klimatizací (tzn., že u domů s klimatizací má každá čtvereční stopa navíc zvýší v průměru cenu domu o tuto částku) a 5.02 dolarů pro domy bez klimatizace. Odpovídající p-hodnota pro test statistické významnosti parametru β_3 je 0.02, tudíž i tento marginální vliv je statisticky významný. Toto zjištění znamená, že růst rozlohy domu má tendenci více zhodnocovat dům vybavený klimatizací než dům bez klimatizace. Samotný parametr β_1 je statisticky nevýznamný (p-hodnota je 0.17), což znamená, že stejně velké domy s klimatizací a bez ní se odlišují jen v závislosti na své rozloze a není zde přítomna žádná dodatečná „fixní prémie“ za toto vybavení.

odpověděl buď *„ano, do práce jezdím automobilem“* nebo *„ne, do práce nejezdím automobilem“*, a na tomto základě se vytvoří příslušná umělá proměnná. Tento typ modelů je obsahem kapitoly 6. Na tomto místě je ale dobré zdůraznit, že i odhad metodou nejmenších čtverců je možný, i když tento odhad není korektní, a jsou tak upřednostňovány jiné estimátory, známé jako probit a logit.

2.4 Shrnutí

Na základě této kapitoly tedy již víme, že:

- jednoduchá regrese kvantifikuje vliv vysvětlující proměnné (X) na závisle proměnnou (Y) prostřednictvním regresní přímky $Y = \alpha + \beta X$;
- odhad parametrů metodou nejmenších čtverců (OLS) α a β je spočívá ve volbě takových odhadů, které nám poskytnou přímku nejlépe vyrovnávající pozorování v rámci bodového grafu;
- odhady metodou nejmenších čtverců jsou označovány jako $\widehat{\alpha}$ a $\widehat{\beta}$ a získáme je minimalizací součtu čtverců reziduí (SSR);
- koeficienty jednoduché regrese můžeme interpratvat jako mezní vlivy, tedy jako měřítko toho, jak se změní Y, když změníme X o jednotku;
- koeficient determinace, R^2 nám říká, jako dobře vyrovnává regresní přímka pozorovaná data;
- interval spolehlivosti poskytuje intervalový odhad příslušného koeficient (např. interval pro parametr β, v rámci kterého se můžeme spoláhat na to, že v něm skutečně hodnota koeficientu β leží);

- testování hypotézy, $\beta = 0$, můžeme využít k rozhodování o tom, zdali příslušná vysvětlující proměnná patří do regrese, tedy jestli skutečně vysvětluje chování závisle proměnné;
- test hypotézy, $\beta = 0$, se provádí porovnáním testové statistiky (tzn. např. t-statistiky) s kritickou hodnotou nebo s využitím příslušné p-hodnoty;
- pokud je p-hodnota testu hypotézy, $\beta = 0$, menší než 0.05, můžeme zamítnou tuto hypotézu na hladině významnosti $5\ \%$;
- model vícenásobné regrese má více než jednu vysvětlující proměnnou, nicméně základní intuice týkající se OLS odhadů, intervalů spolehlivosti atd. je stejná jako pro jednoduchý regresní model;
- v případě modelu vícenásobné regrese je interpretace regresních koeficientů podmíněna podmínkami *ceteris paribus* (např. β_j měří marginální vliv vysvětlující proměnné X_j na závisle proměnnou Y za podmínky, že ostatní vysvěltující proměnné zůstávají konstantní);
- pokud v regresi opomeneme důležitou vysvětlující proměnnou, mohou být odhadnuté parametry zavádějící (tento jev označujeme jako zkreslení při opomenutí důležité vysvětlující proměnné), přičemž tento problém je horší tím více, čím silněji jsou opomenuté proměnné korelovány se zahrnutými vysvěltujícími proměnnými;
- pokud jsou vysvětlující proměnné spolu silně korelovány, mohou být odhady koeficientů a statistické testy zavádějící, což označujeme jako problém multikolinearity;
- se v praxi často setkáváme s umělými vysvětlujícími proměnnými, které nabývají hodnoty 0 nebo 1;
- zacházení s umělými vysvětlujícími proměnnými je stejné jako v případě neumělých vysvětlujících proměnných;
- regrese zahrnující pouze umělé vysvětlující proměnné implicitně třídí pozorování do různých skupin (např. domy s klimatizací a domy bez této vymožeností), což napomáhá k interpretaci výsledků odhadů příslušných koeficientů;
- regrese zahrnující umělé i ne-umělé vysvětlující proměnné implicitně třídí pozorování do různých skupin a říká nám, že každá skupina má regresní přímku s různou úrovňovou konstantou (nicméně všechny tyto regresní přímky mají stejný sklon);
- regrese zahrnující umělé, ne-umělé a vzájemně propojené vysvětlující proměnné (tj. součin umělých a neumělých proměnných) nám implicitně rozděluje pozorování do různých skupin a říká nám, že každá skupina má regresní přímku s odlišnou úrovňovou konstantou i sklonem.

Měli bychom tak již znát a umět vysvětlit obsah následujících klíčových pojmů:

- ❀ jednoduchá regrese
- ❀ vysvětlovaná (závisle) proměnná
- ❀ regresní koeficienty (parametry)
- ❀ regresní přímka
- ❀ metoda nejmenších čtverců
- ❀ regresní součet čtverců
- ❀ interval spolehlivosti parametru
- ❀ nulová a alternativní hypotéza
- ❀ hladina významnosti
- ❀ t-statistika (t-test)
- ❀ F-statistika (F-test)
- ❀ multikolinearita
- ❀ vícenásobná regrese
- ❀ vysvětlující proměnné
- ❀ odhad parametru
- ❀ součet čtverců reziduí
- ❀ celkový součet čtverců
- ❀ koeficient determinace
- ❀ testování hypotéz
- ❀ hladina spolehlivost
- ❀ kritická hodnota
- ❀ p-hodnota
- ❀ koeficient determinace
- ❀ umělé proměnné

Kapitola 3

Lineární regresní model jediné vysvětlující proměnné

V této kapitole se dozvíme:

- ☞ jaké základní statistické metody a techniky jsou využívány v ekonometrii;
- ☞ co je to estimátor a co samotný odhad;
- ☞ jak jsou formulovány (pro případ jednoduché regrese) klasické předpoklady, na jejichž základě jsou odvozeny veškeré potřebné statistické vlastnosti OLS estimátoru a následné procedury testování hypotéz a konstrukce intervalů spolehlivosti;
- ☞ jaké jsou vlastnosti OLS estimátoru;
- ☞ co nám o OLS estimátoru říká Gaussův-Markovův teorém;
- ☞ je to odhad metodou maximální věrohodnosti;
- ☞ jak odvodit intervaly spolehlivosti pro parametr β a jak odvodit a interpretovat test hypotézy, že $\beta = 0$, a to při známém rozptylu náhodných složek, σ^2;
- ☞ jak modifikovat interval spolehlivosti a test hypotézy o parametru při neznámém rozptylu náhodných složek, σ^2;
- ☞ jak využít asymptotickou teorii pro zkoumání vlastností estimátoru pro velikost vzorku jdoucí k nekonečnu, tedy $N \to \infty$;

3.1 Úvod

V předchozí kapitole byl zaveden model vícenásobné regrese s k vysvětujícími proměnnými, který si můžeme zapsat jako

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i,$$

kde index i označuje jednotlivá pozorování, kdy $i = 1, \ldots, N$. Máme tedy celkem N pozorování. Tento model můžeme využít k analýze vlivu vysvětlujících proměnných, $X_1, \ldots, X_k$, na závisle proměnnou, Y. V netechnickém úvodu do regrese byl zaveden OLS estimátor, založený na minimalizaci součtu čtverců reziudí. Zavedli jsme si pojem intervalů spolehlivost a testování hypotéz, které jsou založeny na fOLS estimátoru. V této kapitole přejdeme z neformálního výkladu předchozí kapitoly k výkladu formálnějšímu.

V ekonometrii se neustále setkáváme s pojmem nejistota. Nejsme si jisti tím, jakých hodnot nabývají regresní koeficienty, $\alpha, \beta_1, \ldots, \beta_k$ a musíme je tudíž odhadovat. Jsme obklopeni nejistotou zda-li skutečně platí jednotlivé hypotézy, např. $\beta_j = 0$, a musíme tedy odvodit techniky k jejich testování. Nejistota se týká i budoucích hodnoty vysvětlované proměnné, Y, a musíme tedy odvodit techniky pro jejich předpověď. Teorie pravděpodobnosti nám poskytuje rámec, ve kterém se s nejistotou dokážeme vypořádat. O tom bude tato kapitola.

Abychom si ukázali základní principy v tom nejjednodušším konceptu, budeme pracovat s jednoduchým regresním modelem bez úrovňové konstanty

$$Y_i = \beta X_i + \epsilon_i.$$

Získané závěry a intuice budou podobné těm, které platí i v případě modelu vícenásobné regrese, nicméně jejich odvození je v tomto případě trochu náročnější, zvláště pokud nechceme používat maticového zápisu.

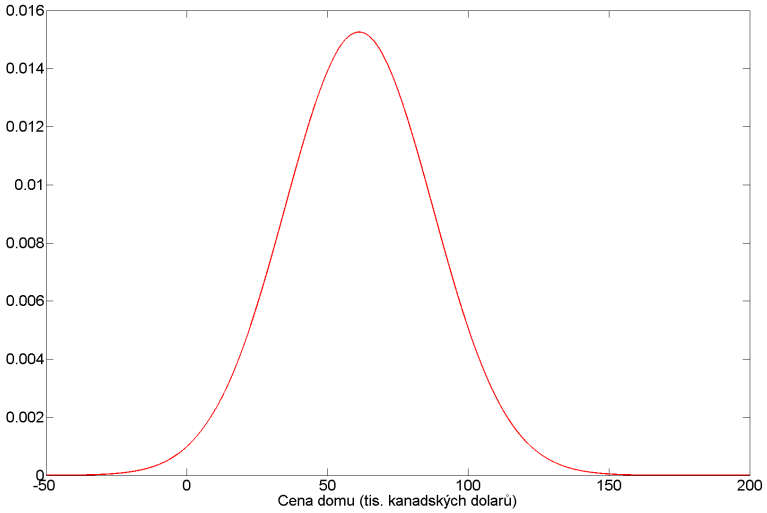
3.2 Základy pravděpodobnosti v kontextu regresního modelu

Ekonometrové začínají svou práci předpokladem, že Y je náhodná veličina. V tomto uvažování nám model (např. regresní model) poskytuje popis toho, jak pravděpodobné jsou hodnoty závisle proměnné. Předpokládejme náš známý příklad, kdy Y je cena domu a X je jeho charakteristika (např. jeho celková rozloha). Předpokládejme, že bychom znali X, ale neznali Y. Řekněme, že pro konkrétní dům je $X = 5000$ čtverečních stop (což je typická hodnota v pozorovaných datech z našeho příkladu). To není dostačující informace pro přesnou znalost ceny domu. Cena domu, Y je tak nejistá. Není však zcela nejistá. Dům s rozlohou $X = 5000$ by se mohl prodávat za zhruba 70000, 60000 nebo 50000 dolarů (což jsou obvyklé ceny v pozorovaných datech), ale určitě se nebude prodávat za 1000 dolarů nebo 1000000 dolarů. Ekonometři využívají teorie pravděpodobnosti k stanovení toho, jaké hodnoty jsou pravděpodobné a jaké ne.

Funkce hustoty pravděpodobnosti (probability density function – p.d.f.) nám shrnuje informace o tom, co víme o náhodné veličině. Obrázek 3.1 je příkladem funkce hustoty pravděpodobnosti pro Y z našeho příkladu cen domů. Shrnuje nám interval pravděpodobných hodnot cen domů, Y, které by mohly odpovídat domům s rozlohou $X = 5000$. Obrázek 3.1 odpovídá normálnímu rozdělení. To je jedno z nejběžnějších rozdělení v ekonometrii. Jedná se o zvonovou křivku, kdy nejpravděpodobnější hodnoty odpovídají hodnotám příslušejícím jejímu vrcholu. Nabývá tedy nejvyšších hodnot pro nejpravděpodobnější cen domů, což jsou hodnoty kolem 50000, 60000 a

70000 dolarů. Nejnižších hodnot nabývá pro málo pravděpodobné ceny domů, což je např. 1000 dolarů nebo 200000 dolarů. Přesný tvar normálního rozdělení závisí na střední hodnotě a rozptylu. Označení náhodné veličiny, jejíž rozdělení odpovídá funkci normální hustoty pravděpodonbosti se střední hodnotou μ a rozptylem σ^2 je

$$Y \sim N\left(\mu, \sigma^2\right).$$

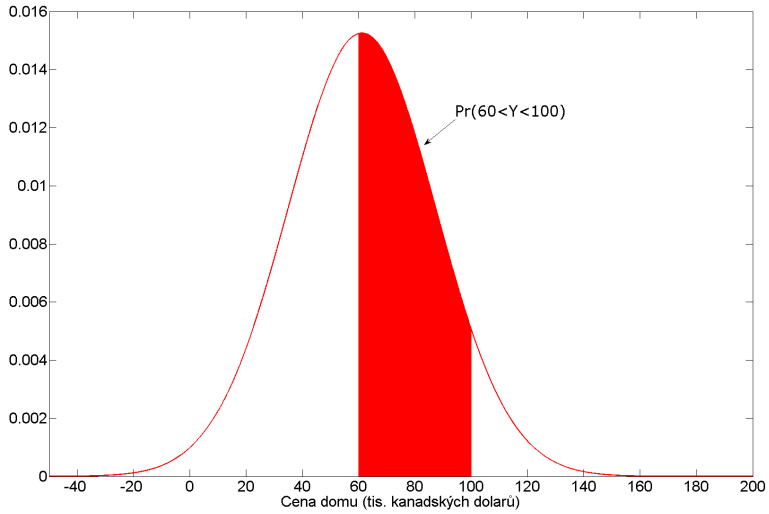


Obrázek 3.1: Funkce normální hustoty pravděpodobnosti ceny domu o rozloze 5000 čtverečních stop.

Obrázek 3.1 byl vykreslen na základě hodnot $\mu = 61.153$ (tzn. střední nebo průměrná hodnota domu o rozloze 5000 čtverečních stop je 61153 dolarů) a $\sigma^2 = 683.812$ (což je charakteristika rozptylu cen domů s rozlohou 5000 čtverečních stop). Tyto hodnoty byly získány na základě příslušné jednoduché regrese Y na X (bez úrovňové konstanty).

Oblast pod křivkou hustoty pravděpodobnosti odpovídá pravděpodobnosti realizace náhodné veličiny v dané oblasti. To nám ilustruje obrázek 3.2, který je téměř identický s obrázkem 3.1. Vybarvená oblast (její plocha) mezi hodnotami 60 a 100 vyjadřuje pravděpodobnost, že dům má hodnotu mezi 60000 dolary a 100000 dolary. V našem příkladu tato pravděpodobnost odpovídá hodnotě 45 %, což můžeme psát jako $Pr(60 \leq Y \leq 100) = 0.45$. Tento údaj snadno získáme výpočtem za použití statistických tabulek či s využitím vhodného ekonometrického software. Z definice hustoty pravděpodobnosti má plocha celé oblasti pod křivkou funkce hustoty pravděpodobnosti hodnotu 1. Logicky, obsah každé jiné oblasti pod touto křivkou bude menší než 1. Pravděpodobnost se musí nacházet mezi hodnotami 0 a 1. Nemůžeme mít pravděpodobnost realizace nějakého jevu 120 % nebo -6 %.

Na chvíli odbočme a popišme si, jak využít statistických tabulek normálního rozdělení. Tabulka standardizovaného normálního rozdělení je většinou dostupná pro *standardizované normální rozdělení*, $N(0, 1)$. To je zcela dostačující pro analýzu obecného



Obrázek 3.2: Funkce normální hustoty pravděpodobnosti ceny domu o rozloze 5000 čtverečních stop a ukázka vybrané oblasti pravděpodobnosti.

normálního rozdělení, $N(\mu, \sigma^2)$, tedy pro libovolnou střední hodnotu μ a rozptyl σ^2. Předpokládejme náhodnou veličinu $Y \sim N(\mu, \sigma^2)$ a dále uvažujme novou náhodnou veličinu

$$Z = \frac{Y - \mu}{\sigma}.$$

Ptotože μ a σ^2 jsou postupně střední hodnotou a rozptylem veličiny Y, můžeme psát $E(Y) = \mu$ a $var(Y) = \sigma^2$. Jaká je střední hodnota a rozptyl náhodné veličiny Z? Uvědomme si, že μ a σ^2 nejsou náhodné veličiny, tudíž z vlastností střední hodnoty (viz příloha B) lze psát:

$$\begin{aligned} E(Z) &= E\left(\frac{Y - \mu}{\sigma}\right) = \frac{E(Y - \mu)}{\sigma} \\ &= \frac{E(Y) - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0. \end{aligned}$$

Podobně z vlastností pro rozptyl náhodné veličiny vyplývá:

$$\begin{aligned} var(Z) &= var\left(\frac{Y - \mu}{\sigma}\right) = \frac{var(Y - \mu)}{\sigma^2} \\ &= \frac{var(Y)}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1. \end{aligned}$$

Náhodná veličina Z tedy pochází ze standardizovaného normálního rozdělení, $N(0, 1)$. Náhodná vleičina Z se označuje rovněž jako *Z-skór*. Konstrukce Z odpovídá standardizaci náhodné veličiny Y.

Jako příklad použití *Z-skóru* předpokládejme barevnou oblast funkce normální hustoty pravděpodobnosti z obrázku 3.2. Bylo řečeno, že $Y \sim N(61.153, 683.812)$. Barevná plocha odpovídá pravděpodobnosti $textPr(60 \leq Y \leq 100)$ a právě tuto pravděpodobnost chceme spočítat. Protože nemáme k dispozici statistické tabulky pro normální rozdělení $N(61.153, 683.812)$, musíme problém převést do podoby zahrnující náš známý *Z-skór*, čímž můžeme využít tabulky standardizovaného normálního rozdělení. Samozřejmě s využitím ekonometrického či statistického software tento problém obvykle řešit nemusíme a program jej vyřeší za nás. Transformaci provedeme velmi snadno:

$$\begin{aligned}
\Pr(60 \leq Y \leq 100) &= \Pr\left(\frac{60-\mu}{\sigma} \leq \frac{Y-\mu}{\sigma} \leq \frac{100-\mu}{\sigma}\right) \\
&= \Pr\left(\frac{60-61.153}{\sqrt{683.812}} \leq \frac{Y-61.153}{\sqrt{683.812}} \leq \frac{100-61.153}{\sqrt{683.812}}\right) \\
&= \Pr\left(-0.04 \leq Z \leq 1.49\right).
\end{aligned}$$

Problém tak byl transformován do podoby zahrnující standardizované normální rozdělení a k výpočtu příslušné pravděpodobnosti snadno využijeme statistické tabulky, abychom získali $\Pr\left(-0.04 \leq Z \leq 1.49\right) = 0.45$. Pokud to není zřejmé, poznamenejme, že oblast pod křivkou hustoty pravděpodobnosti můžeme rozdělit na dvě části, tedy $\Pr\left(-0.04 \leq Z \leq 1.49\right) = \Pr\left(-0.04 \leq Z \leq 0\right) + \Pr\left(0 \leq Z \leq 1.49\right)$. Tabulky distribuční funkce normálního rozdělenínám přímo říkají, že $\Pr\left(0 \leq Z \leq 1.49\right) = 0.4319$. Protože je normální rozdělení symetrické, musí platit $\Pr\left(-0.04 \leq Z \leq 0\right) = \Pr\left(0 \leq Z \leq 0.04\right)$, což podle tabulek odpovídá hodnotě 0.0160. Součet obou pravděpodobností dává hodnotu 0.4479, což jsme zaokrouhlili na hodnotu 0.45. Tento postup odpovídá tabulkám z Koopovy učebnice [17].

Našim cílem je odvození vlastností OLS estimátoru, intervalů spolehlivosti a testování hypotéz. Doposud jsme předpokládali, že vysvětlovaná proměnná Y je náhodná veličina pocházející z normálního rozdělení. Na tomto základě jsme si zavedli funkci hustoty pravděpodobnosti, kterou lze využít k vyjádření nejistoty spojené s hodnotami Y (např. s jakou pravděpodobností budou hodnoty Y ležet v nějakém intervalu). Protože budeme hojně využívat oprátorů očekávané hodnoty a rozptylu, je dobré si v případě obtíží zopakovat základy pravděpodobnosti a statistiky uvedené v příloze B.

3.3 Klasické předpoklady regresního modelu

Doposud jsme si nic neřekli o odhadu koeficientu β, intervalech spolehlivosti a ani o testování hypotéz. K tomu si ale musíme uvést předpoklady, které formálně definují normální lineární regresní model. Tyto předpoklady se označují jako tzv. *klasické předpoklady* a jsou východiskem pro studium regresního modelu. Připoměňme si, že máme k dispozici N pozorování závisle proměnné, $Y_1, Y_2, \ldots, Y_N$. Jazykem pravděpodobnosti je každé toto pozorování realizací náhodné veličiny. Množina předpokladů, které zcela definují funkci hustoty pravděpodobnosti pro všechna tato pozorování nám definují příslušný model. Klasické předpoklady definují normální lineární regresní model.

Klasické předpoklady pro každé z $i = 1, \ldots, N$ pozorování jsou:

1. $E(Y_i) = \beta X_i$.
2. $var(Y_i) = \sigma^2$.
3. $cov(Y_i, Y_j) = 0$ pro $i \neq j$.
4. Y_i má normální rozdělení.
5. X_i je pevně dáno, není to tedy náhodná veličina.

Kompaktní vyjádření těchto předpokladů je

$$Y_i \sim N(\beta X_i, \sigma^2),$$

pro $i = 1, \ldots, N$ přičemž Y_i, Y_j jsou vzájemně nekorelovány pro $i \neq j$.

V kapitole 2 jsme popisovali regresní model jako vyrovnání bodového grafu pozorování přímkou. Klasické předpoklady tuto neformální intuici formalizují a je dobré si každý z předpokladů rozebrat.

První předpoklad, $E(Y_i) = \beta X_i$, odpovídá předpokladům linearity modelu. Poznamenejme, že zatím pracujeme s nejjednodušší variantou modelu s jedinou vysvětlující proměnnou a bez úrovňové konstanty. V modelu vícenásobné regrese je tento předpoklad nahrazen výrazem $E(Y_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$. Očekáváme tedy, že Y_i leží na regresní přímce.[7] Samozřejmě, pohledem na skutečná data bude spíše neobvyklé, že by Y_i leželo přímo na regresní přímce, nicméně náš předpoklad nám říká, že v průměru tomu tak bude. Některá pozorování budou ležet nad přímkou, některá nad ní, v průměru však regresní přímku tvořit budou.

Druhý předpoklad je ten, že všechna pozorování mají stejný rozptyl. V kapitole 5 tento předpoklad uvolníme. Nicméně, na příkladu si ilustrujme, kdy by tento předpoklad mohl být narušen. Přepdokládejme náš příklad s cenami domů, kdy vysvětlující proměnná je celková rozloha domu. V našich datech máme velké domy (s velkou rozlohou a pozemkem okolo) a malé domy. Předpokládejme, že malé domy si jsou velmi podobné (jedná se například o malé bungalovy s podobnou specifikací). V takovémto případě budou mít malé domy podobnou cenu. Pokud by se prodalo několik deítek těchto domů, kupující by rychle zjistil, že jejich cena je přibližně např. 30000 dolarů a nebyl by ochoten zaplatit o mnoho více, stejně jako prodávající by nebyl ochoten prodat za cenu výrazně nižší. Variabilita cen malých domů ($var(Y)$) tak bude pro malé domy malá. Předpokládejme dále, že velké domy budou mnohem diverzifikovanější. Ceny velkých domů se budou od sebe mnohem více ličit, jejich variabilita tak bude velká. Tento příklad by odpovídal situaci, kdy je předpoklad o shodném rozptylu narušen. Variabilita pro Y_i nebude stejná, bude se lišit pro malé a velké domy. Tento předpokald je nazýván předpokladem o *homoskedasticitě*, tedy konstantním rozptylu. Jeho porušení odpovídá situaci, kdy nastává případ tzv. *heteroskedasticity*, tedy rozdílného rozptylu.

[7] V případě vícerovnicového modelu se jedná o regresní rovinu pro dvě vysvětlující proměnné, či jistý druh regresní nadroviny pro případ více než dvou vysvětlujících proměnných, což ale není snadné si vizuálně představit. Každopádně při zafixování ostatních vysvětlujících proměnných se při zobrazení Y vzhledem k jedné z vysvětlujících proměnných o regresní přímku jednat bude.

Třetí předpoklad znamená vzájemnou nekorelovanost pozorování. Připomeňme si, že

$$corr(Y_i, Y_j) = \frac{cov(Y_i, Y_j)}{\sqrt{var(Y_i)var(Y_j)}},$$

a tedy $cov(Y_i, Y_j) = 0$ znamená, že $corr(Y_i, Y_j) = 0$. Tento předpoklad je většinou splněn pro průřezová data.[8] Předpokládejme např. data z dotazníkových šetření (např. šetření pracovníků a jejich charakteristiky). Dotazníková šetření obvykle náhodně zvolí pracovníky, kterým budou položeny otázky. Náhodný výběr by měl zajistit nekorelovanost jejich odpovědí. V případě časových řad uvidíme, že tento předpoklad nebývá splně. Např. při analýze míry nezaměstnanosti se setkáváme se skutečností, že ekonomická aktivita (hospodářský cyklus) se vyvíjí nahoru a dolů relativně pozvolně a s určitou setrvačností. To znamená, že míra nezaměstnanosti roku 1996 bude obvykle korelována s mírou nezaměstnanosti roku 1997. Pokud byl rok 1996 rokem nízké nezaměstnanosti, bude na tom obvykle i rok 1997 podobně (výraznější je tato argumentace u čtvrtletních dat, tedy u sousedících čtvrtletí). Nezaměstnanost tak může být korelována v čase, hovoříme tak o *autokorelaci*. V kapitole 5 si ukážeme, jak se vypořádat s uvolněním tohoto předpokladu.

Čtvrtý předpoklad říká, že Y pochází z normálního rozdělení. Najít motivaci pro tento předpoklad není snadné, je třeba se spokojit s tím, že předpoklad normality je při empirické aplikaci obvykle (ale ne vždy) rozumný. V příloze B a v přílohách k některým jednotlivým kapitolám je vysvětleno, jak lze využít v některých modelech asymptotickou teorii pro opuštění předpokladu o normalitě.

Poslední předpoklad je ten, že vysvětlující proměnná je pevně daná a nejedná se tedy o náhodnou veličinu. V experimentálních vědách se jedná o rozumný předpoklad. Nechá se proběhnout nějaký experiment při zvolených hodnotách vysvětlujících proměnných, a po té se pozorují jeho výsledky. Protože jsou hodnoty vysvětlujících proměnných vybrány, X není náhodná veličina. Ekonomický výzkum obvykle není experimentální vědou a tento předpoklad tak nemusí být v určitém kontextu rozumný. Příloha 2: Využití asymptotické teorie v jednoduchém regresním modelu ukazuje, jak využít asymptotitcké teorie k nahrazení tohoto předpokladu předpokladem, že X_i je náhodná veličina nekorelovaná s chybovou složkou regrese. Za tohoto předpokladu budou všechny odvozené výsledky OLS odhadu z této kapitoly stále platit. Pokud však X_i je náhodná veličina korelovaná s chybovým členem, OLS výsledky již platit nebudou. V kapitole 5 je tento případ diskutován a je ukázáno proč by měl být místo OLS estimátoru využit estimátor instrumentálních proměnných.

Klasické předpoklady jsme v našem případě prezentovali v kontextu toho, jaké vlastnosti požadujeme od Y_i. V ekonoemtrii je však obvyklý způsob uvažování předpokladů v kontextu chyby regrese, tedy náhodných složek. Ekvivalentně tak můžeme s využitím vlastností operátorů střední hodnoty, očekávání a kovariance vyjádřit klasické předpoklady jako:

1. $E(\epsilon_i) = 0$.
2. $var(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$. Konstantní rozptyl chyb (homoskedasticita).

[8] Existuje však literatura, věnovaná problému korelovanosti průřezových dat, kdy se hovoří o prostorové korelaci *spatial correlation*.

3. $cov(\epsilon_i, \epsilon_j) = 0$ pro $i \neq j$. Náhodné složky jsou nekorelovány.

4. ϵ_i je normálně rozdělené.

5. X_i je pevně dáno, není to tedy náhodná veličina.

Výsledek $var(\epsilon_i) = E(\epsilon_i^2)$ vyplývá ze samotné definice rozptylu, $var(\epsilon_i) = E(\epsilon_i^2) - [E(\epsilon_i)]^2$, a ze skutečnosti, že střední hodnota náhodné složky je nulová (tj. $E(\epsilon_i) = 0$). Intuice těchto předpokaldů je podobná té, která byla diskutována výše. Např. u prvního předpokladu uvažujeme lineární model, kdy přímka vyrovnání neprochází logicky všemi body bodového grafu. Některá rezidua jsou tak kladná, jiná záporná, ale v průměru je jejich střední hodnota nulová. Předpoklad dvě odpovídá tomu, že rezidua budou mít stejný rozptyl bez ohledu na hodnotu vysvětlující proměnné X. Třetí předpoklad říká, že případná chyba, která nastane např. při vyplňování jednoho dotazníku v rámci výběrového šetření nijak neovlivní případnou chybu v případě jiného dotazníku. Čtvrtý předpoklad, předpoklad normality, nemá zřejmou intuici, nicméně jej lze obejít s využitím asymptotické teorie. Podobně je na tom i předpoklad pátý.

3.4 Vlastnosti OLS estimátoru pro parametr β

V kapitole 2 jsme si zavedli estimátor metody nejmenších čtverců jako takový, jehož výsledkem je nejlepší regresní přímka vyrovnání bodového grafu pozorování. V rámci jednoduchého regresního modelu bez úrovňové konstanty

$$Y_i = \beta X_i + \epsilon_i$$

získáme OLS estimátor minimalizací součtu čtverců reziduí

$$SSR = \sum_{i=1}^{N} \widehat{\epsilon}_i^2.$$

Nalezení příslušného vzorce je velmi snadné, neboť se jedná o jednoduchý minimalizační problém. Řešení je

$$\widehat{\beta} = \frac{\sum_{i=1}^{N} X_i Y_i}{\sum_{i=1}^{N} X_i^2}.$$

Z definice má OLS jednu velmi pěknou vlastnost: dává nám regresní přímku vyrovnání, která nejlépe vyrovnává data ve smyslu co možná nejmenšího součtu čverců reziduí. Má však i jiné atraktivní vlastnosti. Abychom si je ukázali a následně diskutovali výsledky pokud jde o konstukci intervalů spolehlivosti a testování hypotéz, musíme si odvodit rozdělení OLS estimátoru.

Je třeba si uvědomit, že $\widehat{\beta}$ je náhodná veličina. Vzorec tohoto estimátoru závisí na Y_i pro $i = 1, \ldots, N$, což jsou všechno normální náhodné veličiny. Klasické předpoklady navíc říkají, že X_i pro $i = 1, \ldots, N$ jsou nenáhodné veličiny. Vztah pro $\widehat{\beta}$ tak naznačuje, že se jedná o lineární funkci normálních náhodných veličin (díky linearitě výsledného vztahu hovoříme o lineárním estimátoru). Základy pravděpodobnosti a statistiky nám říkají, že výsledkem lineární kombinace normálních náhodných veličin je

opět normální náhodná veličina. Z toho samozřejmě vyplývá, že $\widehat{\beta}$ je náhodná veličina odpovídající normálnímu rozdělení. Normální náhodné veličiny jsou jednoznačně určeny svou střední hodnotou a rozptylem. Známe-li tyto charakteristitky, jsme schopni plně popsat pravděpodobnostní rozdělení OLS estimátoru. Proč je dobré toto rozdělení znát? Odpověď je snadná, budeme díky němu schopniodvoditintervaly spolehlivosti a postupy testování hypotéz odpovídající těm z kapitoly 2.

Před samotný odvozením střední hodnoty a rozptylu $\widehat{\beta}$ si nejprve odvodíme alternativní vyjádření OLS estimátoru, které využijeme v některých dalších odvozeních. Tuto rovnici označíme jako (*). Ve vztahu pro $\widehat{\beta}$ nahradíme Y_i odpovídajícím výrazem na pravé straně regresního modelu, tedy výrazem $X_i\beta+\epsilon_i$ a jednotlivé členy uspořádáme, čímž získáváme vztah

$$\begin{aligned}\widehat{\beta} &= \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{\sum X_i(X_i\beta+\epsilon_i)}{\sum X_i^2} \qquad (*)\\ &= \beta + \frac{\sum X_i\epsilon_i}{\sum X_i^2}.\end{aligned}$$

Tento vzorec nám říká, že estimátor $\widehat{\beta}$ můžeme zapsat jako součet skutečné hodnoty odhadovaného parametru, β, a část zahrnující chybové členy a hodnoty vysvětlující proměnné. Tento výraz zahrnuje neznámé chyby, tudíž se nejedná o výraz vhodný k praktickému vyčíslení OLS odhadu, nicméně jedná se o užitečný výraz pro řadu teoretických odvození.

Vlastnost 1: Střední hodnota OLS estimátoru je β:

$$E\left(\widehat{\beta}\right) = \beta.$$

Důkaz:

$$\begin{aligned}E\left(\widehat{\beta}\right) &= E\left(\beta + \frac{\sum X_i\epsilon_i}{\sum X_i^2}\right) = \beta + E\left(\frac{\sum X_i\epsilon_i}{\sum X_i^2}\right)\\ &= \beta + \frac{1}{\sum X_i^2}E\left(\sum X_i\epsilon_i\right) = \beta + \frac{1}{\sum X_i^2}\sum X_i E(\epsilon_i)\\ &= \beta.\end{aligned}$$

Důkaz využívá rovnici (*), vlastnosti operátoru střední hodnoty a první z klasických předpokladů o nulové střední hodnotě náhodné složky. Při odvození je třeba nezapomenout na to, že β a X_i jsou nenáhodné veličiny a lze je tak brát v odvození jako konstanty.

Vlastnost 1 je velmi atraktivní. Jakýkoliv estimátor nabízí jen odhad koeficientu β. Nemůžeme tak očekáva, že $\widehat{\beta}$ bude přesně roven parametru β. Tato vlastnost nám říká, že OLS odhad bude v průměru roven skutečné hodnotě β. Někdy bude OLS odhad vyšší, někdy nižší, ale v průměru to bude ten správný odhad.

Vlastnost 1 nám umožňuje zavést důležitý koncept zvaný *nevychýlenost* resp. *nestrannost*. Řeknem, že estimátor je *nestranný* nebo *nevychýlený (unbiased)*, pokud je jeho střední (očekávaná) hodnota rovna tomu co jím odhadujeme. Protože je v našem

případě $E(\widehat{\beta}) = \beta$, můžeme říct, že $\widehat{\beta}$ je nestranný (nevychýlený) estimátor koeficientu β (jeho konkrétní hodnota je pak nestranný či nevychýlený odhad). Nestrannost je velmi žádoucí vlastnost a ekonometři se snaží hledat estimátory, které jsou nevychýlené. Opakem nestranného estimátoru je vychýlený estimátor, jehož očekávaná (střední) hodnota není rovna sktutečné hodnotě toho, co odhadujeme.

Vlastnost 2: Rozptyl OLS estimátoru je při splnění klasických předpokladů roven

$$var\left(\widehat{\beta}\right) = \frac{\sigma^2}{\sum X_i^2}.$$

Důkaz:

$$\begin{aligned} var\left(\widehat{\beta}\right) &= var\left(\beta + \frac{\sum X_i\epsilon_i}{\sum X_i^2}\right) = var\left(\frac{\sum X_i\epsilon_i}{\sum X_i^2}\right) \\ &= \left(\frac{1}{\sum X_i^2}\right)^2 var\left(\sum X_i\epsilon_i\right) = \left(\frac{1}{\sum X_i^2}\right)^2 \sum X_i^2 var(\epsilon_i) \\ &= \left(\frac{1}{\sum X_i^2}\right)^2 \sigma^2 \sum X_i^2 = \frac{\sigma^2}{\sum X_i^2}. \end{aligned}$$

Důkaz využívá rovnici (*), vlastnosti operátoru rozptylu a druhý z klasických předpokladů o konstantním rozptylu náhodné složky, který je roven σ^2. Při odvození je opět třeba nezapomenout na to, že β a X_i jsou nenáhodné veličiny a lze je tak brát v odvození jako konstanty (pokud by byla vysvětlující proměnná náhodnou veličinou, musí platit předpoklad, že je nekorelovaná s náhodnou složkou, tedy $E(X_i\epsilon_i) = 0$ pro všechna i).

Vlastnost 2 nám kvantifikuje variabilitu náhodné veličiny, $\widehat{\beta}$. Je vcelku dobré, pokud mají nestranné estimátory malou variabilitu. Zmínili jsme se, že $\widehat{\beta}$ je odhad skutečné hodnoty parametru β, někdy je nižší, někdy vyšší, ale v průměru nabývá té správné hodnoty. Intuitivně je určitě lepší mít odhady, které jsou jen o něco málo vyšší nebo nižší, než odhady, které které jsou mnohem vyšší nebo mnohem nižší. Estimátor s malým rozptylem bude mít právě prvně zmíněnou vlastnost (tzn. bývá jen o něco málo vyšší nebo nižší).

Ekonometrické programy nám vedle OLS odhadu vrátí i jeho směrodatnou odchylku. Tyto směrodatné odchylky jsou rovny $\sqrt{var(\widehat{\beta})}$. Je potřeba zdůraznit, že pro řešení ekonometrických problémů je možno využít různé druhy estimátorů. Lze ukázat, že pro jednoduchý lineární regresní model bez úrovňové konstanty je nestranný estimátor rovněž:

$$\tilde{\beta} = \frac{\sum_{i=1}^N Y_i}{\sum_{i=1}^N X_i},$$

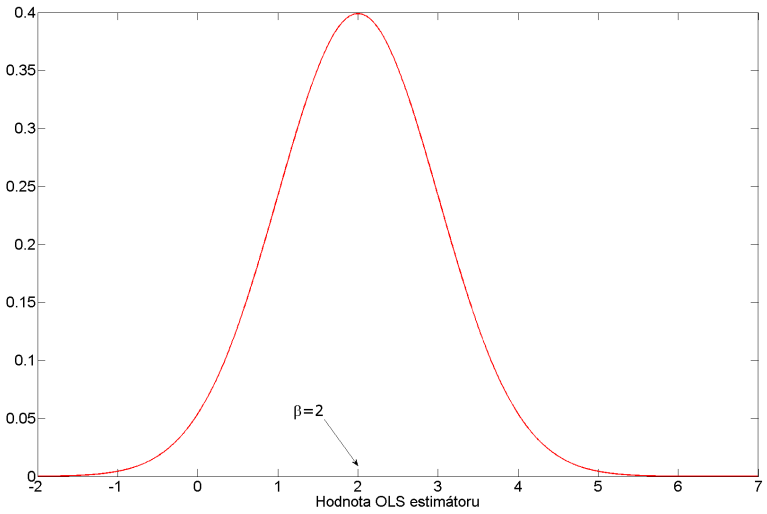
za předpokladu, že $\sum_{i=1}^N X_i \neq 0$. Proč tedy raději použít OLS estimátor než tento uvedený? Lze ukázat, že tento estimátor má větší variabilitu než $\widehat{\beta}$. Pokud tedy volíme mezi dvěma nestrannými estimátory, je dobré zvolit ten, který má nižší rozptyl. Hovoříme tak o situaci, kdy jeden estimátor je *vydatný (efficient)* oproti jinému estimátoru, pokud má menší rozptyl.

Vlastnost 3: Při splnění klasických předpokladů odpovídá rozdělení OLS estimátoru:

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum X_i^2}\right).$$

Tato vlastnost vyplývá z vlastností nestrannosti a vydatnosti OLS estimátoru a ze závěru o normalitě $\widehat{\beta}$. Tato vlastnost je důležitá zejména pro odvození intervalů spolehlivosti a testování hypotéz.

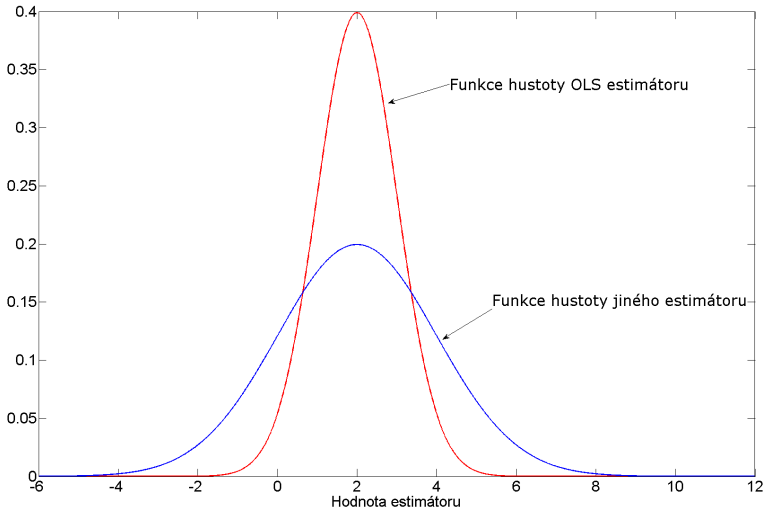
Obrázek 3.3 je příkladem funkce hustoty pravděpodobnosti pro $\widehat{\beta}$. Funkce hustoty pravděpodobnosti je zpodobněním nejistoty spojené s náhodnou veličinou. Oblast, kde je funkce hustoty nejvyšší odpovídá nejpravděpodobnějším hodnotám realizace náhodné proměnné. Obrázek 3.3 je vykreslen na základě předpokladu, že $\widehat{\beta} \sim N(2,1)$ a je zde patrné, že vrchol funkce hustoty je v hodbotě střední hodnoty náhodné veličiny, což odpovídá hodnotě $\beta = 2$. Je tedy velmi pravděpodobné, že odhad, $\widehat{\beta}$, bude v okolí své skutečné hodnoty. Víme však, že plocha pod křivkou vyjadřuje pravděpodobnost spojenou s různými intervaly realizace. Slušnou pravděpodobnost tak má i hodnota odhadu blízko 1 nebo 3. Ze statistických tabulek normálního rozdělení můžeme spočítat, že $\Pr(1.0 \leq \widehat{\beta} \leq 3.0) = 0.68$. Existuje tedy 68% pravděpodobnost, že $\widehat{\beta}$ bude v okolí své skutečné hodnoty, $\beta = 2$, kdy toto okolí odpovídá odchylce ± 1. Stejně tak je $\Pr(0 \leq \widehat{\beta} \leq 1) = 0.14$, tedy je 14% šance, že $\widehat{\beta}$ bude ležet v intervalu 0 až 1.



Obrázek 3.3: Funkce hustoty pravděpodobnosti OLS estimátoru.

Pro ilustraci vydatnosti a role rozptylu při rozhodování mezi dvěma estimátory předpokládejme, že $\widehat{\beta}$ je z $N(2,1)$ a stejně tak předpokládejme další nevychýlený estimátor $\tilde{\beta} \sim N(2,4)$, Jedná se o nestranný estimátor, protože $E(\tilde{\beta}) = 2$, což je v tomto příkladu skutečná hodnota parametru β. Tento další estimátor však má větší rozptyl než OLS estimátor, protože $var(\tilde{\beta}) = 4$. Funkce hustot pravděpodobnosti pro oba estimátory jsou vykresleny v obrázku 3.4. Protože platí, že $var(\tilde{\beta}) > var(\widehat{\beta})$, je funkce

hustoty nového estimátoru více rozptýlená. Tento nový estimátor tak s mnohem větší pravděpodobností poskytne odhad, který bude mnohem více vzdálen od skutečné hodnoty, $\beta = 2$. Pro OLS estimátor dostaneme jen s malou pravdpěodobností záporný odhad, protože $\Pr(\widehat{\beta} \leq 0) = 0.02$. S novým estimátorem je $\Pr(\tilde{\beta} \leq 0) = 0.16$, a existuje tak mnohem větší šance, že získáme zápornou hodnotu odhadu parametru.



Obrázek 3.4: Funkce hustot pravděpodobnosti OLS estimátoru a méně vydatného estimátoru.

Doposud jsme si ukázali, že OLS estimátor je nestranný a má určitý rozptyl. Nestrannost (nevychýlenost) jsme si definovali jako žádoucí vlastnost, naopak vysoký rozptyl jako vlastnost méně žádoucí. Volba mezi nevychýlenými estimátory probíhá právě na základě co nejmenšího rozptylu. To ale není zase tak jednoduché. Někdy může být obtížné dokázat, že je estimátor nevychýlený a i když to dokážeme, může být zcela nemožné naléztestimátor, který bude mít nejmenší možnou variabilitu oproti konkurenčním estimátorům. V regresním model, při splnění klasických předpokladů je však nalezení takovéhoto estimátoru velmi snadné, a to díky následujícímu teorému.

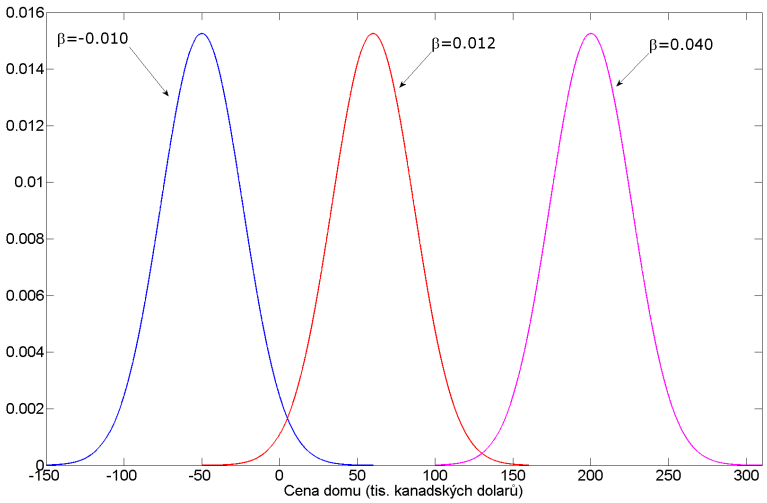
Vlastnost 4: *Gaussův-Markovův teorém.*

V regresním modelu při splnění klasických předpokladů je estimátor metody nejmenších čtverců nejlepší, lineární a nevychýlený estimátor. Setkáváme se tak s označením, že OLS estimátor je BLUE (Best Linear Unbiased Estimator), tedy nejlepší (nejmenší variabilita), lineární (je lineární funkcí pozorování závisle proměnné) a nestranný (nevychýlený) estimátor. Příloha 1: Důkaz Gaussova-Markovova teorému nabízí důkaz tohoto teorému. Důkaz teorémů nevyužívá z pěti klasických předpokladů předpoklad o normalitě. OLS estimátor je tedy BLUE i v případě, kdy náhodné složky nemají normální rozdělení.

Zdá se, že jsme si představili dostatek argumentů, proč je OLS estimátor dobrý estimátor (při splnění klasických předpokladů). Je to estimátor, který minimalizuje součet čtverců reziduí, je nestranný, a má nejmenší možný rozptyl oproti jiným lineárním,

nestranným estimátorům. Ukážeme si ještě jeden argument pro jeho využívání, který nám umožní zavést další mimořádně důležitý koncept ekonometrie, kterým je *odhad metodou maximální věrohodnost (maximum likelihood estimation)*.

Klasické předpoklady nám říkají, že Y_i je z normálního rozdělení, $N(\beta X_i, \sigma^2)$. Problémem odhadu je určení pravděpodobných hodnot β a σ^2. Volba β ovlivňuje střední hodnotu normálního rozdělení. Na obrázku 3.1 bylo vykresleno normální rozdělelní pro cenu domu (Y) s rozlohou 5000 čtverečních stop. Obrázek jsme interpretovali tak, že se jednalo o vyjádření neurčitosti spojené s cenou takovéhoto domu. Cena domu byla nepravděpodobněji přibližně 60000 dolarů(střední hodnota byla 61153 dolarů), nicméně notná dávka pravděpodobnosti byla spojena s cenami směrem k 100000 nebo 20000 dolarů. Jak tedy můžeme vědět, že se jedná o rozumné shrnutí nejistoty spojené s cenou takového domu? Odpověď je snadná, stačí se podívat na skutečné prodejní ceny domů s rozlohou přibližně 5000 čtverečních stop. V našich datech se většina těchto domů prodávala za 60000 dolarů, nicméně některé se prodaly dráže, některé levněji. Obrázek 3.1 tak byl vcelku rozumný, neboť rozdělení náhodné veličiny Y (cena domu s rozlohou 5000 čtverečních stop) koncentruje většinu své pravděpodobnosti do oblastí skutečné realizace pozorovaných hodnot Y.[9] Myšlenka, že specifikovaná funkce hustoty pravděpodobnosti pro Y by měla být v souladu s tím, co pozorujeme, stojí v pozadí odhadů metodou maximální věrohodnosti.



Obrázek 3.5: Tři různé funkce normální hustoty pravděpodobnosti domu o rozloze 5000 čtverečních stop.

Jiný způsob motivace zavedení maximálně věrohodného odhadu je zobrazen na obrázku 3.5. Protože Y_i je z $N(\beta X_i, \sigma^2)$, různé hodnoty β spolu přinášejí různá rozdělení pro Y. Obrázek 3.5 znázorňuje tři funkce hustot normálního rozdělení pro různé

[9] V teorii pravděpodobnosti pracujeme s náhodnými veličinami, tj. např. naše Y, tak i s realizací této náhodné veličiny, což odpovídá našim pozorovaným hodnotám Y. Pro jednoduchost budeme používat stejné značení pro náhodnou veličinu i realizaci, neboť z kontextu je patrné, jestli hovoříme o náhodné veličině nebo její realizaci.

volby β a $X_i = 5000$. Tyto hustoty byly generovány pro $\beta = -0.010$, 0.012 a 0.040. Kterou z nich tedy brát jako odhad skutečné hodnoty β? Která z hustot tak nejlépe reprezentuje nejistotu spojenou s cenou domu o rozloze 5000 čtverečních stop? První funkce hustoty (pro $\beta = -0.010$) má koncentrovanou většinu své pravděpodobnosti do oblasti záporných cen domů (střední hodnota rozdělení je -50000 dolarů). To je zřejmě nesmyslný závěr. Třetí funkce hustoty pravděpodobnosti (pro $\beta = 0.040$) již tak nesmyslná není, nicméně většina pravděpodobnsoti je soustředěna do oblastí velmi vysokých cend domu (střední hodnota je v tomto případě 200000). Říká nám, že je takřka jisté, že cena domu bude větší než 150000 dolarů, což je v rozporu s tím, co pozorujeme v datech. Druhá funkce hustoty pravděpodobnosti (pro $\beta = 0.012$) je stejná jako ta z obrázku 3.1 a již jsme hovořili o tom, že se jedná o rozumný závěr, který koresponduje s našim pozorováním. Pokud tedy máme na výběr z těchto tří možností hodnot β, je hodnota 0.012 nejvěrohodnější. Hodnota $\beta = 0.012$ by tak měla být zvolena jako odhad. To je právě to, co dělá estimátor metody maximální věrohodnosi. „Podívá se“ na všechny možné hodnoty pro parametr β a vybere tu hodnotu, která vede k nejpravděpodobnějšímu (nejvěrohodnějšímu) rozdělení závisle proměnné.

Formálně v sobě odhad metodou maximální věrohodnosti obsahuje maximalizace *věrohodnostní funkce (likelihood function)*. Věrohodnostní funkce je funkce sdružené hustoty pravděpodobnosti pro všechna pozorování. Při platnosti klasických předpokladů máme k dispozici $Y_1, \ldots, Y_N$, které jsou všechny normální náhodné veličiny, každá se střední hodnotou βX_i a rozptylem σ^2, přičemž jsou všechny navzájem nekorelované. Označíme si funkci hustoty pravděpodobnosti každé náhodné veličiny jako $p(Y_i|X_i, \beta)$, kdy zahrnutí X_i, β jasně říká, že tato funkce hustoty závisí (je podmíněna) hodnotami vysvětlující proměnné a koeficientem β. Sdružená hustota pravděpodobnosti všech pozorování může být zapsána jako

$$p\left(Y_1, \ldots, Y_N\right) = \prod_{i=1}^{N} p\left(Y_i|X_i, \beta\right).$$

Tento výsledek vyplývá ze skutečnosti, že funkce sdružené hustoty pravděpodobnosti je prostým součinem jednotlivých dílčích hustot pravděpodobnosti, pokud jsou náhodnéveličiny navzájem nezávislé. Klasický předpoklad o vzájemné nekorelovanosti náhodných veličin znamená pro normální náhodné veličiny, že jsou rovněž i nezávisle (nezávislost je definována v příloze B). Jakmile do funkce sdružené hustoty pravděpdobnosti $p\left(Y_1, \ldots, Y_N\right)$ dosadíme pozorovaná data (např. pozorované ceny domů a hodnoty jejich rozlohy pro 546 domů z našeho příkladu), získáme věrohodnostní funkci, kterou označíme jako $L(\beta)$. Tuto funkci jsme zapsali jako funkci parametru β, abychom jasně zdůraznili, že závisí na parametru β, který chceme odhadnout.

Odhad metodou maximální věrohodnosti zahrnuje nalezení takové hodnoty β, která maximalizuje věrohodnostní funkci $L(\beta)$. Myšlenka stojící v pozadí principu maximální věrohodnosti je využitelná s celou řadou modelů, nejen regresních modelů. V případě linárního regresního modelu je odvození příslušného estimátoru snadné.

Vlastnost 5: Estimátor metody maximální věrohodnosti (ML estimátor) parametru β je identický jako OLS estimátor při splnění klasických předpokladů.

Důkaz: Využitím vztahu pro funkci normální hustoty pravděpodobnosti (viz příloha B) získáváme

$$p\left(Y_i|X_i\beta\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(Y_i - \beta X_i\right)^2\right].$$

Věrohodnostní funkce je součin této funkce hustoty pro každé pozorování. S využitím vlastností exponenciální funkce dostáváme

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(Y_i - \beta X_i\right)^2\right] \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - \beta X_i\right)^2\right]. \end{aligned}$$

Úkolem je je nelézt hodnotu β, která maximalizuje věrohodnostní funkci. To může vypadat složitě, ale dá se použít jednoduchý „trik“. Místo věrohodnostní funkce budeme pracovat s (přirozeným) logaritmem této věrohodnostní funkce. Pro většinu modelů se tím práce výrazně zjednodušší. Maximu logaritmu věrohodnostní funkce bude stejné jako maximum věrohodnostní funkce nelogaritmované.[10] V tomto případě je logaritmus věrohodnostní funkce, který označíme jako $l(\beta)$, následující:

$$\begin{aligned} l(\beta) &= \ln\left[L(\beta)\right] \\ &= \ln\left\{\frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - \beta X_i\right)^2\right]\right\} \\ &= \ln\left\{\frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}}\right\} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - \beta X_i\right)^2. \end{aligned}$$

S využitím diferenciálního počtu nalezneme hodnotu β, která maximalizuje $l(\beta)$. Ale i bez toho se můžeme v našem případě obejít. Podíváme-li se na výraz pro $l(\beta)$, vidíme, že β se vyskytuje pouze v členu $\sum_{i=1}^{N}\left(Y_i - \beta X_i\right)^2$. Vzhledem k zápornému členu, kterým jej násobíme (tj. $\frac{}{1}2\sigma^2$), musíme pro maximalizaci $l(\beta)$ minimalizovat člen $\sum_{i=1}^{N}\left(Y_i - \beta X_i\right)^2$. Ale tento člen není nic jiného, než součet čtverců náhodných chyb z kapitoly 2, který přejde do součtu čtverců reziduí (SSR), nahradíme-li parametr β jeho odhadem. Výsledkem minimalizace tohoto výrazu je OLS estimátor. Tím pádem ale říkáme, že i estimátor metody maximální věrohodnosti musí minimalizovat součet čtverců chyb. Tudíž oba estimátory musí být totožné. Platí tedy, že pro lineární regresní model a při splnění klasických předpokladů je ML estimátor roven OLS estimátoru. To je další důvod pro tvrzení, že OLS estimátor je dobrý estimátor koeficinetu β.

3.5 Odvození intervalu spolehlivosti pro parametr β

Intervaly spolehlivosti napomáhají zopovědět otázku přesnosti OLS odhadů. Intervaly spolehlivosti poskytují odhady intervalu, ve kterých s danou pravděpodobností leží sku-

[10] To vyplývá s obecného matematického pravidla, že hodnota x, která maximalizuje funkci $f(x)$ je stejná jako hodnota, která maximalizuje funkci $g\left[f(x)\right]$, kde $g[\cdot]$ je rostoucí funkce.

tečná hodnota parametru β. V kapitole 2 bylo ukázáno jak tyto intervaly prakticky interpretovat. Na tomto místě se zaměříme na jejich odvzoení. V prvním kroku budeme předpokládat, že rozptyl chyb, σ^2, je znám. Uvolnění tohoto předpokladu bude obsahem části 3.7. Postup konstrukce intervalu spolehlivost je vcelku jasny. Nejprve je třeba znát pravděpodobnostní rozdělení odhadu parametru, $\widehat{\beta}$. To závisí na skutečné hodnotě β. Na tomto základě se člen β osamostatní mezi dvě nerovnostní znaménka ohraničující kraje intervalu spolehlisovsti. Tento postup je obecný a týká se jakéhokoli modelu i parametru.

Odvození začíná využitím vlastnosti, že náhodná veličina $\widehat{\beta}$ má normální rozdělení se střední hodnotou β a specifickou hodnotou rozptylu. Na tomto základě je konstruován *Z-skór* (viz část 3.2):

$$Z = \frac{\widehat{\beta} - E(\widehat{\beta})}{\sqrt{var(\widehat{\beta})}} = \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum x_i^2}}}.$$

Náhodná veličina Z odpovídá standardizaci náhodné veličiny $\widehat{\beta}$, má tedy nulvoou střední hodnotu a jednotkový rozptyl, $Z \sim N(0,1)$. Na tomto základě stačí využít tabulky pro standardizované normální rozdělení, pro stanovení příslušného tvrzení o pravděpodobnosti realizace této náhodné veličiny, např.

$$\Pr\left[-1.96 \leq Z \leq 1.96\right] = 0.95.$$

Tento konfidenční interval je konstruován s pravděpodobností 95 % a jedná se tedy o 95% interval spolehlivosti. Pro jeho konstrukci stačí osamostatnit parametr β:

$$\begin{aligned}
&\Pr\left[-1.96 \leq \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum x_i^2}}} \leq 1.96\right] \\
&= \Pr\left[-1.96\sqrt{\frac{\sigma^2}{\sum x_i^2}} \leq \widehat{\beta} - \beta \leq 1.96\sqrt{\frac{\sigma^2}{\sum x_i^2}}\right] \\
&= \Pr\left[\widehat{\beta} - 1.96\sqrt{\frac{\sigma^2}{\sum x_i^2}} \leq \beta \leq \widehat{\beta} + 1.96\sqrt{\frac{\sigma^2}{\sum x_i^2}}\right] \\
&= 0.95.
\end{aligned}$$

Poznamenejme opět, že $\widehat{\beta}$ je náhodná veličina, nicméně β je dle klasické ekonometrické interpretace nenáhodná veličina.[11] Předchozí rovnici tak nelze interpretovat jako pravděpodobností výpověď o parametru β a nenazýváme tudíž tento interval jako „pravděpodobnostní interval“, ale jako „interval spolehlivosti“.

Máme tedy odvozen 95% interval spolehlivosti pro β při splnění klasických předpokladů a předpokladu, že známe σ^2 v podobě

$$\widehat{\beta} - 1.96\sqrt{\frac{\sigma^2}{\sum x_i^2}} \leq \beta \leq \widehat{\beta} + 1.96\sqrt{\frac{\sigma^2}{\sum x_i^2}}$$

[11] Jen pro zajímavost, v současnosti dochází k rozmachu Bayesiánská ekonometrie, která oproti klasické ekonometrii chápe β jako náhodnou veličinu.

Alternativní způsob vyjádření je

$$\widehat{\beta} \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_i^2}}$$

nebo

$$\left[\widehat{\beta} - 1.96\sqrt{\frac{\sigma^2}{\sum x_i^2}}, \widehat{\beta} + 1.96\sqrt{\frac{\sigma^2}{\sum x_i^2}}\right].$$

Hodnota $95\ \%$ je označována jako *hladina spolehlivosti (cofidence level)*. Konstrukci při jiných hladinách spolehlivosti provedene analogicky se stanovením jiných hodnot pravděpodobnosti (např. $90\ \%$) a tedy i jiných čísel ze statistických tabulek. Odvození 90% intervalu spolehlivosti probíhá stejným způsobem, tedy konstrukcí *Z-skóru*, kdy s využitím statistických tabulek můžeme psát $\Pr[-1.64 \leq Z \leq 1.64] = 0.95$.. Výsledkem bude stejný vzorec s rozdílem, že „1.96“ bude nahrazeno hodnotou „1.64“.

Pro praktické využití je tento vztah pro interval spolehlivosti obtížně použitelný, protože obvykle neznáme rozptyl chyb, σ^2. Ukážeme si však, jak jej lze nahradit příslušným odhadem a jak se změní výsledná podoba doposud odvozených vztahů. Nejprve se ale věnujme testování hypotézo parametru β.

3.6 Testování hypotéz o parametru β

I v tomto případě budeme nejprve předpokládat, že známe hodnotu rozptylu chyb, σ^2. Většina ekonometrických prgramů nám vrátí informaci týkající se testu hpotézy, že $\beta = 0$ a tomuto problému jsme se věnovali neformálně již v kapitole 2. V této části se zaměříme na formální podrobnosti.

Abychom si ukázali, že postup testování hypotéz je možno uplatnit v jakémkoli modelu, uvedeme si zde obecné kroky v jakémkoli testu hypotézy s konkrétním odvozením pro testování hypotéz o parametru β.

Krok 1. Specifikace nulové hypotézy, H_0 a alternativní hypotézy H_1.

Hypotézy zahrnující parametr β lze zapsat jako $H_0 : \beta = \beta_0$, kdy β_0 je známo (obvykle $\beta_0 = 0$). Alternativní hypotéza je obvykle $H_1 : \beta \neq \beta_0$, ale je možno uvažovati jednostrannou alternativu, což se projeví pouze v jiné kritické hodnotě.

Krok 2. Specifikace testové statistiky.

Pro hypotézu z předchozího kroku je obvyklá testová statistika

$$Z = \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum X_i^2}}}.$$

Krok 3. Nalezení rozdělení testové statistiky za předpokladu, že H_0 platí.

Při diskuzi v rámci odvození intervalu spolehlivosti jsme došli k závěru, že

$$Z = \frac{\widehat{\beta} - \beta_0}{\sqrt{\frac{\sigma^2}{\sum X_i^2}}} \sim N(0,1).$$

Krok 4. Volba hladiny významnosti.

Obvyklá volba je $5\ \%$ (tj. 0.05).

Krok 5. Na základě kroků 3 a 4 získáme kritickou hodnotu.

Protože je Z z $N(0,1)$ a $\Pr[-1.96 \leq Z \leq 1.96] = 0.95$, je kritická hodnota právě 1.96. Poznamenejme, že stanovení pravděpodobnosti 0.95 pro kritickou hodnotu odpovídá jedna mínus hladina významnosti. Protože se jedná o oboustranný test hypotézy (stejně jako oboustranný interval spolehlivosti), kritická hodnota odpovídá 97.5% (0.975) kvantilu standardizovaného normálního rozdělení, což je $1 - 0.05/2$. Při oboustranném testu musí pravděpodobnost realizace v intervalu mezi oběma kritickými hodnotami dát hodnotu $95\ \%$, což znamená, že oblast nad kritickou hodnotou odpovídá 2.5% pravděpodobnosti (a stejně tak i pod mínus kritickou hodnotou), což dohromady dává $5\ \%$.

Krok 6. Spočítáme testovou statistiku z kroku 2 a porovnáme ji s kritickou hodnotou z kroku 5. Zamítáme H_0 (ve prospěch alternativní hypotézy), pokud absolutní hodnota testové statistiky je větší než kritická hodnota (v opačném případě H_0 nezamítáme).

V našem případě zamítáme H_0, pokud $|Z| > 1.96$.

Stručně řečeno, H_0 nezamítáme v případě, kdy vypočtená hodnota testové statistiky je konzistentní stím, co by mělo být, pokud je H_0 platí. Formálně tedy odvozujeme rozdělení testové statistiky za předpokladu platnosti H_0. Díky tomu můžeme stanovit pravděpodobnost $\Pr[-1.96 \leq Z \leq 1.96] = 0.95$. Jinými slovy, pokud je H_0 pravdivá, můžeme si být s 95% pravděpodobností jistí, že Z bude ležet mezi -1.96 a 1.96. Pokud námi spočtená hodnota Z v tomto intervalu neleží, bereme to za jako důkaz v neprospěch H_0 a zamítáme hypotézu, že $\beta = \beta_0$.

3.7 Modifikace při neznámém rozptylu chyb σ^2

V praktických aplikací velmi zřídka známe hodnotu σ^2, a vzniká nám otázka, co dělat, když je rozptyl náhodných složek neznámý. Odpověď je vcelku jasná, nahradíme tento rozptyl jeho odhadem. Pokud to uděláme, změní se nám rozdělení testové statistiky.

Obvyklý estimátor pro σ^2 je výběrový rozptyl, s^2. Připomeňme si, že OLS rezidua jsou definována jako

$$\widehat{\epsilon}_i = Y_i - \widehat{\beta} X_i.$$

OLS estimátor σ^2 je dán jako

$$s^2 = \frac{\sum \widehat{\epsilon}_i^2}{N-1}.$$

Lze ukázat, že se jedná o nestranný estimátor a tedy

$$E\left(s^2\right) = \sigma^2.$$

Formální důkaz je trochu náročnější a nebudeme ho uvádět, nicméně intuice obtížná není. V rámci klasických předpokladů jsme předpokládali $E\left(\epsilon_i^2\right) = \sigma^2$, což naznačuje, že ϵ_i^2 by mohl být dobrý estimátor pro σ^2. Protože máme N různých chyb, proč je všechny v rámci estimátoru nevyužít? To nám naznačuje použít aritmetický průměr čtverců chyb jakožto estimátoru pro σ^2:

$$\frac{\sum \epsilon_i^2}{N}.$$

Bohužel, ϵ_i nejsou přímo pozorovatelné. Nicméně, můžeme nahradit tyto chyby regrese příslušnými rezidui, a získat estimátor v podobě

$$\frac{\sum \widehat{\epsilon}_i^2}{N}.$$

Ve skutečnosti lze ukázat, že toto je estimátor σ^2 metodou maximální věrohodnosti. Nevychýlený OLS estimátor je velmi podobný, jen N je nahrazeno $N-1$, což zajišťuje nestrannost estimátoru s^2. V rámci modelu vícenásobné regrese se vzorec pro s^2 ještě upraví. Pokud model obsahuje k vysvětlujících proměnných a úrovňovou konstantu, potom

$$s^2 = \frac{\sum \widehat{\epsilon}_i^2}{N-k-1} = \frac{SSR}{N-k-1}.$$

V rámci odvození intervalů spolehlivosti pro β při neznámém rozptylu σ^2 stačí nahradit σ^2 jeho odhadem s^2. Další odvození se nijak nezmění, pouze rozdělení estimátoru již není nadále normální. Změní se na Studentovo t-rozdělení. Toto rozdělení je vysvětleno v příloze B. Je velmi podobné normálnímu rozdělení. Podobně jako u normálního rozdělení, při konstrukci *Z-skóru* má výsledná veličina nulovou střední hodnotu a jednotkový rozptyl. Rozdílem oproti normálnímu rozdělení je to, že statistické tabulky Studentova t-rozdlění závisejí na tzv. *stupních volnosti (degrees of freedom)*. Pro naše potřeby stačí znát to, že je to snadno spočítatelná veličina, která nám řekne na jaký řádek tabuelk se pdoívat(respektive, co zadat to příslušné funkce v ekonometrickém programu).

Prvním krokem v odvození intervalu spolehlivosti bylo tvrzení, že

$$Z = \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{\sum X_i^2}}} \sim N(0,1).$$

Po nahrazení σ^2 výrazem s^2 získáváme

$$Z = \frac{\widehat{\beta} - \beta}{\sqrt{\frac{s^2}{\sum X_i^2}}} \sim t_{N-1},$$

kde t_{N-1} je Studentovo t-rozdělení s $N-1$ stupni volnosti.

Jako příklad předpokládejme, že $N = 22$. Pokud bychom znali σ^2, využili bychom tabulek k nalezení $\Pr[-1.96 \leq Z \leq 1.96] = 0.95$, z nichž odvodíme onterval spolehlivosti:

$$\widehat{\beta} \pm 1.96\sqrt{\frac{\sigma^2}{\sum x_i^2}}.$$

S neznámý rozptylem, σ^2, nahrazeným s^2 je třeba nalézt příslušný kvantil Studentova rozdělení s 21 stupní volnosti (opět, pro oboustranný interval spolehlivosti hledáme 97.5% kvantil). Díky tomu, můžeme psát $\Pr[-2.08 \leq Z \leq 2.08] = 0.95$, z čehož odvodíme interval spolehlivosti

$$\widehat{\beta} \pm 2.08\sqrt{\frac{s^2}{\sum x_i^2}}.$$

Analogicky řešíme problém testování hypotéz. Kritické hodnoty však musíme hledat v tabulkách Studentova t-rozdělení. Při známém σ^2 platilo

$$Z = \frac{\widehat{\beta} - \beta_0}{\sqrt{\frac{\sigma^2}{\sum X_i^2}}} \sim N(0,1).$$

Při neznámém rozptylu použijeme testovou statistiku

$$t = \frac{\widehat{\beta} - \beta_0}{\sqrt{\frac{s^2}{\sum X_i^2}}} \sim t_{N-1},$$

kde t_{N-1} je Studentovo t-rozdělení s $N-1$ stupni volnosti. V kapitole 2 jsme tuto statistiku nazývali t-statistikou (právě z důvodu, že má Studentovo t-rozdělení).

V našem příkladu tedy máme $N = 22$. Při známém σ^2jsme s použitím tabulek našli $\Pr[-1.96 \leq Z \leq 1.96] = 0.95$, díky čemuž máme kritickou hodnot 1.96, což využijeme v testování hypotézy na hladině významnosti $5\ \%$. Při neznámém rozptylu, σ^2, použijeme s^2 a z tabulek Studentova t-rozdělení s 21 stupní volnosti zjistíme, že $\Pr[-2.08 \leq t \leq 2.08] = 0.95$, tedy kritická hodnota je 2.08.

Tímto jsme dokončili výklad ohledně testování hypotézy $H_0 : \beta = \beta_0$. Většina relevantních počítačových programů prezentuje v rámci testování hypotéz kromě hodnot testových statistik i p-hodnoty. Obdržíme tedy i p-hodnotu pro $H_0 : \beta = 0$. Díky tomu nemusíme hledat kritické hodnoty ve statistických tabulkách. Připomeňme si, že p-hodnota je rovna nejmenší hladině významnosti při které můžeme zamítnout H_0. Pokud pracujeme s $5\ \%$ hladinou významnosti, můžeme zamítnout H_0 v případě, kdy je p-hodnota menší než 0.05.

3.8 Shrnutí

Na základě této kapitoly (a násedujících příloh) tedy již víme:

- jaké základní statistické metody a techniky jsou využívány v ekonometrii;

- jak ekonometři formulují pravděpodobnostní předpoklady o tom, jak jsou generována data;
- že estimátor je funkční předpis pro odhad např. neznámých parametrů modelu, a to na základě pozorovaných dat;
- významnými nástroji jsou operátory očekávané hodnoty a rozptylu, bez kterých se fakticky při odvozování požadovaných vlastností neobejdeme;
- že pro případ jednoduchého regresního modelu bez úrovňové konstanty je možno odvodit vcelku přehledným způsobem důelžité závěry bez použití maticové algebry (výsledky pro model vícenásobné regrese jsou intuitivně podobné, ale příslušná odvození jsou bez maticové algebry dosti náročné a nepřehledné);
- jakým způsobem je využíváno normální rozdělení (charakterizované střední hodnotou a rozptylem) v kontextu jednoduchého regresního modelu;
- jak jsou formulovány (pro případ jednoduché regrese) klasické předpoklady, na jejichž základě jsou odvozeny veškeré potřebné statistické vlastnosti OLS estimátoru a následné procedury testování hypotéz a konstrukce intervalů spolehlivosti;
- jaké jsou vlastnosti OLS estimátoru, což zahrnuje i důkaz toho, že se jedná o nestranný estimátor, kdy jeho pravděpodobnostní rozdělení odpovídá normálnímu rozdělení, tzn. $\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum X_i^2}\right)$;
- že nám Gaussův-Markovův teorém říká, že OLS je při splnění klasických předpokladů BLUE (nejlepší, lineární, nestranný estimátor);
- že existuje i odhad metodou maximální věrohodnosti a že lze dokázat, že estimátor metody maximální věrohodnosti parametru β zcela odpovídá OLS estimátoru;
- jak odvodit interval spolehlivosti pro parametr β za předpokladu známého rozptylu náhodných složek, σ^2;
- jak odvodit a interpretovat test hypotézy, že $\beta = 0$ při známém rozptylu náhodných složek, σ^2;
- jak vypadá OLS estimátor pro rozptyl náhodných složek, σ^2;
- jak modifikovat interval spolehlivosti a test hypotézy při neznámém rozptylu náhodných složek, σ^2;
- že pro regresní model existuje něco jako asymptotická teorie, kterou můžeme využít pro zkoumání vlastností estimátoru pro velikost vzorku jdoucí k nekonečnu, tedy $N \to \infty$;
- asymptotickou teorii lze využít k uvolnění předpokladu o normálním rozdělení náhodných složek a o nenáhodném charakteru vysvětlující proměnné;

- že za předpokladu nekorelovanosti vysvětlující proměnné s náhodnoou chybou (složkou), je odhad parametru, $\widehat{\beta}$ asymptoticky normální a tudíž i veškeré doposud odvozené intervaly spolehlivosti a postupy testování hypotéz zůstávají aproximativně v platnosti.

Měli bychom tak již znát a umět vysvětlit obsah následujících klíčových pojmů:

- estimátor
- odhad
- klasické předpoklady
- homoskedasticita
- heteroskedasticita
- autokorelace náhodných složek
- nestrannost estimátoru
- vydatnost estimátoru
- linearita estimátoru
- maximálně věrohodný odhad
- Gaussův-Markovův teorém
- asymptotická teorie
- konzistentní estimátor
- asymptotická normalita

Příloha 1: Důkaz Gaussova-Markovova teorému

Předpokládejme, že pracujeme s jednoduchým regresním modelem bez úrovňové konstanty a jsou splněny klasické předpoklady. Chceme najít estimátor s minimálním rozptylem mezi skupinou všech estimátorů, které jsou nestranné (nevychýlené) a lineární v závisle proměnné. Označme lineární a nestranný estimátor jako β^*. Protože se jedná o lineární estimátor v závisle proměnné, musí mít podobu

$$\beta^* = c_1 Y_1 + \ldots + c_N Y_N$$

pro konstanty $c_1, \ldots, c_N$. Protože je β^* nevychýlený, musí platit $E(\beta^*) = \beta$. S využitím vlastností operátoru střední hodnoty a prvního a pátého klasického předpokladu lze odvodit

$$\begin{aligned} E(\beta^*) &= E(c_1 Y_1 + \ldots + c_N Y_N) \\ &= c_1 E(Y_1) + \ldots + c_N E(Y_N) \\ &= c_1 \beta X_1 + \ldots + c_N \beta X_N \\ &= \beta \sum_{i=1}^{N} c_i X_i. \end{aligned}$$

Podíváme-li se na tento výraz, musí pro nestranný estimátor β^* platit $\sum_{i=1}^{N} c_i X_i = 1$.

Rozptyl β^* lze vypočítat s vyžitím vlastností operátoru rozptylu a druhého a třetího

z klasických předpokladů:

$$\begin{aligned} var\left(\beta^*\right) &= var\left(c_1 Y_1 + \ldots + c_N Y_N\right) \\ &= c_1^2 var\left(Y_1\right) + \ldots + c_N^2 var\left(Y_N\right) \\ &= c_1^2\sigma^2 + \ldots + c_N^2\sigma^2 \\ &= \sigma^2 \sum_{i=1}^{N} c_i^2. \end{aligned}$$

Využitím těchto výrazů je otázka nalezení lineárního nestranného estimátorů s minimální rozptyle vcelku jasný problém diferenciálního počtu. Chceme nalézt koeficienty $c_1, \ldots, c_N$, které budou minimalizovat $\sigma^2 \sum_{i=1}^{N} c_i^2$ při omezení $\sum_{i=1}^{N} c_i X_i = 1$. Jedná se tedy o problém minimalizace s omezením, který lze řešit standardním způsobem (např. s využitím metody Lagrangeových multiplikátorů). Řešením je

$$c_j = \frac{X_j}{\sum_{i=1}^{N} X_i^2},$$

pro $j = 1, \ldots, N$. Dosazením do výrazu pro β^* získáme

$$\begin{aligned} \beta^* &= c_1 Y_1 + \ldots + c_N Y_N \\ &= \frac{X_1 Y_1}{\sum X_i^2} + \ldots + \frac{X_N Y_N}{\sum X_i^2} \\ &= \frac{\sum X_i Y_i}{\sum X_i^2} = \widehat{\beta}. \end{aligned}$$

Jinými slovy, lineární, nestranný estimátor s minimálním rozptylem je estimátor metody nejmenších čtverců.

Příloha 2: Využití asymptotické teorie v jednoduchém regresním modelu

Jedním z klasických předpokladů je, že Y_i (nebo ekvivalentně ϵ_i) je normálně rozdělená náhodná veličina. V této příloze si ukážeme, jak lze využít asymptotickou teorii pro uvolnění tohoto předpokladu. Formálně provedeme veškerá odvození bez předpokladu normality náhodných chyb. Přesněji, nebudeme využívat žádné předpoklady pokud jde o funkci hustoty pravděpodobnosti chyb regresního modelu. V empirické praxi zřídka kdy víme, jestli jsou náhodné chyby náhodně rozděleny. Možnost abstrahovat od tohoto předpokladu je tak vcelku důležitá. Využijeme asymptotické metody, které jsou založeny na tom, co se stane, když se velikost vzorku bude blížit nekonečnu. Základní definice jsou obsahem přílohy B. Ukážeme si, že OLS estimátor je *konzistentní* a *asymptoticky normální*. Připomeňme si, že naše odvození intervalů spolehlivosti a testování hypotéz začínalo konstatováním, že $\widehat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum X_i^2})$. V této příloze dojdeme ke stejným závěrům, až na to, že budou platit aproximativně. Můžeme tak těchto závěrů

využít k aproximativnímu odvození intervalů spolehlivosti a testů hypotéz, a to v analogickém duchu, jak tomu bylo v této kapitole. Ačkoli tedy nepředpokládáme normalitu náhodných složek, dospějeme k identickým intervalům spolehlivosti a postupům testování hypotéz. Můžeme tedy říct, že předpoklad normality nehraje žádnou roli. Ovšem pozor, výsledky platí jen asymptoticky, tedy pro $N \to \infty$. V praxi samozřejmě nebudeme mít nekonečnou velikost vzorku, budeme mít $N = 50$, 100 nebo 1000. Z tohoto důvodu musí být veškeré závěry brány aproximativně (čím vyšší velikost vzorku, tím samozřejmě roste i přesnost výsledků).

Asymptotická teorie nám umožňuje uvolnit i další z klasickcýh předpokladů, a to ten, že X_i je nenáhodná veličina. V této příloze budeme předpokládat, že X_i je *nezávisle a identicky rozdělená* náhodná veličina *independent and identically distributed – i.i.d.*, která je nezávislá na náhodných složkách. Střední hodnotu X_i označíme jako μ_X a rozptyl X_i jako σ_X^2. Všechny ostatní klasické předpoklady stále platí.

V této kapitole jsme pracovali s OLS estimátorem v podobě

$$\widehat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2}$$

a odvodili jsme si jiný způsob jeho vyjádření v podobě

$$\widehat{\beta} = \beta + \frac{\sum X_i \epsilon_i}{\sum X_i^2}.$$

Tento výraz využijeme v v následujícím odvození.

Vlastnost 1: $\widehat{\beta}$ je konzistentní estimátor parametru β.

Důkaz:

$$\begin{aligned}
\operatorname{plim}\left(\widehat{\beta}\right) &= \operatorname{plim}\left(\beta + \frac{\sum X_i \epsilon_i}{\sum X_i^2}\right) \\
&= \beta + \operatorname{plim}\left(\frac{\sum X_i \epsilon_i}{\sum X_i^2}\right) \text{ dle Slutského teorému} \\
&= \beta + \operatorname{plim}\left(\frac{\frac{1}{N}\sum X_i \epsilon_i}{\frac{1}{N}\sum X_i^2}\right) \\
&= \beta + \frac{\operatorname{plim}\left(\frac{1}{N}\sum X_i \epsilon_i\right)}{\operatorname{plim}\left(\frac{1}{N}\sum X_i^2\right)} \text{ dle Slutského teorému.}
\end{aligned}$$

Nyní lze využít *zákon velkých čísel* (viz příloha B) k vyhodnocení limity v pravděpodobnosti $\operatorname{plim}\left(\frac{1}{N}\sum X_i \epsilon_i\right)$. Základní myšlenku zákona velkých čísel lze shrnout do tvrzení, že *„průměr konverguje ke střední hodnotě“*. Výraz podobný $\frac{1}{N}\sum X_i \epsilon_i$ odpovídá průměru. Lze ukázat, že podmínky použití zákona velkých čísel jsou pro proměnnou $X_i \epsilon_i$ splněny. Platí, že

$$\operatorname{plim}\left(\frac{1}{N}\sum X_i \epsilon_i\right) = E(X_i \epsilon_i) = 0,$$

což vyplývá z předpokladu o nezávislosti chyby regrese a vysvětlující proměnné. Protože nula dělená kladným číslem zůstává nulou ($\frac{1}{N}\sum X_i^2$ je kladné protože součet čverců je vždy nezáporný, pokud je v součtu alespoň jedna nenulová hodnota). Důkaz lze zakončit závěrem, že $\text{plim}(\widehat{\beta}) = \beta$. Pro úplnost si ale odvoďme výraz ve jmenovateli. Můžeme využít zákon velkých čísel pro tvrzení, že $\text{plim}\left(\frac{1}{N}\sum X_i^2\right) = E(X_i^2)$. Z definice rozptylu, $var(X_i) = E(X_i^2) - [E(X_i)]^2$, můžeme psát $E(X_i^2) = var(X_i) + [E(X_i)]^2 = \sigma_X^2 + \mu_X^2$. Tedy

$$\text{plim}\left(\widehat{\beta}\right) = \beta + \frac{0}{\sigma_X^2 + \mu_X^2} = \beta,$$

čímž je dokázána konzistence OLS estimátoru.

Vlastnost 2: OLS estimátor je asymptoticky normální. Pro $N \to \infty$ platí

$$\sqrt{N}(\widehat{\beta} - \beta) \sim N\left(0, \frac{\sigma^2}{\sigma_X^2 + \mu_X^2}\right).$$

Důkaz:
Rovnice (*) může být přepsána jako

$$\sqrt{N}(\widehat{\beta} - \beta) = \sqrt{N}\frac{\sum X_i\epsilon_i}{\sum X_i^2} = \sqrt{N}\frac{\frac{1}{N}\sum X_i\epsilon_i}{\frac{1}{N}\sum X_i^2}.$$

Čitatel i jmenovatel jsme přenásobili zlomkem $\frac{1}{N}$, což samozřejmě nijak celý výraz nezmění. Cítatel a jmenovatel je však díky tomu přepsán do podoby průměrů a můžeme tak využít nástroje asymptotické teorie. Centrální limitní větu (viz příloha B) lze využít pro tvrzení, že čitatel výrazu je asymptoticky normální. Tzn., že pro $N \to \infty$,

$$\sqrt{N}\frac{1}{N}\sum X_i\epsilon_i \sim N\left(0, var(X_i\epsilon_i)\right).$$

Využitím definice rozptylu, vlastností operátoru střední hodnoty (stále předpokládáme, že X_i a ϵ_i jsou nezávislé), skutečnosti, že náhodné chyby mají nulovou střední hodnotu, a s využitím předchozího odvození $E(X_i^2) = \sigma_X^2 + \mu_X^2$ dostáváme

$$\begin{aligned} var(X_i\epsilon_i) &= E(X_i^2\epsilon_i^2) - [E(X_i\epsilon_i)]^2 \\ &= E(X_i^2)E(\epsilon_i^2) - [E(X_i)E(\epsilon_i)]^2 \\ &= (\sigma_X^2 + \mu_X^2)\sigma^2 - [\mu_X 0]^2 \\ &= (\sigma_X^2 + \mu_X^2)\sigma^2. \end{aligned}$$

Ukázali jsme si rovněž, že

$$\text{plim}\left(\frac{1}{N}\sum X_i^2\right) = (\sigma_X^2 + \mu_X^2).$$

S využitím Cramerova teorému můžeme zkombinovat výsledky centrální limitní věty s výrazy pro $var\,(X_i\epsilon_i)$ a plim $\left(\frac{1}{N}\sum X_i^2\right)$, čímž dostáváme

$$\sqrt{N}(\widehat{\beta}-\beta) \xrightarrow{N} N\left(0, \frac{(\sigma_X^2+\mu_X^2)\sigma^2}{(\sigma_X^2+\mu_X^2)^2}\right).$$

Vzájemným vykrácením členů v části rozptylu normálního rozdělení dostáváme

$$\sqrt{N}(\widehat{\beta}-\beta) \xrightarrow{N} N\left(0, \frac{\sigma^2}{(\sigma_X^2+\mu_X^2)}\right)$$

a vlastnost 2 je dokázána.

Využití asymptotických výsledků v praxi

Vlastnost 2 nám říká, že pro $N \to \infty$ konverguje $\sqrt{N}(\widehat{\beta}-\beta)$ k normálnímu rozdělení. Pokud je N velké, budou tyto výsledky platit aproximativně. Využitím vlastností operátorů střední hodnoty a rozptylu, můžeme vztah z vlastnosti 2 převést do aproximativní podoby:

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{N(\sigma_X^2+\mu_X^2)}\right).$$

Problém je, že v praxi výraz $(\sigma_X^2+\mu_X^2)$ neznáme. Víme ale, že $\left(\frac{1}{N}\sum X_i^2\right) = \sigma_X^2+\mu_X^2$. Z tohoto důvodu je $\left(\frac{1}{N}\sum X_i^2\right)$ konzistentní estimátor pro $\sigma_X^2+\mu_X^2$. Dosazením tohoto estimátoru získáme aproximativní výsledek

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum X_i^2}\right).$$

To je ale výsledek, z kterého jsme vycházeli při odvozování intervalů spolehlivosti a testů hypotéz při splnění klasických předpokladů! Všechny výsledky z velé této kapitoly tak platí (aproximativně) i v tomto případě a nemusíme zde tato odvození opakovat. Místo toho si zopakujme hlavní závěry: pokud uvolníme předpoklad o normalitě náhodných složek, získáme identické intervaly spolehlivosti i testy hypotéz, které byly odvozeny při splnění všech klasických předpokladů (v tomto případě ale výsledky platí aproximativně).

Kapitola 4

Lineární regresní model více vysvětlujících proměnných

V této kapitole se dozvíme:

- ☞ další podrobnosti týkající se problému zkreslení při nezahrnutí důležité vysvětlující proměnné či multikolinearity, zejména s ohledem na formálnější důkaz důsledku opomenutí těchto problémů;
- ☞ jak modifikovat koeficient determinace, R^2, do podoby korigovaného koeficientu determinace, který zohledňuje počet použitých vysvětlujících proměnných;
- ☞ že existují další postupy testování hypotéz, speciálně vhodné pro model vícenásobné regrese;
- ☞ jak testovat hypotézy zahrnující kombinaci více regresních koeficientů;
- ☞ že jeden z testů zahrnuje využití F-statistiky a je určen právě pro testování vícenásobných hypotéz o regresních koeficientech;
- ☞ že existují testy založené na věrohodnostním poměru, které je možno aplikovat mnohem šířeji, než jen na test hypotéz zahrnující kombinaci regresních parametrů (na tomto místě si je ale ukážeme v kontextu regresního modelu);
- ☞ na čem je založen test věrohodnostního poměru, Wladův test a test Lagrangeových multiplikátorů;
- ☞ jak volit odpovídající funkční podobu regresního modelu;
- ☞ že nelineární vztahy mezi proměnnými lze modelovat v kontextu standardní lineární regrese, a to vhodnou nelineární transformací původních vysvětlujících proměnných.

4.1 Úvod

V kapitole 2 jsme si zavedli model vícenásobné regrese s k vysvětlujícími proměnnými v podobě

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i,$$

kde i je index označující jednotlivá pozorování, kdy $i = 1, \ldots, N$. V kapitole 2 jsme se zaměřili na interpretaci výsledků, které nám dávají ekonometrické počítačové programy. Probrali jsme interpretaci regresníchkoeficientů, intervalů spolehlivost, postup testování hypotéz a koeficient determinace, R^2, jakožto měřítko kvality vyrovnání.

V této kapitole se zaměříme na rigoróznější výklad témat diskutovaných v kapitole 2. Důvod je ten, že formální přístup umožňuje detailnější proniknutí do dané problematiky a je také možné zabývat se problémy, které bez využití matematiky pokrýt nelze. Konkrétně půjde o problematiku multikolinearity a zkreslení při nezahrnutí podstatně vysvětlující proměnné. V modelu vícenásobné regrese se nám nabízí širší paleta možností testování hypotéz. Popíšeme si dva přístupy k testování hypotéz. První z nich zahrnuje F-statistiku a jeužitečný při testování vícenásobných hypotéz oregresních koeficientech. Druhý přístup je založen na testu věrohodnostních poměrů a lze jej využít rovněž pro test vícenásobných hypotéz o regresních koeficientech, nicméně možnosti jeho využití jsou mnohem širší (je aplikovatelný i v jiných typech modoelů, než je lineární regresní model). V této kapitole se krátce zmíníme i o problému volby vhodné funkční formy regrese.

4.2 Model vícenásobné regrese – základní výsledky

V kapitole 3 byly odvozeny teoretické výsledky na základě splnění klasických předpokladů. Tyto předpoklady jsou v zásadě totožné i v případě modelu vícenásobné regrese

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i.$$

Klasické předpoklady jsou následující:

1. $E(\epsilon_i) = 0$. Nulová střední hodnota náhodných složek.
2. $var(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$. Konstantní rozptyl náhodných složek (homoskedasticita).
3. $cov(\epsilon_i, \epsilon_j) = 0$ pro $i \neq j$. ϵ_i a ϵ_j jsou vzájemně nekorelované.
4. ϵ_i má normální rozdělení.
5. $X_{1i}, \ldots, X_{ki}$ jsou pevně daná, jedná se o nenáhodné veličina.

Vlastnosti OLS estimátoru jsou stejné jako v případě jednoduché regrese. OLS estimátor je nestranný a intervaly spolehlivosti a testy hypotéz jsou odvozovány podobně. Gaussův-Markovův teorém stále říká, že při splnění klasických předpokladů je OLS estimátor nejlepší lineární nevychýlený estimátor. Výsledné vzorce jsou bez použití

maticové algebry poněkud složitější a méně přehledné, intuice stojící v pozadí a jejich interpretace se však nemění.

Pro ilustraci uvažujme model s úrovňovou konstantou a dvěma vysvětlujícími proměnnými

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

OLS estimátor získáme minimalizací součtu čtverců reziduí. Výsledkem je

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\left(\sum x_{1i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{1i} \sum x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}, \\
\widehat{\beta}_2 &= \frac{\left(\sum x_{2i} y_i\right)\left(\sum x_{1i}^2\right) - \left(\sum x_{1i} y_i\right)\left(\sum x_{1i} \sum x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}, \\
\widehat{\alpha} &= \overline{Y} - \widehat{\beta}_1 \overline{X}_1 - \widehat{\beta}_2 \overline{X}_2,
\end{aligned}$$

kde proměnné s čárou nahoře označují příslušné výběrové střední hodnoty (např. $\overline{X}_2 = \frac{\sum X_{2i}}{N}$) a malými písmeny jsou označovány odchylky od odpovídajících průměrů, tedy

$$\begin{aligned}
y_i &= Y_i - \overline{Y}, \\
x_{1i} &= X_{1i} - \overline{X}_1, \\
x_{2i} &= X_{2i} - \overline{X}_2.
\end{aligned}$$

Už jen pro dvě vysvěltující proměnné jsou vzorce bez použití maticového vyjádření poněkud komplikované. NIcméně, interpretace koeficientů zůstává jednoduchá, tzn. *β_j vyjadřuje mezní vliv j-té vysvětlující proměnné na Y, při neměnných ostatních vysvětlujících proměnných*. Intervaly spolehlivost a postupy testování hypotéz rovněž zůstávají nezměněny, povze příslušné vzorce mají drobné odlišnosti. Jejich interpretace však zůstává zachována. Lze ukázat, že OLS estimátor je nestranný (tj. $E\left(\widehat{\beta}_j\right) = \beta_j$ pro $j = 1, 2, 3$), atd. Z těchto důvodů se k příslušným odvozením nebudeme vracet.

Pokud jde o ony drobné odlišnosti. Nestranný estimátor pro rozptyl náhodných složek, σ^2, je

$$s^2 = \frac{\sum \widehat{\epsilon}_i^2}{N - k - 1},$$

kde

$$\widehat{\epsilon}_i = Y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{1i} - \ldots - -\widehat{\beta}_k X_{ki}$$

jsou OLS rezidua.

V případě $k = 2$ můžeme rozptyl OLS odhadů zapsat jako

$$\begin{aligned}
var\left(\widehat{\beta}_1\right) &= \frac{\sigma^2}{(1 - r^2)\sum x_{1i}^2}, \\
var\left(\widehat{\beta}_2\right) &= \frac{\sigma^2}{(1 - r^2)\sum x_{2i}^2},
\end{aligned}$$

kde r je (výběrový) koeficient korelace mezi X_1 a X_2. V praxi ve většině případů nahrazujeme σ^2 příslušným odhadem. Rozptyl odhadu parametrů využíváme při konstrukci intervalů spolehlivosti a v rámc itestováníhypotéz. Jako příklad si uvedeme postup využití t-testu pro testování hypotézy $\beta_2 = 0$ pro případ $k = 2$. V případě, že σ^2 známe, jsou kroky následující:

Krok 1. Specifikace nulové hypotézy H_0 a alternativní hypotézy H_1.

V našem příkladu je $H_0 : \beta_2 = 0$ a $H_1 : \beta_2 \neq 0$.

Krok 2. Specifikace testové statistiky.

Pro hypotézu z kroku 1 je obvyklou testovou statistikou

$$Z = \frac{\widehat{\beta}_2 - \beta_2}{\sqrt{var\left(\widehat{\beta}_2\right)}} = \frac{\widehat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{(1-r^2)\sum x_{2i}^2}}}.$$

Krok 3. Specifikace rozdělení testové statistiky za předpokladu platnosti nulové hypotézy.

S využitím odvození analogických těm z kapitoly 3

$$Z = \frac{\widehat{\beta}_2}{\sqrt{\frac{\sigma^2}{(1-r^2)\sum x_{2i}^2}}} \sim N(0,1).$$

Krok 4. Volba hladiny významnosti.

Provedeme obvyklou volbu $5\ \%$ (0.05).

Krok 5. Využitím kroků 3 a 4 získáme kritickou hodnotu.

Protože Z odpovídá $N(0,1)$ a $\Pr[-1.96 \leq Z \leq 1.96] = 0.95$, je kritická hodnota 1.96. Hladina spolehlivosti je v tomto případě 0.95, což je 1 mínus námi zvolená hladina významnosti 0.05.

Krok 6. Výpočet testové statistiky z kroku 2 a její porovnání s kritickou hodnotou z kroku 5. H_0 zamítáme v případě, kdy je absolutní hodnota testové statistiky větší než kritická hodnota (v opačném případě nezamítáme).

V našem příkladu zamítáme H_0 pokud $|Z| > 1.96$.

V případě, že je rozptyl, σ^2, neznámý, nahrazujeme jej jeho odhadem, s^2. V takové případě má testová statistika Studentovo t-rozdělení. Rovnice z kroku 3, tedy je

$$t = \frac{\widehat{\beta}_2}{\sqrt{\frac{\sigma^2}{(1-r^2)\sum x_{2i}^2}}} \sim t_{N-k-1},$$

kde t_{N-k-1} je Studentovo t-rozdělení s $N - k - 1$ stupni volnosti. Veškeré kroky odvození jsou tedy identické jako v případě jednoduché regrese.

V kapitole 2 bylo uvedeno oblíbené měřítko kvality modelu s hlediska jeho souladu s daty v podobě koeficientu determinace, R^2. V případě vícenásobné regrese je jeho definice stejná:

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum \widehat{\epsilon}_i^2}{\sum \left(Y_i - \overline{Y}\right)^2},$$

a lze jej interpretovat jako podil variability závisle proměnné, které je vysvětlena (variabilitou či chováním) vysvětlujících proměnných.

S použitím koeficientu determinace v modelu vícenásobné regrese vzniká několik problémů, které si vyžadují zavedení jeho modifikované verze. Tato modifikovaná verze se nazývá *korigovaný koeficient determinace* a označuje se jako $\overline{R}^2$. Problémem koeficientu determinace je to, že s přidáním další vysvětlující proměnné do modelu vždy zvýší (resp. nikdy nesníží) jeho hodnotu, a to i v případě nevýznamnosti této nové vysvětlující proměnné (respektive příslušného koeficientu). Proč tomu tak je? Koeficient determinace měří kvalitu vyrovnání. S přidánímdalší vysvětlující proměnné tak neexistuje možnost, že by nové vyrovnání mohlo být horší než vyrovnání před přidáním této proměnné. V regresi může být v extrémním případě koeficient nově přidané proměnné nulový, čímž zůstane původní koeficient determinace nezměněn. Obecně tedy dochází po přidání nové vysvětlující proměnné vždy k lepšímu (nebo nejhůře stejnému) vyrovnání než před jejím přidáním. Formálně, metoda nejmenších čtverců hledá hodnoty koeficientů, které vedou k minimalizaci součtu čtverců reziduí, SSR. Přidáním nové vysvětlující proměnné dostává metoda OLS novou dimenzi v rámci které může dále minimalizovat SSR, což znamená, že SSR se sníží a R^2 vzroste.

Jedním ze způsobů, jak rozhodnout, která vysvětlující proměnná by měla být zahrnuta v regresi, je využití postupu testování hypotéz. Díky němu z regrese vyloučíme nevýznamné proměnné a zahrneme proměnné, u kterých je příslušný parametr statisticky významný. Bylo by lákavé využít k tomuto rozhodování i koeficient determinace, R^2, a zahrnout tak ty proměnné, které zvyšují jeho hodnotu a tím pádem ikvalitu vyrovnání dat modelem. Předchozí odstavec nám však napověděl, že toto není dobrá strategie, neboť bychom tímto zahrnuli do regrese i proměnné bez statisticky významného vlivu na chování vysvětlované veličiny. Korigovaný koeficient determinace, $\overline{R}^2$ však tímto nedostatkem netrpí a můžeme ho mnohdy k řešení problému zahrnutí té či oné vysvěltující proměnné využít. Korigovaný koeficient determinace s přídáním další vysvěltující proměnné vždy nevzroste. Pokud ale vzroste, znamená to, že si můžeme být téměř jisti, že tato vysvětlující proměnné by měla být v reresi přítomná. Pokud máme dva regresní modely se stejnouzávisle proměnnou, ale s odlišnými vysvětlujícími proměnnými, potom model s vyšší hodnotou $\overline{R}^2$ je ten lepší.

Korigovaný koeficient determinace, $\overline{R}^2$, je definován jako

$$\overline{R}^2 = 1 - \frac{\frac{SSR}{N-k-1}}{\frac{TSS}{N-1}} = 1 - \frac{s^2}{\frac{1}{N-1}\sum \left(Y_i - \overline{Y}\right)^2}.$$

Poznamenejme, že s^2 je odhad rozptylu náhodných složek, σ^2, a $\frac{1}{N-1}\sum \left(Y_i - \overline{Y}\right)^2$ je výběrový rozptyl. Korigovaný koeficient determinace tak zahrnuje část, která měří velikost rozptylu náhdoných složek relativně k variabilitě vysvětlované proměnné. Má tedy

podobnou motivace, která stojí v pozadí definice koeficientu determinace, R^2. Nelze jej však interpretovat jednoduše tak, že odpovídá podílu variability závisle proměnné, kterou lze vysvětlit chováním vysvětlujících proměnných. Lze ukázat, že s přidáním nové vysvětlující proměnné nemusí vzrůst a lze jej tak využít pro účely volby mezi různými modely.

Abychom si shrnuli dosavadní poznatky, máme k dispozici dvě měřítka kvality vyrovnání, které se obvykle v praxi využívají, koeficient determinace (R^2) a jeho korigovanou variantu ($\overline{R}^2$). Koeficient fdeterminace má zcela zřejmou interpretaci v podobě měřítka míry variability závisle proměnné, která je vysvětlena chováním vysvětlujících proměnných. Nelze jej však využít k rozhodnutí, který z „konkurenčních" modelů vybrat. Korigovaný koeficient determinace nemá tak pěknou a jednoznačnou interprataci, nicméně jej lze využít k rozhodování o tom, který z modelů (vysvětlujících chování stejné veličiny) je lepší.

4.3 Otázky volby vysvětlujících proměnných

V kapitole 2 jsme se zabývali otázkou jaké vysvětlující proměnné zahrnout nebo nezahrnout do našehoregresního modelu. Na jedné straně zde byla snaha zahrnout co možná nejvíce vysvětlujících proměnných, který by byly schopnyvysvětlit chování závisle proměnné, na druhé straně stála skutečnost, že zahrnutí pnerelevantních proměnných (tedy těch bez síly vysvětlit chování závisle proměnné) vede k méně přesným výsledkům odhadu. Tento druhý požadavek vedl k závěru o zahrnutí co možná nejmenšího počtu proměnných. V našem rozhodování však s úspěchem využíváme postupy testování hypotéz, které nám pomohou zodpovědět otázku, která z vysvětlujících proměnných je či není statisticky významné (prostřednictvím statistické významnosti nebo nevýznamnosti odpovídajícího regresního koeficientu).

V této části se zaměříme na formálnější diskuzi problému zkreslení při nezahrnutí důležité vysvětlující proměnné a při zahrnutí irelevantnívysvětlující proměnné. Blížesi rozebereme i problematiku multikolinearity vysvětlujících proměnných.

4.3.1 Problém nezahrnutí relevantní vysvětlujících proměnných

abychom si ukázali, co se stane, pokud opomeneme zahrnout relevantní vysvětlující proměnnou, budeme předpokládat, že skutečný model má podobu regresního modelu s dvěma vysvětlujícími proměnnými:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,$$

přičemž uvažujeme splnění klasických předpokladů. Výsledky z předchozí části kapitoly ukazovaly, že správný OLS odhad parametru β_1 je

$$\widehat{\beta}_1 = \frac{\left(\sum x_{1i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{1i} \sum x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2},$$

kdy malými písmenky označené proměnné jsou odpovídající odchylky od průměrů.

Předpokládejme, že jsme chybně opomněli zařadit do regrese druhou vysvěltující proměnnou a pracovali jsme tak s modelem

$$Y_i = \alpha + \beta_1 X_{1i} + \epsilon_i.$$

Jednoduchým rozšířením odvození z kapitoly 3 lze ukázat, že OLS odhad parametru β_1 je

$$\tilde{\beta}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2},$$

a tento estimátor parametru β_1 označíme jako $\tilde{\beta}_1$, abychom si ho odlišili od správného OLS estimátoru, $\widehat{\beta}_1$. Je zřejmé, že oba estimátory se od sebe odlišují, což už je varovným signálem, že opomenutím důležité vysvětlující proměnné děláme něco špatně. Ve skutečnosti je $\tilde{\beta}_1$ vychýlený estimátor (není již tedy nestranný).

Abychom si tuto skutečnost ukázali, použijeme několik pomocných výpočtů. Nejprve poznamenejme, že

$$\begin{aligned}\overline{Y} &= \frac{\sum Y_i}{N} = \frac{\sum (\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i)}{N} \\ &= \alpha + \beta_1 \overline{X}_1 + \beta_2 \overline{X}_2 + \bar{\epsilon}.\end{aligned}$$

Tuto rovnici můžeme využít k vyjádření

$$\begin{aligned}y_i &= Y_i - \overline{Y} \\ &= (\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i) - \left(\alpha + \beta_1 \overline{X}_1 + \beta_2 \overline{X}_2 + \bar{\epsilon}\right) \\ &= \beta_1 x_{1i} + \beta_2 x_{2i} + (\epsilon_i - \bar{\epsilon}).\end{aligned}$$

Tento výraz pro y_i můžeme dosadit do vztahu pro $\tilde{\beta}_1$:

$$\begin{aligned}\tilde{\beta}_1 &= \frac{\sum x_{1i} (\beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i - \bar{\epsilon})}{\sum x_{1i}^2} \\ &= \frac{\beta_1 \sum x_{1i}^2}{\sum x_{1i}^2} + \frac{\beta_2 \sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i} (\epsilon_i - \bar{\epsilon})}{\sum x_{1i}^2} \\ &= \beta_1 + + \frac{\beta_2 \sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i} (\epsilon_i - \bar{\epsilon})}{\sum x_{1i}^2}.\end{aligned}$$

Tvrzení o vychýlenosti estimátoru, $\tilde{\beta}_1$ snadno ukážeme využitím operátoru střední hodnoty pro obě strany rovnice, tedy

$$\begin{aligned}E\left(\tilde{\beta}_1\right) &= E\left(\beta_1 + \frac{\beta_2 \sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i} (\epsilon_i - \bar{\epsilon})}{\sum x_{1i}^2}\right) \\ &= \beta_1 + \frac{\beta_2 \sum x_{1i} x_{2i}}{\sum x_{1i}^2},\end{aligned}$$

kdy odvození využívá vlastnosti operátoru střední hodnoty, kdy střední hodnota konstanty je konstanta (z klasických předpokladů vyplývá nenáhodnost vysvětlujících proměnných), a střední hodnota náhodných složek je nulová. Platí tedy, že $E\left(\tilde{\beta}_1\right) \neq \beta_1$

a pokud tedy nezhrneme do regrese relevantní vysvětlující proměnnou, dostaneme vychýlený odhad koeficientu zahrnuté vysvětlující proměnné. Tuto skutečnost označujeme jako *zkreslení při nezahrnutí relevantních vysvětlujících proměnných (omitted variables bias)*.

Z předchozího výrazu vidíme, že toto zkreslení nenastává v případě, kdy je $\beta_2 = 0$ nebo $\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2}$. První případ nás moc nemusí zajímat, protože pokud je $\beta_2 = 0$, potom X_2 není obsažena ve skutečné regresní rovnici a tudíž žádná důležitá vysvětlující proměnné nebyla opomenuta. Druhý případ je zajímavější. Výraz $\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2}$ je úzce spojen s korelací mezi X_1 a X_2, kterou označíme jako r. Pokud se podíváme na výraz pro korelaci z kapitoly 1, můžeme vidět, že $\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2} = 0$ pokud je $r = 0$. Zkreslení při nezahrnutí důležité proměnné tedy nenastává v případě, pokud je nezahrnutá vysvětlující proměnná nekorelována se zahrnutou vysvětlující proměnnou. Tento závěr jsme vyslovili již v kapitole 2, ale na tomto místě jsme si to formálně dokázali.

Poznamenejme ještě, že $\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2} > 0$ pokud je $r > 0$ a $\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2} < 0$ pokud je $r < 0$. Na tomto základě můžeme hovořit o směru zkreslení. Pokud je $\beta_2 > 0$, můžeme říct, že *„pokud je opomenutá vysvětlující proměnná pozitivně korelována se zahrnutou vysvětlující proměnnou, bude odhad koeficientu zahrnuté proměnné zkreslen směrem nahoru“*. Pokud je $\beta_2 < 0$, potom lze říct, že *„pokud je opomenutá vysvětlující proměnná pozitivně korelována se zahrnutou vysvětlující proměnnou, bude odhad koeficientu zahrnuté proměnné zkreslen směrem dolů“*.

Protože ale skutečnou hodnotu β_2 neznáme, musíme být s tímto typem tvrzení opatrní. Nicméně, tyto poznatky s úspěchem využijeme při interpretaci empirických výsledků. Předpokládejme příklad s cenami domů, kdy závisle proměnná byla cena domu a vysvětlující proměnné byly příslušné charakteristiky domu včetně jeho celkové rozlohy. Mezi důležité charakteristiky domu určitě patří jeho umístění. Nechť $X_1 =$ rozloha domu a $X_2 =$ atraktivita polohy. Předpokládejme, že nemáme data o X_2 (což je v empirických studiích tohoto typu běžné). V tomto případě pravděpodobně nastane problém zkreslení při nezahrnutí relevantní proměnné, a to díky nepřítomnosti „atraktivity polohy“, která určitě je důležitá pro určení ceny domu. Předpokládejme, že její vliv na cenu domu je pozitivní ($\beta_2 > 0$). Budeme rovněž předpokládat, že atraktivita polohy je pozitivně svázána s rozlohou domu (atraktivní poloha je spjata s velkými domy na velkých pozemcích) a tedy $\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2} > 0$. Této skutečnosti můžeme využít k tvrzení, že parametr u X_1 bude vychýlen (zkreslen) směrem nahoru. Je třeba aale znovu upozornit, že argumenty o směru zkreslení závisí na korektnosti analýzy ohledně předpokládaného znaménka skutečné hodnoty parametru a na korelaci nezahrnuté proměnné s proměnnou (proměnnými) v regresi zahrnutými.

4.3.2 Zahrnutí irelevantních vysvětlujících proměnných

Předpokládejme nyní opačnou situaci, kdy skutečný model bude mít podobu

$$Y_i = \alpha + \beta_1 X_{1i} + \epsilon_i.$$

Budeme opět předpokládat platnost klasických předpokladů. Náš model však budeme nyní chybně specifikovat s irelevantní proměnnou v podobě

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

Využijeme tedy estimátor

$$\tilde{\beta}_1 = \frac{\left(\sum x_{1i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{1i} \sum x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}$$

místo korektního estimátoru

$$\widehat{\beta}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2}.$$

Gaussův-Markovův teorém nám říká, že $\widehat{\beta}_1$ je nejlepší lineární nestranný estimátor. Má tedy menší variabilitu než jakýkoli jiný nestranný estimátor. Pokud tedy ukážeme, že $\tilde{\beta}_1$ je nestranný, potom s odkazem na Gaussův-Markovův teorém můžeme říct, že $var\left(\tilde{\beta}_1\right) > var\left(\widehat{\beta}_1\right)$, což dokazuje, že zahrnutí irelevantní vysvětlující proměnné vede k méně přesným odhadům.

Je tedy důležité dokázat nestrannost $\tilde{\beta}_1$. Stejně jako v předchozí části si vyjádříme y_i, tnetokrát však jako

$$y_i = \beta_1 x_{1i} + \left(\epsilon_i - \overline{\epsilon}\right).$$

Tento výraz dosadíme do vztahu pro $\tilde{\beta}_1$, čímž dostaneme

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\left(\sum x_{1i}\left[\beta_1 x_{1i} + \left(\epsilon_i - \overline{\epsilon}\right)\right]\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i}\left[\beta_1 x_{1i} + \left(\epsilon_i - \overline{\epsilon}\right)\right]\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2} \\
&= \frac{\beta_1\left[\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2\right]}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2} \\
&\quad + \frac{\left(\sum x_{1i}\left(\epsilon_i - \overline{\epsilon}\right)\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i}\left(\epsilon_i - \overline{\epsilon}\right)\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2} \\
&= \beta_1 + \frac{\left(\sum x_{1i}\left(\epsilon_i - \overline{\epsilon}\right)\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i}\left(\epsilon_i - \overline{\epsilon}\right)\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}.
\end{aligned}$$

Aplikací operátoru střední hodnoty na obě strany této rovnice získáme

$$\begin{aligned}
E\left(\tilde{\beta}_1\right) &= E\left[\beta_1 + \frac{\left(\sum x_{1i}\left(\epsilon_i - \overline{\epsilon}\right)\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i}\left(\epsilon_i - \overline{\epsilon}\right)\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}\right] \\
&= \beta_1 + E\left[\frac{\left(\sum x_{1i}\left(\epsilon_i - \overline{\epsilon}\right)\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i}\left(\epsilon_i - \overline{\epsilon}\right)\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}\right] \\
&= \beta_1,
\end{aligned}$$

kdy poslední výraz lze odvodit s využitím skutečnosti, že střední hodnota náhodných složek je nulová a vysvětlující proměnné jsou brány jako fixní, nenáhodné veličiny.

Zahrnutí irelevantní proměnné tedy nezpůsobí vychýlenost OLS odhadu, nicméně na základě Gaussova-Markovova teorému budou výsledné odhady méně přesné, než by mohly být. OLS estímator tak ztrácí na své vydatnosti.

Závěrem předchozí části bylo poselství, že bychom se měli snažit do regrese zahrnout všechny ty vysvětlující proměnné, které mohou ovlivňovat chování vysvětlované proměnné. Závěr této části je, že bychom neměl v regresi nechávat irelevantní-proměnné, neboť to snižuje přesnost odhadů všech koeficientů (tedy i těch, které jsou relevantní). V praxi tedy začínáme s co možná největším počtem vysvětlujících proměnných a na základě příslušného testování hypotéz o statistické významnosti parametrů postupně vyřazujeme ty proměnné, které jsou irelevantní, tedy ty, které nemají patřičnou vysvětlující sílu pro vysvětlení chování závisle proměnné veličiny. Samozřejmě, v rámci postupného vyřazování opakujeme příslušný odhad a testy hypotéz o statistické významnosti parametrů.

4.3.3 Multikolinearita

Posledním problémem volby vysvětlujících proměnných je multikolinearita. Intuitivně o ní byla zmínka v kapitole 2. Multikolinearita nastává v případě, kdy jsou vysvětlující proměnné navzájem silně korelovány. Volně řečeno, pokud jsou dvě proměnné silně korelovány, nesou v sobě zhruba tutéž informaci. OLS estimátor tak má problém v odhadu dvou oddělených mezních vlivů pro dvě takto silně korelované proměnné. Příslušné jednotlivé koeficienty jsou tak nepřesně odhadnuty, a to dokonce i v případě, kdy obě vysvětlující proměnné mohou mít společně velkou vysvětlující sílu. Obvyklým řešením problému multikolinearity je vypuštění jedné z vyscoe korelovaných vysvětlujících proměnných. V této části si tedy ukážeme další technické detaily týkající se problémumultikolinearity.

V části 4.2 byly ukázány vztahy pro rozptyly OLS estimátorů v modelu vícenásobné regrese se dvěma vysvětlujícími proměnnými:

$$var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{(1-r^2)\sum x_{1i}^2},$$
$$var\left(\widehat{\beta}_2\right) = \frac{\sigma^2}{(1-r^2)\sum x_{2i}^2}.$$

Tyto vztahy vstupují do odvození intervalů spolehlivosti a do postupů testování hypotéz. Všimněme si, že zde vystupuje korelační koeficient, r. V extrémním případě perfektní multikolinearity ($r = 1$ nebo $r = -1$) tak tyto rozptyly nejsme schopni vypočítat (vyžadovalo by to dělení nulou). Ve skutečnosti by nebylo možno v tomto případě vypočítat ani odhady příslušných parametrů, neboť i zde by nastal případ dělení nulou, a v příslušném ekonometrické programu bychom obdrželi chybové hlášení.

Jen zřídka se setkáváme s případem perfektní multikolinearity. Častější je případ vysoké (ale ne dokonalé) vzájemné korelace vysětlujících proměnných. Koeficient korelace, r je tak blízký hodnotám 1 nebo -1. V tomto případě je výraz $(1-r^2)$ blízký nule a rozptyly odhadů parametrů jsou tak obrovské. V případě modelu vícenásobné

regrese získáváme podobné vztahy týkající se odvození intervalů spolehlivosti i testování hypotéz o statistické významnosti parametrů jako v případě jednoduché regrese. Například t-statistika pro test $H_0 : \beta_2 = 0$ je

$$t = \frac{\widehat{\beta}_2}{\sqrt{var\left(\widehat{\beta}_2\right)}}.$$

Podíváme-li se na tento výraz, ukazuje se, že v případě silné multikolinearity jsou rozptyly odhadů parametrů ($\widehat{\beta_1}$ a $\widehat{\beta_2}$) veliké, tím pádem t-statistiky budou malé a intervaly spolehlivosti široké. Formálně jsme si tak ukázali, že multikolinearita vede k nepřesným odhadům rozptylů odhadů parametrů a testy statistické významnosti nám v těchto případech budou ukazovat na nevýznamnost parametrů β_1 a β_2.

Multikolinearita však nebude mít vliv na koeficient determinace, R^2 a tím pádem i na kvalitu vyrovnání. Vztah pro R^2 je

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{(N-k-1)s^2}{TSS}.$$

Korelace mezi dvěma vysvětlujícími proměnnými do tohoto vztahu nevstupuje. Regrese tak může dobře vystihnout chování dat (s^2 tak může být relativně malé, vzhledem k celkové variabilitě vysvětlované proměnné) i v případě přítomnosti multikolinearity.

Důsledkem multikolinearity tedy je to, že některé nebo dokonce všechny vysvětlující proměnné budou nevýznamné, ačkoli model bude dobře vysvětlovat chování vysvětlované proměnné (koeficient determinace bude vysoký). Tyto závěr platí i pro model s více než dvěma vysvětlujícími proměnnými. Existují sofistikované metody pro testování přítomnosti multikolinearity, nicméně v praxi je zcela dostačující a adekvátní metoda prozkoumání korelační matice vysvětlujících proměnnách. Pokud je korelace mezi dvěma našimi vysvětlujícími proměnnými příliš vysoká, máme zde problém multikolinearity. Co je ale myšleno „příliš vysokou“ korelací? Neexistuje nějaké pevné pravidlo, nicméně dobrým vodítkem může být to, že pokud zjistíme korelaci $|r| > 0.9$ pro nějaké dvě vysvětlující proměnné, máme zde co do činění s multikolinearitou. Řešení pak obvykle spočívá ve vypuštění jedné z proměnných, které ji způsobují.

4.4 Testování hypotéz v modelu vícenásobné regrese

V této části se vrátíme k obecnému modelu vícenásobné regrese v podobě

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i.$$

Stále předpokládáme, že jsou splněny klasické předpoklady. Jak jsme viděli, testování statistické významnosti jednoho parametru (tedy testování toho, jestli jedna vysvětlující proměnná má dostatečnou vysvětlující sílu) provádíme pomocí t-testu hypotézy $H_0 : \beta_j = 0$ (kde β_j označuje některý z koeficientů). V řadě situací nás však může zajímat test hypotézy, který zahrnuje více než jeden parametr. Zde si ukážeme dva přístupy. První přísutp, pomocí F-testů, je vhodný pro test hypotéz zahrnující jakýkoliv

počet lineárních kombinací regresních koeficientů. Druhý přístup, který využívá testy věrohodnostních poměrů, může dělat totéž, nicméně, lze jej využít i pro testy nelineárních omezení, a to i v rámci jiných než regresních modelů.

4.4.1 F-testy

V kapitole 2 byl ukázán test hypotézy o nulové hodnotě koeficientu determinace, $R^2 = 0$, pro ověření toho, jestli vysvětlující proměnné společně statisticky významně vysvětlují chování závisle proměné. Tento test je ekvivalentní testu hypotézy

$$H_0 : \beta_1 = \ldots = \beta_k = 0.$$

Tato hypotéza zahrnuje řadu omezení (tzn. $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$ až $\beta_k = 0$.), spadá tedy do kontextu diskutovaném v této části. Je třeba zdůraznit, že testování hypotézy $H_0 : \beta_1 = \ldots = \beta_k = 0$ není totéž co testování k samostatných hypotéz $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$ až $H_0 : \beta_k = 0$.

V kapitole 2 byla zavedena F-statistika pro testování této hypotézy. Pro model vícenásobné regrese s k vysvětlujícími proměnnými a úrovňovou konstantou má podobu

$$F = \frac{R^2}{1 - R^2} \frac{N - k - 1}{k}.$$

V rámci testování pracujeme s rozdělením testové statistiky za předpokladu, že je nulová hypotéza pravdivá. Na tomto základě získáme kritickou hodnotu testu z odpovídajícího rozdělení. V tomto případě má F-statistika rozdělení $F_{k,N-k-1}$. Statistické tabulky F-rozdělení jsou součástí většiny statistických a ekonometrických učebnic a lze tak s nimi snadno pracovat. Podobně lze kritické hodnoty získat i v rámci ekonometrických či statistických programů. Tyto programy obvykle poskytují p-hodnoty příslušných statistických testů. Pokud je tato hodnota menší než námi stanovená hladina významnosti (obvykle 0.05), zamítáme nulovou hypotézu, H_0, na odpovídající hladině významnosti (obvykle $5\ \%$).

Toto je příklad F-testu. Formálně lze F-testy využít k testování jakékoliv hypotézy, kterou lze zapsat jako lineární kombinaci regresních koeficientů. Pro ilustraci si uvedeme příklad modelu vícenásobné regrese se třemi vysvětlujícími proměnnými. Původní regresní model budeme označovat jako *neomezený model (unrestricted model)*. Regresní model se zahrnutím restrikcí vyplývajících z formulované hypotézy pak označíme jako *omezený model (restricted model)*. Neomezený model je tak v našem případě model

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i.$$

Příkladem hypotézy, kterou můžeme testovat s využitím F-testu je

$$H_0 : \beta_1 = \beta_2 = 0.$$

Tato hypotéza v sobě obsahuje dvě omezení, konkrétně $\beta_1 = 0$ a $\beta_2 = 0$, což jsou lineární funkce regresních parametrů. Jakoukoliv lineární funkci regresních koeficientů lze zapsat jako $a\beta_1 + b\beta_2 + c\beta_3 = d$ pro nějaké konstanty a, b, c a d. Nulové omezení typu $\beta_2 = 0$ je lineární funkce, kdy $a = c = d = 0$ a $b = 1$.

Výsledný omezený model má podobu

$$Y_i = \alpha + \beta_3 X_{3i} + \epsilon_i.$$

F-testy lze využít k testování mnohem obecnějších hypotéz, jako např.

$$H_0 : \beta_1 = 0, \quad \beta_2 + \beta_3 = 1.$$

Druhé z omezení lze zapsat jako $\beta_2 = 1 - \beta_3$. Omezený model tak bude mít podobu

$$Y_i - X_{2i} = \alpha + \beta_3 (X_{3i} - X_{2i}) + \epsilon_i.$$

Tento omezený model je jednoduchý regresní model se závisle proměnnou $Y - X_2$, úrovňovou konstantou a vysvětlující proměnnou $(X_{3i} - X_{2i})$.

Obecně lze pro jakoukoli množinu lineárních omezení kladených na neomezenh model implementovat do nového omezeného modelu, který bude stále lineární regresní model, ovšem s jinou závisle proměnnou a (nebo) jinými vysvětlujícími proměnnými. Pro testování těchto hypotéz lze použít následující testovou statistiku:

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(N-k-1)},$$

kde SSR je nám dobře známý součet čtverců reziduí, kdy dolní index UR a R rozlišují mezi součtem čtverců reziduí „neomezeného (unrestricted)“ a „omezeného (restricted)“ regresního modelu. Počet testovaných omezení je q (např. $q = 2$ v příkladu výše). Protože platí, že $SSR_R > SSR_{UR}$, je statistika F kladná (model s méně omezeními může vždy dosáhnout nižšího součtu čtverců reziduí, SSR). Velké hodnoty F naznačují, že H_0 není korektní. Pro specifikaci toho, co myslíme „velkou“ hodnotou F pro zamítnutí hypotézy H_0 musíme specifikovat pravděpodobnostní rozdělení této statistiky a na tomto základě pak zjistit kritickou hodnotu. Postup je tedy identický s jakýmkoliv jiným testem hypotéz. SPočítáme testovou statistiku (v našem případě F) a porovnáme ji s odpovídající kritickou hodnotou. Pokud je F větší než kritická hodnota, zamítáme H_0 (v opačném případě H_0 nezamítáme). Lze odvodit, že F má Fischerovo-Snedecerovo rozdělení, $F_{q,N-k-1}$. Jinými slovy, kritické hodnoty získáme z F-rozdělení s q stupni volnost v čitateli a $N - k - 1$ stupni volnosti ve jmenovateli. V praxi pak obvykle využijeme p-hodnotu poskytovanou ekonometrickými programovými balíčky.

F-statistika je někdy zapisována pomoci koeficientů determinace, R^2, neomezeného a omezeného modelu:

$$F = \frac{\left(R^2_{UR} - R^2_R\right)/q}{\left(1 - R^2_{UR}\right)/(N-k-1)}.$$

Tento výraz lze však korektně použít jen v případě, kdy jsou závisle proměnné neomezeného a omezeného modelu stejné. To znamená, že toto vyjádření F-statistiky nelze využít např. v případě testování omezení $\beta_1 = 0$, $\beta_2 + \beta_3 = 1$.

4.4.2 Testy věrohodnostních poměrů

Myšlenku maximálně věrohodného odhadu jsme si zavedli v rámci diskuze nad tím, proč je OLS estimátor dobrý estimátor. Přístup s využitím maximální věrohodnosti lze použít i v rámci testování hypotéz. *Test věrohodnostního poměru (likelihood ratio test)* je poněkud komplikovanější než F-test nebo t-test, má však jednu velkou výhodu. Testy věrohodnostního poměru lze využít v řadě jiných situacích než jen pro testování omezení kladených na regresní koeficienty. To je rozdíl oproti F-testům. Stejně tak je lze využít i mimo rámec regresních modelů. V této části se však zaměříme na jeho aplikaci pro model vícenásobné regrese.

V kapitole 3 byla odvozena věrohodnostní funkce jednoduchého regresního modelu. Tuto funkci lze samozřejmě zobecnit i pro případ vícenásobné regrese do podoby:

$$\begin{aligned} & L\left(\alpha, \beta_1, \ldots, \beta_k, \sigma^2\right) \\ & = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(Y_i - \alpha - \beta_1 X_{1i} - \ldots - \beta_k X_{ki}\right)^2\right] \\ & = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - \alpha - \beta_1 X_{1i} - \ldots - \beta_k X_{ki}\right)^2\right]. \end{aligned}$$

Analogickým způsobem jako v kapitole 3 lze ukázat, že estimátor regresních koeficientů *metodou maximální věrohodnosti (Maximum Likelihood Estimator – MLE)* odpovídá estimátoru metodou nejmenších čtverců (tyto odhady označíme postupně jako $\widehat{\alpha}$, $\widehat{\beta}_1$,..., $\widehat{\beta}_k$) a MLE pro rozptyl σ^2 je

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{\sum\left(Y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{1i}, \ldots, \widehat{\beta}_k X_{ki}\right)^2}{N} \\ &= \frac{\sum \widehat{\epsilon}_i^2}{N}. \end{aligned}$$

Všimněme si, že odhad rozptylu náhodných složek již není nestranný (pro konečný počet pozorování). Opět budeme rozlišovatmezi omezeným a neomezeným modelem. Pro zjednodušení značení budeme používat horní index U pro označení odhadů neomezeného modelu a R pro označení odhadů omezeného modelu. V tomto případě je

$$L\left(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U, \widehat{\sigma}^{2U}\right)$$

hodnota věrohodnostní funkce vyhodnocená v neomezených maximálně věrohodných (ML) odhadech.

Předpokládejme, že testované hypotézy zahnrují omezení regresních koeficientů. V předchozí části jsme viděli, že pokud pracujeme s lineárními omezeními koeficientů, můžeme omezený model přepsat do podoby nového regresního modelů s odlišnou vysvětlovanou proměnnou a (nebo) odlišnými vysvětlujícími proměnnými. Aplikace metody nejmenších čtverců na tento nový regresní model nám poskytne maximálně věrohodné odhady parametrů tohoto omezeného modelu. Věrohodnostní funkce vyhodnocená v odhadech omezeného modelu metodou maximální věrohodnosti je

$$L\left(\widehat{\alpha}^R, \widehat{\beta}_1^R, \ldots, \widehat{\beta}_k^R, \widehat{\sigma}^{2R}\right).$$

Pro ilustraci předpokládejme regresní model se třemi vysvětlujícími proměnnými a předpokládejme dále, že chceme testovat hypotézu

$$H_0 : \beta_1 = 0, \quad \beta_2 + \beta_3 = 1.$$

V předchozí části jsme si ukázali, že zohledněním omezení z nulové hypotézy získáme omezený model v podobě

$$Y_i - X_{2i} = \alpha + \beta_3 (X_{3i} - X_{2i}) + \epsilon_i.$$

OLS odhady tohoto jednoduchého regresního modelu nám vrátí hodnoty $\widehat{\alpha}^R$ a $\widehat{\beta}_3^R$. Jaké jsou ale hodnoty $\widehat{\beta}_1^R$ a $\widehat{\beta}_2^R$? K jejich výpočtu využijeme omezení plynoucí z H_0, tedy $\widehat{\beta}_1^R$ a $\widehat{\beta}_2^R = 1 - \widehat{\beta}_3^R$. Za předpokladu lineárních omezení koeficientů může získat ML odhady omezeného modelu díky použití metody nejmenších čtverců pro odpovídající transformovaný model. Testy věrohodnostního poměru můžeme aplikovat i pro hypotézy zahrnující nelineární restrikce, jako např. $H_0 : \beta_1 = \beta_2^3$,, $\beta_3 = \frac{1}{\beta_2}$, nebo zcela obecně $H_0 : g(\beta_1, \ldots, \beta_k) = 0$, kde $g(\cdot)$ je množina až k nelineárních funkcí. S nelineárními omezeními koeficientů však regresní model není nadále lineární. Techniky odhadu nelineárního regresního modelu budou krátce diskutovány v následující části této kapitoly. Nicméně, ekonometrické programy umožňují provést odhady nelineárních regresních modelů vcelku snadným způsobem.

Nyní již víme, jak získat ML odhady omezeného a neomezeného modelu. K testování nulové hypotézy zahrnující lienární restrikce regresních koeficientů tedy potřebujeme zvolit testovou statistiku její rozdělení za předpokladu správnosti této nulové hypotézy. Na tomto základě je pak odvozena kritická hodnota či p-hodnota daného testu. *Věrohodnostní poměr (likelihoo ratio)* je definován jako

$$\lambda = \frac{L\left(\widehat{\alpha}^R, \widehat{\beta}_1^R, \ldots, \widehat{\beta}_k^R, \widehat{\sigma}^{2R}\right)}{L\left(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U, \widehat{\sigma}^{2U}\right)}.$$

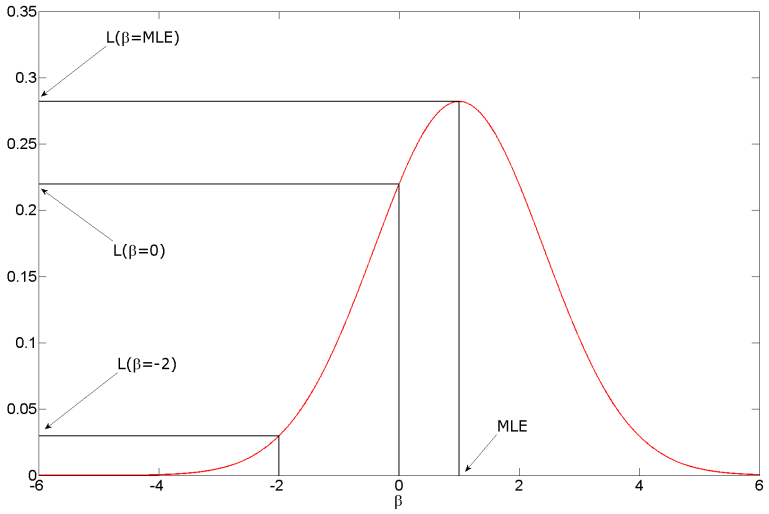
Příslušná testová statistika je $-2\ln(\lambda)$. Rozdělení této statistiky je aproximativně chí-kvadrát , χ^2:

$$-2\ln(\lambda) \sim \chi_q^2,$$

kde q je počet omezení obsažených v H_0. O aproximativním rozdělení hovoříme z toho důvodu, že toto rozdělení je přesné jen pro nekonečnou velikost vzorku (pozorování). Přístup s využitím věrohodnostního poměru je velmi obecný a lze jej použít v jakékoli třídě modelů (tedy nejen v rámci regresních modelů). Statistika $-2\ln(\lambda)$ má totiž aproximativně chí-kvadrát rozdělení v jakémkoliv případě, kdy testujeme hypotézy omezující nějakým způsobem náš model. V 4.6 této kapitoly jsou uvedeny další dva testy využívající věrohodnostních poměrů, konkrétně, *Waldův test* a *test Lagrangeových multiplikátorů*. Jejich odvození a vysvětlení je o něco málo obtížnější než u námi diskutovaného testu věrohodnostního poměru.

Nebudeme řešit důkaz toho, proč má $-2\ln\lambda$ aproximativně χ^2-rozdělení. Zaměříme se však na intuitivní vysvětlení toho, proč je tento test rozumný a jak ho lze využít v praxi. Základní myšlenka testování založeném na věrohodnostním poměru

je ta, že zavedení restrikcí vede k nižší hodnotě věrohodnostní funkce. Odhad metodou maximální věrohodnosti zahrnuje maximalizaci funkce. Z matematického hlediska jsme schopni nalézt vyšší (nebo alespoň stejnou) hodnotu maxima v případě, kdy můžeme hledat maximum na množině všech možných hodnot příslušných parametrů, než v případě, kdy si jejich hodnoty omezíme. Z tohoto důvodu bude vždy platit $L\left(\widehat{\alpha}^R, \widehat{\beta}_1^R, \ldots, \widehat{\beta}_k^R, \widehat{\sigma}^{2R}\right) \leq L\left(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U, \widehat{\sigma}^{2U}\right)$ a tedy $0 \leq \lambda \leq 1$. Pokud je však nulová hypotéza, H_0, pravdivá, měla by nastat situace, že λ bude velmi blízko 1 a tedy testová statistika $-2\ln(\lambda)$ by měla být malá. Na druhé straně, pokud jsou omezení nekorektní, vede jejich zavedení k razantnímu snížení věrohodnostní funkce a λ by měla být velmi malá a statistika $-2\ln(\lambda)$ bude naopak velká. Jako v případě každé testové statistika nám kritická hodnota rozhodne o tom, jaké hodnoty testové statistiky je možno pokládat za „dostatečně velké“ k zamítnutí nulové hypotézy. V našem případě hledáme kritické hodnoty v tabulkách pro rozdělení χ^2.



Obrázek 4.1: Věrohodnostní funkce.

Předchozí odstavec si můžeme ilustrovat na příkladu jednoduchého regresního modelu se známým rozptylem, bez úrovňové konstanty a jediným koeficientem, β. Neomezená věrohodnostní funkce je tak $L(\beta)$. Předpokládejme test hypotézy $H_0 : \beta = 0$. Za platnosti této hypotézy je omezená věrohodnostní funkce $L(\beta = 0)$. Obrázek 4.1 ukazuje příklad této věrohodnostní funkce (jedná se o věrohodnostní funkci odpovídající funkci normální hustoty pravděpodobnosti se střední hodnotou 1 a rozptylem 2, nicméně pro naši ilustraci to nemá žádný význam). ML odhad je hodnota parametru β která přináší nejvyšší hodnotu věrohodnostní funkce. Leží tedy vrcholem křivky a je označen jako MLE. Hodnota věrohodnostní funkce v tomto bodě je označena jako $L(\beta = MLE)$. Z definice se jedná o nejvyšší hodnotu, jakou může věrohodnostní funkce nabýt. V bodě $\beta = 0$ je její hodnota nižší. Na obrázku je označen jako

$L(\beta = 0)$. Věrohodnostní poměr je

$$\lambda = \frac{L(\beta = 0)}{L(\beta = MLE)}.$$

I kdyby byla skutečná hodnota parametru β nulová, nemůžeme očekávat, že MLE bude přesně nulový, protože odhad téměř nikdy nebude to co skutečná (neznámá) hodnota parametru. V takovém případě nám nestačí pro zamítnutí hypotézy $H_0 : \beta = 0$ zjištění, že ML odhad je nenulový. V tomto přípdě je $\lambda = 0.773$ a testová statistika $-2\ln(\lambda) = 0.515$. Protože v rámci H_0 uvažujeme jedno omezení, musíme hledat kritickou hodnotu našeho testu v rozdělení χ^2_1, tedy s jedním stupněm volnosti. Použijeme-li 5% hladinu významnosti, kritická hodnota je 3.84 (uvažujeme v tomto případě jednostranný test a tudíž hodnota 3.84 odpovídá 95% kvantilu daného rozdělení). Protože testová hodnota je menší než kritická, $0.515 < 3.84$, nezamítáme nulovou hypotézu, že $\beta = 0$.

Předpokládejme nyní hypotézu $H_0 : \beta = -2$. V tomto případě je věrohodnostní funkce omezeného modelu $L(\beta = -2)$ a toto značení je i na obrázku 4.1. Hodnota věrohodnostní funkce je zde mnohem nižší než hodnot v MLE. V tomto případě je věrohodnostní poměr

$$\lambda = \frac{L(\beta = -2)}{L(\beta = MLE)} = \frac{0.031}{0.282} = 0.110.$$

Testová statistika je $-2\ln(\lambda) = 4.416$. Protože je na hladině významnosti $5\ \%$ kritická hodnota rovna 3.84, můžeme zamítnout nulovou hypotézu, že $\beta = -2$.

Při výpočtu věrohodnostního poměru vždy přímo vyhodnocujeme věrohodnostní funkci v ML odhadech pro neomezený a omezený model. Pro případ modelu vícenásobné regrese lze věrohodnostní poměr zapsat v jednodušší podobě. Obecně lze věrohodnostní funkci modelu vícenásobné regrese vyhodnocenou v ML odhadech omezené nebo neomezené varianty zapsat v podobě

$$\begin{aligned} &L\left(\widehat{\alpha}, \widehat{\beta}_1, \ldots, \widehat{\beta}_k, \widehat{\sigma}^2\right) \\ &= \frac{1}{(2\pi\widehat{\sigma}^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^{N}\left(Y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{1i} - \ldots - \widehat{\beta}_k X_{ki}\right)^2\right]. \end{aligned}$$

Pokud však využijeme výraz pro ML odhad rozptylu, $\widehat{\sigma^2}$, vidíme, že se nám celý člen v hranatých závorkách zredukuje na konstantu $-\frac{N}{2}$. Příslušná mocnina se nám tak v každém věrohodnostním poměru vykrátí (bude vždy stejná), podobně to je i s dalším konstantním členem, $\frac{1}{(2\pi)^{\frac{N}{2}}}$. Jediná část věrohodnostní funkce, která je v rámci výpočtu věrohodnostního poměru jedinečná a důležitá je

$$\begin{aligned} L\left(\widehat{\alpha}, \widehat{\beta}_1, \ldots, \widehat{\beta}_k, \widehat{\sigma}^2\right) &\propto \frac{1}{(\widehat{\sigma}^2)^{\frac{N}{2}}} \\ &\propto \frac{1}{(SSR)^{\frac{N}{2}}}, \end{aligned}$$

kde samozřejmě

$$SSR = \sum \hat{\epsilon}_i^2$$

je součet čtverců reziduí. Symbol $\propto$ čteme jako *„je proporcionální"*. Pokud dále budeme pomocí horních indexů rozlišovat neomezený (U) a omezený model (R) můžeme věrohodnostní poměr zapsat jako:

$$\lambda = \frac{\frac{1}{(SSR^R)^{\frac{N}{2}}}}{\frac{1}{(SSR^U)^{\frac{N}{2}}}} = \left(\frac{SSR^U}{SSR^R}\right)^{\frac{N}{2}}.$$

Věrohodnostní poměr tak lze spočítat s využitím součtu čtverců reziduí neomezeného a omezeného regresního modelu. Protože již víme, jak využít metodu nejmenších čtverců k odhadu omezeného a neomezeného regresního modelu, snadno tak získáme i příslušné součty čtverců reziduí. Je rovněž zřejmé, jak jednoduchý je v tomto případě výpočet v rámci ekonometrických programů.

4.5 Volba funkční podoby modelu vícenásobné regrese

4.5.1 Nelinearita v regresi

Doposud jsme pracovali s lineární podobou modelu vícenásobné regrese:

$$Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \epsilon_i.$$

Někdy je však žádoucí uvažovat o vztahu mezi vysvětlovanou proměnnou a vysvětlujícími proměnnými v nelineární podobě, tedy

$$Y_i = f\left(X_{1i}, \ldots, X_{ki}, \alpha, \beta_1, \ldots, \beta_k\right) + \epsilon_i,$$

kde $f(\cdot)$ je nějaká nelineární funkce vysvětlujících proměnných a parametrů. Pro označení parametrů použijeme stejná řecká písmena jako doposud. V tomto případě však jejich interpretace může být jiná než v rámci lineárního regresního modelu a nemusí tak označovat úrovňovou konstantu či sklony (mezní vlivy).

K odhadu nelineárních regresních modelů se využívá metoda maximální věrohodnosti. Pokud budeme předpokládat, že rezidua splňují klasické předpoklady, můžeme věrohodnostní funkci psát ve tvaru

$$L\left(\alpha, \beta_1, \ldots, \beta_k\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(Y_i - f\left(X_{1i}, \ldots, X_{ki}, \alpha, \beta_1, \ldots, \beta_k\right)\right)^2\right].$$

Maximálně věrohodné estimátory obdržíme maximalizací této funkce. Oproti lineárnímu regresnímu modelu však tyto estimátory nelze obecně získat v algebraickém vyjádření. To znamená, že po vyjádření prvních parciálních derivací, jejich položením rovným nule a řešením nezískáme snadno vyhodnotitelné řešení pro $\widehat{\alpha}, \widehat{\beta_1}, \ldots, \widehat{\beta_k}$. V

tomto případě se využívá k nalezení příslušných odhadů numerických optimalizačních algoritmů. Většina ekonometrických balíčků však odhad nelineárních regresních modelů zvládne.

V řadě případů dokážeme nelineární funkci transformovat na lineární a využít tak techniky lineární regrese. Jedinou podmínkou tak je, aby regresní funkcebyla lineární v parametrech (nikoli ve vysvětlujících proměnných). Jako příklad si uveďmě Cobb-Douglasovu produkční funkci, která vyjadřuje závislost výstupu, Y, na zapojení různých vstupů, $X_1, \ldots, X_k$, a to následovně:

$$Y_i = \alpha_1 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \ldots X_{ki}^{\beta_k}.$$

Pokud zlogaritmujeme obě strany rovnice, získáme

$$\ln(Y_i) = \alpha + \beta_1 \ln(X_{1i}) + \ldots + \beta_k \ln(X_{ki}),$$

kde $\alpha = \ln(\alpha_1)$. Přidáním náhodné složky získáme lineární regresní model, pouze vysvětlovaná proměnní je $\ln(Y)$ a vysvětlující proměnné jsou $\ln(X_1), \ldots, \ln(X_k)$. Můžeme tak snadno provést regresi logaritmu vysvětlované na logaritmy vysvěltujících proměnných a využít všechny doposud diskutované závěry modelu vícenásobné regrese. Pokud náhodné složky splňují klasické předpoklady, říká nám Gaussův-Markovův teorém, že OLS estimátor je BLUE, intervaly spolehlivosti lze konstruovat standardním způsobem využitím dříve prezentovaných vztahů a stejně tak je možno provádět testování hypotéz. Tato regrese je často nazývána jako *log-lineární*, neboť je lineární v logaritmech proměnných.

Jeden z důvodů, proč používáme logaritmy proměnných je to, že zde existuje jednoduchá interpretace regresních koeficientů. Připomeňme si, že regresní koeficienty lineárního regresního modelu můžeme interpretovat způsobem *„jestliže se X_j zvýší o jednotku, potom Y má tendenci zvýšit se o β_j jednotek (za předpokladu, že se hodnoty ostatních vysvětlujících proměnných se nemění)“*. Interpretace regresních koeficientů probíhala v jednotkách odpovídajících jednotkám vysvětlované a vysvětlujících proměnných (např. dolary, tuny apod.). Pokud však jak závisle proměnná, tak i j-tá vysvětlující proměnná je vyjádřena v logaritmech, nehrají zde jednotky žádnou roli. Příslušné koeficienty jsou interpretovány jako *elasticity*, tedy *jestliže se X_j zvýší o jedno procento, potom má Y tendenci zvýšit se o β_j procent (za předpokladu, že se hodnoty ostatních vysvětlujících proměnných nemění)“*. Místo o změnách jednotek hovoříme o procentní změně.

Logaritmování proměnných je obvyklý způsob jejich transformace, nicméně je si zde třeba dávat pozor, protože logaritmus nuly a záporných čísel není definová. Tomu se je třeba vyvarovat. Není však problém, pokud jedna část proměnných bude v rovnici uváděna v logaritmech a jiná část ne. Například v ekonomii práce je obvyklé pracovat s modelem, kde závisle proměnná je logaritmus mzdy každého jednotlivce a vysvětlující proměnné jsou počet let vzdělání (X_1) a počet let pracovních zkušeností (X_2) těchto jednotlivců:

$$\ln(Y_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

Protože je závisle proměnná vyjádřena v logaritmech jedná se na první pohled o nelineární regresi. Lineární regrese je to však po transformaci proměnných a máme zde náš

lineární regresní model. V případě, kdy je závisle proměnná vyjádřena v logaritmu a vysvětlující proměnná (či proměnné) nejsou, můžeme příslušné koeficienty (např. β_1) interpretovat způsobem, *„jestliže se X_1 zvýší o jednotku, zvýší se závisle proměnná o β_1 procent (za předpokladu, že se hodnoty ostatních vysvětlujících proměnných nemění)“*.

Pokud si tento příklad dále rozšíříme, můžeme uvažovat situaci, kdy pracovní zkušenost nemá na mzdu lineární vliv. Můžeme tak předpokládat, že nový pracovníci se v zaměstnání stále a stále zlepšují, což se promítá i do jejich (vyšší) mzdy. V určitém okamžiku již ale pracovníci zvládají svou práci natolikdobře, že dodatečný rok zaměstnání již nezvyšuje jejich produktivitu a růst jejich mzdy se tak může zpomalovat či zastavit. Jednoduchým způsobem jak toto chování v regresním modelu zohlednit je zadefinovat novou vysvětlující proměnnou odpovídající druhé mocnině zkušenosti, tedy

$$\ln(Y_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{2i}^2 + \epsilon_i.$$

Přestože je vysvětlovaná proměnná a jedna z vysvětlujících proměnných nelineární transformace původních dat (logaritmická a kvadratická), stále se jedná o regresi lineární ve závisle proměnné i vysvětlujících proměnných a můžeme tak využívat všechny doposud diskutované techniky spojené s odhadem takovéhoto modelu.

Další typ nelineární transformace dovoluje využít i vztah mezi vysvětlujícími proměnnými. Uvažujme model se dvěma vysvěltujícími proměnnými

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

Uvažujme, co se stane, pokud bychom dodali třetí vysvětlující proměnnou $X_1 X_2$:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i.$$

Opět zde máme lineární regresní model se třemi vysvětlujícími proměnnými. Zajímavá je však otázka intepretace parametrů. Jaký je mezní vliv X_1 na Y, za předpokladu neměnnosti ostatních vysvětlujících proměnných? V původním modelu by tento mezní vliv vyjadřoval parametr β_1. Po dodání třetího členu vzájemného propojení obou dvou proměnných však můžeme model přepsat do podoby

$$Y_i = \alpha + [\beta_1 + \beta_3 X_{2i}] X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

Mezní vliv X_1 na Y je tedy $[\beta_1 + \beta_3 X_{2i}]$. Mezní vliv již není konstantní a liší se v závislosti na hodnotě X_2. Podobně můžeme získat mezní vliv X_2 na Y, což je $[\beta_2 + \beta_3 X_{1i}]$. Obvykle je tento mezní vliv (či marginální efekt) vyhodnocován a prezentován v průměru pozorovaných dat, tedy např. $\left[\beta_1 + \beta_3 \overline{X}_{2i}\right]$.

Pro ilustraci, proč může být zkoumání takovéto interakce mezi vysvětlujícími proměnnými zajímavé, předpokládejme příklad vlivu vzdělání na mzdu. Provedeme tedy regresi s následujícími proměnnými:

- $Y =$ logaritmus mzdy;
- $X_1 =$ počet let vzdělání;
- $X_2 =$ skóre při testu inteligence.

V regresi bez členu kombinujícího obě vysvětlující proměnné je mezní vliv X_1 na Y roven β_1 a tento parametr je označován jako „výnosy ze vzdělání (the return to schooling)“. Při tomto odhadu by byly výnosy ze vzdělání pro všechny stejné. Pokud však do modelu dodáme novou proměnnou odpovídající součinu vysvětlujících proměnných, budou výnosy ze vzdělání rovny $[\beta_1 + \beta_3 X_{2i}]$. Tento model nám tak umožňuje analyzovat, jestli se výnosy ze vzdělání liší pro různé skupiny lidí, konkrétně v našem případě, jestli inteligentní studenti mají větší užitek ze vzdělání než studenti méně inteligentní. V regresi bez členu vzájemného ovivňování bychom tuto hypotézu nebyli schopni testovat. V regresi s tímto členem pak bude stačit sledovat statistickou významnost parametru β_3.

4.5.2 Jak rozhodnout o podobě nelineární závislosti?

Vzniká nám otázka jak rozhodnout, která případná forma nelinearity je nejvhodnější. Velkou roli určitě hraje ekonomická teorie, která nám poskytne ekonomický model, najehož základě pak vytvoříme odpovídající model ekonometrický (tedy model pro praktickou ekonometrickou analýzu). Pokud ale ekonomická teorie není v tomto ohledu jednoznačná, nabízí se řada možností, jak mezi různými specifikacemi modelu rozhodnout.

Základní problematika volby konkrétníá nelineární podoby vztahu mezi vysvětlovanou a vysvětlujícími proměnnými je obdobná problematice volby vysvětlujících proměnných diskutovaných již dříve v této kapitole. Předpokládejme, že nás zajímá, který z následujících dvou modelů je korektnější:

$$Y_i = \alpha + \beta_1 X_{1i} + \epsilon_i,$$
$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

V tomto případě s úspěchem využijeme t-testo pro testování hypotézy $H_0 : \beta_2 = 0$. Pokud je na tomto základě β_2 statisticky nevýznamný, dáme přednost první regrese. Alternativně můžeme provést obě regrese a porovna korigované koeficienty determinace, $\overline{R}^2$. Model s vyšší hodnotou bude naším favoritem. Tento postup je uplatnitelný bez ohledu na to, jakou podobu má X_2, tedy bezohledu na to, jestli $X_2 = X_1^2$ nebo $X_2 = \ln(X_1)$ nebo $X_2 = \frac{1}{X_1}$ nebo jiná lineární funkce X_1.

Předpokládejme, že v předchozím případě zhrnuje korektní model člen $X_2 = X_1^2$ (jedná se tedy o regresní křivku paraboly, tedy polynomu druhého stupně), ale my jej oponeme a provedem regresi bez tohoto členy, tedy pouze s X_1. V takovém případě opomíjíme důležitou vysvětlující proměnnou a dostaneme se tak do situace zkreslení při opomenutí relevantní vysvětlující proměnné. Dalším problémeme může být i problém multikolinearity, neboť X_1 může být silně korelována s nelineární funkcí X_1. Obecně je zde nejlepší radou zkoušet různé regrese s různými nelineárními transformacemi vysvětlujících proměnných využít testování hypotéz o statistické významnosti parametrů nebo korigovaných koeficientů determinace k rozhodování o nejlepší funkční podobě.

Je třeba zdůraznit, že korigovaný koeficient determinace je možné pro rozhodování mezi jednotlivými modely použít jen v případě transformací vysvětlujících proměnných. Modely tak mají nalevé straně stejnou vysvětlovanou proměnnou. Nelze jej po-

užít pro případ různých transformací závisle proměnné. Připomeňme si, že vztahy pro R^2 a $\overline{R}^2$ závisejí na členu, který odpovídá rozptylu vysvětlované veličiny, tedy $var\,(Y)$ (tzn. $\frac{1}{N-1}\sum\left(Y_i-\overline{Y}\right)^2$ je odhad tohoto rozptylu). Při rozhodování mezi dvěma regresemi s odlišnými vysvětlovanými proměnnými jsou i rozptyly závisle proměnné zcela odlišné a nesrovnatelné. Použití korigovaného koeficientu determinace tedy vyžaduje aplikaci na modely se stejnou závisle proměnnou, Y.

Jako příklad uvažujme situaci, kdy chceme rozhodnout mezi dvěma následujícími modely:

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{1i} + \epsilon_i, \\ \ln(Y_i) &= \alpha + \beta_1 X_{1i} + \epsilon_i. \end{aligned}$$

V tomto případě nám korigovaný koeficient determinace, $\overline{R}^2$, v našem rozhodování nepomůže. Jiná situace by byla, pokud bychom volili mezi

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{1i} + \epsilon_i, \\ Y_i &= \alpha + \beta_1 \ln(X_{1i}) + \epsilon_i. \end{aligned}$$

V tomto případě bude využití korigovaného koeficientu determinace k rozhodnutí o tom, jestli vysvětlující proměnnou brát nebo nebrát v logaritmu, dobrým postupem. Alternativní postup může být i práce s regresním modelem

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 \ln(X_{1i}) + \epsilon_i.$$

S využitím t-testu pak rozhodneme o tom, jestli jsou β_1 a (nebo) β_2 statisticky významné. Tento postup však může vést k problému multikolinearity, neboť X_1 a $\ln(X_1)$ bývají často silně korelovány.

Diskuze nad otázkou, jak rozhodnout mezi modely s různě transformovanými vysvětlovanými proměnnými není až tak jednoduchá a vyžaduje statistické metody nad rámec znalostí potřebných doposud. Existuje však jeden speciální a důležitý případ, pro který vyplatí ukázat příslušný test. Tento test lze použít pro zodpovězení otázky, jestli použít lineární nebo log-lineární regresi. Modely, mezi kterými rozhodujeme, jsou:

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \epsilon_i, \\ \ln(Y_i) &= \alpha + \beta_1 \ln(X_{1i}) + \ldots + \beta_k \ln(X_{ki}) + \epsilon_i. \end{aligned}$$

První model budeme označovat jako lineární regresi a druhý jako log-lineární regresi. Ve druhé regresi ani nezáleží na tom, jestli jsou všechny nebo jen část vysvětlujících proměnných vyjádřena v podobě logaritmů. Důležité je, že v první regresi je závisle proměnné nelogaritmovaná a ve druhé naopak logaritmována je.

Problémem zde je to, že obě vysvětlované proměnné nejsou přímo srovnatelné. Můžeme ale použít novou proměnnou

$$Y_i^* = \frac{Y_i}{\tilde{Y}},$$

kde $\tilde{Y}$ je geometrický průměr nelogaritmované závisle proměnné (tzn. $\tilde{Y} = (Y_1 \cdot Y_2 \cdot \ldots \cdot Y_N)^{\frac{1}{N}}$). Lze ukázat, že v tomto případě můžeme v regresích použít tuto novou proměnnou a její logaritmus, díky kterým již obě závisle proměnné srovnatelné budou.

Nejprve tedy provedeme lineární a log-lineární regresi s použitím Y^* respektive $\ln(Y^*)$ jakožto závisle proměnných. Nechť dále SSR_{LIN} a SSR_{LOG} jsou součty čtverců reziduí těchto dvou modelů (dolní indexy hovoří vcelku jasně, pro který z modelů je ten či onen součet čtverců určen). Předpokládejme, že $SSR_{LIn} > SSR_{LOG}$ (tj. lineární regrese má větší součet čtverců). Testová statistika pro test hypotézy, že lineární a log-lineární regrese vyrovnávají data stejně je

$$LL_1 = \frac{1}{2N} \ln\left(\frac{SSR_{LIN}}{SSR_{LOG}}\right).$$

Jako v případě jakékoli testové statistiky je potřeba odvodit její rozdělení za předpokladu platnosti nulové hypotézy, čímž následně jsme schopni získat kritické hodnoty. V tomto případě platí, že tímto rozdlěením je χ^2_1. S využitím statistických tabulek pro chí-kvadrát rozdělení platí, že kritická hodnota na hladině významnosti $5\ \%$ je (pro případ jedno stranného testu) 3.841. V praxi tedy spočítáme LL_1 a pokud je tato hodnota větší než 3.841, zamítáme hypotézu, že neexistuje rozdíl mezi lineární a log-lineární regresí a preferujeme v tomto případě log-lineární variantu. Pokud je LL_1 menší než 3.841, nejsme schopni zamítnout nulovou hypotézu a můžeme tedy pracovat s kteroukoli z obou regresí.

Pokud by platilo, že $SSR_{LOG} > SSR_{LIN}$, má testová statistika podobu

$$LL_2 = \frac{1}{2N} \ln\left(\frac{SSR_{LOG}}{SSR_{LIN}}\right).$$

Tato testová statistika má rovněž χ^2_1 rozdělení a pro test na hladině významnosti $5\ \%$ bychom spočítali LL_2 a v případě, že by její hodnota byla větší než 3.841, zamítli bychom hypotézu, že není rozdíl mezi oběma regresemi a preferovali bychom v tomto případě model s lineární specifikací. Na druhé straně, pokud by hodnota LL_2 byla menší než 3.841, nemůžeme zamítnou hypotézu o rozdílu mezi oběma regresemi a mohli bychom tak s nimi zcela rovnoceně pracovat.

4.6 Shrnutí

Tato kapitola z velké části rozšířila poznatky z kapitoly 2 ve více formálnějším duchu. Na základě této kapitoly (včetně příloh) tedy již víme:

- že problém zkreslení při nezahrnutí důležité vysvětlující proměnné vede ke snaze o zahrnutí co možná nejvíce vhodných vysvěltujících proměnných do regrese;
- že problém zahrnutí irelevantních vysvětlujících proměnných naopak zdůrazňuje nutnost vyřadit všechny irelevantní vysvětlující proměnné;
- že k rozhodování o tom, které proměnné zařadit a které ne nám napomáhájí postupy testování hypotéz (např. standardní test statistické významnosti parametru);
- že F-testy jsou užitečné pro testování hypotéz zahrnující kombinace více regresních koeficientů;

- že testy založené na věrohodnostním poměru lze rovněž úspěšně využít k testování hypotéz o kombinaci více regresních koeficientů, nicméně jejich použití je mnohem širší;
- že mezi testy založené na věrohodnostním poměru patří Waldův test a test Lagrangeových multiplikátorů (jejichž použití se liší v závislsoti na tom, jestli je snadnější odhadnout omezený nebo neomezený model);
- jak vybrat vhodnou funkční podobu regresního modelu;
- že řada nelineárních vztahů mezi vysvětlovanou a vysvětlujícími proměnnými lze zapsat v podobě standardního modelu vícenásobné regrese zahrnující nelineární transformace proměnných;
- veškeré předchozí odvození (vztahuící se k testování hypotéz a ke skutečnosti, že OLS je BLUE) platí i pro proměnné regresního modelu, které jsou nelineární transformací pvodních proměnných;
- že v praxi experimentujeme s různě specifikovanými regresemi, zahrnujícími různé nelineární transformace proměnných a na základě testů hypotéz nebo korigovaného koeficientu determinace, $\overline{R}^2$, pak dokážem rozhodnout o tom nejvhodnějším;
- že je třeba dát si pozor v rámci výběru mezi různými lineárními transformacemi závisle proměnné, kdy byl ukázán případ testování toho, jestli použít logaritmovanou nebo nelogaritmovanou závisle proměnnou.

Měli bychom tak již znát a umět vysvětlit obsah následujících klíčových pojmů:

- Korigovaný koeficient determinace, $\overline{R}^2$
- Omezený model
- Nelinearita v regresi
- Linearizace modelu
- Waldův test
- F-test pro omezení koeficientů
- Neomezený model
- Log-lineární model
- Test věrohodnostního poměru
- Test Lagrangeových multiplikátorů

Příloha: Waldův test a test Lagrangeových multiplikátorů

Test věrohdnostního poměru je velmi mocný a užitečný nástroj. Existují však další dva obvykle používané test: Waldův test a test Lagrangeových multiplikátorů. K jejich důkladnému vysvětleníje potřeba poněkud rozšířenější statistické teorie a matematika (zejména maticová algebra). Ovšem řada ekonoemtrických balíčků tyto testy obsahuje a podobně je možné, že se s nimi setkáme při studiu odborných článkůa publikací,

které je rovněž využívají. V této příloze si tedy naznačíme alespoň základní intuici s nimi spojenou a krátce si je popíšeme, abychom byli schopniporozumět výstupu, který nám může nabídnout ekonometrický software (případně různé odborné práce).

Určitou nevýhodou testu věrohodnostního poměru je potřeba odhadu jak omezeného, tak i neomezeného modelu. Obvykle se jedná o zanedbatelnou nevýhodu, např. v případě lineárních omezení, kdy je omezený model možné snadno přepsat do podoby nového regresního modelu, který je snadné odhadnout. V případě nelineárních restrikcí však může být odhad omezeného modelu obtížný. Oproti tomu existují i případy, kdy je situace opačná. Je obtížné odhadnout neomezenou variantu modelu, ale model s omezeními již problém odhadnout není. Výhodou Waldova testu je ta, že vyžaduje odhad pouze neomezeného modelu, výhoda testu Lagrangeových multiplikátorů je naopak potřeba odhadu pouze omezeného modelu. Abychom se vyhnuli použití maticové algebry a sofistikovanějších statistických metod, budeme si použití těchto testů ilustrovat na jednoduchých případech.

Waldův test si ukážeme na případu modelu vícenásobné regrese s jediným omezením koeficientů, odpovídajících hypotéze $H_0 : g(\alpha, \beta_1, \beta_2, \ldots, \beta_k) = c$ pro nějakou funkci $g(\cdot)$ a nějakou konstantu c. Rozšíření na případ více omezení je konceptuálně podobné, nicméně matematicky obtížnější. Odhad neomezeného regresního modelu poskytuje ML odhady $\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U$. Myšlenka Waldova testu je taková, že v případě správnosti hypotézy H_0 by měly být odhady v blízkosti hodnot splňujících omezení. Mělo by tedy platit, že $g(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U)$ nebude příliš vzdálené od hodnoty c. Waldova testová statistika měří to, jestli je rozdíl $g(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U) - c$ dostatečně malý. Podobně jako u t-statistiky je zde výraz „dostatečně malý“ brán relativně k nejistotě spojené s estimátorem, která je vyjádřena skrze jeho rozptyl (či směrodatnou odchylku). Waldova statistika je

$$W = \frac{\left[g\left(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U\right) - c\right]^2}{var\left[g\left(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U\right)\right]}.$$

Vysvětlili jsme si, jak spočítat výraz v čitateli. Výraz ve jmenovateli je obtížnější. Rozptyl, $var[g(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U)]$, je rozptyl funkce $g(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U)$ (nebo jeho odhad). V jednoduchých případech získat tento rozptyl není obtížné. Předpokládejme, že $g(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U) = \widehat{\beta}_1^U + \widehat{\beta}_2^U$. S využitím vlastností operátoru rozptylu, platí

$$var\left(\widehat{\beta}_1^U + \widehat{\beta}_2^U\right) = var\left(\widehat{\beta}_1^U\right) + var\left(\widehat{\beta}_2^U\right) + 2cov\left(\widehat{\beta}_1^U, \widehat{\beta}_2^U\right).$$

Rozptyl a kovarianci OLS odhadů nám běžně spočítají ekonometrické programy a lze je tedy snadno získat. Pro model vícenásobné regrese se dvěma vysvětlujícími proměnnými jsou vztahy pro $var(\widehat{\beta}_1^U)$ a $var\left(\widehat{\beta}_2^U\right)$ prezentovány v úvodu této kapitoly. Pro případ nelineárních restrikcí kladených na koeficienty vyžaduje získání rozptylu $var[g(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U)]$ komplikovanější statistické metody. Pro naše potřeby nám ale stačí, že tento rozptyl jsou schopné spočítat příslušné ekonometrické balíčky.

Jako v případě jakéhokoli jiného statistického testu je třeba odvodit rozdělení příslušné statistiky za předpokaldu platnosti H_0, na jejímž základě se získají odpovídající

kritické hodnoty či p-hodnoty. stejně jako v případě testu věrohodnostního poměru je rozdělení Waldovy statistiky aproximativně (tedy asymptoticky) chí-kvadrát

$$W \sim \chi^2_q,$$

kde q je počet omezení v rámci nulové hypotézy, H_0 (v našem příkladě je $q = 1$).

Test Lagrangeových multiplikátorů si budeme ilustrovat pro případ, kdy je neomezený model jednoduchý regresní model s jediným koeficientem, β, přičemž omezený model je definován v rámci hypotézy $H_0 : \beta = c$. Test Lagrangeových multiplikátorů zahrnuje odhad pouze omezeného modelu. V našem příkladu je to jednoduché, neboť platí $\widehat{\beta}^R = c$. Motivace testu je taková, že v případě platnosti H_0 by maximálně věrohodný odhad omezeného modelu neměl být příliš vzdálen od ML odhadu neomezeného modelu (v našem příkaldu by tedy c nemělo být příliš vzdálené od $\widehat{\beta}$, tedy OLS odhadu). Diferenciální počet nám však říká, že v maximu věrohodnostní funkce (tedy přesněji, v maximu jakékoli funkce jedné proměnné) je první derivace funkce nulová (což odpovídá směrnici tečny v bodě). Pokud je tedy H_0 pravdivá, měla by být derivace věrohodnostní funkce vyhodnocená v $\widehat{\beta}^R$ blízko nule.

Tato intuice je obsažena v příslušné statistice testu Lagrangeových multiplikátorů. Statistika má podobu:

$$LM = \frac{\left[d \ln L\left(\widehat{\beta}^R\right)\right]^2}{I\left(\widehat{\beta}^R\right)}.$$

Tato testová statistika je opět formálním vyjádřením řešení intuitivní otázky, jak hodně vzdálený nule je sklon tečny věrohodnostní funkce při zohlednění restrikcí. Čitatel přímo počítá směrnici tečny v tomto bodě. Jako v případě jiných statistik však chceme velikost této odchylky vyjádřit relativně vzhledem k nejistotě spojenou s tímto odhadem. Jmenovatel LM měří tuto nejistotu, kdy $I(\cdot)$ je označována jako *informační matice*. Pro naše potřeby nám může stačit, že ji můžeme interpretovat jako rozptyl první derivace věrohodnostní funkce.

LM statistika má rozdělení, které je aproximativně (asymptoticky) chí-kvadrát:

$$LM \sim \chi^2_q,$$

kde q je počet restrikcí v kontextu H_0 (pro náš jednoduchý příklad je $q = 1$).

Je nutné zdůraznit, že statistiky pro test věrohodnostního poměru, Waldův test a test Lagrangeových multiplikátorů jsou asymptoticky ekvivalentní. To znamená, že s růstem velikosti vzorku (počtu pozorování) k nekonečnu, budou se jejich hodnoty rovnat. Jejich použití závisí na tom, jestli je obtížné odhadnout omezený nebo neomezený model. V případě testování nelineárních restrikcí regresních koeficientů se často používá Waldův test, protože je obtížné identifikovat omezený model se zahrnutím nelineárních restrikcí. V další kapitole bude diskutováno rozšíření klasických předpokladů (jedno z nich je případ heteroskedasticity). Neomezený model tak nemusí splňovat klasické předpoklady a v závislosti na přesné podobě modelu nemusí být snadno odhadnutelný. Obvyklý způsob jak testovat možnosti těchto rozšíření je formulace nulové hypotézy, kde tato rozšíření nebudou přítomna. Jinými slovy, nulová hypotéza H_0 tak předpokládá splnění klasických předpokladů. V tomto případě je odhad omezeného modelu snadny a využívá se tak testu Lagrangeových multiplikátorů.

Kvůli nárokům na matematické a statistické znalosti tato příloha neposkytla vyčerpávající odvození a diskuzi nad Waldovým testem a testem Lagrangeových multiplikátorů. Dosavadní výklad by však měl být dostačující pro intuitivní pochopení těchto testů, což by mělo umožnit jejich praktické využívání v empirické praxi s využitím odpovídajících ekonometrických programů. Stejně tak bychom měli být schopni pochopit výsledky odborných prací, které tyto koncepty využívají.

Kapitola 5

Lineární regresní model a uvolnění klasických předpokladů

V této kapitole se dozvíme:

- ☞ jaké důsledky má nesplnění některých z klasických předpokladů;
- ☞ že ve většině případů jejich nesplnění vede ke ztrátě vydatnosti OLS estimátoru případně až k vychýlenosti či nekonzistenci OLS odhadů;
- ☞ jak testovat možnosti nesplnění některých z klasických předpokladů;
- ☞ jaké možností řešení problému nesplnění některých z klasických předpokladů se nám nabízejí;
- ☞ že řešením je využití buď estimátoru zobecněné metody nejmenších čtverců, což většinou odpovídá metdoě OLS aplikované na vhodně transformovaný model;
- ☞ že problém korelovanosti vysvětlujících proměnných s náhodnou složkou můžeme řešit s využitím metody instrumentálních proměnných.

5.1 Úvod

Doposud jsme se bavili o regresním modelu při splnění klasických předpokladů. Některé, nebo dokonce všechny z těchto předpokladů jsou naprosto nezbytné pro odvození výsledků z předchozích kapitol. Odvození OLS estimátoru tyto předpoklady nevyžadovalo. Jediné na čem bylo založeno byl předpoklad lineárnízávisloti mezi závisle proměnnou a vysvěltujícími proměnnými. Byl odvozen na základě minimalizace součtu čtverců reziduí, což odpovídalo záměru o nalezení regresní rovnice přímky (či

nějaké nadroviny) nejlépe prokládající pozorovaná data. Pro ukázání toho, že OLS stimátor má žádoucí vlasnosti však tyto klasické předpoklady (či část z nich) vyžadovalo (např. důkaz Gaussova-Markovova teorému dokazujícího, že OLS estimátor je BLUE vyžadoval všechny klasické předpoklady kromě předpokladu normality náhodných složek). Odvození intervalů spolehlivosti a všech postupů při testování hypotéz vyžadovalo splnění všech klasických předpokladů. Krátce řečeno, statistická odvození vyžadují nějaké přepdoklady, a doposud jsme veškerou diskuzi nad výsledky vedly v rámci užitečné množiny předpokladů, které jsme nazvali klasickými předpoklady. Tyto výsledky můžeme brát jako jakési srovnávací měřítko (benchmark) pro další rozšíření.

V empirické praxi je pravděpodobné, že některé nebo snad dokonce všechny klasické předpoklady nejsou splněny. Je tedy vcelku důležité využít specifické postupy testování hypotéz pro ověření jejich splnění a v případě, že tyto předpoklady skutečně splněny nejsou, je nutné vyvinout vhodné estimátory, které pro tyto případy použitelné jsou. Tomu bude věnována tato kapitola. Začíná obecnou teoretickou diskuzí před analýzou jednotlivých případů. Tyto případy spadají dodvou kategorií. První kategorie se týká použití tzv. *estimátoru (metody) zobecněných nejmenších čtverců (General Least Squares estimator – GLS)*, řešeny jsou otázky *heteroskedasticity* a *autokorelace* náhodných složek. Druhá kategorie případů je spjata s použitím tzv. *estimátoru (metody) instrumentálních proměnných (Instrumental Variables estimator – IV estimator)*.

5.2 Základní teoretické výsledky

Výsledky předchozích kapitol byly odvozeny pro model vícenásobné regrese v podobě

$$Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \epsilon_i.$$

Tyto výsledky jsou korektní při splnění klasických předpokladů:

1. $E(\epsilon_i) = 0$. Nulová střední hodnota náhodných složek.
2. $var(\epsilon_i) = \sigma^2$. Konstantní rozptyl náhodných složek.
3. $cov(\epsilon_i, \epsilon_j) = 0$ pro $i \neq j$. Vzájemná nekorelovanost náhodných složek.
4. ϵ_i má normální rozdělení.
5. Vysvětlující proměnné jsou fixní, tedy nenáhodné veličiny.

Nulová střední hodnota náhodných složek

První předpoklad nám implikuje, že pracujeme se správným regresním modelem. Při splnění tohoto předpokladu tedy platí

$$E\left(Y_i\right) = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}.$$

Pokud jsme zvolili špatné vysvětlující proměnné, bude tento předpoklad nesprávný. O volbě vysvětlujících proměnných byla řeč v kapitolách 2 a 4. Nemá tedy smysl dále řešit uvolnění tohoto předpokladu. Ve skutečnosti, pokud je v regresním modelu přítomna

úrovňová konstanta, je tento předpoklad vždy splněn.[12] Pokud by v regresi nebyla přítomna úrovňová konstanta a průměrná hodnota reziduí bude nenulová, vznikne nám několik nepříjemností. První z nich je ta, že koeficient determinace, R^2 definovaný jako $1 - \frac{SSR}{TSS}$, může být negativní. To znamená, že výběrový průměr $\overline{Y}$ dokáže „vysvětlit" více variability v Y než variabilita vysvětlujících proměnných. Druhá, a mnohem větší nepříjemnost je ta, že regrese bez úrovňové může vést k vážným zkreslením (vychýlením) odhadů koeficientů. Koeficient determinace a korigovaný koeficient determinace se navíc stávají mírami bez jakéhokoli významu. To je důsledkem toho, že střední hodnota (výběrový průměr) závisle proměnné již nebude roven průměru (střední hodnotě) vyrovnaných hodnot modelu (nepracujeme tak se správným modelem, jak bylo zmiňováno v úvodu). Celkový součet čtverců TSS již nelze rozložit v případě modelu bez úrovňové konstanty na regresní součet čtverců (RSS) a součet čtverců reziduí (SSR), tedy $TSS \neq RSS + SSR$. Z tohoto důvodu je žádoucí pracovat v modelu vždy s úrovňovou konstantou, a to i v případě, kdy se ukáže její statistická nevýznamnost.

Předpoklad normality a jeho testování

Čtvrtý klasický předpoklad, o normalitě náhodných složek, lze (aproximativně) uvolnit s využitím asymptotické teorie. Asympotitcká teorie nám říká, co se stane, když se velikost našeho vzorku bude blížit nekonečnu, jak tedy budou vypadat intervaly spolehlivosti, postup testování hypotéz pro nekonečně velké N. Tyto výsledky můžeme využít aproximativně i pro reálná data, pokud N nebude příliš malé. Asympotické teorii jsou věnovány vybrané přílohy jednotlivých kapitol (zejména Příloha 2: Využití asymptotické teorie v jednoduchém regresním modelu) a samostatná část přílohy B. V regresním modelu jsou asymptotické výsledky totožné s doposud odvozenými výsledky. Např. v kapitole 3, věnované jednoduchému regresnímu modelu, jsme na základě využití klasických předpokladů odvodili, že OLS estimátor má normální rozdělení:

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum X_i^2}\right).$$

Tyto výsledky pak byly využity k odvození konfidenčních intervalů a postupů pro testování hypotéz. I v případě, kdy náhodné složky nemají normální rozdělení, můžeme ukázat, že $\widehat{\beta}$ je aproximativně rozdělen podle $N(\beta, \frac{\sigma^2}{\sum X_i^2})$. Výsledky pro intervaly spolehlivost a testování hypotéz z kapitoly 3 zůstávají v platnosti i bez předpokladu normality (předpoklad 4), i když samozřejmě jen aproximativně.

Pokud pracujeme s malými vzorky (tedy s malým počtem pozorování), tak je určitě žádoucí, otestovat splnění předpokladu normality náhodných složek. Jednotlivé náhodné složky jsou samozřejmě pro nás neznámé, a tak musíme využít jejich odhad, což jsou rezidua, (tedy rozdíl mezi pozorovanou hodnotou a hodnotou vyrovnanou, tedy predikovanou modelem na základě našeho odhadu parametrů).

Jarqueův-Berův test normality Jeden z nejběžněji používaným testem normality je *Jarqueův-Berův test*, nazvaný podle svých autorů, kterými v roce 1980 byli Carlos M.

[12] To si můžeme sami vyzkoušet, pokud po provedení regrese spočítáme průměrnou hodnotu reziduí, jakožto odhad jejich střední hodnoty.

Jarque a Anil K. Bera. Tento test využívá vlastnosoti normálně rozdělené náhodné veličiny, která je jednoznačně charakterizována svými prvními dvěma momenty, což je střední hodnota a rozptyl (přesněji řečeno, rozptyl je tzv. druhý centrovaný moment). Standardizované třetí a čtvrté momenty rozdělení se nazývají *šikmost (skewness)* a *špičatost (kurtosis)*. Šikmost nám udává, do jaké míry je rozdělení nesymetrické kolem své střední hodnoty a špičatost nám říká, jak tlusté jsou konce tohoto rozdělení a jak je tedy zašpičatělé (jedná-li se o unimodální rozdělení, tedy s jediným vrcholem). Normální rozdělení není zešikměné, šikmost je tedy nulová a má koeficient špičatosti roven 3. S touto hodnotou je srovnávána špičatost jiných rozdělení. Můžeme tak definovat přesah špičatosti, který je roven koeficientu špičatosti mínus tři. Normální rozdělení bude mít v tomto případě hodnotu přesahu špičatosti rovnou nule. Jedná se o symetrické rozdělení a označuje se jako „mesokurtické".

Tzv. „leptokurtické" (špičatější) rozdělení je symetrické rozdělení, které má tlustší konce a je více zašpičatělé kolem své střední hodnoty než normálně rozdělená náhodná veličiny s toutéž střední hodnotou a rozptylem. „Platokurtické" (plošší) rozdělení je méně zašpičatělé a má tenčí konce než normální rozdělení (opět se stejnou střední hodnotou a rozptylem).

Pánové Bera a Jarque tuto myšlenku formalizovali do podoby testu toho, zdali jsou koeficienty šikmosti a převisu špičatosti společně nulové. Pokud tedy označujeme náhodné složky jako ϵ a jejich rozptyl jako σ^2, lze ukázat, že koeficienty šikmosti ($skew$) a špičatosti ($kurt$) můžeme vyjádřit jako

$$skew = \frac{E\left(\epsilon^3\right)}{\left(\sigma^2\right)^{3/2}} \quad kurt = \frac{E\left(\epsilon^4\right)}{\left(\sigma^2\right)^2}.$$

Protože je střední hodnota náhodných složek nulová, tak v čitatelích skutečně vystupují příslušné centrované momenty (obecný r-tý centrovaný moment je definován jako $E[Y - E(Y)]^r$) a ve jmenovatelích dochází k jejich celkové standardizaci. Špičatost normálního rozdělení je 3, tedy převis špičatosti $(kurt - 3)$ je nula.

Jarqueova-Berova testová statistika je dána jako

$$JB = N\left[\frac{skew^2}{6} + \frac{(kurt-3)^2}{24}\right],$$

kde N je celková velikost vzorku. Testová statistika má asymptiticky rozdělení chíkvadrát se dvěma stupní volnosti, χ^2_2. To vše samozřejmě při platnosti nulové hypotézy, H_0, že výběr pochází z normálního rozdělení (a alternativní hypotézy, že tomu tak není).

Koeficienty šikmosti a špičatosti nahradíme příslušnými odhady. S využitím reziduí, kdy tedy místo ϵ použijeme příslušné OLS odhady, $\widehat{\epsilon}$, a za použití odhadů rozptylu náhodných složek, $\widehat{\sigma}^2$, získáme odpovídající odhady šikmosti a špičatosti, $\widehat{skew}$ resp. $\widehat{kurt}$. Je však důležité zdůraznit, že při výpočtu $\widehat{\sigma}^2$ nepoužíváme OLS odhad tohoto rozptylu, ale ML odhad, tedy $\frac{1}{N}\sum\widehat{\epsilon}_i^2$. Na tomto základě jsme schopni spočítat hodnotu příslušné testové statistiky, najít odpovídající kritické hodnoty (při zvolené hladině významnosti) a rozhodnout o zamítnutí nebo nezamítnutí nulové hypotézy. Vysoké hodnoty šikmosti a (nebo) vysoké či příliš nízké hodnoty špičatosti vedou logicky k vyšším

hodnotám testové statistiky a vedou tak obvykle k zamítnutí nulové hypotézy. Většina ekonometrických programů je schopna tento test provést a obvykle vrací i příslušnou p-hodnotu tohoto testu.

Co dělat v případě nesplnění předpokladu normality? Nutno říct, že není zcela jasně řečeno, co při nesplnění tohoto předpokladu dělat. Jenou z možností je použít odhadovou metodu, která tento předpoklad nevyžaduje, nicméně její implementace může být obtížná a mnohdy si nemusíme být zcela jisti, jaké má vlastnosti. Pro ne-normálně rozdělené náhodné složky lzse využít estimátor metody maximální věrohodnosti. Ve finanční ekonomii se často pracuje s výnosy nějykých aktiv (např. akcií) a odpovídající regresní model mají obvykle ne-normální rozdělení. V regresích pro finanční aplikace má rozdělení náhodných složek tlustší konce oproti normálnímu rozdělení (ve výnosnosti aktiv tak pozorujeme větší volatilitu, než by nám napovídalo odpovídající normální rozdělení). Rozdělení s tlustšími konci je Studentovo t-rozdělení. Můžeme tak pracovat s regresním modelem, pro který platí veškeré klasické předpoklady až na to, že náhodné složky budou mít Studentovo t-rozdělení. Tento předpoklad nám umožňuje odvodit věrohodnostní funkci a získat tak ML estimátor. Věrohodnostní funkce samozřejmě nebude stejná jako např. ta, odvozená v kapitole 7. Nicméně pro získání příslušného estimátoru je použit stejný postup. Možností pro využitá rozdělení je nepřeberně mnoho, a existuje tak i spousta možností pro věrohodnostní funkce. Obecně však platí, že jakékoli rozdělení náhodných složek může být podchyceno v rámci maximálně věrohodného přístup (pro tuto metodu jsou však preferována rozdělení unimodální, tedy s jediným vrcholem).

Nicméně, obvykle je preferováno použití OLS estimátoru, protože má dobře zanalyzované vlastnosti při různých situacích, které mohou nastat. Pro velké vzorky je nesplnění předpokladu normality nevýznamné, neboť s využitím asymptotické teorie jsme schopni dosáhnout aproximativně stejných výsledků jako pro případ splnění předpokladu normality.

Zamítnutí předpokladu normality může zapříčinit existence jednoho či dvou extrémních reziduí. Takováto pozorování leží na chvostech rozdělení, kdy jejich čtvrtá mocnina, která vstupuje do výpočtu koeficientu špičatosti, způsobuje jeho velkou hodnotu. Takováto pozorování, která nespadají do obvyklého chování ostatních pozorování se nazývají *odlehlá pozorování (outliers)*. Poud je toto důvodem nesplnění předpokaldu normality, lze s úspěchem využít umělé proměnné, které tato pozorování efektivně „odstraní".

Pro ilustraci tohoto postupu předpokládejme model měsíčních výnosu nějakých aktiv v období let 1980-1990. V rámci odhadu tohoto modelu jsme získali rezidua, kdy reziduum pro pozorování z října roku 1987 ukazovalo výraznou výchylku oproti ostatním reziduím, což je znak odlehlého pozorování. Můžeme tak definovat novou proměnnou (D_1), která bude rovna 1, pokud se jedná o pozorování z října roku 1987 a nula pro ostatní pozorování. Následně se bude pracovat s modelem, který bude zahrnovat i tuto novou proměnnou, tedy např:

$$Y_1 = \alpha + \beta_1 X_{1i} + \beta_2 X2i + \beta_3 D_{1i} + \epsilon_i.$$

Tato umělá proměnná, která bude mít hodnotu 1 pouze pro jediné pozorování bude mít ten efekt, jako bychom danou proměnnou vyhodili z našich pozorování, přičemž

výsledné reziduum pro toto pozorování bude nulové. Výsledný odhad koeficientu u této umělé proměnné bude roven právě reziduu, které bychom získali v rámci regrese bez umělé proměnné.

Argument proti použití tohoto přístupu bývá ten, že takto používané umělé proměnné k odstranění odlehlých pozorování uměle zlepšují charakteristiky modelu a tím jej zkreslují. Odstranění odlehlých pozorování vede ke snížení směrodatných odchylek, snížení součtu čtverců reziduí SSR a tím ke zvýšení koeficientu determinace, R^2 (což znamená lepší soulad modelu s daty). Odstranění pozorování je však problematické z věcného důvodu, pokud z pohledu statistiky předpokládáme, že každé pozorování v sobě nese nějakou informaci o problému, který zkoumáme.

Argument pro „odstranění“ odlehlých pozorování je založen na tom, že tyto hodnoty mohou silně zkreslit odhady koeficientů, neboť při minimalizaci součtu čtverců reziduí, mohou tato pozorování silně zkreslit odhady parametrů (regresní přímka či nadrovina se může vychýlit směrem k těmto odlehlým pozorováním).

Na jedné straně zde tedy stojí potřeba odstranění odlehlých pozorování, které mohou nežádoucím způsobem ovlivnit OLS odhady a způsobit ne-normalitu reziduí, na druhé straně zde stojí fakt, že každé pozorování představuje užitečný kousek informace. Obvykle ptedy zahrneme příslušné umělé proměnn do modelu tehdy, pokud je zde jak jejich statistická potřeba, tak i rozumné věcné zdůvodnění, plynoucí z našich znalostí o zkoumané problematice, tedy například, že se v daném časové okamžiku či období stalo něco mimořádného, co ovlivnilo vysvětlovanou proměnnou. V našem příkladu to molo být černé pondělí 19. října v roce 1987, spojené s krachem na finančních trzích po celém světě. Další příklady zahrnují jednorázové vládní krize, paniky na finančních trzích apod.

Další klasické předpoklady

V další části kapitoly se budeme věnovat předpokladům 2, 3 a 5. Porušení předpokladu 5 je spojeno se zavedením estimátoru instrumentálních proměnných (IV estimátoru). Jak uvidíme, nesplnění předpokladů 2 a 3 vede k možnosti využití tzv. GLS estimátoru, což je estimátor *zobecněné metody nejmenších čtverců (Generalized Least Squares)*. Budeme předpokládá dva speciální případy jeho použití: případ heteroskedasticity (což je nesplnění předpokladu 2 o konstantním rozptylu náhodných složek) a případ autokorelovaných náhodných složek (nesplnění předpokladu 3). Ještě před tím si ale udělejme přehled hlavních závěrů a obecných postupů, které budeme využívat při řešení obou zmiňovaných případů:

- Při splnění klasických předpokladů nám Gaussův-markovův teorém říká, že *OLS je BLUE*. Při nesplnění předpokladů 2 a 3 však OLS estimátor zůstává nestranný (nevychýlenný), ale ji není nejlepší, nemá tedy minimální rozptyl.

- Většina postupů řešení těchto problémů spočívá v transformaci původního modelu na model nový, který již klasické předpoklady splňuje.

- V takovém případě je estimátor OLS opět BLUE a můžeme využít veškeré postupy, techniky a odvození z předešlých kapitol (aplikovaných na transformovaný model).

- OLS estimátor používaný na takto transformovaný model se nazývá GLS estimátor.

Pro ilustraci budeme většinu základních ekonometrických postupů ilustrovat na příkaldu jednoduchého regresního modelu (bez úrovňové jkonstanty). To snižuje požadavky na použitý matematický aparát (vyhneme se tak zejména maticovému zápisu). Intuitivně podobná (ale matematicky komplikovanější) odvození lze provést i pro model vícenásobné regrese.

5.3 Heteroskedasticita

Klasický předpoklad číslo 2 (v našem přehledu) nám říká, že náhodné složky (chyby) mají mít stejný rozptyl. Tento požadavek označujeme jako požadavek na *homoskedasticitu* (náhodných složek). Abychom si lépe ilustrovali implikace tohoto předpokladu, uvažujme množinu pozorování týkajících se cen domů. Náhodná složka nám říká zdali je cena domu nadhodnocena nebo podhodnocena vzhledem k cenám podobných domů. Jinými slovy, regresní přímka zachycuje obecné chování v datech a náhodné složky nám měří jak daleko je cena konkrétního domu vzdálena od této přímky. Homoskedasticita neznamená, že všechny náhodné složky budou stejné pro každý dům, říká nám jen, že budou pocházet ze stejného rozdělení. To je určitě rozumný předpoklad. Může však nastat situace, kdy např. malé domy budou obecně charakterizovány tendencí mít méně rozptýlené náhodné chyby než domy větší. To může být způsobeno tím, že malé domy budou navzájem velmi podobné (např. se jedná o řadu bungaovů, stavěnou v podobnou dobu a v podobném stylu) a velké domy budou naopak navzájem dosti rozdílné. V takovém případě bude prodejce malého domu s mnohem menší pravděpodobností nadhodnocovat nebo podhodnocovat cenu svého domu (např. ho může napadnout, že se podívá na ceny ostatních podobných domů, bungalovů, v okolí, které se pordávají ve shruba stejnou dobu). Prodejci velkých domů, kteří mohou porovnávat svůj dům jen s velmi omezeným počtem podobných domů, budou mnohem pravděpodobněji cenu svého domu nadhodnocovat nebo podhodnocovat. V takovém případě je nerealistický předpoklad, že rozptyl chyb pro ceny malých domů bude stejný jako u domů velkých.

Heteroskedasticita nastává v případě, kdy se rozptyl náhodných složek liší v rámci jednotlivých pozorování. Obecný způsob, jak umožnit existenci heteroskedasticity, je nahrazení předpokladu 2 výrazem[13]

$$var\left(\epsilon_i\right) = \sigma^2\omega_i^2,$$

kde $i = 1, \ldots, N$. V tomto výrazu má ω_i^2 dolní index i, který říká, že rozptyl náhodných složek může být pro jednotlivá pozorování různý.

[13] Některé publikace člen σ^2 neuvádějí a spokojují se se zápisem $var(\epsilon_i) = \omega_i^2$. To samozřejmě na věci nic nemění a závěry jsou totožné jako v této kapitole. Člen σ^2 je jen faktor společný pro všechny rozptyly chyb a nezáleží tak na tom, jestli jej zahrneme do jedinho výrazu pro rozptyl nhodných složek, nebo jej „vytkneme“ před část celkového rozptylu, která je pro jednitlivé náhodné složky specifická.

5.3.1 Teoretické výsledky při známém rozptylu chyb $\sigma^2\omega_i^2$

V této části se zaměříme na vlastnosti OLS estimátoru a jeho porovnání s novým estimátorem parametru β pro případ, kdy rozptyly náhodných složek známe. V praxi tomu tak obvykle samozřejmě nebývá, a proto se v další části zaměříme na případ, kdy je tyto rozptyly nutné odhadovat.

Vlastnosti OLS estimátoru při existenci heteroskedasticity

Připomeňme si, že pracujeme s jednoduchým regresním modelem v podobě

$$Y_i = \beta X_i + \epsilon_i,$$

kdy jsou splněny všechny klasické přepdoklady až na předpoklad o homoskedasticitě. V kapitole 3 jsme si ukázali dvě možnosti zápisu OLS estimátoru:

$$\widehat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2} = \beta + \frac{\sum X_i \epsilon_i}{\sum X_i^2}.$$

Při splnění klasických předpokladů jsme si ukázali, že

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum X_i^2}\right),$$

což bylo dále využito k odvození intervalů spolehlivosti a postupům testování hypotéz.

Při existenci heteroskedasticity se řada odvození a důkazů nijak nemění. Rozptyl náhodných složek se nevyskytuje jak při odvození nestrannosti OLS estimátoru, tak při ilustraci toho, že $\widehat{\beta}$ má normální rozdělení. Není tak nutné tyto důkazy znovu opakovat a stačí nám spokojit se s konstatováním, že OLS estimátor zůstává i při přítomnosti heteroskedasticity nestranný (tzn. $E(\widehat{\beta}) = \beta$) a má stále normální rozdělení.

Nicméně, rozptyl OLS estimátoru již přítomností heteroskedasticity ovlivněn je. Konkrétně, pro naše předpoklady (tzn. splněné klasické předpoklady až na přítomnost heteroskedasticity) je rozptyl estimátoru dán jako

$$var(\widehat{\beta}) = \frac{\sigma^2 \sum X_i^2 \omega_i^2}{\left(\sum X_i^2\right)^2}.$$

Důkaz využívá vlastností operátoru rozptylu a vyjádření $\widehat{\beta}$ v rovnici (*) z kapitoly 3 a

je následující:

$$\begin{aligned} var\left(\widehat{\beta}\right) &= var\left(\beta + \frac{\sum X_i\epsilon_i}{\sum X_i^2}\right) \\ &= var\left(\frac{\sum X_i\epsilon_i}{\sum X_i^2}\right) \\ &= \frac{1}{\left(\sum X_i^2\right)^2} var\left(\sum X_i\epsilon_i\right) \\ &= \frac{1}{\left(\sum X_i^2\right)^2} \sum X_i^2 var(\epsilon_i) \\ &= \frac{\sigma^2}{\left(\sum X_i^2\right)^2} \sum X_i^2\omega_i^2. \end{aligned}$$

Výsledný vztah možná nevypadá zrovna pěkně, nicméně důležitý závěr je ten, že při existenci heteroskedasticity je rozptyl OLS estimátoru odlišný od rozptylu při splnění všech klasických předpokladů. Pro praktickou aplikaci je podstané to, že v případě existence heteroskedasticity vede její ignorování a použití OLS estimátoru (v nějakém ekonometrickém programu) k nekorektnímu použití vzorce pro $var(\widehat{\beta})$. Software použije vzorec pro případ splnění všech klasických předpokladů na místo správného vzorce $var(\widehat{\beta}) = \frac{\sigma^2}{\left(\sum X_i^2\right)^2} \sum X_i^2\omega_i^2$. Vztah pro $var(\widehat{\beta})$ vstupuje do konstrukce intervalů spolehlivosti a testových statistik. Každý na tomto základě prezentovaný interval spolehlivosti a každý test hypotéz tak bude zcela nekorektní a zpochybnitelný!

Shrňme si dosavadní poznatky, estimátor metody nejmenších čtverců zůstává nestranný při přítomnsoti heteroskedasticity (odhad je tedy v pořádku), ovšem vše ostatní (intervaly spolehlivosti, testy hypotéz apod.) již korektní nejsou. Jediným případem, kdy je použití metody nejmenších čtverců přijetelné je tehdy, pokud nám počítač spočítá rozptyl odhadu parametru korektním způsobem, tedy využije (pro případ jednoduché regrese) vztah $var(\widehat{\beta}) = \frac{\sigma^2}{\left(\sum X_i^2\right)^2} \sum X_i^2\omega_i^2$. K tomuto se vrátíme později při diskuzi nad tzv. heteroskedasticitě konzistentním estimátorem. V praxi má problém s OLS estimátorem ten důsledek, že se mnozí upínají k použití estimátoru zobecněných nejmenších čtverců. Zdůrazněme, že tyto závěry platí i pro lineární regresní model s více vysvětlujícími proměnnými (jen podoba příslušných vzorců důkazů je bez maticového zápisu mnohem složitější).

Estimátor metody zobecněných nejmenších čtverců v případě heteroskedasticity

Princip řešení heteroskedasticity byl naznačen v části 5.2. Jde o to, transformovat původní model trpící neduhem heteroskedasticity modelem novým, který již všechny klasické přepdoklady splňovat bude. Původní regresní model je ve tvaru (pro $i = 1, \ldots, N$)

$$Y_i = \beta X_i + \epsilon_i,$$

kde náhodné složky jsou heteroskedastické, ale v dalších ohledech již klasické předpoklady splňují. Předpokládejme nyní transformovaný model, který vznikne vydělením

obou stran rovnice členy ω_i:

$$\frac{Y_i}{\omega_i} = \beta\frac{X_i}{\omega_i} + \frac{\epsilon_i}{\omega_i}$$

což při úspornějším značení můžeme zapsat v podobě

$$Y_i^* = \beta X_i^* + \epsilon_i^*,$$

kde $Y_i^* = \frac{Y_i}{\omega_i}$, $X_i^* = \frac{X_i}{\omega_i}$ a $\epsilon_i^* = \frac{\epsilon_i}{\omega_i}$.

Lze snadno ověřit, že tento transformovaný model bude splňovat všechny klasické předpoklady. Z vlasntostí operátoru střední hodnoty přímo plyne, že transformované náhodné složky splňují předpoklad 1 a 3 (nulová střední hdnota a nekorelovanost). Předpoklad o homoskedasticitě lze dokázat následovně:

$$\begin{aligned} var\left(\epsilon_i^*\right) &= var\left(\frac{\epsilon_i}{\omega_i}\right) \\ &= \frac{1}{\omega_i^2} var\left(\epsilon_i\right) \\ &= \frac{\sigma^2\omega_i^2}{\omega_i^2} = \sigma^2. \end{aligned}$$

Ukázali jsme si tedy, že transformovaný model splňuje klasické přepdoklady. Veškeré výsledky pro metodu nejmenších čtverců jsou tak aplikovatelné s využitím transformovaného modelu, tedy s využitím Y^* jako vysvětlované proměnné a X^* jeko proměnné vysvětlující. na tomto zákaldě můžeme říct, že OLS estimátor (transformovaného modelu) je BLUE a z něj vycházející intervaly spolehlivosti a postupy testování hypotéz jsou zcela korektní.

Použití OLS estimátoru na transformováný model vede k příkladu estimátoru označovaného jako estimátor zobecněné metody nejmenších čtverců. Ze vztahů pro OLS estimátor aplikovaný na transformovaná data vyplývá

$$\widehat{\beta}_{GLS} = \frac{\sum X_i^* Y_i^*}{\sum X_i^{*2}},$$

kdy dolní index GLS dodáváme proto, aby bylo zřejmé, že se jedná o GLS estimátor a nikoli OLS estimátor s původními, netransformovanými daty, který označujeme jako $\widehat{\beta}$. GLS estimátor můžeme zapsat i s využitím původních dat jako

$$\widehat{\beta}_{GLS} = \frac{\sum \frac{X_i}{\omega_i}\frac{Y_i}{\omega_i}}{\sum\left(\frac{X_i}{\omega_i}\right)^2} = \frac{\sum \frac{X_i Y_i}{\omega_i^2}}{\sum \frac{X_i^2}{\omega_i^2}}.$$

GLS estimátor odpovídá pro případ heteroskedasticity estimátoru *metody nejmenších vážených čtverců (weighted least-square – WLS)*. Každé pozorování je váženo, kdy váhy jsou inverzne proporcionální standardní odchylce náhodných složek. Pozorování s velkou variabilitou chyb (tzn. méně spolehlivá data) získávají v estimátoru menší váhu a pozorování s malým rozptylem náhodné složky (více polehlivá data) získávají váhu

větší. Oproti OLS estimátoru, který nerozlišuje mezi spolehlivými a nespolehlivými pozorováními. GLS estimátor má tu dobrou vlastnost, že více spolehlivým pozorováním dává větší váhu. Spolehlivost a nespolehlivost je zde chápána v duchu velikosti rozptylu náhodných složek.

Přestože jsme doposud pracovali sjednoduchým regresním modelem, rozšíření pro případ modelu vícenásobné regrese je téměř automatické a intuitivní. GLS estimátor získáme tak, že vydělíme každé pozorování vysvětlované proměnné a všech vysvětlujících proměnných odpovídající hodnotou ω_i a na takto transformovaný model aplikujeme metodu nejmenších čtverců.

GLS estimátor je zcela ekvivalentní OLS estimátoru pro transformovaný model a lze tak využít závěry předchozích kapitol (aplikovaných na transformovaný model). Místo X_i a Y_i ve vzorcích kapitoly 3 použijeme X_i^* a Y_i^*. Protože transformovaný model splňuje všechny klasické přepdoklady, můžeme (pro případ jednoduché regrese) říct, že

$$\widehat{\beta}_{GLS} \sim N\left(\beta, \frac{\sigma^2}{\sum X_i^{*2}}\right).$$

GLS estimátor je tak nestranný s rozptylem

$$\begin{aligned} var(\widehat{\beta}_{GLS}) &= \frac{\sigma^2}{\sum X_i^{*2}} \\ &= \frac{\sigma^2}{\sum \left(\frac{X_i^2}{\omega_i^2}\right)}. \end{aligned}$$

Připomeňme si, že rozptyl OLS estimátoru je $var(\widehat{\beta}) = \frac{\sigma^2 \sum X_i^2 \omega_i^2}{\left(X_i^2\right)^2}$. Při existenci heteroskedasticity lze vidět, že rozptyl GLS estimátoru je odlišný.

Gaussův-Markovův teorém nám navíc říká, že pří splnění klasických předpokladů je OLS estimátor BLUE. Tady máme estimátor $\widehat{\beta}_{GLS}$, který je ekvivalentní OLS estimátoru aplikovaného na transformovaný model, který splňuje klasické předpoklady. V případě heteroskedasticity nám tak automaticky vychází, že $\widehat{\beta}_{GLS}$ je BLUE. Implikací je pak to, že

$$var(\widehat{\beta}_{GLS}) \leq var(\widehat{\beta}),$$

kde $\widehat{\beta}$ je OLS stimátor aplikovaným na původní, netransformovaný model. Oba estimátory jsou nestranné (nevychýlené), ale GLS má menší rozptyl, je tedy více efektivní (vydatnější).

Skutečnost, že

$$\widehat{\beta}_{GLS} \sim N\left(\beta, \frac{\sigma^2}{\sum X_i^{*2}}\right)$$

lze využít pro odvození konfidenčních intervalů a testování hypotéz analogickým způsobem jako v předchozích kapitolách. Vzorce jsou stejné, jen původní proměnné X_i a Y_i nahrdíme transformovanými protějšky, X_i^* a Y_i^*.

5.3.2 Odhad pro případ neznámých rozptylů náhodných chyb

Předchozí odvození předpokládala, že známe ω_i^2. V praxi se obvykle setkáme s případem, kdy je ω_i^2 neznámé. V tomto případě připadají v úvahu dva postupy. Pokud jsme schopni nalézt odhad ω_i^2, potom lze tento odhad zahrnout do vztahu pro GLS estimátor. Prakticky se to provádí transformací původního modelu (k čemu se hned dostaneme). Alternativní způsob spočívá ve využití *heteroskedasticitě konzistentní estimátor (heteroskedasticity consistent estimator – HCE)*. Oba přístupy si nyní popíšeme.

GLS a transformace modelu

V řadě případů může být heteroskedasticita vztažena k jedné z vysvětlujících proměnných. Pracujeme tedy stále s modelem vícenásobné regrese

$$Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

při splnění všech klasických předpokladů, s výjimkou, že

$$var(\epsilon_i) = \sigma^2 \omega_i^2 = \sigma^2 f(Z_i),$$

kde Z je některá z vysvěltujících proměnných a $f(\cdot)$ je kladná funkce (oborem hodnot je tedy kladné číslo). Požadavek na kladný obor hodnot je dán tíém, že rozptyl je vždy větší než nula. Obvykle tedy volíme

$$f(Z_i) = Z_i^2$$

nebo

$$f(Z_i) = \frac{1}{Z_i^2}.$$

První z výrazů odpovídá situaci, kdy předpokládáme, že *rozptyl náhodných složek roste přímo úměrně s velikostí vysvětlující proměnné*, druhý výraz pak představujeopačnou situaci, tedy *rozptyl složek je nepřímo úměrný příslušné vysvětlující proměnné*. V příkladu s cenami domů jsme uvažovali možnost, že velké domy budou mít vyšší rozptyl cen. Pokud tomu tak skutečně je, dává nám smysl volba $f(Z_i) = Z_i^2$, kde Z_i odpovídá velikosti domu. Další v praxi používané funkce jsou $f(Z_i) = \exp(Z_i)$ nebo $f(Z_i) = \exp(-Z_i)$ pro rozptyly náhodných složek, které jsou přímo úměrné resp. nepřímo úměrné proměnné Z_i. V praxi obvykle zkoušíme různé volby Z, která je jednou z vysvětlujících proměnných $X_1, \ldots, X_k$.

Předpokládejme, že jsme schopni nalézt proměnnou Z. Jak v tomto případě využít GLS? Za předpokladu známé heteroskedasticity nám metoda GLS říká, že bychom měli transformovat pozorovaná data do podoby modelu

$$\frac{Y_i}{\omega_i} = \alpha \left(\frac{1}{\omega_i}\right) + \beta_1 \left(\frac{X_{1i}}{\omega_i}\right) + \ldots + \beta_k \left(\frac{X_{ki}}{\omega_i}\right) + \left(\frac{\epsilon_i}{\omega_i}\right)$$

a následně použít metodu OLS na takto transformovaný model. Protože ale předpokládáme, že $\omega_i^2 = f(Z_i)$, můžeme tuto transformaci snadno provést.

Uvažujme případ, kdy je $f(Z_i) = Z_i^2$ (rozptyl náhodných složek je tedy přímo uměrný Z). V tomto případě odpovídá transformovaný model modelu

$$\frac{Y_i}{Z_i} = \alpha\left(\frac{1}{Z_i}\right) + \beta_1\left(\frac{X_{1i}}{Z_i}\right) + \ldots + \beta_k\left(\frac{X_{ki}}{Z_i}\right) + \left(\frac{\epsilon_i}{Z_i}\right).$$

GLS estimátor tak získáme vydělením všech proměnných proměnnou Z_i a následným odhadem modelu metodou OLS s využitím takto transformovaných dat. Tento nový model má koeficienty totožné s parametry původního modelu. Parametr β_k transformované regrese je stejný jako parametr β_k původní regrese má tedy interpretaci jakožto mezní vliv X_k na Y (za předpokladu neměnnosti ostatních vysvětlujících proměnných). Protože nemůžeme dělit nulou, není možné použít tuto transformaci pro proměnné, jejichž hodnota je pro některé z pozorování nulová. Toto pravidlo tak vylučuje využití umělých proměnných pro transformaci. Na druhé straně, pokud je heteroskedasticita charakterizována funkcí $f(Z_i) = \exp(Z_i)$m jsou i nulové hodnoty proměnné Z_i akceptovatelné.

Pokud bychom předpokládali případ, kdy $f(Z_i) = \frac{1}{Z_i}$ (rozptyly náhodných složek jsou nepřímo úměrné Z), má transformovaný model podobu

$$Y_i Z_i = \alpha Z_i + \beta_1 X_{1i} Z_i + \ldots + \beta_k X_{ki} Z_i + \epsilon_i Z_i$$

a GLS estimátor tak získáme násobením našich vysvětlujících i vysvětlovaných proměnných proměnnou Z_i a následným využitím metody OLS.

V praxi obvykle nevíme, která z vysvětlujících proměnných by měla být použita a jestli rozptyl náhodných složek je k této proměnné vztažen přímo nebo nepřímo úměrně. Za této situace je dobré vyzkoušet různé volby Z (všechny možné vysvětlující proměnné, které jsou z tohoto pohledu smysluplné) a provést tak různé transformace původních proměnných. Použitím testů heteroskedasticity, ke kterým se zanedlouho dostaneme, uvidíme, jestli zde problém heteroskedasticity existuje a jestli je možné odstranit odpovídající transformací. Pokud námtesty naznačují přítomnost heteroskedasticity, můžeme experimentovat z různými transformacemi a dívat se, jestli některou z nich heteroskedasticitu z modelu odstraníme.

Heteroskedasticita je obvyklá pro průřezová data. V řadě případů tento problém vyřeší logaritmování proměnných. Log-lineární model byl diskutován v kapitole 4, části 4.5. Díky možnosti eliminovat problém heteroskedasticity je tento model hojně využíván v empirické praxi.

Heteroskedasticitě konzistentní estimátor

Transformace modelu může problém heteroskedasticity vyřešit. Mnohdy nám nicméně tento postup nemusí přinést kýžených výsledků. Nemusí senám podařit najít vhodnou proměnnou Z nebo může existovat více proměnných, jejichž kombinace problém heteroskedasticity způsobuje. V tomto případě model nejsme schopni transformovat a využit tak GLS estimátor.

Ani v tomto případě nemusíme zoufat. OLS odhad zůstává stále nestranný a jedná se o druhý nejlepší estimátor z hlediska vydatnosti. Nicméně, v tomto případě již nelze

využít dříve odvozeného vztahu pro rozptyl odhadu parametru. Nejsou splněny klasické předpokaldy a použití tohoto vztahu vede k nekorektním intervalům spolehlivosti a postupům testování hypotéz. Správný vzorec pro rozptyl odhadu parametru je

$$var(\widehat{\beta}) = \frac{\sigma^2 \sum X_i^2 \omega_i^2}{\left(\sum X_i^2\right)^2}.$$

Co tedy můžeme udělat je to, využít OLS estimátor a použít tento správný vztah pro odvození rozptylu odhadu parametru. Problém je ale ten, že neznáme $\sigma^2\omega_i^2$. Řešením je použití odpovídajícího odhadu. Pro neformální motivaci, odkud tento odhad vzít, si připoměňme, že

$$var(\epsilon_i) = E\left(\epsilon_i^2\right) = \sigma^2\omega_i^2.$$

OLS estimátor nám však dává rezidua

$$\widehat{\epsilon}_i$$

pro $i = 1, \ldots, N$. Protože je $\widehat{\beta}$ nestranný, platí, že i rovněž $\widehat{\epsilon}_i$ je nevychýlenný odhad náhodné složky ϵ_i. Samozřejmě, $\widehat{\epsilon}_i^2$ není totéž co ϵ_i^2 a ani totéž co $E(\epsilon_i^2)$. Je však rozumné domnívat se, že $\widehat{\epsilon}_i^2$ je dobrým estimátorem $E(\epsilon_i^2)$ a tedy i $\sigma^2\omega_i^2$. Odhad rozptylu odhadu parametru, $var(\widehat{\beta})$ je tedy

$$\widehat{var(\widehat{\beta})} = \frac{\sum X_i^2 \widehat{\epsilon}_i^2}{\left(\sum X_i^2\right)^2}.$$

Tento odhad rozptylu můžeme použít ve vztahu pro intervaly spolehlivosti a postupech testování hypotéz, čímž získáme platné statistické postupy a výsledky. Výše uvedený vztah je příkladem *heteroskedasticitě konzistentního estimátoru (heteroskedasticity consistent estimator – HCE)*. Konkrétně se jedná o tzv. *Whiteův estimátor*[14] Existují i další typy HCE. Většina ekonometrických programů nám dává možnost volby jejich použití a nemusíme se tak starat o to, jakou podobu má přítomná heteroskedasticita a jak ji řešit. Nevýhodou použití HCE je to, že OLS estimátor je méně vydatný než odpovídající GLS estimátor (tzn. že OLS estimátor má větší rozptyl). Tím není myšleno nic jiného, než to, že odhady budou mít tendenci být méně přesné, intervaly spolehlivosti tak budou širší a t-testy budou mnohem pravděpodobněji indikovat nevýznamnost parametrů (a odpovídajících proměnných).

Další estimátory

Výše uvedené dva přístupy jsou nejběžněji používaným způsobemřešení problém heteroskedasticity. Možné je však použití jiných estimátorů a některé ekonometrické balíčky jejich použití umožňují. Jejich odvození by šlo nad rámec potřebnýchznalostí. Základní intuice stojící v pozadí však bude dostačující pro jejich využití v praxi.

Ukázali jsme si, jak lze využít OLS rezidua k odhadu rozptylu náhodných složek. Tyto odhady však lze snadno dosadit do původníhovztahu GLS estimátoru. Výsledkem je estimátor zvaný *přípustný GLS estimátor (feasible GLS estimator – FGLS)*.

[14] Smozřejmě Whiteův estimátor je obecně vyjádřitelný pro případ modelu vícenásobné regrese, což však vyžaduje použití maticového zápisu.

Jeho použití je efektivní v případě známé podoby heteroskedasticity (např. vímě, že $var(\epsilon_i) = \sigma^2\omega_i^2 = \sigma^2 f(Z_i)$). V tomto případě existují metody, jak získat odhad $\widehat{\omega}_i^2$, který lze využít ke konstrukci FGLS estimátoru

$$\widehat{\beta}_{FGLS} = \frac{\sum \frac{X_i Y_i}{\widehat{\omega}_i^2}}{\sum \frac{X_i^2}{\widehat{\omega}_i^2}}$$

a k odhadu rozptylu tohoto estimátoru.

Podobně lze heteroskedasticitu ve známé podobě řešit s využitím odhadu metodou maximální věrohodnosti. To zahrnuje využití metod nelineární maximalizace. Tyto postupy jsou ale automaticky implementovány ve většině ekonometrických programů a lze je tak snadno využít.

Pro případ možnosti, kdy existuje více vysvětlujících proměnných způsobujících heteroskedasticitu je postup takový, že se v prvním kroku provede OLS odhad regresního modelu a na základě získaných reziduí je provedena regrese čtverce těchto reziduí na (kladnou) funkci vysvětlujících proměnných, u kterých se předpokládá, že se podle nich řídí rozptyl náhodných složek. Získané vyrovnané hodnoty jsou pak odhadem jednotlivých rozptylů náhodných složek vstupujících do odpovídající transformace modelu a následně se opět aplikuje metoda OLS což odpovídá aplikaci FGLS estimátoru. Obvykle se předpokládá závislost rozptylu náhodných složek v podobě

$$var(\epsilon_i) = \sigma^2\omega_i^2 = \exp(\alpha_0 + \alpha_1 Z_{1i} + \alpha_p Z_{pi}),$$

kde $Z_1, \ldots, Z_p$ je obvykle výběr některých nebo všech vysvětlujících proměnných v regresi. Volbou exponenciální funkce máme jistotu, že rozptyl nabývá opravdu kladných hodnot. Po získání OLS odhadů reziduí, $\widehat{\epsilon}_i$ provedeme regresi

$$\ln\left(\widehat{\epsilon}_i^2\right) = \alpha_0 + \alpha_1 Z_{1i} + \alpha_p Z_{pi} + \nu_i,$$

kde ν_i je náhodná složka s obvyklými vlastnostmi. Odhad parametrů získáme aplikací metody nejmenších čtverců a na tomto základě získáme odhady rozptylů

$$\widehat{var(\epsilon_i)} = \exp(\widehat{\alpha}_0 + \widehat{\alpha}_1 Z_{1i} + \widehat{\alpha}_p Z_{pi}).$$

Tyto odhady použijeme k transformaci původních proměnných modelu a aplikujeme metodu OLS na tento transformovaný model, ve kterém již rezidua splňují předpoklad homoskedasticity (pokud jsme zvolili správný funkční tvar pro charakter variability rozptylu náhodných složek).

5.3.3 Testování heteroskedasticity

Víme tedy, že při splnění všech klasických předpokladů je OLS estimátor to lepší, co můžeme použít (je tedy BLUE). V případě existence heteroskedasticity bychom měli použít GLS nebo HCE. Je tedy dobré vědět, jestli heteroskedasticita je v modelu přítomna nebo není. Existuje celá řada testů heteroskedasticity. Ukážemesi tedy tři nejběžněji používané.

Goldfeldův-Quandtův test

GOldfeldův-Quandtův test je dobrým testem v případě, kdy tušíme, že heteroskedasticita závisí na některé vysvětlující proměnné, Z, která je obvykle jednou z použitých vysvětlujících proměnných $X_1, \ldots, X_N$.Myšlenka testu je taková, že pokud rozdělíme naše data na dvě části, jednu vysokými hodnotami Z a druhou s nízkými hodnotami Z a provedem dvě odělené regrese s využitím těchto dvou skupin dat, potom bychom měli v případě existence heteroskedasticity získat dva ůrzné rozptyly náhodných složek (respektive reziduí). Test tak zahrnuje následující kroky:

1. Seřadíme data podle velikosti Z (tzn. první pozorování bude to s nejnižší hodnotou Z, druhé bude to s druhou největší hodnotou Z atd.).

2. Vynecháme prostřednich d pozorování (neexistuje pevné pravidlo jak velké by d mělo být, obvyklá volba je $d = 0.2N$, tedy $20\ \%$ pozorování).

3. Provedeme dvě oddělené regrese, jednu s využitím pozorování s nízkými hodnotami Z a druhou s použitím pozorování s vysokými hodnotami Z.

4. Spočítáme součet čtverců reziduí (SSR) pro každou s obou regresí (získáme tak SSR_{LOW} a SSR_{HIGH}).

5. Spočítáme Goldfeldovu-Quandtovu testovou statistiku

$$GQ = \frac{SSR_{HIGH}}{SSR_{LOW}}.$$

Při platnosti nulové hypotézy o homoskedasticitě má GQ statistika Fisherovo-Snedecorovo rozdělení, $F_{0.5(N-d-2k-2),0.5(N-d-2k-2)}$. Počet stupňů volnosti je tedy $0.5(N - d - 2k - 2)$, tedy počet pozorování v každém vzorku při zohlednění (odečtení) počtu odhadovaných parametrů $(k + 1)$. Pro získání kritických hodnot tak lze využít tabulky F-rozdělení. Hypotézu o homoskedasticitě zamítáme (a tedy uvažujeme existenci heteroskedasticity) v případě, kdy je GQ větší než kritická hodnota. Tento test je užitečný v případě, kdy chceme provést GLS transformaci, protože v případě nalezení heteroskedasticity pro určitou volbu Z můžeme tuto proměnnou využít k transformaci modelu. Konkrétně tak model bude mít podobu

$$\frac{Y_i}{Z_i} = \alpha\left(\frac{1}{Z_i}\right) + \beta_1\left(\frac{X_{1i}}{Z_i}\right) + \ldots + \beta_k\left(\frac{X_{ki}}{Z_i}\right) + \left(\frac{\epsilon_i}{Z_i}\right).$$

Tento test je dobrý v případě, kdy se rozptyl náhodných složek vyvíjí přímo úměrně na Z (např. v našem příkladu cen domů může platit, že „ceny velkýchdomů mají tendenci mít vyšší variabilitu chyb“). Poku máme pocit, že rozptyl chyb je nepřímo úměrný Z (např. „velké domy majítendenci mít malý rozptyl náhodných složek“), potom je test proveden analogicky s odpovídající změnou řazení. Data z prvního kroku seřadíme podle $\frac{1}{Z}$ a pokud nám Goldfeldův-Quandtův test naznačí heteroskedasticitu, budeme pracovat s transformovaným modelem

$$Y_i Z_i = \alpha Z_i + \beta_1 X_{1i} Z_i + \ldots + \beta_k X_{ki} Z_i + \epsilon_i Z_i.$$

I v rámci transformovaného modelu je třeba přítomnost heteroskedasticity otestovat. Pokud i v transformovaném modeluje heteroskedasticita přítomna, nelze využít metodu OLS jakožto GLS estimátor. Tato transformace totiž problém heteroskedasticity nevyřešila a musíme se pokusit o jiný druh transformace. Pokud ji nejsme schopni nalézt, je třeba použít HCE.

V praxi obvykle použijeme různé volby Z. Uvažujme příklad s cenami domů, kde máme dvě vysvětlující proměnné: rozlohu, X_1, a počet ložnic, X_2. Může nastat případ, že heteroskedasticita je spojena s rozlohou domu, počtem ložnic, oběma proměnnými nebo žádnou. Přepdokládejme, že je heteroskedasticita spojena pouze s rozlohou domu. Goldfeldův-Quandtův test s použitím $Z = X_1$ nebo $Z = \frac{1}{X_1}$ by měl korektně naznačit přítomnost heteroskedasticity. Pokud však použijeme $Z = X_2$, test nám pravděpodobně hypotézu o homoskedasticitě nezamítne a my tak nekorektně budeme pracovat s představou nepřítomnosti tohoto problému. Pokud si tedy nejsme zcela jisti, že heteroskedasticity (pokud existuje) je spojene s jednou konkrétní proměnnou, měli bychom provést Goldfeldův-Quandtův test pro každou možnou proměnnou (popř. inverzní hodnotu této proměnné).

Pokud jde o technickou stránku tohoto testu, stojí za povšimnutí, že tento test je variantou standardního dvou výběrového statistického testu hypotézy o shodě rozptylů. V našem případě tedy testujeme nulovou hypotézu, že rozptyly obou výběrů si jsou rovny, oproti alternativní hypotéze, že rozptyl v první části vzorku je větší než ve druhé. V ideálním případě by jejich podíl měl být roven jedné. Protože ale nemáme nekonečně velké výběry musíme statisticky rozhodnout, jakou hodnotu budeme statisticky považovat za blízkou jedničkové hodnotě a jakou už ne. A právě toto dilema za nás „řeší“ kritické hodnoty rozdělení testové statistiky při dané volbě hladiny významnosti. Protože uvažujeme pravostrannou alternativu testu hypotézy, kritická hodnota odpovídá hodnotě 95% kvantilu (při hladině významnosti 0.05) F-rozdělení. Pravostranná alternativa je v tomto případě používaná nejčastěji. Samozřejmě můžeme zvolit i levostrannou alternativu, kdy alternativní hypotéza odpovídá předpokladu, že rozptyl první části vzorku je menší než ve druhé. V této situaci ale musíme vzít kritickou hodnotu odpovídající 5% kvantilu Fisherova rozdělení. Test hypotézy může mít i oboustrannou alternativu, kdy kritické hodnoty nejsou symetrické jako u t-testu, ale jsou vzájemně inverzní (a odpovídají pro hladinu významnosti $5\ \%$ 97.5% kvantilu a 2.5% kvantilu).

Breuschův-Paganův test

Goldfeldův-Quandtův test je dobrý v případě, kdy se sama o sobě nabízí logická volba Z nebo když víme, že je heteroskedasticita spojena s jednou z vysvětlujících proměnných a jsme dost trpěliví v experimentování s odpovídající volbou Z. Breuschův-Paganův test je dobrýv případě, kdy existuje více vysvětlujících proměnných (a jejich kombinace), které mohou ovlivnit rozptyl náhodných složek, tedy

$$var\epsilon_i = \sigma^2 f(\gamma_0 + \gamma_1 Z_{1i} + \ldots + \gamma_p Z_{pi}),$$

kde $f(\cdot)$ je kladná funkce. V pozdadí tedy stojí myšlenka, že rozptyl chyb by měl záviset ne některé nebo všech proměnných $Z_1, \ldots, Z_p$ (které jsou obvykle totožné s vysvětlujícími proměnnými v regresi). Test heteroskedasticity je založen na hypotéze,

že $\gamma_1 = \gamma_2 = \ldots = \gamma_p = 0$, neboť právě v tomto případě $Z_1, \ldots, Z_p$ neovlivňují rozptyl náhodných složek. Můžeme tedy získat test věrohodnostního poměru pro nulovou hypotézu, $H_0 : \gamma_1 = \gamma_2 = \ldots = \gamma_p = 0$. Konkrétně zde jde o test Lagrangeových multiplikátorů, který má tu výhodu, že vyžaduje OLS odhad dvou rovnic. Tento test zde nebudeme detailně popisova, stačí nám jen jeho praktická implementace.

Breuschův-Paganův test zahrnuje následující kroky:

- Provedeme OLS regrese původního modelu (ignorujeme tedy možnost heteroskedasticity), získáme reziduí, $\widehat{\epsilon}_i$, a spočítáme

$$\widehat{\sigma}^2 = \frac{\sum \widehat{\epsilon}_i^2}{N}.$$

- Provedeme druhou regresi rovnice

$$\frac{\widehat{\epsilon}_i^2}{\widehat{\sigma}^2} = \gamma_0 + \gamma_1 Z_{1i} + \ldots + \gamma_p Z_{pi} + \nu_i.$$

- Spočítáme Breuchovu-Paganovu statistiku s využitím regresního součtu RSS z této druhé regrese:

$$BP = \frac{RSS}{2}.$$

- Tato testová statistika má chí-kvadrát rozdělení s p stupní volnosti, $\chi^2(p)$, na základě kterého snadno získáme potřebné kritické hodnoty.

Většina ekonometrických programů má tento test implementován a můžeme jej tedy snadno použít.

Whiteův test heteroskedasticity

Whiteův test heteroskedasticity je podobný Breuschovu-Paganovu testu, s tou výjimkou, že druhá regrese má trochu jinou podobu a je použita i jiná testová statistika. Stejěn jako s Breuschovým-Paganovým testem vycházíme z OLS regrese s původními proměnnými. Získaná rezidua (jejich čtverce) jsou využita ve druhé regresi. V této druhé regresi je funkční vztah mezi čtverci reziduí a vysvětlujícími proměnnými obdobný tomu využívanému v Breuschovu-Paganovu testu. Mnohem častěji jsou však jako proměnné vysvětlující rozptyl náhodných složek brány čtverce a křížové součiny ostatních vysvětlujících proměnných v původní regresi. Whiteův test tedy zahrnuje následující kroky:

- Provedeme OLS regresi na původních datech a modelu (při abstrahování od možné heteroskedasticity) a získáme odpovídající rezidua, $\widehat{\epsilon}_i$.

- Provedeme druhou OLS regresi rovnice

$$\widehat{\epsilon}_i^2 = \gamma_0 + \gamma_1 Z_{1i} + \ldots + \gamma_p Z_{pi} + \nu_i$$

  a získáme koeficient determinace, R^2, této regrese.

- Spočítáme Whiteovu testovou statistiku

$$W = NR^2.$$

- Testová statistika má chí-kvadrát rozdělení s p stupni volnosti, $\chi^2(p)$.

Proměnné $Z_1, \ldots, Z_p$ mohou být libovolné, obvykle jsou však založeny na vysvětlujících proměnných původní regrese. V tomto případě se využívají prakticky všechny vysvětlující proměnné, $Z_1 = X_1, \ldots, Z_k = X_k$. Navíc, Whiteův test využívá nejčastěji jedinečné čtverce a vzájemné (křížové) součiny vysvětlujících proměnných. Máme-li dvě vysvětlující proměnné, X_1 a X_2, měli bychom použít $Z_1 = X_1$, $Z_2 = X_2$, $Z_3 = X_1^2$, $Z_4 = X_2^2$ a $Z_5 = X_1 X_2$. Z technického hlediska je tak nutné ošetřit případ možné multikolinearity těchto proměnných (v původní regresi se může vyskytovat X_1 a X_1^2, z čehož vyplývá, že v regresi v rámci Whiteova testu bychom měli $Z_1 = X_1$, $Z_2 = X_1^2$, $Z_3 = X_1^2$, $Z_4 = X_1^4$ a $Z_5 = X_1^3$). Příslušné proměnné způsobující multikolinearitu však je možné vynechat.

Výhodou Breuschova-Paganova testu a Whiteova testu je to, že jej lze provést pouze jednou. Stačí zvolit množinu proměnných, které by mohly způsobovat heteroskedasticitu (obvykle se jedná o všechny vysvětlující proměnné regrese). Nevýhodou je zde to, že pokud nám tyto testy indikují přítomnost heteroskedasticity, nedávají nám žádné vodítko, jak bychom měli model transformovat a využít tak GLS estimátor. Vše co se dozvíme je to, že existuje problém heteroskedasticity a je vztažen k jedné (nebo několika) proměným, $Z_1, \ldots, Z_p$. Tyto výhody a nevýhody jsou přesně opačné oproti Goldfeldovu-Quandtovu testu. Goldfeldův-Quandtův test vyžaduje výběr jedné proměnné Z (popř. je možné vyzkoušet různé varianty volby Z). Pokud však shledáme, že jedna z proměnných Z způsobuje heteroskedasticity, jedná se o vodítko, jak transformovat model a provést odhad metodou zobecněných nejmenších čtverců.

5.3.4 Doporučení pro empirickou praxi

Pokud se domníváme, že v našem modelu a datech může být problém heteroskedasticity, začínáme tím, že provedeme některé z testů heteroskedasticity:

- Jestliže testy indikují přítomnost heteroskedasticity, je dobré zkusit model transformovat pro odstranění tohoto problému.

- Někdy dostačují jednoduché postupy jako logaritmování některých nebo všech proměnných v modelu (kapitola 4, část 4.5 řeší otázku interpretace koeficientů pro případ modelu s některými nebo všemi proměnnými vyjádřenými v logaritmech).

- Někdy pro řešení tohoto problému dostačuje násobení nebo dělení všech proměnných nějakou proměnnou Z.

- Vždy když provedeme některou z transformací, je třeba provést testy heteroskedasticity pro ověření toho,jestli byl problém vyřešen. Nejlepší jsou v tomto ohledu Whiteův test a Breusch-Paganův test.

- Pokud nedokážeme nalézt transformaci k odstranění problému heteroskedasticity, využijme HCE.
- Pamatujme, že v při existenci heteroskedasticity jsou testy hypotéz nekorektní. Před samotným testováním hypotéz (např. jestli je některá z vysvětlujících proměnných nevýznamná) je tedy potřeba počkat, až problém vyřešíme nebo až použijeme HCE.

Pro ilustraci slouží příklady 5.1 (5.2) a 5.3.

Příklad 5.1. *Vysvětlení specifik výdajů na vzdělaní mezi zeměmi*
Ekonomové zabývající se problematikou vzdělávání mohou řešit otázku, proč některé země vydávají větší částky na vzdělávání než ostatní.Pro zodpovězení této otázky máme k dispozici průřezová data pro 38 zemí:

- $EDUC =$ vládní výdaje navzdělávání (mil. amerických dolarů);
- $GDP =$ hrubý domácí produkt (mil. amerických dolarů);
- $POP =$ populace (mil. obyvatel).

Předpokládejme, že chceme provést regresi proměnné $EDUC$ na GDP a POP, přičemž ale máme obavy o splnění předpokladu homoskedasticity náhodných složek. V tomto příkladu si ukážeme, jak postupovat, a zaměříme se na Goldfeldův-Quandtův test. Goldfeldův-Quandtův test vyžaduje uspořádat všechny proměnné podle proměnné, kterou označíme jako Z. Následně rozdělíme pozorování do dvou skupiny a vynecháme prostředních d pozorování. V tomto příkladě zkusíme různé volby Z, vždy však zvolíme $d = 8$ (což je asi $20\,\%$ celého vzorku). GQ testová statistika má rozdělení $F_{0.5(N-d-2k-2),0.5(N-d-2k-2)}$, což je pro náš případ $F_{12,12}$. S využitím statistických tabulek pro F-rozdělení vyplývá, že kritická hodnota je 2.69.
Nejprve provedeme Goldfeldův-Quandtův test pro $Z = POP$. Získáme hodnotu statistiky $GQ = 0.51$, což je menší hodnota než hodnota kritická. Případ heteroskedasticity přímo úměrné proměnné vyjadřující velikost populace tak nelze přijmout. V dalším kroku můžeme zkusit Goldfeldův-Quandtův test pro $Z = \frac{1}{Z}$ a zjišťujeme, že $GQ = 1.96$, což je opět méně než kritická hodnota.[a] V rámci testu jsme tedy nemohli zamítnou nulovou hypotézu o homoskedasticitě. Není tedy statisticky prokazatelné, že by zde existovala heteroskedasticita nepřímo úměrně vztažená k velikosti populace.
Zkusíme-li volbu hrubého domácího produktu $Z = GDP$, získáme $GQ = 11.64$, což je výrazně více než kritická hodnota. Zamítáme nulovou hypotézu ohomoskedasticitě ve prospěch alternativní hypotézy o heteroskedasticitě, která je přímo úměrně vztažena k velikosti HDP, tedy proměnné GDP.
(pokračování v příkladu 5.2)

[a] Zvídavý čtenář snadno zjistí, že tuto statistiku nemusíme znovu počítat, neboť se logicky jedná o inverzní hodnotu statistiky spočtené před tím.

Příklad 5.2. *Vysvětlení specifik výdajů na vzdělaní mezi zeměmi (dokončení příkladu 5.1)*
V dalším kroku se pokusíme transformovat model k eliminaci heteroskedasticity. Skutečnost, že nám Goldfeldův-Quandtův test ukázal, že heteroskedasticita je způsobena přímo úměrně velikosti HDP, měli bychom se pokusit transformovat model vydělenímvšech proměnných právě proměnnou GDP. Závisle proměnná tedy bude $\frac{EDUC}{GDP}$, což odpovídá podílu výdajů na vzdělání k celkovému HDP, a vysvětlující proměnné budou $\frac{1}{GDP}$, $\frac{GDP}{GDP}$ a $\frac{POP}{GDP}$. Platí samozřejmě, že $\frac{GDP}{GDP} = 1$ a tato proměnná je tak úrovňovou konstantouv transformovaném modelu. Aplikací Goldfeldova-Quandtova testu v této regresi (řazení je stále s využitím $Z = GDP$) získáme hodnotu $GQ = 0.71$ což jeméně než kritická hodnota. Hypotézu o homoskedasticitě v tomto modelu nemůžeme zamítnou, což znamená, že heteroskedasticita není v tomto transformovaném regresním modelu problém.[a] OLS estimátor pro tento transformovaný model je ekvivalentní GLS estimátoru.
Transformovaný model by měl být využit k prezentaci konečných výsledků v rámci našeho empirického výzkumu. Přípomeňme si, že regresní koeficienty nejsou dotčeny touto transformaci. Pokud máme původní model (regresní přímku)

$$EDUC = \alpha + \beta_1 GDP + \beta_2 POP$$

bude transformovaný model (regresní přímka) v podobě

$$\begin{aligned}\frac{EDUC}{GDP} &= \alpha\frac{1}{GDP} + \beta_1\frac{GDP}{GDP} + \beta_2\frac{POP}{GDP} \\ &= \alpha\frac{1}{GDP} + \beta_1 + \beta_2\frac{POP}{GDP}\end{aligned}$$

Koeficienty však stále zůstávají stejné(α, β_1 a β_2) a mají rovněž stejnou interpretaci. Například, β_2 je stále mezní vliv obyvatelstva na výdaje na vzdělávání (za předpokladu neměnnosti ostatních vysvětujících proměných). Poznamenejme je, že β_1 (mezní vliv HDP na výdaje na vzdělávání) je úrovňovou konstantou v transformovaném modelu.

[a] V praxi je v této regresi vhodné provést Whiteův nebo Breuschův-Paganův test, abychom si byli zcelajisti neexistencí problému heteroskedasticity

5.4 Regresní model s autokorelací náhodných chyb

Třetí klasický předpoklad nám říká, že chyby regrese pro různá pozorování jsou vzájemně nekorelovaná. To je obvykle rozumný předpoklad v případě průřezových dat. Při práci s časovými řadami však tento předpoklad již tak rouzumný není. V případě časových řad jsou pozorování v jednom čase obvykle silně korelovány se sousedními pozorováními. Příkladem z oblasti makroekonomie je případ hospodářského cyklu, který v sobě obnáší skutečnost, že proměnné mají tendenci mít v jednotlivých fázích cyklu podobné hodnoty a vývoj (např. hospodářský růst je nízký či záporný v obdobích recese a naopak vysoký v obdobích expanze). Korelace mezi pozorováními má obvykle

Příklad 5.3. *Vysvětlení cen domů*
V předchozích kapitolách jsme využili OLS metody pro data $N = 546$ domů prodaných v kanadském Windsoru. Závisle proměnná, Y, je prodejní cena domu v kanadských dolarech. Vysvětlující proměnné jsou:

- $X_1 =$ rozloha domu (ve čtverečních stopách);
- $X_2 =$ počet ložnic;
- $X_3 =$ počet koupelen;
- $X_4 =$ počet pater(bez suterénu);
- $D_1 = 1$ pokud má dům příjezdovou cestu $(= 0$ pokud ne$)$;
- $D_2 = 1$ pokud má dům relaxační místnost $(= 0$ pokud ne$)$;
- $D_3 = 1$ pokud má dům sklep $(= 0$ pokud ne $)$;
- $D_4 = 1$ pokud má dům centrální vytápění$(= 0$ pokud ne$)$;
- $D_5 = 1$ pokud má dům klimatizaci $(= 0$ pokud ne$)$.

Po provedení regrese se všemi těmito proměnnými můžeme využít Goldfeldův-Quandtův test pro testování heteroskedasticity. Některé varianty však mohou způsobovat komplikace (seřazení dat při použití umělé proměnné nebo trasfromace modelu pomocí umělé proměnné). V tomto příkladu využijeme Whiteův a Breusch-Paganův test- Tytotesty vyžadují volbu proměnných, které ovlivňují rozptyl náhodných složek (proměnné $Z_1, \ldots, Z_p$). Použijeme-li stejné proměnné,jako jsou vysvětlující, získámě $BP = 112.93$ a $W = 44.97$. Kritické hodnoty obou těchto testů bereme z rozdělení $\chi^2(9)$. Kritická hodnota pro hladinu významnosti $5\ \%$ je 16.92. Protože jsou obě testové statistiky větší než kritická hodnota, zamítáme v obou případech nulovou hypotézu o homoskedasticitě ve prospěch hypotézy o existenci heteroskedasticity. Odpovídající OLS odhady jsou tak nestrann, ale intervaly spolehlivost a testy hypotéz jsou nekorektní. To se může týkat i výsledků regresí z kapitoly 2.
Na tomto základě můžeme zkoušet model různě transformovat. Pokud takovou transformaci nejsme schopni nalézt, musíme použít HCE. V našem příkladu však sačí jednoduchá transformace, kdy vezmeme logaritmy všech proměnných (s výjimkou umělých). Získáme tak log-lineární specifikaci. Hodnoty následných testů heteroskedasticity jsou $BP = 12.96$ a $W = 11.45$. Kritická hodnota pro hladinu významnosti $5\ \%$ je stále 16.92 a v obou případech tak nejsme schopni zamítnou hypotézu o homoskedasticitě. Logaritmická transformace tak v tomto příkladu dokázala problém heteroskedasticity vyřešit.

za následek korelaci mezi různými náhodnými členy, což znamená porušení třetího klasického předpokladu.

Protože tento problém nastává u časových řad, budeme používat pro indexaci jednotlivých pozorování značení $t = 1, \ldots, T$. Metodám práce s časovými řadami jsou věnovány kapitoly 7 a 8. V případě časových řad nemusí být použití regrese vždy korektní a mohou tak existovat adekvátnější metody vzhledem ke zkoumanému problému. V této části budeme předpokládat model vícenásobné regrese:

$$Y_t = \alpha + \beta_1 X_{1t} + \ldots + \beta_k X_{kt} + \epsilon_t.$$

Předpokládáme splnění všech klasických předpokladů, s výjimkou toho, který nám říká, že $cov(\epsilon_t, \epsilon_{t-s}) \neq 0$ pro některá $s \neq 0$. Pokud je například $s = 1$, je chyba regrese v jednom čase korelována s chybou v předchozím obdobím.

V této části se budeme zabývat důležitým tématem zvaným *autokorelace*, který implikuje $cov(\epsilon_t, \epsilon_{t-1}) \neq 0$. V rámci odvození výsledků budeme sledovat tutéž logiku jako v případě heteroskedastity. Při splnění všech klasických předpokladů nám Gaussův-Markovův teorém říká, že OLS je BLUE. Při nesplnění předpokladu o nekorelovanosti náhodných složek však toto tvrzení neplatí. OLS zůstává nestranným estimátorem, nicméně, není již nejlepším (nejvydatnějším), nemá tedy nejnižší rozptyl. To nás vede k využití GLS estimátoru. Tento estimátor si odvodíme vhodnou transformací původního regresního modelu do podoby modelu nového, který bude splňovat všechny klasické předpoklady. OLS estimátor pro tento transformovaný model již bude BLUE a lze tak využít veškerou doposud vysvětlovanou teorii pokud jde o odvození vlastností tohoto estimátoru, intervaly spolehlivosti, postupy testování hypotéz atd. Všechny závěry z předchozích kapitol budou platné – pro transformovaný model. OLS estimátor využitý pro takto transformovaný model je GLS estimátor.

5.4.1 Vlastnosti autokorelovaných náhodných chyb

Budeme pracovat s modelem vícenásobné regrese, pro který budou splněny všechny klasické předpoklady s výjimkou toho, že náhodné chyby budou generovány na základě *autoregresního procesu řádu 1 (AR(1))*:

$$\epsilon_t = \rho \epsilon_{t-1} + u_t,$$

kde u_t splňuje klasické předpoklady. To znamená, že $E(u_t) = 0$, $var(u_t) = \sigma_u^2$ a $cov(u_t, u_{t-s}) = 0$ (pro $s \neq 0$). Poznamenejme, že jsme hodnotě rozptylu, σ_u^2, přiřadili dolní index, aby bylo explicitně jasné, že se jedná o rozptyl složky u_t (ne chyby v regresním modelu). Předpokládáme $-1 < \rho < 1$. Již na tomto místě si můžeme říct, že tato podmínka nám zajišťuje *stacionaritu* chybových členů a nemusíme tak řešit problémy spojené s *jednotkovými kořeny (unit roots)* a *kointegraci* (jejich definice jsou obsahem kapitol 7 a 8),

V našem výkladu se zaměříme na případ $AR(1)$ procesu, nicméně případ obecného $AR(p)$ procesu náhodných chyb je jednoduchým rozšířením do podoby:

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + \ldots + \rho_p \epsilon_{t-p} + u_t.$$

Klasické předpoklady jsou stanoveny s ohledem na regresní chybový člen, ϵ_t, a ne „nový“ chybový člen, u_t. Známe vlastnosti posledně zmiňovaného členu, ale ne členu

před tím. Musíme se tedy zaměřit na odvození vlastností regresních chyb, ϵ_t. Mezi hlavní charakteristiky náhodných veličin patří střední hodnota, rozptyl a vzájemné kovariance. A právě tyto charakteristiky si odvodíme v této části regrese. Pro tato odvození je obvyklé zapsat si $AR(1)$ proces náhodných složek různými způsoby. Standardní způsob odvození vlastností $AR(1)$ modelu je předpoklad, že platí pro všechny časové okamžiky: tedy začíná v čase $-\infty$ a probíhá až do budoucího času ∞. My samozřejmě pozorujeme tento proces pro čas $t = 1, \ldots, T$. Protože však $AR(1)$ specifikace platí v každém časovém okamžiku, můžeme psát

$$\epsilon_{t-s} = \rho\epsilon_{t-s-1} + u_{t-s},$$

pro jakékoli t a s. Můžeme si zvolit $s = 1$ a využít toho k nahrazení ϵ_{t-1}, které se nachází na pravé straně původní rovnice pro $AR(1)$ proces. Pokud to uděláme, budeme mít na pravé straně ϵ_{t-2}. Ale znovu, pokud si zvolíme $s = 2$, zísáme výraz pro ϵ_{t-2}, který využijeme opět v původní rovnici $AR(1)$ procesu. Opakovaným nahrazováním výrazů na pravé straně původní $AR(1)$ rovnice jsme schopni vyjádřit tuto pravou stranu pouze pomocí členů $u_t, u_{t-1}, \ldots$. Následující rovnice nám tento postup jasně ukazuje:

$$\begin{aligned}
\epsilon_t &= \rho\epsilon_{t-1} + u_t \\
&= \rho(\rho\epsilon_{t-2} + u_{t-1}) + u_t \\
&= \rho^2\epsilon_{t-2} + \rho u_{t-1} + u_t \\
&= \rho^3\epsilon_{t-3} + \rho^2\rho_{t-2} + \rho u_{t-1} + u_t \\
&= \ldots \\
&= u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + \ldots \\
&= \sum_{i=0}^{\infty} \rho^i u_{t-i}.
\end{aligned}$$

Tento výraz můžeme využít k důkazu toho, že chyby regrese mají nulovou střední hodnotu:

$$\begin{aligned}
E(\epsilon_t) &= E\left(\sum_{i=0}^{\infty} \rho^i u_{t-i}\right) \\
&= 0
\end{aligned}$$

protože $E(u_{t-i}) = 0$ pro jakýkoliv časový okamžik.

Odvození rozptylů a kovariancí je poněkud komplikovanější a vyžaduje znalosti o vlastnosti nekonečné řady. Pokud máme číslo c menší než 1 v absolutní hodnotě, potom

$$\sum_{i=0}^{\infty} c^i = \frac{1}{1-c}.$$

Využitím vlasností operátoru rozptylu a díky skutečnosti, že u_t splňuje klasické

předpoklady, můžeme odvodit

$$\begin{aligned} var\left(\epsilon_t\right) &= var\left(\sum_{i=0}^{\infty}\rho^i u_{t-i}\right) \\ &= \sum_{i=0}^{\infty}\rho^{2i} var(u_{t-i}) = \sigma_u^2\sum_{i=0}^{\infty}\rho^{2i} \\ &= \frac{\sigma_u^2}{1-\rho^2}, \end{aligned}$$

kde poslední výraz vyplývá z vlastnosti nekonečné řady pro $c = \rho^2$. Důležitým závěrem pro rozptyl náhodné složky, $var(\epsilon_t)$ je to, že je konstantní, je tedy neměnný pro celé časové období. Autokorelované náhodné chyby jsou tedy homoskedastické.

Odvození kovariancí mezi dvěma náhodnými složkami je opět lehce komplikované, nicméně vyžaduje pouze předchozí výrazy pro regresní chyby a vlastnosti operátoru očakávání. S využitím definice operátoru kovariance a díky důkazu toho, že $E(\epsilon_t) = 0$, můžeme psát:

$$\begin{aligned} cov(\epsilon_t, \epsilon_{t-s}) &= E(\epsilon_t\epsilon_{t-s}) - E(\epsilon_t)E(\epsilon_{t-s}) \\ &= E(\epsilon_t\epsilon_{t-s}) \\ &= E\left[\left(\sum_{i=0}^{\infty}\rho^i u_{t-i}\right)\left(\sum_{i=0}^{\infty}\rho^i u_{t-s-i}\right)\right]. \end{aligned}$$

Nyní již stačí jen vzájemně pronásobit členy sum uvnitř operátoru očekávání. Na první pohled se to zdá být téměř nemožné. Využijeme-li skutečnost, že $cov(u_t, u_{t-s}) = E(u_t, u_{t-s}) = 0$ (pro $s \neq 0$), zbavíme se většiny členů. Například u_t se objevuje pouze v první závorce, ale ne ve druhé. Každý člen zahrnující u_t tak může být odstraněn, protože bude mít vždy podobu $E(u_t u_{t-j})$ pro různé kladné hodnoty j a příslušné hodnoty budou nulové. Odstraněním těchto členů tak získáme:

$$\begin{aligned} cov(\epsilon_t, \epsilon_{t-s}) &= E(\rho^s u_{t-s}^2 + \rho^{s+2}u_{t-s-1}^2 + \ldots) \\ &= E\left(\sum_{i=0}^{\infty}\rho^{s+2i}u_{t-s-i}^2\right) \\ &= \sum_{i=0}^{\infty}\rho^{s+2i}E(u_{t-s-i}^2) \\ &= \sigma_u^2\sum_{i=0}^{\infty}\rho^{s+2i} = \sigma_u^2\rho^s\sum_{i=0}^{\infty}\rho^{2i} \\ &= \frac{\sigma_u^2\rho^s}{1-\rho^2} \\ &= \rho^s var(\epsilon_t). \end{aligned}$$

Detaily tohoto odvození nemusí být na první pohled zřejmé. Nicméně myšlenka v pozadí je jednoduchá: protože $cov(\epsilon_t, \epsilon_{t-1}) \neq 0$, nejsou splněny klasické předpoklady a nemůžeme s odkazem na Gaussův-Markovův teorém říct, že OLS je BLUE.

Musíme najít GLS estimátor, který již bude BLUE. Než se k tomu dostaneme, vyplatí se zmínit některé zajímavé implikace vycházející z předchozích odvození. První z nich je to, že předchozí odvození závisí na tom, jestli je $|\rho| < 1$. Výsledky pro nekonečnou sumu, využívané pro výpočet rozptlu a kovariance, neplatí v případě, kdy $|\rho| \geq 1$. Výsledky v rámci mezikroků (jako je vztah $var(\epsilon_t) = \sigma^2 \sum_{i=0}^{\infty} \rho^{2i}$) platí i v případě, že $|\rho| \geq 1$. V tomto případě je však $var(\epsilon_i)$ nekonečno. K těmto otázkám se vrátíme v dalších kapitolách věnovaných analýze časových řad, když se budeme věnovat problému nestacionarity. V této kapitole budeme vždy předpokládat, že $|\rho| < 1$. V tomto případě vídíme, že náhodné chyby regrese se stávají méně a méně korelované v průběhu času. Pokud tomu nevěříme, stačí porovnat $cov(\epsilon_t, \epsilon_{t-1}) = \rho var(\epsilon_t)$ a $cov(\epsilon_t, \epsilon_{t-2}) = \rho^2 var(\epsilon_t)$. Pokud je ρ kladné a menší než jedna, musí platit, že $\rho > \rho^2$ a tedy i $cov(\epsilon_t, \epsilon_{t-1}) > cov(\epsilon_t, \epsilon_{t-2})$. Tím jak se stává s větší a větší, bláží se výraz ρ^s nule, a tedy $cov(\epsilon_t, \epsilon_{t-s})$ se bude blížit nule pro dostatečně velká s. Tato vlastnost je typická pro většinu časových řad (zejména v makroekonomii). Existuje tedy obvykle silná korelace mezi náhodnou složkou regrese v současnosti a regresní chybou v předchozím období, nícméně tato korelace se vytrácí, pokud budeme přepdokládáat vztah mezi náhodnou složkou dnes a náhodnou složkou v mnohem vzdálenější minulosti. Tato vlasntost charakterizuje všechny autoregresní stacionární proces, nejen $AR(1)$ proces.

5.4.2 GLS estimátor pro model s autokorelací náhodných složek

Ukázali jsme si, že model vícenásobné regrese s autokorelovanými náhodnými složkami nesplňuje klasické předpoklady. Platí, že OLS je v tomto případě nestranný, nicméně se již nejedná o nejlepší estimátor. Vhodný GLS estimátor bude nestranný a navíc bude mít menší rozptyl. Připomeňme si, že GLS metodu lze chápat jako OLS metodu aplikovanou na vhodně transformovaný model. V případě náhodných chyb, generovaných $AR(1)$ procesem, se odpovídající transformace nazývá „kvazi-diferencování“. Co je tím míněno lze ukázat v rámci standardnícho modelu vícenásobné regrese v podobě

$$Y_t = \alpha + \beta_1 X_{1t} + \ldots + \beta_k X_{kt} + \epsilon_t.$$

Tento model bude platit v každém časovém okamžiku, můžeme ho tedy vzít pro čas $t-1$ a násobit obě strany rovnice parametrem ρ:

$$\rho Y_{t-1} = \rho\alpha + \rho\beta_1 X_{1t-1} + \ldots + \rho\beta_k X_{kt-1} + \rho\epsilon_{t-1}.$$

Pokud odečteme tuto rovnici od původní regresní rovnice dostaneme

$$Y_t - \rho Y_{t-1} = \alpha - \rho\alpha + \beta_1(X_{1t} - \rho X_{1t-1}) + \ldots + \beta_k(X_{kt} - \rho X_{kt-1}) + \epsilon_t - \rho\epsilon_{t-1}$$

což můžeme psát jako

$$Y_t^* = \alpha^* + \beta_1 X_{1t}^* + \ldots + \beta_k X_{kt}^* + u_t.$$

Ovšem, u_t je chybou této nové regrese a tato chyba již splňuje klasické předpoklady. Metoda OLS aplikovaná na tento model tak bude již BLUE. Jedná se GLS estimátor pro model vícenásobné regrese s autokorelovanými náhodnými složkami (konkrétně

generovanými $AR(1)$ procesem). Gaussův-Markovův teorém nám říká, že tento estimátor bude mít menší rozptyl než OLS estimátor.

Je třeba si uvědomit, že GLS estimátor v sobě zahrnuje potřebu regrese Y^* na $X_1^*, \ldots, X_k^*$. Proměnné v této regresi jsou „kvazi-diferencované“:

$$\begin{aligned} Y_t^* &= Y_t - \rho Y_{t-1}, \\ X_{1t}^* &= X_{1t} - \rho X_{1t-1}, \\ &\text{atd.} \end{aligned}$$

Případ pro $\rho = 1$ odpovídá „diferenciaci“ proměnných. Transformace v našem pojetí není totožná, a proto hovoříme o „kvazi-diferencování“.

Vzniká nám tu ovšem jeden menší problém. Pokud máme data od $t = 1, \ldots, T$, potom $Y_1^* = Y_1 - \rho Y_0$ zahrnuje Y_0 (a totéž platí i pro vysvětlující proměnné). Nicméně my nejsme schopni pozorovat *počáteční podmínku*, či pozorování jako je Y_0. Problém počáteční podmínky lze řešit různými způsoby. Nejobvyklejší a nejjednodušší strategií (kterou budeme využívat) je vzít data od $t = 2, \ldots, T$ a použít $t = 1$ pro proměnné jako počáteční podmínky. Tímto postupem je problém vyřešen, i když za cenu ztráty pozorování. Existují i mnohem sofistikovanější metody jako je odhad metodou maximální věrohodnosti, které v sobě tuto ztrátu pozorování neobsahují.

Pokud bychom znali ρ, mohli bychom „kvazi-diferencovat“ data a provést OLS odhad s využitím transformovaných dat. Výsledný estimátor (což je GLS estimátor) je nejlepší, lineární, nestranný estimátor. Intervaly spolehlivosti a postupy testování hypotéz jsou obdobné jako v předchozích kapitolách (využívají transformovaná data). V praxi však ρ známe jen velmi zřídka, ne-li vůbec. Musíme tak ρ nahradit jeho odhadem, $\widehat{\rho}$. Existuje řada způsobů jak $\widehat{\rho}$ získat. Jedním z nich je tzv. *Cochranova-Orcuttova procedura*.

Cochranova-Orcuttova procedura

Cochranova-Orcuttova procedura začíná využitím metody OLS metody (která dává nestranné odhady) pro odhad původní regrese a následně využívá OLS rezidua k odhadu ρ. Postupuje se v následujících krocích:

1. Provedení regrese Y_t na úrovňovou konstantu, $X_1, \ldots, X_k$ s využitím metody OLS a získání příslušných reziduí, $\widehat{\epsilon}_t$.

2. Provední regrese $\widehat{\epsilon}_t$ na $\widehat{\epsilon}_{t-1}$ využitím OLS a použití odhadnutého koeficientu jako $\widehat{\rho}$.

3. Kvazi-diferencování všech proměnných k získání:

$$\begin{aligned} Y_t^* &= Y_t - \widehat{\rho} Y_{t-1}, \\ X_{1t}^* &= X_{1t} - \widehat{\rho} X_{1t-1}, \\ &\text{atd.} \end{aligned}$$

4. Provedení regrese Y_t^* na úrovňovou konstantu, $X_1^*, \ldots, X_k^*$ metodou OLS, čímž dostáváme GLS odhady koeficientů.

Motivace tohoto postupu spočívá na myšlence, že

$$\epsilon_t = \rho\epsilon_{t-1} + u_t$$

vypadá jako regresní model, kdy závisle proměnná je ϵ_t a vysvětlující proměnná je ϵ_{t-1}. Proměnné ϵ_t a ϵ_{t-1} nejsou pozorovány, nicméně tyto náhodné složky můžeme nahradit odpovídajícími rezidui, $\widehat{\epsilon}_t$ a $\widehat{\epsilon}_{t-1}$, čímž jsme schopni reálně odhadnout danou regresi.

Cochranova-Orcuttova procedura je obvykle zobecňována do podoby *iterační Cochranovy-Orcuttovy procedury*. Motivace tohoto estimátru vzniká na základě skutečnosti, že Cochranova-Orcuttova procedura používá k odhadu reziduí (používaných pro odhad parametru ρ) nevydatný OLS estimátor (a původní data). Nicméně, Cochranova-Orcuttova procedura může být využita sama o sobě k odhadu reziduí. Proč tedy nevyužít tato rezidua k lepšímu odhadu parametru ρ, který zase můžeme využít ke kvazi-diferencování dat a k získaní nového GLS estimátoru. To je vcelku rozumný postup a lze jej provést opakovaně, tedy *iterovaně*. Formálně probíhá iterační Cochranova-Orcuttova procedura v následujících krocích:

1. Provedení regrese Y_t na úrovňovou konstantu, $X_1, \ldots, X_k$ s využitím metody OLS a získání příslušných reziduí, $\widehat{\epsilon}_t$.

2. Provední regrese $\widehat{\epsilon}_t$ na $\widehat{\epsilon}_{t-1}$ využitím OLS a použití odhadnutého koeficientu jako $\widehat{\rho}$.

3. Kvazi-diferencování všech proměnných k získání:

$$\begin{aligned} Y_t^* &= Y_t - \widehat{\rho}Y_{t-1}, \\ X_{1t}^* &= X_{1t} - \widehat{\rho}X_{1t-1}, \\ &\text{atd.} \end{aligned}$$

4. Provedení regrese Y_t^* na úrovňovou konstantu, $X_1^*, \ldots, X_k^*$ metodou OLS, čímž dostáváme GLS odhady koeficientů $\widehat{\alpha}, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$.

5. Vytvoření nových reziduí použitím GLS odhadu z kroku 4, $\widehat{\epsilon}_t = Y_t - \widehat{\alpha} - \widehat{\beta}_1 X_{1t} - \ldots - \widehat{\beta}_k X_{kt}$.

6. Návrat ke kroku 2 opakování postupu stále dokola dokud se odhad, $\widehat{\rho}$ nepřestane měnit (nepřestane měnit s danou tolerancí).

Většina ekonometrických programů dovoluje provést Cochranovu-Orcuttovu proceduru v podstatě automaticky. Mezi další populárními metody odhadu patří i metoda maximální věrohodnosti. Metoda maximální věrohodnosti zahrnuje nelineární maximalizační metody, čímž se ale nebudeme v našem případě zabývat. Opět je ale dobré upozornit, že i tuto metodu lze snadno použít v mnoha ekonometrických programech.

Konečně bychom měli zmínit i další obvyklý přístup k odhadu. V rámci diskuze nad problémem heteroskedasticity bylo řečeno, že metoda OLS může být použita (protože poskytuje nestranné odhady), nicméně nemůžeme korektně použít standardní vztahy

pro konfidenční intervaly a postupy testování hypotéz, protože rozptyl odhadu parametrů, $var(\widehat{\beta})$ (nacházející se v příslušných vztazích), je ovlivněn přítomností heteroskedasticity. Použití korektního výrazu pro $var(\widehat{\beta})$ by umožnilo korektní použití OLS metody. Příslušný estimátor, který tuto korekci prováděl byl nazýván heteroskedasticitě konzistentní estimátor (HEC). Byl sice méně vydatný než GLS estimátor, nicméně se jednalo o druhé nejlepší řešení problému, když jsme nebyli schopni implementovat GLS estimátor. Podobná úvaha existuje i pro případ regresního modelu s autokorelovanými náhodnými složkami. Existují *autokorelaci konzistentní estimátory*, které umožňují korektní použití OLS metody i při existenci autokorelovaných náhodných složek. Opět se těmto estimátorům nebudeme podrobně věnovat, stačí nám vědět, že většina ekonometrických balíčků tyto estimátory zahrnuje. Nejpopulárnějším z těchto estimátorů je *Neweyho-Westův estimátor*. Obvykle se pracuje s tzv. *heteroskedasticitě a autokorelaci konzistentními estimátory (HAC)*.

5.4.3 Testování autokorelace náhodných chyb

Pokud je $\rho = 0$, nejsou náhodné chyby korelovány a využití standardních OLS metod je tudíž korektní. OLS je BLUE a lze využít veškeré doposud odvozené postupy konstrukce intervalů spolehlivosti a postupů testování hypotéz. Pokud je však $\rho \neq 0$, existuje GLS estimátor, jako např. Cochranův-Orcuttův estimátor, který je lepší. TO může být pro nás motivací pro test nulové hypotézy: $H_0 : \rho = 0$ oproti alternativní hypotéze $H_1 \neq 0$. Existuje řada testů tohoto typu, a na tomto místě si popíšeme ty nejpopulárnější. Jedná se o testystandardně obsažené v ekonometrických programech.

Testování založené na věrohodnostním poměru

V kapitole 4 je ukázka toho, jak využít věrohodnostní poměr v případě, kdy máme k dispozici jak omezený, tak i neomezený model, které chceme porovnat. Do kategorie těchto testů spadá test věrohodnostního poměru pro $H_0 : \rho = 0$ oproti alternativě $H_1 : \rho \neq 0$. Test věrohodnosntího poměru vyžaduje maximálně věrohodný odhad jak omezeného, tak i neomezeného modelu. Maximálně věrohodný estimátor regresního modelu s autokorelovanými náhodnými složkami není příliš obtížný (a opět je součástí řady ekonometrických balíčků), nicméně je komplikovanější v tom smyslu, že vyžaduje použití nelineárních optimalizačních metod (opět ekonometrické programy to „udělají“ za nás). Nicméně z důvodu potřeby odhadu obou modelů není tento věrohodnostní test přítomnosti autokorelované náhodné složky obvykle používán.

Alternativou je zde test Lagrangeových multiplikátorů. To vyžaduje odhad pouze omezeného modelu. V tomto případě je omezený model nám známý model vícenásobné regrese splňující klasické předpoklady. Test Lagrangeových multiplikátorů tak lze snado provést. V této části kapitoly jsme se zaměřili na regresní model s náhodnou složkou generovanou $AR(1)$ procesem, nicméně test Lagrangeových multiplikátorů lze snadno provést pro obecnější případ. Případ $AR(p)$ procesu náhodných složek odpovídá situaci, kdy

$$\epsilon_t = \rho_1 \epsilon_{t-1} + \rho_2 \epsilon_{t-2} + \ldots + \rho_p \epsilon_{t-p} + u_t.$$

Ukážeme si tedy, jak provést test sdružené hypotézy $H_0 : \rho_1 = 0, \rho_2 = 0, \ldots, \rho_p = 0$ v kontextu modelu vícenásobné regrese. Test Lagrangeových multiplikátorů hypotézy H_0 zahrnuje následující kroky:

1. Provedení regrese Y_t na úrovňovou konstantu a $X_1, \ldots, X_k$ použitím OLS metody a získání příslušných reziduí, $\widehat{\epsilon}_t$.
2. Provedení regrese $\widehat{\epsilon}_t$ na úrovňovou konstantu a $X_1, \ldots, X_k, \widehat{\epsilon}_{t-1}, \ldots, \widehat{\epsilon}_{t-p}$ využitím metody OLS a získání koeficientu determinace, R^2.
3. Spočítání testové statistiky
$$LM = TR^2.$$

Přestože si na tomto místě nebudeme ukazovat důkaz, platí, že při platnosti nulové hypotézy, H_0, má LM statistika (aproximativně) $\chi^2(p)$ rozdělení. Kritické hodnoty testu tak získáme ze statistických tabulek chí-kvadrát rozdělení. Tento test je obvykle označován jako *Breuschův-Godfreyho test*.

Boxův-Piercův test a Ljungův test

Existují další dva oblíbené (a úzce propojené) testy založené na myšlence, že pokud nejsou náhodné složky autokorelován, potom by korelace mezi různými náhodnými složkami měla být nulová. My samožejmě náhodné složky nejsme schopni pozorovat, nicméně jejich odhady jsou rezidua. Pokud tedy získáme rezidua $\widehat{\epsilon}_t$, pro $t = 1, \ldots, T$, na základě OLS regrese Y na úrovňovou konstantu a $X_1, \ldots, X_k$, potom můžeme odhadou korelace mezi ϵ_t a ϵ_{t-s} jako

$$r_s = \frac{\sum_{t=s+1}^{T} \widehat{\epsilon}_t \widehat{\epsilon}_{t-s}}{\sum_{t=s+1}^{T} \widehat{\epsilon}_t^2}.$$

V tomto případě probíhají sumy od $s + 1$ do T, protože $\widehat{\epsilon}_{s-1}, \ldots, \widehat{\epsilon}_0$ nepozorujeme. Jedná se tedy o postup, kdy nastavujeme počáteční hodnoty náhodných složek na nulu.

Boxova-Pierceova testová statistika nazývaná též *Q-statistikou* je

$$Q = T \sum_{j=1}^{p} r_j^2,$$

kde výběr p označuje to, že testujeme existenci $AR(p)$ procesu pro náhodné složky. *Ljungova testová statistika* je

$$Q^* = T(T+2) \sum_{j=1}^{p} \frac{r_j^2}{T-j}.$$

Obě tyto testové statistiky, obsažené v ekonometrických programech, mají (aproximativně) $\chi^2(p)$ rozdělení při platnosti $H_0 : \rho_1 = 0, \rho_2 = 0, \ldots, \rho_p = 0$. Kritické hodnoty testu tak získáváme ze statistických tabulek chí-kvadrát rozdělení.

V případě těchto testů je potřeba upozornit na jedno důležité úskalí. V některých případech může být jednou z vysvětlujících proměnných *zpožděná vysvětlovaná proměnná*. Pokud tomu tak je, pak Boxův-Piercův a Ljungův test nejsou vhodné a neměly by být používány. Test Lagrangeových multiplikátorů však je možno použít.

Durbinův-Watsonův test

Dalším populárním testem pro test přítomnosti náhodných chyb generovaných $AR(1)$ procesem využívá *Durbinovu-Watsonovu statistiku*. Stejně jako v předchozích testech začínáme OLS regresí vysvětlované veličin, Y, na úrovňovou konstantu a vysvětlující proměnné, $X_1, \ldots, X_k$, díky níž získáme OLS rezidua, $\widehat{\epsilon}_t$. Testová statistika je

$$DW = \frac{\sum_{t=2}^{T}(\widehat{\epsilon}_t - \widehat{\epsilon}_{t-1})^2}{\sum_{t=1}^{T}\widehat{\epsilon}_t^2}.$$

Lze ukázat, že $0 \leq DW \leq 4$. Pro jistou intuici si všimněme, že v případě pozitivní korelace náhodných složek, by sousední rezidua, $\widehat{\epsilon}_t$ a $\widehat{\epsilon}_{t-1}$, měla mít tendenci být sobě podobná, tudíž i výraz $(\widehat{\epsilon}_t - \widehat{\epsilon}_{t-1})^2$ by měl být malý a DW bude blízko 0. V opačné situci, kdy existuje negativní korelace mezi sousedními náhodnými složkami, budou $\widehat{\epsilon}_t$ a $\widehat{\epsilon}_{t-1}$ od sebe vzdáleny hodně, a DW bude velké (blízko 4). Přesněji, hodnoty sousedních reziduí budou podobné, až na znaménko. Žádná autokorelace odpovídá situaci s průměrnými hodnotami DW (okolo hodnoty 2). Jinou možností motivace těchto výsledků je ta vlastnost, že DW je aproximativně rovna $2(1 - \widehat{\rho})$, kde $\widehat{\rho}$ je odhad ρ (viz diskuze týkající se Cochranovy-Orcuttovy procedury). Jedná se o odhad korelačního koeficientu mezi sousedními náhodnými složkami.

Durbinovu-Watsonovu statistiku můžeme použít neformálně způsobem, že pokud je např. blízo nule, odpovídá to pozitivní korelaci. Lze ji však použít i v rámci formálního testování hypotéz. Bohužel rozdělení DW (které je nutné pro získání kritických hodnot testu) není standardní rozdělení (jakými jsou Studentovo t-rozdělení nebo chí-kvadrát rozdělení). Řada ekonometrických balíčků má dokonce výstup v podobě p-hodnot DW testu. Víme, že v případě, že p-hodnota je menší než 0.05 (naše hladina významnosti), potom můžeme zamítnou H_0. V tomto případě to znamená, že rezidua jsou autokorelována.

Pokud nám náš software nedává p-hodnoty Durbin-Watsonova testu, musíme využít odpovídající statistické tabulky. Ty lze získat z většiny ekonometrických učebnic, či skrze web (Google nám na dotaz k této statistice vrátí tucty stránek s dostupnými tabulkami). Rovněž některé ekonometrické programy mají v sobě zahrnutu možnost získat příslušné kritické hodnoty. Z těchto tabulek získáme hodnotu dolní meze (d_L) a horní meze (d_U). Tyto hodnoty závisí na počtu pozorování, T, a počtu vysvětlujících proměnných, k, a jsou tedy různé pro jednotlivé aplikace. Při daných hodnotách DW statistiky můžeme hodnoty d_L a d_U využít pro naše závěry na základě tabulky 5.1.

Všimněme si, že pro některé hodnoty DW nám test nedává jednoznačné výsledky. Statistické tabulky nám tak nemusí jednoznačně zamítnout nebo nezamítnout nulovou hypotézu, $H_0 : \rho = 0$. Z tohoto důvodu je Durbinův-Watsonův test méně populární než předchozí testy (testuje navíc jen autokorelaci prvního řádu). Durbinův-Watsonův test (podobně jako Boxův-Pierceův a Ljungův test) není vhodný v případě,kdy jednou z vysvětlujících proměnných je zpožděná závisle proměnná.

Durbinova h-statistika

Posledním testem autokorelace náhodných složek, kterému se budeme věnovat, je *Durbinův h-test*. Jedná se o test vyvinutý Durbinem pro případ, kdy je jedna z vysvětlu-

Tabulka 5.1: Vyhodnocení Durbinova-Watsonova testu.

Hodnota testu	Vyhodnocení
$4 - d_L < DW < 4$	Zamítáme H_0; závěr $\rho < 0$
$4 - d_U < DW < 4 - d_L$	Výsledek neurčitý
$2 < DW < 4 - d_U$	Příjmeme H_0; závěr $\rho = 0$
$d_U < DW < 2$	Příjmeme H_0; závěr $\rho = 0$
$d_L < DW < d_U$	Výsledek neurčitý
$0 < DW < d_L$	Zamítáme H_0; závěr $\rho > 0$

jících proměnných zpožděná vysvětlovaná proměnná. V praxi bývá vcelku obvyklé použití zpožděných vysvětlovaných proměnných jako vysvětlující proměnné. Příklad takového modelu je

$$Y_t = \alpha + \beta Y_{t-1} + \gamma X_t + \epsilon_t.$$

Parametr β jsme použili jako koeficient příslušný zpožděnné vysvětlované proměnné. Durbinův h-test v sobě zahrnuje nejdříve provedení OLS regrese Y na úrovňovou konstantu, zpožděnou závisle proměnnou a X (případně další vysvětlující proměnné v regresi) a získání DW statistiky a rozptylu odhadu parametru β (tj. $var(\widehat{\beta})$). Následně je spočítána testová statistika (*Durbinovo* h):

$$h = \left(1 - \frac{DW}{2}\right)\sqrt{\frac{T}{1 - Tvar(\widehat{\beta})}}.$$

Kritické hodnoty tohoto testu lze získat ze statistických tabulek standardizovaného normálního rozdělení, protože (aproximativně) platí

$$h \sim N(0, 1).$$

Poznamenejme, že tento test nebude fungovat pokud je $Tvar(\widehat{\beta}) > 1$, protože v tomto případě se marně budeme pokoušet o odmocninu ze záporného čísla (výsledek z oblasti komplexních čísel se nepočítá).

5.4.4 Doporučení pro empirickou praxi

Pokud se domníváme, že v našem modelu a datech může být problém autokorelace (což je u časových řad pravidlem), začínáme tím, že provedeme některé z testů autokorelace:

- Jestliže testy indikují přítomnost autokorelace, je dobré zkusit model transformovat či jinak specifikovat pro odstranění tohoto problému.
- Nabízí se nám použití Cochranovy-Orcuttovy procedury (jako jedna z variant transformace)pro případ existence $AR(1)$ náhodných složek. Přestože se jedná o rozumný přístup řešení, část ekonometrů se staví proti tomuto postupu, protože v případě, kdy náhodné složky neodpovídají $AR(1)$ procesu, vede tato transformace k zavádějícím výsledkům.

Příklad 5.4. *Vliv nákupu počítačů na tržby*
Vypořádání se s autokorelovanými náhodnými složkami nemusí být úplně snadné, přestože obvykle není potřeba velkých experimentů k nalezení GLS tranformace (oproti případu heteroskedasticity). Jednoduše použijeme jeden či více testu autokorelace náhodných složek a pokud testy indikují autokorelaci můžeme použít GLS metodu (jako např. Cochraneovu-Orcuttovu proceduru pro případ autokorelace náhodných složek v podobě $AR(1)$ procesu). Problém je ale právě nalezení správného $AR(p)$ procesu náhodných složek. Pro ilustraci testů využijeme data jedné společnosti za 98 měsícl, která analyzovala to, jestli investice do počítačů zvýšily produktivitu pracovníků a zvýšily tak tržby (prodeje) společnosti. Data lze nalézt v souboru `computer.xls` (data doprovázející knihu Koop [17]) resp. `computer.gdt` (na záložce `Koop` v gretlu je položka `comupte1.gdt` s podobnými, ale nikoli stejnými daty). Konkrétně máme proměnné:

- Y = procentní změna v tržbách vzhledem k předchozímu měsíci;
- X = procentní změna v počtu nakoupených počítačů vzhledem k předchozímu měsíci.

S využitím metody OLS získáme regresní přímku (rovnici vyrovnání v podobě)

$$Y = \underset{(1.55)}{0.28} + \underset{(5.59)}{0.95},$$

kde čísla v závorkách odpovídají t-statistikám testu hypotéz o nulovosti každého z parametrů. Protože je t-statistika pro parametr sklonu rovna 5.59 a 5% kritická hodnota Studentova t-rozdělení je 1.98, zdá se být vliv investic do počítačů silně významný. Pokud jsou však náhodné chyby autokorelovány, nemusí být tento závěr korektní. Na tomto základě provedeme nekolik testů pro testování náhodných složek v podobě $AR(p)$ procesů. Pro ilustraci se zaměříme jen na $p = 1$, nicméně v praxi bychom obvykle zkoušeli několik různých hodnot pro p. LM statistika je 67.10, Boxova-Piercova statistika je 65.01 a Ljungova statistika je 67.02. Příslušná 5% kritická hodnota pro všechny tyto testy odpovídá $\chi^2(1)$ rozdělení a je rovna 3.84. Protože všechny tyto testové hodnoty jsou výrazně vyšší než kritická hodnota, všechny testy zamítají nulovou hypotézu o neautokorelovanosti náhodných složek. Máme zde tedy důkaz toho, že náhodné složky jsou autokorelovány.
Posledním testem je Durbinova-Watsonova statistika, která je rovna hodnotě 0.35. Pohledem do tabulek pro Durbinovu-Watsonovu statistiku (při využití 5% hladiny významnosti), s tím, že máme jendoduchý regresní model ($k = 1$) a počet pozorování $T = 98$, získáváme $d_l = 1.65$ a $d_U = 1.69$. Protože je $0 < DW < d_L$, můžeme říct, že náhodné složky jsou pozitivně korelovány.
(pokračování v příkladu 5.5)

- Po každé transformaci je vhodné použít některý z testů pro zjištění, jestli byl problém vyřešen.

- Pokud nedokážeme nalézt transformaci k odstranění problému autokorelace, využijme raději HAC estimátor. Tento přístup je nejvíce doporučován, vzhledem k rozšířenému názoru, že je lepší smířit se s autokorelací a méně efektivním estimátorem, než používat transformaci, která neodpovídá charakteru náhodných složek a může tak vést k zavádějícím výsledkům.

Příklad 5.5. *Vliv nákupu počítačů na tržby (dokončení příkladu 5.1)*
Protože jsme nalezli dostatek důkazů pro přítomnost autokorelace, jsou předchozí OLS výsledky nevěrohodné. Využijeme tak Cochranův-Orcuttův estimátor. S jeho využitím získáme regresní odhad v podobě

$$Y = \underset{(0.62)}{0.05} + \underset{(9.59)}{0.55},$$

kde čísla v závorkách odpovídají hodnotám t-testu statistické nevýznamnosti parametrů. To jsou výsledky, které bychom mohli prezentovat v konečné zprávě. Mezní vliv investice do počítačů na tržby společnosti je odlišný od původní OLS regrese, nicméně se potvrzuje, že výsledek je stále silně statisticky významný. Úrovňovou konstantu neřešíme, protože má jinou interpretaci v původní regresi a Cochranově-Orcuttově regresi. Pro případ úrovňové konstanty můžeme získat i hodnotu 0.30, pokud bychom nepoužili jako vysvětlující proměnnou „jedničky“ ale transformované hodnoty $(1 - \widehat{\rho})$. K podobným výsledkům vede i iterativní Cochranova-Orcuttova procedura. Nicméně i po provedení této procedury testy poukazují na autokorelaci náhodných složek. Pokud nechceme dále experimentovat s typy $AR(p)$ procesu náhodných složek a příslušnými transformacemi modelu, je dobré vrátit se k původní regresi a použít heteroskedasticitě a autokorelaci konzistentní (HAC) estimátor.

5.5 Metoda instrumentálních proměnných

Poslední předpoklad, který si v této kapitole uvolníme je předpoklad, že vysvětlující proměnné jsou pevně daná čísla, tedy nenáhodné veličiny. To je realistický předpoklad v experimentálních vědách, kde si experimentátor volí hodnoty vysvětlujících proměnných (např. množství podávaného léku v experimentu z medicíny) a potom pozoruje náhodný výsledek tohoto experimentu (např. zdravotní stav dobrovolníků v rámci tohoto experimentu z oblasti medicíny). V tomto případě si sám výzkumník volí hodnoty vysvětlujících proměnných, nejedná se tak o náhodné veličiny. Ekonomie je však obvykle není za experimentální vědu považována. Předpoklad pevně daných vysvětlujících proměnných tak může být nerealistický. V této části si ukážeme, že chápání vysvětlujících proměnných jakožto náhodných veličin nezpůsobí žádné výžnější problémy, pokud vysvětlující proměnné nebudou korelovány s náhodnou složkou. To tedy znamená, že když vysvětlující proměnné budou náhodné a nezávislé na chybách regrese, potom zůstává OLS estimátor dobrým estimátorem veškeré doposud odvozené

výsledky zůstávají v platnosti. Pokud však vysvětlující proměnné je s náhodnou složkou korelována, metoda OLS by neměla být použita. V tomto případě jsou totiž OLS odhady vychýlené a vzniká nám tak potřeba jiného estimátoru. Estimátor používaný v těchto případech je nazýván *estimátor resp. metoda instrumentálních proměnných (instrumental variables – IV)*.

Bohužel je faktem, že pokud vysvětlujícím proměnným „umožníme“ náhodnost, mnohá z odvození předchozích kapitol zahrnující operátory očekávání a rozptyli (např. výpočet střední hodnoty a rozptylu OLS estimátoru) se stanou složitějšími a mnohdy i neuskutečnitelnými, pokud nedodáme další předpoklady. Z těchto důvodů jsou nejrelevantnější asymptotické výsledky. To znamená, že výsledky odvozujeme za předpokladu nekonečně velké velikosti vzorku a tyto výsledky pak můžeme vztáhnout na konečné velikosti vzorku s určitou mírou aproximace. V rámci této kapitoly však asymptotickou teorii využívat nebudeme a půjde nám spíše o intuitivní a neformální přístup k problému metody instrumentálních proměnných. Příloha: Asymptotické výsledky pro OLS a IV estimátor nabízí formální, asymptotické důkazy. Nicméně v této části budeme muset využít jeden důležitý asymptotický koncept, a to koncept *konzistentního estimátoru*. Konzistence je podobná vlastnosti nestrannosti a pro naše potřeby stačí vědět, že když budeme mít nekonečně velký vzorek a konzistentní estimátor, tak příslušný odhad bude roven skutečné hodnotě neznámého parametru. Opakem konzistence je nekonzistentnost estimátoru, což je obdoba vychýlennosti.

Pro jednoduchost budou veškerá odvození a důkazy (včetně těch v příloze) vedeny v rámci jednoduchého regresního modelu., nicméně výsledky platí obecně pro model vícenásobné regrese. Budeme tedy pracovat s regresním modelem (a vrátíme se ke značení pozorování $i = 1, \ldots, N$)

$$Y_i = \beta X_i + \epsilon_i,$$

nicméně provedeme jednu změnu v klasických předpokladech. V této části budou klasické přepdoklady následující:

1. $E(\epsilon_i) = 0$, tedy nulová střední hodnota.
2. $var(\epsilon_i) = \sigma^2$, tedy konstantní rozptyl (homoskedasticita).
3. $cov(\epsilon_i, \epsilon_j) = 0$ pro $i \neq j$, tedy vzájemná nekorelovanost ϵ_i a ϵ_j.
4. ϵ_i má normální rozdělení.
5. X_i je náhodná proměnná.

Vyjádření „vysvětlující proměnné jsou náhodné veličiny“ je nedostatečné. Stejně jako jsme vyslovili předpokaldy o náhodných složkách (např. normálně rozděleny, vzájemně nekorelovány), musíme vyslovit i předpoklady o náhodné veličině X_i, na jehož základě si pak můžeme představit vlastnosti různých estimátorů. Budeme rozlišovat dva případy: první, kdy vysvětlující proměnná bude nezávislá na náhodné chybě, a druhý, kdy tomu tak nebude.

5.5.1 Náhodná vysvětlující proměnná nezávislá s náhodnou složkou

Protože je nyní X_i náhodná veličina, musíme si stanovit určité předpoklady o jejím rozdělení, zejména pokud jde o její střední hodnotu a rozptyl. Na tomto místě budeme předpokládat, že X_i jsou $i = 1, \ldots, N$ je i.i.d. náhodné veličiny, kdy

$$E(X_i) = \mu_X,$$
$$var(X_i) = \sigma_X^2.$$

Klíčovým pro naše odvození je to, že vysvětlující proměnná a regresní chyba jsou vzájemně nezávislé. Nezávislost je podobná nekorelovanosti, nicméně se jedná o silnější předpoklad. Pro naše potřeby je však nejvýznamnější implikací to, že ϵ_i je nekorelováno s jakoukoli funkcí X_i.

V kapitole 3 jsme pracovali s jednoduchým regresním modelem při splnění všech klasických přepdokladů a dostali jsme se k závěru, že OLS estimátor má následující rozdělení:

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum X_i^2}\right).$$

S využitím asymptotické teorie (viz Příloha: Asymptotické výsledky pro OLS a IV estimátor) lze ukázat, že tento výsledek platí aproximativně i v tomto případě. Je však obtížné dokázat, že platí přesně. Skutečnost, že $E(\widehat{\beta}) = \beta$ znamená, že OLS je nestranný. Tento výsledek lze dokázat s využitím nezávislosti X_i a ϵ_i. OLS estimátor tedy má podobu

$$\widehat{\beta} = \frac{\sum_{i=1}^N X_i Y_i}{\sum_{i=1}^N X_i^2}$$

a můžeme odvodti

$$\widehat{\beta} = \beta + \frac{\sum X_i \epsilon_i}{\sum X_i^2}.$$

Abychom ukázali nestrannost OLS estimátoru, musíme aplikovat operátor střední hodnoty na obě strany rovnice:

$$\begin{aligned}
\widehat{\beta} &= \beta + E\left(\frac{\sum X_i \epsilon_i}{\sum X_i^2}\right) \\
&= \beta + E\left(\sum \left[\frac{X_i}{\sum X_i^2}\right] \epsilon_i\right) \\
&= \beta + \sum E\left(\frac{X_i}{\sum X_i^2}\right) E(\epsilon_i) \\
&= \beta,
\end{aligned}$$

protože $E(\epsilon_i) = 0$. V rámci důkazu je klíčovým krokem využití vlastnosti nezávislsoti vysvětlujících proměnných a náhodné složky. Z toho vyplývá, že i jakákoli funkce vysvětlujících proměnných je nezávislá na náhodné složce. Pokud stále ještě nemáme představu, jak jsme tohoto faktu využili, připomeňme si obecné pravidlo pro dvě náhodné veličiny, A a B, kdy platí že $E(AB) \neq E(A)E(B)$. POkud jsou však A a B

nekorelovány, potom $E(AB) = E(A)E(B)$. Tuto skutečnost jsme využili tak, že A je pro nás funkce vysvětlujících proměnných, $\frac{X_i}{\sum X_i^2}$, a B odpovídá náhodným složkám. Tím pádem jsme vyseparovali člen $E(\epsilon_i)$ a využili jsme toho, že tato střední hodnota je rovna 0. Tím je důkaz nestrannosti proveden. V případě, který bude součástí následující části kapitoly však tento důkaz již nebude fungovat.

Na tomto místě si nebudeme dokazovat další aspekt, který zahrnuje tvrzení, že $\widehat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum X_i^2})$. Důkaz aproximativní platnosti tohoto tvrzení je opět obsahem přílohy této kapitoly. Pro empirickou praxi je důležitým závěrem k zapamatování to, že i když uvolníme klasický předpoklad o tom, že vysvěltující proměnné jsou pevně dané hodnoty, získáme tytéž výsledky, jako v případě OLS estimátoru při splnění všech klasických přepdokladů, i když tyto výsledky budou platit jen aproximativně. To vše samozřejmě za předpokladu nekorelovanosti všech vysvětlujících proměnných s členem chyb regrese.

5.5.2 Vysvětlující proměnná korelovaná s náhodnou složkou

V tomto případě budeme uvažova všechny klasické předpoklady z minulého příkladu, až na to, že budeme předpokládat, že vysvětlující proměnná a chyba regrese jsou vzájemně korelovány, tedy

$$cov(X_i, \epsilon_i) \neq 0.$$

V tomto případě platí, že OLS estimátor je vychýlený a potřebujeme jiný estimátor. Tímto estimátorem je estimátor metody instrumentálních proměnných (IV estimátor).

V příloze k této kapitole jsou obsažena asymptotická odvození, ve kterých je ukázáno, že OLS je nekonzistentní (což je asymptotický koncept podobný vychýlenosti). Důkaz samotné vychýlenosti začíná podobně jako v předchozím případě. Můžeme dojít až do následujícího bodu důkazu:

$$E(\widehat{\beta}) = \beta + E\left(\frac{\sum X_i \epsilon_i}{\sum X_i^2}\right),$$

nicméně od tohoto bodu se nedostaneme dále než k závěru, že není žádný důvod domnívat se, že $E(\frac{\sum X_i \epsilon_i}{\sum X_i^2}) = 0$. Ve skutečnosti tomu tak ani není. Pokud bychom ignorovali člen ve jmenovateli, $\sum X_i^2$, mohli bychom čitatel zapsat jako $E(\sum X_i \epsilon_i) = \sum E(X_i \epsilon_i) = \sum cov(X_i, \epsilon_i) \neq 0$. Toto je snad dostačující pro intuitivní chápání toho, proč $cov(X_i, \epsilon_i) \neq 0$ implikuje vychýlenost OLS estimátoru. Hlavně je třeba porozumět tomu, že v případě korelace vysvětlující proměnné a náhodné složky jsou OLS odhady vychýlené a neměly by být používány.

Ukážeme si příklady toho, proč by mohlo k této korelaci dojít. Nicméně, nejdříve si zaveďme nový estimátor, který by měl být v tomto případě používán.

Estimátor metody instrumentálních proměnných

Instrumentální proměnná (instrument), Z, je náhodná veličina, která je nekorelována s chybou regrese, ale která je korelována s vysvětlující proměnnou. IV estimátor je dán

jako

$$\widehat{\beta}_{IV} = \frac{\sum_{i=1}^{N} Z_i Y_i}{\sum X_i Z_i}.$$

Nebudeme si tento vzorec nijak více dokazovat. Lze jej snadno spočítat a díky jeho dostupnosti ve většině ekonometrických programech je snadné i jeho praktické používání. Většina odvození a výsledků vztahujících se k IV estimátoru je asymptotických. Důležité je pro nás to, že tento estimátor má atraktivní vlastnosti. Je konzistentní a jeho rozptyl lze odhadnout a využít při konstrukci intervalů spolehlivosti a postupech testování hypotéz ve stejném duchu jako v případě OLS estimátoru.

Pro podrobnosti týkající se IV estimátoru si zadefinujeme následující značení. Instrumentální proměnná je náhodná veličina a její střední hodnotu a rozptyl budeme značit jako

$$E(Z_i) = \mu_Z,$$
$$var(Z_i) = \sigma_Z^2.$$

První, a to zcela klíčovou vlastností instrumentální proměnné je to, že je nekorelovaná s chybou regrese, tedy

$$cov(Z_i, \epsilon_i) = 0.$$

Druhou vlastností je to, že je korelována s vysvětlující proměnnou, tedy

$$cov(X_i, Z_i) = E(X_i Z_i) - \mu_Z \cdot \mu_X = \sigma_{XZ} \neq 0.$$

Asymptotická odvození z přílohy této kapitoly nám implikují, že aproximativně lze pracovat s následujícím rozdělením IV estimátoru:

$$\widehat{\beta}_{IV} \sim N\left(\beta, \frac{(\sigma_Z^2 + \mu_Z^2)\sigma^2}{N(\sigma_{XZ} + \mu_X \mu_Z)^2}\right).$$

Tento výsledek je porovnatelný s výsledkem OLS estimátoru, $\widehat{\beta}$, z předchozího případu, který měl aproximativně normální rozdělení, $N(\beta, \frac{\sigma^2}{\sum X_i^2})$, i když jeho výraz je poněkud vypadá trochu komplikovaněji. V praxi lze neznámé střední hodnoty a rozptyly nahradit odpovídajícími výběrovými charakteristikami. Lze tak nahradit μ_X výběrovým průměrem, $\overline{X}$, σ_Z^2 výběrovým rozptylem, $\frac{(Z_i - \overline{Z})^2}{N-1}$, atd. Nebudeme si zde uvádět další podrobnosti, jak technicky tyto odhady zvládnout, stačí nám vědět, že ekonometrické programy mají příslušné procedury odhadu metodou IV v sobě obsaženy a korektně nám tak spočítají potřebné vztahy pro konfidenční intervaly a postupy testování hypotéz.

Asi zatím nemáme představu o tom, jak získat instrumentální proměnnou či proměnné v praxi. Tuto problematiku zakrátko zmíníme.

Použití IV estimátoru v praxi

V praxi existují dvě otázky, které jsme doposud nezmínili. První z nich je, „Co když máme model vícenásobné regrese zahrnující více než jednu vysvětlující proměnnou?“,

a druhá pak, „Co když máme více instrumentálních proměnných, než je potřeba?“. Odpověď na první otázku je jednoduchá: potřebujeme nejméně jeden instrument pro každou vysvětlující proměnnou, která je korelována s náhodnou složkou. Pokud tedy máme model vícenásobné regrese se třemi vysvětlujícími proměnnými a dvě z nich jsou korelovány s náhodnou složkou, potom potřebujeme dvě instrumentální proměnné. Pro tento případ se vztah pro IV estimátor stává mnohem komplikovanější (bez využití maticové algebry), nicméně ekononometrické programy tento IV estimátor pro nás spočítají. Dobré je zdůraznit i to, že i když je jen jedna vysvětlující proměnná v modelu vícenásobné regrese korelována s náhodnými složkami, neměli bychom používat OLS odhad, protože by vedl k vychýleným odhadům všech koeficientů.

Pro ukázku toho, jako postupovat v případě více instrumentálních proměnných než je potřeba, vraťme se k regresnímu modelu s jedinou vysvětlující proměnnou, X. Proměnná X je korelována s náhodnou složkou, tudíž potřebujeme IV estimátor. Existují dvě instrumentální proměnné: Z_1 a Z_2. Mohli bychom použít jednu z nich, nicméně rozumnější postup je ten, že využijeme *zobecněny estimátor instrumentálních proměnných (generalized instrumental variables estimator – GIVE)*. To v sobě zahrnuje provedení výchozí regrese vysvětlující proměnné na všechny instrumenty:

$$X_i = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + u_i,$$

kde používáme značení γ_0, γ_1 a γ_2 pro regresní koeficienty a u_i pro náhodné chyby této regrese, protože tím chceme zdůraznit, že se jedná o odlišnou regresi než tu zahrnující parametr β. OLS odhad této výchozí regrese nám poskytuje vyrovnané hodnoty:

$$\widehat{X_i} = \widehat{\gamma}_0 + \widehat{\gamma}_1 Z_{1i} + \widehat{\gamma}_2 Z_{2i}.$$

Lze ukázat, že proměnná $\widehat{X}$ je nekorelována s náhodnými složkami původní regrese, a jedná se tedy o vhodný instrument. To je právě to, co GIVE využívá jako instrument. GIVE je tedy dán jako

$$\widehat{\beta}_{GIVE} = \frac{\sum_{i=1}^{N} \widehat{X}_i Y_i}{\sum_{i=1}^{N} X_i \widehat{X}_i}.$$

Tento estimátor je konzistentní a využívá informaci z obou instrumentů tím nejefektivnějším způsobem. Opět, ty nejlepší ekonometrické balíčky nám GIVE lehce spočítají.

Hausmanův test

Pokud je jedna nebo více vysvětlujících proměnných v modelu vícenásobné regrese korelovaných s náhodnou složkou, je OLS estimátor vychýlený a je tak potřeba použít IV estimátor. Pokud však tomu tak není a žádná z vysvětlujících proměnných není korelována s chybami regrese, potom (při splnění klasických předpokladů) je OLS estimátor BLUE a je tedy vydatnější než IV estimátor. V tomto případě tak není využití IV estimátoru žádoucí. To nás vede k požadavku testování toho, jestli je vysvětlující proměnná korelována s chybou regrese. A právě to je účelem *Husmanova testu*. Jeho formální odvození přesahuje rámec ekonometrické teorie prezentované v tomto text. Nicméně, většina ekonometrických programů nám tento test provede automaticky. Z tohoto důvodu si tento test nebudeme odvozovat ani detailně rozebírat. Bude nám stačit diskutovat tento test z neformálního a intuitivního hlediska.

Nechť tedy H_0 odpovídá hypotéze, že vysvětlující proměnné v modelu vícenásobné regrese jsou nekorelovány s náhodnou složkou. Základní myšlenka, která stojí v pozadí Hausmanova testu je ta, že pokud hypotéza H_0 platí, jsou OLS a IV estimátory konzistentní a měly by nám tak vracet přibližně podobné výsledky. Pokud však H_0 neplatí, potom je OLS estimátor nekonzistentní a IV estimátor konzistentní a výsledné odhady by se tedy navzájem měly lišit. Hausmanův test tedy tuto myšlenku formalizuje a využívá testovou statistiku, která měří rozdíly mezi $\widehat{\beta}$ a $\widehat{\beta}_{IV}$.

Přestože tento test lze snadno provést s využitím implementovaných funkcí ekonometrických programů, je užitečné poznamenat, že jej lze provést s využitím OLS metod. Ukážeme si to na příkladu jednoduchého regresního modelu s jediným instrumentem. Regresní model tak má podobu

$$Y_i = \alpha + \beta X_i + \epsilon_i.$$

Hausmanův test není založen přímo na této regresi, ale na regresi se zahrnutou instrumentální proměnnou, Z:

$$Y_i = \alpha + \beta X_i + \gamma Z_i + \epsilon_i.$$

Z tohoto nám vyplývá, že Hausmanův test je ekvivalentní standardnímu t-testu hypotézy $H_0 : \gamma = 0$. Pokud odhadneme regresi Y na X a Z s využitím metody nejmenších čtverců a následně zjistíme, že Z je statisticky významné, můžeme zamítnou hypotézu o nekorelovanosti vysvětlující proměnné a náhodné složky. To znamená, že zamítneme H_0 a dojdeme k závěru, že IV estimátor by měl být použit. Nebudeme se pouštět do podrobného rozboru toho, proč je tento postup ekvivalentní Hausmanovu testu, poznamenejme jen, že je tento test takto můžeme snadno provést.

Podívejme se nyní na to, jak bude Hausmanův test implementován v modelu vícenásobné regrese s využitím GIVE v případě, kdy máme tři vysvětlující proměnné (X_1, X_2 a X_3), z nichž dvě jsou potenciálně korelovány s náhodnou složkou (X_2 a X_3). Pro X_2 máme dvě instrumentální proměnné (Z_1 a Z_2), kdy pro X_3 máme tři instrumentální proměnné (Z_3, Z_4 a Z_5). V tomto případě je postup Hausmanova testu takový, že provedeme výchozí regresi (viz diskuze týkající se GIVE) k získání vyrovnaných hodnot $\widehat{X}_2$ a $\widehat{X}_3$ a následně „vylepšíme“ původní regresi pomocí těchto proměnných. Konkrétně tak Hausmanův test zahrnuje následující kroky:

1. Provedeme OLS regresi X_2 na úrovňovou konstantu, Z_1 a Z_2. Získáme vyrovnané hodnoty $\widehat{X}_2$.

2. Provedeme OLS regresi X_3 na úrovňovou konstantu, Z_3, Z_4 a Z_5. Získáme vyrovnané hodnoty $\widehat{X}_3$.

3. Provedeme OLS regresi Y na úrovňovou konstantu, X_1, X_2, X_3, $\widehat{X}_2$ a $\widehat{X}_3$.

4. Provedeme F-test hypotézy, že koeficienty u $\widehat{X}_2$ a $\widehat{X}_3$ jsou současně rovny nule.

5. Pokud F-test zamítá nulovou hypotézu, pokračujeme v odhadu s využitím zobecněné metody instrumentálních proměnných. Pokud nulovou hypotézu nezamítáme, použijeme OLS k odhadu původní regrese.

Hausmanův test tak může být proveden s využitím standardních OLS technik. Tento test jsme si ukázali v kontextu konkrétního příkladu, nicméně by nám to mělo poskytnou dostatečnou ilustraci toho, jak tento test pracuje obecně. Stručně řečeno, pro každou vysvětlující proměnnou která je potenciálně korelována s náhodnou složkou bychom měli nalézt odpovídající instrument nebo, v případě, kdy existuje více instrumentů, je třeba získat vyrovnané hodnoty z regrese vysvětlující proměnné na své instrumenty. Konečně pak provedeme regresi zahrnující původní vysvětlující proměnné spolu s instrumenty či vyrovnanými hodnotami. Pokud jsou tyto instrumenty a/nebo vyrovnané hodnoty statisticky významné (na základě t-test resp. F-testu), potom nám to dává vodítko pro použití IV estimátoru.

Sarganův test

V předchozí části jsme si popsali postup pro testování toho, jestli je nebo není nutné provést odhad s použitím instrumentálních proměnných, *pokud máme k dispozici platné instrumenty jako je* Z. Ovšem otázkou samozřejmě je to, jak poznáme, že Z je skutečně platný instrument? Připomeňme si, že podmínkou toho, aby Z bylo platným instrumentem je to, že Z musí být korelováno s X (což si snadno ověříme provedením regrese X na Z a pohledem na t-statistiku) a současně nekorelováno s chybou regrese. Právě nekorelovanost s náhodnou složkou není zase tak snadno ověřitelná, protože chyby regrese nejsme schopni pozorovat. Otázka testování toho, že máme skutečně platné instrumenty je tak komplikovaná a obtížná. Ve skutečnosti dokonce platí, že pokud máme případ, kdy máme k dispozici pouze jeden potenciální instrument ke každé vysvětlující proměnné, nebude existovat způsob testování toho, jestli je instrument platný nebo ne. Abychom si to intuitivně přiblížili, uvažujme jednoduchý regresní model

$$Y_i = \beta X_i + \epsilon_i$$

a předpokládejme, že zde existuje jeden potenciální instrument, Z. Aby se jednalo o platný instrument, musí být nekorelovaný s chybou regrese, tedy $cov(Z, \epsilon_i) = 0$. Protože je náhodná složka nepozorovatelná, měli bychom se pokusit o její nahrazení rezidui a vytvořit testovou statistiku založenou na $cov(Z_i, \widehat{\epsilon}_i)$. Ale jaká rezidua bychom měli použít? Rezidua získaná na základě IV estimátoru (tzn. $\widehat{\epsilon}_i^{IV} = Y_i - \widehat{\beta}_{IV} X_i$) vypadají vcelku rozumně, nicméně jsou zcela nevhodná, protože mohou být nekonzistentní. Ve skutečnosti je totiž možné, že Z není platný instrument (což je nakonec to, co chceme testovat) a $\widehat{\beta}_{IV}$ je tak nekonzistentní. OLS rezidua jsou opět nevhodná, protože mohou být nekonzistentní stejně tak (a neexistuje způsob, jak použít Hausmanův test k ověření této možnosti, protože nevíme, jestli je Z platným instrument). Ve skutečnosti tak neexistuje v tomto případě způsob testování validity instrumentů. To je vcelku výžný problém spojený s metodami instrumentálních proměnných. Jak uvidíme dále, v mnoha případech nám ekonomická teorie sama nabízí dobré instrumentální proměnné. Nicméně, v ideálním světe by měl existovat statistický test, který bychom mohli použít k potvrzení toho, že se jedná o dobré instrumenty. V případě, kdy máme jen jeden potenciální instrument ke každé vysvěltující proměnné však jednoduše tohoto ideálu nedosáhneme.

Pokud však (jako v případě zobecněného IV estimátoru) máme k dispozici více

instrumentů než vysvětlujících proměnných korelovaných s náhodnou složkou, potom testy k ověření platnosti instrumentů existují. Odvození těchto testů vyžaduje ekonometrickou teorii přeshující znalosti základního kurzu ekonometrie. Většina ekonometrických programů však opět tyto testy dokáže provést a my je tak můžeme snadno využít v praxi. Populárním testem je *Sarganův test*. Přestože si ho neodvodíme, můžeme si velmi snadno popsat jak se v praxi provádí. Předpokládejme model vícenásobné regrese s k vysvětlujícími proměnnými, které jsou všechny potenciálně korelovány s náhodnou složkou. Celkem máme k dispozici r instrumentálních proměnných, kdy $r > k$. Sarganův test zahrnuje následující kroky:

1. Provedení regrese Y na úrovňovou konstantu, $X_1, \ldots, X_k$ využitím zobecněného IV estimátoru a získání IV reziduí, $\widehat{\epsilon}_i^{IV}$.

2. Provedení OLS regrese IV reziduí, $\widehat{\epsilon}_i^{IV}$, na úrovňovou konstantu a všechny unstrumenty, $Z_1, \ldots, Z_r$ a získání koeficientu determinace, R^2 z této regrese.

3. Sarganova testová statistika je NR^2 a kritické hodnoty lze získat na základě tabulek $\chi^2(r-k)$ rozdělení.

Všimněme si, že tato testová statistika má podobu „velikost vzorku krát R^2 z konkrétní regrese“. Testy s podobnou statistikou jsme viděli již dříve (např. test autokorelace náhodných složek). Tato podoba je charakteristická pro testy Lagrangeových multiplikátorů a i v případě Sarganova testu se o tento druh testu jedná. V předchozím textu jsem nebyli schopni nabídnout důkazy pro řadu výsledků týkající se instrumentálních proměnných. Nicméně pro jeden z nich si stačí všimnout, že kritické hodnoty Sarganova testu zahrnují $\chi^2(r-k)$ rozdělení. Pokud je $r = k$, tak tento test nebude fungovat, protože neexistuje rozdělení $\chi^2(0)$. Ovšem, pokud je $r = k$, znamená to, že máme stejný počet instrumentů jako je počet vysvěltujících proměnných. To je případ, který jsme zmiňovali. Pokud je $r = k$, neexistuje způsob testování validity instrumentů.

5.5.3 Příčiny korelace vysvětlující proměnné a náhodné složky

Doposud jsme si popsali důsledky korelovanosti vysvětlující proměnné s náhodnou složkou, naznačili jsme si co dělat (tj. použít IV estimátor) a ukázali jsme si test toho, jestli je vysvětlující proměnná korelovaná s náhodnou složkou (tj. použití Hausmanova testu). Zatím jsme si ale neřekli, *proč* by měla být vysvětlující proměnná korelovaná s chybou regrese. Existuje řada důvodů, proč by tomu tak být mohlo. V této části si tak ukážeme nekolik příkladů toho, jak může tento druh korelace vzniknout. Nabídne nám to i určité rady ohledně důležité otázky, jak bychom měli vybírat instrumentální proměnné. Je snadné říct, „musíme se podívat po instrumentální proměnné, která je korelovaná s vysvětlující proměnnou, ale nekorelovaná s chybou regrese“, ale může být dost težké takovou proměnnou v praxi najít. V kontextu časových řad je obvyklé jako instrument použito pozorování z minulých období (např. X_{t-1}). Protože jsou časové řady často autokorelovány, znamená to, že instrument bude korelován s vysvětlující proměnnou. Za přepdokladu, že chyba regrese nebude autokorelována (což bychom měli testovat), instrument by neměl být s chybou regrese korelován. V jiných kontextech však výběr instrumentů může být obtížným problém.

Chyba měření ve vysvětlující proměnné

Předpokládejme, že chceme provést regresi

$$Y_i = \beta X_i + \epsilon_i$$

a tato regrese splňuje klasické přepdoklady. Ovšem, tuto regresi nemůžeme provést, protože X_i přímo nepozorujeme, místo toho pozorujeme

$$X_i^* = X_i + \nu_i,$$

kde ν_i je i.i.d. s nulovou střední hodnotou, rozptylem σ_ν^2 a je nezávislá na ϵ_i. Jinými slovy, X pozorujeme s určitou chybou. Tato situace může nastat v řadě případů. Ekonomové například pracují s daty z výběrových šetření, která závisejí na schopnosti jednotlivců vyplnit správně dotazníky. Tito jednotlivci však často mohou dělat při vyplňování chyby a do těchto dat nám tak může vstoupit chyba měření.

Obvykle nebývá chyba měření v závisle proměnné problém (ciž si snadno můžeme odvodit). Chyba měření ve vysvětlujících proměnných však může vést k regresi, kdy vysvětlující proměnné jasou korelovány s náhodnou složkou. Proč tomu tak je? Stačí si nahradit v původní regresi X_i výrazem $X_i - \nu_i$, čímž získáme model:

$$\begin{aligned} Y_i &= \beta(X_i^* - \nu_i) + \epsilon_i \\ &= \beta X_i^* + \epsilon_i^*, \end{aligned}$$

kde $\epsilon_i^* = \epsilon_i - \beta\nu_i$. Tento nový regresní model si můžeme odhadnout, protože máme data o X_i^*. Jaká je však v této nové regresi korelace mezi vysvětlující proměnnou, X_i^* a náhodnou složkou, ϵ_i^*? Využijeme definici kovariance a platnosti operátoru střední hodnoty, čímž získáme

$$\begin{aligned} cov(X_i^*, \epsilon_i^*) &= E(X_i^* \epsilon_i^*) \\ &= E\left[(X_i + \nu_i)(\epsilon_i - \beta\nu_i)\right] \\ &= -\beta\sigma_\nu^2. \end{aligned}$$

Tato kovariance není nulová, pokud je $\beta \neq 0$ (v případě, kdy je $\beta = 0$ nám vysvětlující proměnná v regresi nevystupuje) nebo pokud je $\sigma_\mu^2 \neq 0$ (v případě $\sigma_\mu^2 = 0$ zde však problém chyby měření není, protože to znamená, že $\nu_i = 0$ pro $i = 1, \ldots, N$). Chyba měření ve vysvětlujících proměnných (ale ne ve vysvětlované proměnné) způsobuje jejich korelovanost s chybou regrese. OLS estimátor tak je vychýlený a v ideálním případě by měl být využit IV estimátor.

Model simultánních rovnic

Metody instrumentálních proměnných jsou často využívány v tzv. *modelech simultánních rovnic*. Abychom porozuměli problému, který nám v modelech tohoto typu nastává, musíme si zde zavést několik pojmů, které by však pro nás neměly být cizí z předchozího studia ekonomie. Proměnná je *endogenní*, pokud je determinována uvnitř modelu, který nás zajímá. Nazveme ji *exogenní*, pokud tomu tak není. Tyto pojmy jsou

úzce svázány s otázkou kauzality. Zmiňovali jsme se, že regresi lze nejsnadněji interpretovat v případě, kdy vysvětlující proměnná kauzálně působí nazávisle proměnnou a ne naopak. Jinými slovy, regresnímodel předpokládá, že Y je determinováno tím, co se stane s X. Nespecifikovali jsem nijak to, jak byla proměnná X generována. V tomto případě je Y závisle proměnná, je endogenní, a X, vysvětlující proměnná, je chápána jako exogenní. Intuitivně řečeno, pokud předpokládámě, že vysvětlující proměnné jsou exogenní, je použití OLS odhadu v pořádku. Pokud jsou některé vysvětlující proměnné endogenní, je dost možné, že jsou korelovány s náhodnou složkou a je potřeba IV odhadu.

Klasická ilustrace tohoto problému je obsahem modelu nabídky a poptávky. Předpokládejme poptávkovou křivku nějakého produktu v podobě

$$Q_D = \beta_D P + \epsilon_D,$$

kde Q_D je poptávané množství, které závisí na ceně, P. Do této regresní rovnice jsme dodaly i chybu modelu a používáme index D k vyjádření rovnice poptávky. Předpokládejme dále nabídkovou křivku danou jako

$$Q_S = \beta_S P + \epsilon_S,$$

kde Q_S je nabízené množství závisející na ceně P. Index S používáme k označení parametrů a proměnných vztažených k rovnici nabídky. V rovnováze platí $Q_D = Q_S$ a tato podmínka je využívána k nalezení rovnovážné ceny a množství. Uvědomme si, že cena i množství jsou determinovány v rámci modelu, a jedná se tak o endogenní proměné. Tyto dvě rovnice (spolu s rovnicí identity, $Q_D = Q_S$) jsou modelem simultánních rovnic.

Tímto jsme si představili problém z hlediska ekonomie, co je ale problémem ekonometrickým? Ekonometra zajímá odhad (na základě dat) parametrů sklonu nabídkové a poptávkové křivky. Co se stane, když získáme údaje o množstvích a cenách a provedeme regresi množství na cenu? Získáme určitě OLS odhady parametrů sklonu. Jedná se však o odhady β_D nebo β_S? Neexistuje způsob, jak se to dozvědět. V praxi OLS pravděpodobně neodhadne ani nabídkovou ani poptávkovou křivku.

K formálnější ilustraci problémů, které nám vznikají v rámci modelů simultánních rovnic (a jak se tyto problémy vztahují k otázce instrumentálních proměnných), přidejme do jedné z rovnic exogenní proměnnou. Předpokládejme, že poptávané množství nějakého zboží závisí také na důchodu, I:

$$Q_D = \beta_D P + \gamma I + \epsilon_D.$$

To může být vcelku rozumné, protože rozhodování spotřebitele o jeho nákupech nezávisí jen naceně statků, ale i na jeho důchodu. Nabídka je určena společnostmi vyrábějících daný produkt a není pravděpodobné, že by závisela na důchodu spotřebitelů, nakupujících jejich produkt. Nabídková křivka tak nebude na důchodu záviset. Důchod je exogenní proměnná a není determinována v rámci modelu nabídky a poptávky.

Budeme předpokládat, že chyby regrese v rovnicích nabídky a poptávky splňují klasické předpoklady (s výjimkou předpokladu, že vysvětlující proměnné jsou nekorelovány s náhodnou složkou). Indexy S a D použijeme k rozlišení toho, jestli se náhodná

složka vztahuje krovnici nabídky nebo poptávky. Například předpoklad, že rovnice poptávky má homoskedastické náhodné složky implikuje

$$var(\epsilon_D) = \sigma_D^2.$$

V praxi bude mít ekonometr k dispozici data pro $i = 1, \ldots, N$ (nebo pro $t = 1, \ldots, T$) pozorování, ale pro přehlednost tyto indexy nebudeme v našem značení používat.

Dvě rovnice původního modelu nabídky a poptávky mají podobu tzv. *strukturální formy* resp. *strukturální formy modelu*. Vycházejí tedy z ekonomické teorie (ekonomické struktury) daného problému a mají endogenní proměnné na obou stranách rovnic (vysvěltující proměnné jsou tedy i endogenní proměnné). Využitím základních matematických operací si můžeme tyto rovnice přeuspořádat tak, abychom měli jen na pravé straně exogenní proměnné a pro každou endogenní proměnnou budeme mít jednu rovnic. Výsledkem je tzv. *redukovaná forma modelu* či jen *redukovaná forma*. Pro model nabídky a poptávky položíme rovnítko (v modelu platí identita rovnováhy mezi nabídkou a poptávkou) mezi pravé strany nabídkové i poptávkové rovnice a uspořádáme si výsledky tak, že na levé straně bude pouze P:

$$\begin{aligned} P &= \frac{-\gamma}{\beta_D - \beta_S} I + \frac{\epsilon_S - \epsilon_D}{\beta_D - \beta_S} \\ &= \pi_1 I + \epsilon_1, \end{aligned}$$

kde poslední rovnice je zjednodušením pro kompaktnost značení, kdy místo výrazu $-\gamma\beta_D - \beta_S$ budeme používat jednoduše koeficient redukované formy π_1 a podobně na tom bude i náhodná složka redukované formy modelu, ϵ_1.

Pokud si označíme rovnovážný výstup jako Q (tzn. $Q_S = Q_D \equiv Q$) a dosadíme získaný výraz pro cenu, P, do rovnice nabídky, můžeme získat druhou rovnici redukované formy:

$$\begin{aligned} Q &= \beta_S(\pi_1 I + \epsilon_1) + \epsilon_S \\ &= \beta_S \pi_1 I + \beta_S \epsilon_1 + \epsilon_S \\ &= \pi_2 I + \epsilon_2. \end{aligned}$$

Je potřeba zdůraznit, že rovnice redukované formy se odlišují od původních strukturálních rovnic. Redukovaná forma v sobě obsahuje jednu rovnici pro každou z endogenních proměnných a vysvětlující proměnná v každé z rovnic je I, což je exogenní proměnná.

Na základě těchto čtyř rovnic (dvě pro redukovanou formu modelu a dvě pro strukturální formu modelu) si můžeme ukázat několik zajímavých problémů týkajících se modelu nabídky a poptávky. Prvním z nich je to, že naivní výzkumník se může podívat na první rovnici strukturální podoby modelu (tj. rovnice poptávkové křivky) a provést regresi Q na P a I, čímž získá odhad poptávkové křivky. Rovnice redukované formy pro P nám však jasně říká, že tento postup nám dá vychýlenné odhady. Můžeme si totiž ukázat, že vysvětlující proměnná P a chyby regrese, ϵ_D, jsou v rovnici tohoto našeho

naivního výzkumníka navzájem korelovány:

$$\begin{aligned} cov(P, \epsilon_D) &= E(P\epsilon_D) \\ &= E\left[(\pi_1 I + \epsilon_1)\epsilon_D\right] \\ &= E\left[\left(\frac{\epsilon_S - \epsilon_D}{\beta_D - \beta_S}\right)\epsilon_D\right] \\ &= \frac{-\sigma_D^2}{\beta_D - \beta_S} \\ &\neq 0, \end{aligned}$$

kde jsme využili skutečnosti, že I je exogenní (a lze ji brát jako nenáhodnou veličinu), a vlastností operátoru střední hodnoty. Protože je vysvětlující proměnn, Pm korelovaná s chybou regrese, ukázali jsme si, že přímý OLS odhad poptávkové rovnice (tzn. odhad strukturální rovnice, kde je jedna z vysvěltujících proměnných endogenní) vede k vychýleným a nekonzistentním výsledkům. Podobné odvození lze provést i pro aplikace přímého OLS odhadu na nabídkovou křivku, což nás rovněž dovede k vychýleným a nekonzistentním odhadům.

Dále předpokládejme OLS odhad rovnic redukované formy. Lze snadno ukázat (což si lze provést jako cvičení), že chyby regrese redukované formy splňují klasické předpoklady. Protože jedinou vysvětlující proměnnou je exogenní proměnná (kterou lze brát jako nenáhodnou veličinu), je OLS nejlepší lineární nestranný estimátor pro model redukované formy. OLS odhady redukované formy, $\widehat{\pi}_1$ a $\widehat{\pi}_2$, jsou tedy dobré odhady a lze je snadno získat. Jaké jsou však odhady strukturálních parametrl, tedy parametrů strukturální podoby modelu? Ty nás mnohdy zajímají více, protože strukturální parametry mají svou věcnou, ekonomickou interpretaci (např. sklony nabídkové a poptávkové křivky). Koeficienty redukované formy nám přímo neposkytují odhady koeficientů strukturální formy modelu, tedy v našem případě odhady parametrů sklonu poptávkové a nabídkové křivky.

Často je však možné, že lze koeficienty redukované formy použít k odhadům koeficientů strukturální podoby modelu. V našem příkladě platí

$$\pi_2 = \beta_S \pi_1.$$

Můžeme tedy použít nevychýlené odhady parametrů redukované formy, $\widehat{\pi}_1$ a $\widehat{\pi}_2$, a získat odhad sklonu nabídkové křivky:

$$\widehat{\beta}_S = \frac{\widehat{\pi}_1}{\widehat{\pi}_2}.$$

Tento postup je nazýván jako *nepřímé nejmenší čtverce (indirect least squares)*. Problém toho, jestli lze z odhadu koeficientů redukované formy odvodit koeficienty strukturální podoby modelu je úzce svázán s otázkou tzv. identifikovatelnosti.

Jak se toto vše týká tématu instrumentálních proměnných? Platí, že metoda nepřímých nejmenších čtverců je ekvivalentní IV odhadu nabídkové křivky s využitím I jako instrumentu (v rámci cvičení je možné si tuto skutečnost odvodit). Problematice metody odhadu simultánních rovnic se nebudeme na tomto místě dále věnovat. Existuje řada odhadových metod, které spadají do skupiny IV estimátorů. Jednou z nich je *dvoustupňová metoda nejmenších čtverců (two-stage least squares – TSLS)*.

K použití IV estimátorů v praxi potřebujeme instrumenty. Náš příklad ukázal, jak tyto instrumenty získat v případě modelu simultánních rovnic. Poznamenejme, že estimátor nepřímých nejmenších čtverců nám poskytl odhad parametru sklonu nabídkové křivky. Sami se můžeme přesvědčit, že neexistuje způsob, jak s využitím estimátoru nepřímých nejmenších čtverců získat odhad parametru β_D, tedy parametru sklonu poptávkové křivky. Proč tomu tak je? Zásadní význam má to, že důchod nevystupuje v křivce nabídky. Obecným pravidlem je, že pokud máme exogenní proměnnou, která není obsažena v některé z rovnicí, lze ji použít jako instrumentální proměnnou pro tuto rovnici. V našem případě neexistuje žádná exogenní proměnná, který by nebyla zahrnuta v rovnici poptávky, a je tak nemožné nalézt IV estimátor, který by nám dovolil odhadnout koeficient sklonu křivky poptávky. Pokud by však křivka nabídky obsahovala exogenní vysvětlující proměnnou, která by nebyla přítomna v rovnici poptávky, bylo by možno provést IV odhad rovnice poptávky. Pokud bychom například měli model týkající se nějakého zemědělského produktu, mohla by být takovou exogenní proměnnou počasí, protože počasí určitě ovlivňuje nabídku takovéhoto produktu, ale asi těžko ovlivní poptávku.

Předchozí diskuze by nám měla umožnit použití ekonometrických programů (a interpretvoat rozumně výsledky) pro modely simultánních rovnic. Téma simultánních rovnice je však velmi obsáhlé téma, a existuje zde spousta dalších otázek, které jsme zde nestihli ani naznačit. Například, co se stane, když se nějaká endogenní proměnná objeví v jedné rovnici, ale ne v jiné (např. cena se objeví jen v rovnici poptávky, ale ne nabídky). Tato situace by nám totiž pomohla v našich odhadech. V extrémním případě, kdyby existovala rovnice bez endogenních vysvětlujících proměnných, by OLS odhad této rovnice vedl ke konzistentním odhadům. Tato skutečnost , spolu s pravidlem, že pokud máme exogenní proměnné, které nejsou obsaženy v některé rovnici, můžeme je použít jako instrumenty pro tuto rovnici, by nám dovolila rozumnou emprickou práci v řadě kontextů. Pokud bychom se však chtěli pouštět do práce s modely simultánních rovnic, nejlepší cestou je seznámit se s tímto tématem v pokročilejších ekonometrických učebnicích.

Příklad korelace vysvětlující proměnné s náhodnou složkou

Ekonomická teorie nám může sama naznačit, že vysvětlující proměnná je korelovaná s náhodnou složkou a pokud máme to štěstí, tak nám třeba i nabídne rozumnou instrumentální proměnnou. V této části kapitoly se podíváme na příklad toho, kdy by tato situace mohla nastat.

Předpokládejme, že nás zajímá odhad výnosů ze vzdělání (returns to schooling) a máme data z průzkumů mnoha jednotlivců, kdy závisle proměnná je Y = důchod, a vysvětlující proměnné jsou X = počet let vzdělávání ve škole a další vysvětlující proměnné (např. pracovní zkušenosti, věk, typ zaměstnání apod.). Tyto další proměnné budeme pro jednoduchost ignorovat. Odhad parametru u proměnné X nás zajímá nejvíce, protože se jedná o měřítko výnosů ze vzdělání. V tomto případě je však pravděpodobné, že X je korelováno s náhodnou složkou a OLS odhad tak bude nekonzistentní. Než se dostaneme k tomu, proč by tak tomu mělo být, zamysleme se nad interpretací náhodných složek. Jednotlivci s pozitivní náhodnou složkou získávají neobvykle vyšší úroveň příjmu. Jejich příjem je tak vyšší než by odpovídalo jejich vzdělání. Jed-

notlivci s negativními chybami pak dostávají neobvykle nižší úroveň příjmu. Jejich příjem kat je nižší než by napovídalo jejich vzdělání. Co by mohlo být korelováno s touto náhodnou složkou? Každý jednotlivec má obvykle své osobní kvality jako jsou inteligence, ambice, úsilí, talent, štěstí nebo podporu rodiny. Nazvěme pro jednoduchost tuto kvalitu „talentem“. Tato kvalita bude pravděpodobně spojena s náhodnou chybou (např. jednotlivci s větším talentem mají tendenci dosáhnout neobvykle vyššího příjmu). Nicméně tato kvalita může ovlivnit i volbu vzdělání daného jednotlivce. Například více talentovaní jedinci půjdou mnohem pravděpodobněji studovat na univerzitu (a získají tak více let vzdělání). Budeme tak pravděpodobně čelit situaci, kdy talentovaní jedinci budou mít tendenci mít jak více let vzdělání, tak i neobvykle vysoké příjmy (tzn. kladné chyby regrese). Náhodná chyba a vysvětlující proměnná by tak měla být ovlivněna talentem. V tomto případě by chyba regrese a vysvětlující proměnná, odpovídající rokům vzdělání, měly být pozitivně korelovány.

To nám samozřejmě naznačuje, že regrese Y na X (a další vysvětlující proměnné) by neměla být odhadována metodou nejmenších čtverců, protože vysvětlující proměnná je korelovaná s náhodnou složkou. Výzkumníci, kteří chtěli odhadnout výnosy ze vzdělání, k tomuto problému zaujali dva hlavní přístupy. Někteří usilovali o nalezení jiných vysvětlujících proměnných, které by byl proxy proměnnými chybějící kvality, tedy např. talentu z předchozího odstavce. Tímto druhem proměnných by mohly být výsledky testu inteligence. Alternativní přístup pak bylo použití metody instrumentálních proměnných. V našem případě chceme proměnnou, která by byla korelována s rozhodováním o počtu vystudovaných let na škole a která by byla nekorelovaná s náhodnou složkou (tzn. nebyla by nijak závislá na faktorech, které by vysvětlovaly proč jednotlivci mají neobvykle vysoké nebo nízké příjmy). Alternativní způsob vyjádření tohoto problému je to, že chceme najít proměnnou, která ovlivňuje počet let vzdělání a která nemá přímý vliv na příjem jendotlivců.

Někteří výzkumníci použili jako instrumenty charakteristiky rodičů či starších sourozenců. Oprávněnost této volby je takové, že pokud má některý z rodičů univerzitní vzdělání, bude daný jednotlivec pocházet pravděpodobně z rodiny, která si vzdělání cení. To zvyšuje šance dostat se na univerzitu. Zaměstnavatel se však neptá na to, jestli rodiče dané osoby chodili na univerzitu, a tak by tato proměnná neměla mít přímý vliv na příjem dotazovaného jednotlivce. Vzdělání jeho rodičů ovlivňuje rozhodování o počtu let vzdělání a nemá přímý vliv na příjem.

Jiní výzkumníci (zejména z USA) využili jako instrument geografické místo, ze kterého jedinci pocházejí. Zdůvodnění je takové, že pokud člověk žije v komunitě kde existuje univerzita nebo vyší odborná škola, je mnohem pravděpodobnější, že tento člověk půjde na univerzitu. (či vyšší odborná škola). Zaměstnavatele opět nebude zajímat, odkud daný člověk pochází a proměnná vyjadřující místo, odkud člověk pochází by tak neměla mít přímý vliv na příjem. Geografické místo, odkud jednotlivec pochází tak ovlivňuje rozhodování o vzdělání, ale neovlivňuje přímo jeho příjem.

Tento příklad byl snad alespoň drobnou ilustrací toho, jak nás samotný ekonomický problém, který analyzuje. může varovat před možností korelace některé z vysvětlujících proměnných s náhodnou složkou a případně i může nabídnout možné instrumenty.

5.6 Shrnutí

Na základě této kapitoly tedy již víme:

- jak testovat mnohá z porušení klasických předpokladů;
- jak postupovat, pokud na nesplnění některého z předpokladu narazíme;
- že předpoklad nulové střední hodntoy náhodných složek je automaticky splněn, pokud pracujeme s modelem s úrovňovou konstantou;
- že předpoklad normality můžeme opomenout, pokud využijeme závěry asymptotické teorie;
- že i při porušení přepdokladu homoskedasticity a nekorelovanosti náhodných složek zůstává OLS estimátor nestranný;
- že použití metody OLS při nesplnění těchto předpokladů nás vede na scestí, protože výsledné intervaly spolehlivosti postupy testování hypotéz jsou nekorektní;
- že existují estimátory (např. heteroskedasticitě nebo autokorelaci konzistentní estimátory), které umožňují korektní využívání metody OLS;
- že i přes použití alternativních estimátorů rozptylu náhodných složek není OLS odhad vydatný;
- že existuje zobecněná metoda nejměnších čtverců (GLS), kterou lze interpretovat jako metodu OLS aplikovanou na vhodně transformovaný model;
- že pokud je vysvětlující proměnná korelována s náhodnou složkou, je OLS estimátor vychýlenný a nekonzistentní a neměl by být využíván;
- že existuje estimátor metody instrumentálních proměnných (IV), díky kterému lze předchozí problém překonat;
- že základní závěry této kapitoly jsou:

  1. V případě práce s malými vzorky je předpoklad normality relativně důležitý, protože při jeho nesplnění jsou veškeré následné testy zcestné.
  2. K testování normality využijeme např. Jarqueův-Berův test, který testuje sdruženou hypotézu o šikmosti a špičatosti výběrového rozdělení reziduí.
  3. Problém s normalitou lze často řešit omezením (zahrnutím) vlivu několika odlehlých pozorování, které za nesplněním předpokladu normality stojí. Řešením je rovněž použití metody maximální věrohodnosti aplikované na model i s „nenormálně“ rozdělenými náhodnými složkami.
  4. Pokud mají chyby regrese různé rozptyly (heteroskedasticity), nebo jsou vzájemně korelované (autokorelace náhodných složek), je OLS estimátor nestranný, ale není již nejlepší. Nejlepší estimátor je GLS estimátor.

5. Pokud je v regresi přítomna heteroskedasticita, lze GLS estimátor spočítat s využitím OLS estimátoru aplikovaného na transformovaný model. Tato transformace vyžaduje vážení každého pozorování a metoda je tak nazývána metodou vážených nejmenších čtverců (WLS). Pokud nejsme schopni nalézt odpovídající transformaci, měli bychom použít heteroskedasticitě konzistentní estimátor.
6. Existuje řada testů heteroskedasticity, zahrnujících GOldfeldův-Quandtův test, Breuschův-Paganův test nebo Whiteův test.
7. V případě časových řad je obvyklé, že náhodné složky jsou vzájemně korelovány. Obvyklým způsobem modelování tohoto faktu je chápání náhodných složek jakožto autoregresních procesů. GLS estimátor lze vypočítat s využitím metody OLS na transformovaný model. Požadovaná transformace zahrnuje kvazi diferencování každé z proměnných. Populární metodou implementace GLS estimátoru je Cochranova-Orcuttova procedura.
8. Existuje řada testů autokorelace náhodných složek, zahrnujících LM test, Boxův-Pierceův test, Ljungův test, Durbinův-Watsonův test a Durbinův h-test.
9. V řadě aplikací je nepravděpodobné chápání vysvětlující proměnné jako pevně dané (nenáhodné) veličiny. Je tak nutné uvolnění předpokladu o její nenáhodnosti.
10. Pokud jsou vysvětlující proměnné náhodné, ale všechny jsou nekorelované s náhodnou složkou, lze použít standardní OLS techniky z předchozích kapitol.
11. Pokud jsou vysvětlující proměnné náhodné veličiny a nektěré z nich jsou korelovány s náhodnýmisložkami regrese, je OLS estimátor nekonzistentní. Oproti tomu IV estimátor je konzistentní.
12. V případě vícenásobné regrese je potřeba nejméně jednoho instrumentu pro každou z vysvětlujících proměnných korelovaných s náhodnou složkou.
13. Pokud máme platný instrument, lze použít Hausmanův test k testování toho, jestli jsou vysvětlující proměnné korelovány s náhodnou složkou.
14. Obecně je obtížné testovat platnost (sílu) instrumentální proměnné. Pokud však máme více instrumentů než je minimálně třeba, lze použít Sarganův test k testování toho, jestli je nějaký instrument platný.
15. Důležitý případ, kdy vysvětlující proměnné mohou být korelovány s náhodnou složkou, nastává v situaci, kdy jsou vysvětlující proměné měřeny s nějakou chybou.
16. Model simultánních rovnic je dalším případem, kdy jsou vysvětlující proměnné korelovány s náhodnou složkou.

Měli bychom tak již znát a umět vysvětlit obsah následujících klíčových pojmů:

- ❀ Šikmost
- ❀ Špičatost

❀ Jarqueův-Berův test
❀ Zobecněná metoda nejmenších čtverců
❀ Heteroskedasticitě konzistentní estimátor
❀ Goldfeldův-Quandtův test
❀ Whiteův test
❀ Autoregresní proces
❀ Iterační Cochranova-Orcuttova procedura
❀ Neweyho-Westův estimátor
❀ Věrohodnostní test (LM test)
❀ Boxův-Pierceův test
❀ Durbinův-Watsonův test
❀ Metoda instrumentálních proměnných
❀ Zobecněný IV estimátor
❀ Sarganův test
❀ Endogenní proměnná
❀ Strukturní forma modelu
❀ Nepřímé nejmenší čtverce
❀ Transformace modelu
❀ Metoda vážených nejmenších čtverců
❀ Test heteroskedasticity
❀ Breuschův-Paganův test
❀ Autokorelace náhodných složek
❀ Cochranova-Orcuttova procedura
❀ Autokorelaci konzistentní estimátor
❀ Testování autokorelace
❀ Breuschův-Godfreyeho test
❀ Ljungův test
❀ Durbinova h-statistika
❀ Instrumentální proměnná (instrument)
❀ Hausmanův test
❀ Model simultánních rovnic
❀ Exogenní proměnná
❀ Redukovaná forma modelu
❀ Dvoustupňová metoda nejměnších čtverců

Příloha: Asymptotické výsledky pro OLS a IV estimátor

Asymptotické vlastnosti v případě nekorelovanosti

Tuto přílohu začneme diskuzí OLS estimátoru v případě, kdy jsou splněny všechny klasické předpoklady, kromě toho, že vysvětlující proměnná bude náhodná veličina nekorelovaná s náhodnou složkou. Konkrétně , X_i pro $i = 1, \ldots, N$ jsou i.i.d. náhodné veličiny, kdy

$$E(X_i) = \mu_X,$$
$$var(X_i) = \sigma_X^2.$$

Navíc jsme předpokládali, že

$$cov(X_i, \epsilon_i) = E(X_i \epsilon_i) = 0.$$

Asymptotické výsledky pro tento případ byly odvozeny v kapitole 3, Příloha 2: Využití asymptotické teorie v jednoduchém regresním modelu. Nebudeme zde tedy tato odvození opakovat. Užitečné je ale připomenout si, že $\widehat{\beta}$ je konzistentní estimátor parametru

β a OLS estimátor je asymptoticky normální. To znamená, že pro $N \to \infty$ máme

$$\sqrt{N}(\widehat{\beta} - \beta) \sim N\left(0, \frac{\sigma^2}{\sigma_X^2 + \mu_X^2}\right).$$

V praxi lze tyto výsledky využít tak, že výraz $\sigma_X^2 + \mu_X^2$ můžeme nahradit výrazem $\frac{1}{N}\sum X_i^2$, protože tento výraz je konzistentním odhadem výrazu $\sigma_X^2 + \mu_X^2$. Pokud využijeme tohoto estimátoru a drobně upravíme pořadí členů v předchozím výrazu, můžeme psát

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum X_i^2}\right).$$

Tento výsledek byl východiskem pro řadu důležitých výpočtů v kapitole 3 (např. konstrukce intervalů spolehlivosti a t-testů). Pokud je tedy splněn předpoklad nekorelovanosti náhodné vysvětlující proměnné s náhodnou složkou, lze využít veškeré nám známe výsledky spojené s metodou nejmenších čtverců.

Asymptotické vlastnosti v případě korelovanosti

V této části přílohy odvodíme asymptotickévlasntosti OLS a IV estimátoru v případě situace, kdy jsou splněny veškeré klasické předpoklady, kromě toho, že $cov(X_i, \epsilon_i) \neq 0$.

Vlastnost 1: $\widehat{\beta}$ je nekonzistentní estimátor β.

Důkaz:

$$\begin{aligned}
\operatorname{plim}\left(\widehat{\beta}\right) &= \operatorname{plim}\left(\beta + \frac{\sum X_i\epsilon_i}{\sum X_i^2}\right) \\
&= \beta + \operatorname{plim}\left(\frac{\sum X_i\epsilon_i}{\sum X_i^2}\right) \text{ dle Slutského teorému} \\
&= \beta + \operatorname{plim}\left(\frac{\frac{1}{N}\sum X_i\epsilon_i}{\frac{1}{N}\sum X_i^2}\right) \\
&= \beta + \frac{\operatorname{plim}\left(\frac{1}{N}\sum X_i\epsilon_i\right)}{\operatorname{plim}\left(\frac{1}{N}\sum X_i^2\right)} \text{ dle Slutského teorému.}
\end{aligned}$$

Nyní lze využít zákon velkých čísel k vyhodnocení limity v pravděpodobnosti jmenovatele, tedy $\operatorname{plim}\left(\frac{1}{N}\sum X_i^2\right) = E(X_i^2) = \sigma_X^2 + \mu_X^2$. Poslední rovnost vyplývá z uspořádání vztahu pro rozptyl: $var(X_i) = E(X_i^2) - [E(X_i)]^2$.

Tedy,

$$\operatorname{plim}\left(\widehat{\beta}\right) = \beta + \frac{\operatorname{plim}\left(\frac{1}{N}\sum X_i\epsilon_i\right)}{\sigma_X^2 + \mu_X^2}.$$

Nyní využijeme zákon velkých čísel pro vyjádření limity v pravděpodobnosti, což je výraz $\operatorname{plim}\left(\frac{1}{N}\sum X_i\epsilon_i\right)$:

$$\operatorname{plim}\left(\frac{1}{N}\sum X_i\epsilon_i\right) = E(X_i\epsilon_i) = cov(X_i, \epsilon_i) \neq 0.$$

Tedy

$$\text{plim}\left(\widehat{\beta}\right) = \beta + \frac{cov(X_i, \epsilon_i)}{\sigma_X^2 + \mu_X^2} \neq \beta,$$

čímž je dokázána nekonzistence OLS estimátoru.

Vlastnost 2: $\widehat{\beta}_{IV}$ je konzistentní estimátor parametru β.

Důkaz: Nejprve nahradíme Y_i výrazem $\beta X_i + \epsilon_i$ a upravíme výraz pro $\widehat{\beta}_{IV}$ do podoby

$$\widehat{\beta}_{IV} = \beta + \frac{\sum Z_i \epsilon_i}{\sum X_i Z_i}.$$

Aplikací limity v pravděpodobnosti na tuto rovnici získáváme

$$\begin{aligned}
\text{plim}\left(\widehat{\beta}_{IV}\right) &= \text{plim}\left(\beta + \frac{\sum Z_i \epsilon_i}{\sum X_i Z_i}\right) \\
&= \beta + \text{plim}\left(\frac{\sum Z_i \epsilon_i}{\sum X_i Z_i}\right) \text{ dle Slutského teorému} \\
&= \beta + \text{plim}\left(\frac{\frac{1}{N}\sum Z_i \epsilon_i}{\frac{1}{N}\sum X_i Z_i}\right) \\
&= \beta + \frac{\text{plim}\left(\frac{1}{N}\sum Z_i \epsilon_i\right)}{\text{plim}\left(\frac{1}{N}\sum X_i Z_i\right)} \text{ dle Slutského teorému.}
\end{aligned}$$

Lze opět využít zákon velkých čísel k vyjádření limity v pravděpodobnosti, tedy výrazu $\text{plim}\left(\frac{1}{N}\sum Z_i \epsilon_i\right)$. Konkrétně nám tento zákon říká, že

$$\text{plim}\left(\frac{1}{N}\sum Z_i \epsilon_i\right) = E(Z_i \epsilon_i) = 0.$$

Zákon velkých čísel použijeme i k vyjádření $\text{plim}\left(\frac{1}{N}\sum X_i Z_i\right) = E(X_i Z_i) = \sigma_{XZ} + \mu_X \mu_Z$, kde pro vyjádření poslední rovnosti použijeme vzorec pro kovarianci. Tedy,

$$\text{plim}\left(\widehat{\beta}_{IV}\right) = \beta + \frac{0}{\sigma_{XZ} + \mu_X \mu_Z} = \beta,$$

čímž je dokázána konzistence IV estimátoru.

Vlastnost 3: IV estimátor je asymptoticky normální

Pro $N \to \infty$ platí

$$\sqrt{N}(\widehat{\beta}_{IV} - \beta) \sim N\left(0, \frac{(\sigma_Z^2 + \mu_Z^2)\sigma^2}{(\sigma_{XZ} + \mu_X \mu_Z)^2}\right).$$

Důkaz: Rovnici odvozenou v předchozí vlastnosti, tedy

$$\widehat{\beta}_{IV} = \beta + \frac{\sum Z_i \epsilon_i}{\sum X_i Z_i},$$

lze přepsat jako

$$\sqrt{N}(\widehat{\beta}_{IV} - \beta) = \sqrt{N}\frac{\sum Z_i\epsilon_i}{\sum X_i Z_i} = \sqrt{N}\frac{\frac{1}{N}\sum Z_i\epsilon_i}{\frac{1}{N}\sum X_i Z_i}.$$

Centrální limitní věta aplikovaná na čitatel nám říká, že pro $N \to \infty$,

$$\sqrt{N}\frac{1}{N}\sum Z_i\epsilon_i \sim N\left(0, var(Z_i\epsilon_i)\right).$$

Využitím definice rozptylu, vlastností operátoru střední hodnoty a s vědomím toho, že předpokládáme vzájemnou nekorelovanost Z_i a ϵ_i, tedy $cov(Z_i, \epsilon_i) = 0$ vede ke skutečnosti, že

$$var(Z_i\epsilon_i) = (\sigma_Z^2 + \mu_Z^2)\sigma^2.$$

V důkazu vlastnosti konzistence IV estimátori jsme si ukázali, že

$$\operatorname{plim}\left(\frac{1}{N}\sum X_i Z_i\right) = \sigma_{XZ} + \mu_X\mu_Z.$$

S využitím Cramerova teorému můžeme říct, že

$$\sqrt{N}(\widehat{\beta}_{IV} - \beta) \xrightarrow{N} N\left(0, \frac{(\sigma_Z^2 + \mu_Z^2)\sigma^2}{(\sigma_{XZ} + \mu_X\mu_Z)^2}\right).$$

Kapitola 6

Modely kvalitativních a omezených proměnných

V této kapitole se dozvíme:

- ☞ jaké typy modelů použít v případě umělé vysvětlované proměnné;
- ☞ jaké možnosti máme pro případ modelů volby mezi více alternativami;
- ☞ jaké modely využít v případě cenzorovaných vysvětlovaných proměnných;
- ☞ jaké modely využít v případě vysvětlovaných proměnných vyjadřujících počet;
- ☞ jak interpretovat výsledky odhadů všech těchto modelů;
- ☞ co je to podíl šancí, pravděpodobnost volby alternativy, prediktivní pravděpodobnost volby, mezní efekt vlivu vysvětlující proměnných, podíl relativních incidencí, apod.

6.1 Úvod

Regrese je mocný nástroj pro měření míry vlivu vysvětlujících proměnných na proměnnou závislou. V předchozích kapitolách byla prezentována široká paleta modelů, které mohou mít intepretaci regresních modelů. Nebylo příliš zdůrazňováno, že závisle proměnné byly fakticky spojité náhodné veličiny, které mohly nabývat jakýchkoli hodnot. Existují však případy, kdy je závisle proměnná omezena jen na určitý omezený obor hodnot. Může se jednat o umělou proměnnou, která bude nabývat hodnot 0 nebo 1. Může to být proměnná vyjadřující počet omezený na hodnoty $0, 1, 2, \ldots$. V těchto případech nejsou regresní metody předchozích kapitol vhodné. Vzniká nám otázka, jak konstruovat podobný typ regresních modelů (tzn. že vysvětlující proměnná bude ovlivňovat proměnnou závislou), které by pokryly námi zmiňované typy vysvětlované proměnné. A tato otázka bude zodpovězena v této kapitole. V první části se budeme

zabývat *modely kvalitativní (diskrétní) volby*, kdy vysvětlovaná proměnné označuje nějaký druh volby. V další části zmíníme případ *modelů omezené vysvětlované proměnné*, které umožňují různé druhy omezení vysvětlované proměnné. Na první pohled může terminologie a metody zmiňované v této kapitole znít nově a exoticky. Nicméně intuice v pozadí je tatáž jako v předchozích kapitolách: snažíme se nalézt závislost mezi vysvětlovanou a vysvětlujícími proměnnými.

Začneme případem, kdy je vysvětlovaná proměnná umělá. S umělými vysvětlujícími proměnými jsme měli tu čest setkat se v kapitole 2. V řadě případů je to ale vysvětlovaná proměnná, která je umělá. Jako příkald si ilustrujme situaci, kdy nás zajímá zkoumání toho, proč si někteří lidé volí pro cestu do práce auto a jiní jezdí veřejnou dopravou. Data, ze kterých v těchto případech vycházíme pocházejí z dotazníkových šetření, kdy jsou lidé dotazování na to, jestli do práce (nebo kamkoli jinam, záleží na problému, který zkoumáme) cestují autem nebo veřejnou dopravou a k tomu udávají i údaje osobního charakteru, jako je vzdálenost do práce, plat, apod. Regresní model vhodný pro tuto ekonomickou analýzu by jako vysvětlující proměnné volil právě tyto osobní charakteristiky. Závisle proměnná však bude umělá a bude nabývat hodnot 1, pokud dotazovaný jezdí do praáce autem a 0 pokud tomu tak není a do práce jezdí raději veřejnou dopravou. V oředchozích kapitolách jsme pracovali s klsickým předpokladem, že závisle proměnná (respektive náhodná složka) má normální rozdělení. Není moc pravděpodobné, že by vysvětlovaná proměnná nabývající hodnot 0 nebo 1 tuto podmínku splňovala. To je jeden z důvodů, proč jsou regresní metody přímo nepoužitelné (nebo alespoň nevhodné a zkreslující). Ve světle tohoto důvodu (a řady dalších) je nutné zabývat se novými metodami, a v souladu s tímto si zavedeme modely zvané *probit* a *logit*.

Probit a logit modely jsou vhodné v případě, kdy je volba omezena na dvě věci (např. auto nebo veřejná doprava, koupím nebo nekoupím daný výrobek apod.). Z tohoto důvadu používáme pojen *modely binární volby*. V řadě případů však mámem ožnost volby z mnoha alternativ. V příkladu řešení otázky volby dopravního prostředku pro cestu do práce můžeme zařadit i další volbu, např. kolo (a rozšířit tak paletu alternativ na tři). Oblast marketingu zahrnuje má podobné příklady. Řada marketingových studií pracuje s daty, kdy se spotřebitel rozhoduje o volbě řady značek jednoho typu výrobku (např. pro kečupy může volit z pěti dostupných alternativ, které se nacházejí na pultech obchodů). Tento typ modelů je nazýván *modely multinomiální (vícenásobné) volby*. Popíšeme si, jak je možné modely probit a logit rozšířit i pro tyto případy. Výsledné modely jsou označovány jako *multinomiální probit* a *multinomiální logit*.

V druhé polovině kapitoly budeme rozebírat modely, kdy je závisle proměnná nějakým způsobem omezena. Nabízí se řada modelů, které v sobě možnost *omezení (oříznutí)* vysvětlované proměnné obsahují (anglicky se hovoří o *censored* či *truncated* proměnných). Dalším případem jsou *modely počítatelných dat (count data models)*, kdy závisle proměnná označuje počet (např. $0, 1, 2, \ldots$). Existuje pak další spousta modelů, které nějakým způsobem vysvětlovanou proměnnou omezují, zmíníme se o nich jen přehledově, abycho získali představu, jaký z těchto modelů je vhodné využít, když máme k dispozici zrovna ten typ dat, který jejich použití vyžaduje. Podrobnosti k nim je možné nalézt v pokročilejších ekonometrických učebnicích.

6.2 Modely diskrétní volby

6.2.1 Modely binární volby

Než se dostaneme k probit a logit modelům, je dobré ukázat ekonomickou motivaci k jejich použití. To nám pomůže i při interpretaci výsledků získaných na základě jejich odhadu.

Probit a logit modely se často (ale ne výlučně) objevují v případech, kdy jednotlivci provádí nějaký druh volby. Předpokládejme, že jednotlivec volí mezi dvěma alternativami. V keonomii bychom mohli tuto situaci formalizovat specifikací užitkové funkce. Nechť U_{ji} označuje užitek, kterého i-tý jednotlivec ($i = 1 \ldots, N$) dosáhne při volbě alternativy j ($j = 0, 1$). Jednotlivec se rozhodne pro volbu 1, pokud $U_{1i} \geq U_{0i}$ a pro volbu 0 v opačném případě. Protože je výraz $U_{1i} \geq U_{0i}$ ekvivalentní vyjádření $U_{1i} - U_{0i} \geq 0$, můžeme danou volbu chápat v závislosti na rozdílu užitků daných alternativ a definujme tento rozdíl jako

$$Y_i^* = U_{1i} - U_{0i}.$$

V příkladu z oblasti ekonomie dopravy tak individuální volba mezi autem a veřejnou dopravou závisí na tom, která z těchto alternativ přinese každému jednotlivci větší užitek. Ekonom zabývající se ekonomií dopravy by řekl, že Y_i^* by mělo záviset na charakteristikách těchto jednotlivců, což může být jejich plat, doba, kterou trvá jízda do práce autem, doba, kterou trvá cesta do práce v prostředku hromadné dopravy apod. Z pohledu ekonometra tato závislost může odpovídat lineárnímu regresnímu modelu:

$$Y_i^* = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i,$$

kde používáme index i k popsání jednotlivých pozorování, $i = 1, \ldots, N$, a $X_1, \ldots, X_k$ jsou vysvětlující proměnné, které mohou ovlivňovat užitek jednotlivců. V předchozích kapitolách jsme pro jednoduchost výkladu pracovali s jednoduchým regresním model, který zjednodušil výsledné vztahy (bez nutnosti použít maticové značení):

$$Y_i^* = \beta X_i + \epsilon_i.$$

Problémem v této regresi je to, že nepozorujeme užitky každého z jednotlivců, tedy i Y_i^* je nepozorovatelné. Ale i tak je dobré si toto vyjádření modelu zapsat, protože probit model je interpretován v duchu tohoto zápisu, kdy se předpokládá, že chyby regrese splňují všechny klasické předpoklady. Nejdůležitější je předpoklad normality. Logit model lze také interpretovat v duchu této regrese, kdy se předpokládá, že náhodné složky splňují všechny klasické předpoklady, s výjimkou jednoho. Tímto předpoklade je to, že náhodné chyby mají tzv. *logistické rozdělení*. Nemusíme se v této fázi zabývat jeho definicí, protože si v dalším textu popíšeme jeho vlastnosti.

Proměnná Y_i^* je nepozorována, nicméně pozorujeme volbu, kterou každý z jednotlivců provedl. A tato volba nám něco málo o Y_i^* řekne. Konkrétně pozorujeme $Y_i = 1$, pokud i-tý jednotlivce zvolil alternativu 1, a $Y_i = 0$, pokud zvolil alternativu 0. Vztah mezi Y_i a Y_i^* je vyjádřen následujícími rovnicemi:

$$\begin{aligned} Y_i &= 1 \quad \text{pokud} \quad Y_i^* \geq 0, \\ Y_i &= 0 \quad \text{pokud} \quad Y_i^* < 0. \end{aligned}$$

Vyjádřeno slovy, jednotlivec se rozhodne pro volbu 1, pokud je užitek spojený s touto alternativou větší než užitek spojený s alternativou 0 (a obráceně).

Máme tedy regresní model pro nepozorovanou veličinu Y_i^* a rovnice propujující pozorované Y_i s Y_i^*. Jaké ekonometrické metody tedy můžeme použít k odhadu regresních koeficientů a jak následně interpretovat výsledky odhadu? V obou případech je užitečné uvažovat v kontextu pravděpodobnosti provedení dané volby. Následující odvození se vztahují k pravděpodobnosti volby 1. Pravděpodobnost volby 0 bude 1 mínus tato pravděpodobnost. Protože jednotlivec zvolí alternativu 1 v případě, kdy jeho užitek z této alternativy je větší než užitek z alternativy 0, je pravděpodobnost volby 1 dán jako

$$\Pr(Y_i = 1) = \Pr(Y_i^* \geq 0) = \Pr(\beta X_i + \epsilon_i \geq 0) = \Pr(\epsilon_i \geq -\beta X_i).$$

Pokud známe rozdělení náhodných složek, je snadné určení této pravděpodobnosti, tedy pravděpodobnosti $\Pr(\epsilon_i \geq -\beta X_i)$.

Uvědomme si, že probit model předpokládá normální rozdělení chyb regrese a logit model předpokládá logistické rozdělení náhodných složek. Normální rozdělení je rozdělení známé a příslušné pravděpodobnosti lze spočítat s využitím statistických tabulek normálního rozdělení nebo s využitím odpovídajících počítačových programů. V příloze B je zavedeno značení distribuční funkce (resp. kumulativní distribuční funkce) která je pro jakoukoli náhodnou veličinu Z a jakoukoli hodnot z dána jako

$$\Pr(Z \leq z).$$

Pokud je Z standardizovaná normální náhodná veličina (tzn. $N(0,1)$), je obvyklé používat pro distribuční funkci značení

$$\Phi(z).$$

S využitím tohoto značení můžeme v případě našeho probit modelu (a předpokladu, že náhodné složky mají standardizované náhodné rozdělení, což je důležité pro identifikovatelnost parametrů) říct, že

$$\Pr(Y_i = 1) = \Pr(\epsilon_i \geq -\beta X_i) = 1 - \Phi(-\beta X_i) = \Phi(\beta X_i),$$

kde poslední výraz vyplývá ze symetrie standardizovaného normálního rozdělení kolem své střední hodnoty, tedy nuly. Protože platí, že $\Pr(Y_i = 0) = 1 - \Pr(Y_i = 1)$, můžeme říct, že $\Pr(Y_i = 0) = \Phi(-\beta X_i)$.

V případě logit modelu, kdy ϵ_i má logistické rozdělení získáváme (důkaz zde nebudeme uvádět)

$$\Pr(Y_i = 1) = \frac{\exp(\beta X_i)}{1 + \exp(\beta X_i)}$$

a tedy

$$\Pr(Y_i = 0) = \frac{1}{1 + \exp(\beta X_i)}.$$

Tyto vztahy lze využít k interpretaci výsledků probit a logit modelů. Připomeňme si, že v rámci regresních modelů interpretujeme koeficienty tak, že nám měří mezní vliv

vysvětlující proměnné na závisle proměnnou. V případě probit a logit modelů není tinterpretace mezních vlivů zcela přímá. Nicméně, pokud máme $\widehat{\beta}$ (což je odhad parametru β), můžeme pracovat v intencích pravděpodobnosti konkrétní volby pro daného jednotlivce. Jako příklad se vraťme k dopravnímu příkladu, kde $Y = 1$ označuje výběr automobilu (jakožto dopravního prostředku do práce) a X je vysvětlující proměnná odpovídající času trvání cesty do práce (měřeno v minutách). Jakmile máme odhad probit modelu, můžeme například odhadnou pravděpodobnost, že jednotlivec s dojezdností 30, 60 a 120 minut bude jezdit do práce automobilem, následovně:

$$\begin{aligned}
\Pr(Y = 1|X = 30) &= 1 - \Phi(-30\widehat{\beta}),\\
\Pr(Y = 1|X = 60) &= 1 - \Phi(-60\widehat{\beta}),\\
\Pr(Y = 1|X = 120) &= 1 - \Phi(-120\widehat{\beta}).
\end{aligned}$$

Tato informace může být užitečná pro tvůrce politických rozhodnutí (např. v oblasti veřejné dopravy). V případě logit modelu, že třeba dát hodnoty vysvětlující proměnné (30, 60 a 120) do vztahů uvedených dříve, které vycházejí z logistického rozdělení.

Můžeme tedy obdržet interpretaci výsledků odhadu, která je podobná interpretaci mezního vlivu vysvětlujících proměnných, kdy v tomto případě používáme vyjádření v kontextu pravděpodobností. V obvyklé interpretaci mezního vlivu v jednoduchém regresním modelu se tedy ptáme: „Jak se změní Y, když změníme X?“, a příslušná odpověď je β. V modelech kvalitativní volby tuto interpretaci modifikujeme do otázky: „Jak se změní *pravděpodobnost volby* 1, pokud změníme X?“ kdy však odpověď není jednoduše β. S využitím diferenciálního počtu lze odvodit následující mezní vlivy. Pro probit model je mezní vliv X na pravděpodobnost volby 1

$$\phi(\beta X)\beta,$$

kde $\phi(\cdot)$ je vztah pro funkci hustoty normálního rozdělení (viz příloha B). Pro logit model je mezní vliv X na pravděpodobnost volby 1

$$\frac{\exp(\beta X_i)}{1 + \exp(\beta X_i)} \frac{1}{1 + \exp(\beta X_i)} \beta.$$

Tyto vztahy pro mezní vlivy vypadjí na první pohled složitě, nicméně důlžeité je, že je lze spočítat v relevantním ekonometrickém programu. Je třeba ale zdůraznit, že v rámci probit a logit modelů (oproti jendoduché regresi) tyto mezní vlivy závisejí na X. Osoba s dojezdností $X = 30$ minut tak má jiný mezní vliv než osoba s dojezdností $X = 60$ minut. S ohledem na tuto skutečnost tak je v počítačových programech obvyklé, že vyhodnocují mezní vlivy pro průměrnou hodnotu vysvětlujících proměnných.

Obvyklá je rovněž prezentace mezních vlivů v rámci tzv. *podílu šancí (odds ratio)*. Podíl šancí je poměr pravděpodobností volby každé z alternativ:

$$\frac{\Pr(Y_i = 1)}{\Pr(Y_i = 0)}.$$

Pro logit model lze podíl šancí zjednodušit následovně:

$$\frac{\Pr(Y_i = 1)}{\Pr(Y_i = 0)} = \frac{\dfrac{\exp(\beta X_i)}{1 + \exp(\beta X_i)}}{\dfrac{1}{1 + \exp(\beta X_i)}} = \exp(\beta X_i).$$

Logaritmust podílu šancí je tedy v tomto případě βX_i. Parametr β tak lze interpretovat (poněkud nešikovně) jako mezní vliv v kontextu logaritmu podílu šancí, tzn. „pokud zvýšíme X o jednotku, logaritmus podílu šancí se změní o β jednotek“.

Doposud jsme hovořili o probit a logit modelech s jedinou vysvětlující proměnnou. Rozšíření pro případ více vysvětlujících proměnných však není nijak složitý. Doposud také nebylo nic řečeno o samotných odhadech a testování probit a logit modelů. Většina ekonometrických programů nám nabízí odhad modelů tohoto typu (včetně rozšířených variant diskutovaných v další části textu), jehož součástí jsou i výstupy analogické těm, jež byly diskutovány v předchozích kapitolách věnovaných regresním modelům. V případě probit nebo logit modelu zahrnujícího více vysvětlujících proměnných jsou standardním výstupem ekonometrických programů odhad koeficientů $\alpha, \beta_1, \ldots, \beta_k$, příslušné p-hodnoty pro testování statistické významnosti parametru a intervaly spolehlivosti. Specializovaným (komerčním) programem pro odhad tohoto typů modelů je LIMDEP. Nicméně i volně dostupný gretl nabízí odhady všech modelů v této části kapitoly. Výstupeme tak jsou odhady metodou maximální věrohodnotsti, $\widehat{\alpha}, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$ a p-hodnoty testu hypotézy $H_0 : \beta_j = 0$ pro každý parametr j. Jsme tak schopni prezentovat bodové odhady všech koeficientů a diskutovat jejich významnost. Důležitým výstupem je možnost prezentace mezních vlivů každé z vysvěltujících proměnných na pravděpodobnost provedení volby 1 pro jednotlivce, jehož vysvětlující proměnné odpovdíají průměrným hodnotám ve vzorku. Gretl počítá mezní vliv j-té vysvětlující proměnné na pravděpodobnost volby 1 s využitím vztahu

$$\phi\left(\widehat{\alpha} + \widehat{\beta}_1 \overline{X}_1 + \ldots + \widehat{\beta}_k \overline{X}_k\right) \widehat{\beta}_j,$$

kde $\overline{X}_j$ je průměrná hodnota j-té vysvětlující proměnné (tzn. $\frac{\sum_{i=1}^{N} X_{ji}}{N}$). Jiným z mnoha výstupů je logaritmus věrohodnostní funkce. který lze využít pro test věrohodnostního poměru. Stejně jako v normálním lineárním regresním modelu může být v probit modelu problém heteroskedasticity. Nicméně i v tomto případě je k dispozicimožnost odhadu analogická heteroskedasticitě konzistentnímu estimátoru z kapitoly 5.

Nejoblíbenějším způsobem odhadu probit nebo logit modelů je použití metod maximální věrohodnosti. Obecné principy tohoto přístupu byly zmíněny v kapitole 3. Připoměňme si, že věrohodnostní funkce je sdružená hustota pravděpodobnosti pro $Y_1, \ldots, Y_N$, vyhodnocená v rámci aktuálního pozorování. Pokud jsou pozorování vzájemně nezávislá (což předpokládáme), lze věrohodnostní funkci zapsat jako

$$L(\beta) = p(Y_1, \ldots, Y_N) = \prod_{i=1}^{N} p\left(Y_i\right).$$

Myšlenka maximálně věrohodného odhadu spočívá v nalezení takové hodnoty β, která bude maximalizovat tuto funkci. Odhad parametru, $\widehat{\beta}$, tak je hodnota vedoucí k funkci hustoty pravděpodobnosti, na základě které byly s největší pravděpodobností generovaná pozorování $Y_1, \ldots, Y_N$.

Tento typ postupů založených na věrohodnostní funkci lze použít v případě probit a logit modelů. Naše předchozí odvození ukazují jak. Z předchozích odvození jsme získali $p(Y_i)$, tudíž i věrohodnostní funkci. V případě probit modelu je $p(Y_i)$ definovaná tak, že $\Pr(Y_i = 1) = \Phi(\beta X_i)$ a $\Pr(Y_i = 0) = \Phi(-\beta X_i)$. Věrohodnostní funkce je

$$L(\beta) = \prod_{i=1}^{N} p(Y_i) = \prod_{i=1}^{N} \Phi(\beta X_i)^{Y_i} \Phi(-\beta X_i)^{1-Y_i}.$$

Pro přiblížení si interpretace tohoto vztahu poznamenejme, že pokud do vztahu dosadíme buď $Y_i = 1$ nebo $Y_i = 0$, budeme mít pro každého jednotlivce ve vztahu pro věrohodnostní funkci buď $\Phi(\beta X_i)$ nebo $\Phi(-\beta X_i)$.

Věrohodnostní funkce pro logit model je založena na podobném postupu odvození:

$$L(\beta) = \prod_{i=1}^{N} \left(\frac{\exp(\beta X_i)}{1 + \exp(\beta X_i)} \right)^{Y_i} \left(\frac{1}{1 + \exp(\beta X_i)} \right)^{1-Y_i}.$$

V případě jednoduchého regresního modeli odpovídal estimátor maximální věrohodnosti OLS estimátoru. Mohli jsme tedy k výpočtu maximálně věrohodného estimátoru využít příslušný vztah (tj. $\widehat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2}$). V rámci probit a logit modelů však žádný podobný vztah neexistuje, nejsme tedy shopni získat *analytické řešení*. V tomto případě je věrohodnostní funkce maximalizována s využitím optimalizačních algoritmů, které jsou součástí příslušných ekonometrických programů. Výsledkem jsou i statistiky analogické t-testu, a tomu odpovídající p-hodnoty, díky čemuž můžeme testovat nevýznamnost daných parametrů (resp. příslušných vysvětlujících proměnných).

Jak jsme již viděli v případě regresního modelu, s metodou maximálnívěrohodnosti je spojena řada postupů testování hypotéz: test věrohodnostního poměru, Waldův test a test Lagrangeových multiplikátorů. Díky nim jsme schopni testovat i v logit nebo probit modelu jakoukoli hypotézu, která nás zajímá. V kapitole 4 byl motivován test věrohodnostního poměru v kontextu modelu vícenásobné regrese. Ten lze využít pro testy sdružených hypotéz jako např. $H_0 : \beta_1 + \beta_2 = 1, \beta_3 = 0$. Pro jeho použití v praxi musíme nelézt hodnotu věrohodnostní funkce neomezeného modelu a modelu omezeného, kdy předpokládáme, že omezení vycházející z testované hypotézy platí (např. $\beta_1 + \beta_2 = 1$ a $\beta_3 = 0$). Testová statistika věrohodnostního poměru je jednoduše funkcí těchto dvou hodnot. Pokud odhadneme omezenou a neomezenou verzi logit či probit modelů, jsou počítačovým výstupem i hodnoty věrohodnostní funkce (popř. hodnoty logaritmů této funkce). Test věrohodnostního poměru tak lze v případě logit a probit modelů snadno provést.

Otázka, která nás může napadnout, může být: „Jak poznáme, jestli máme použít logit nebo probit?“. Upřímně řečeno, ve většině empirických aplikací je to jedno. Logit i probit modely mají tendenci vracet podobné výsledky. Pro řešení tohoto dilematu však existují různé druhy testů, které lze najít v pokročilých ekonometrických publikacích

(např. Greene [10]), z nichž není problém pochopit jejich použití v prax. Otázka použití rozšířených variant probit nebo logit modelů je však mnohem důležitější v případě, kdy existuje více možností volby, čemuž je věnována následující část kapitoly. Příklad 6.1.

Příklad 6.1. *Proč mít či nemít mimomanželský poměr?*
Rozhodování o tom, jestli člověk bude mít nebo nebude mít mimomanželský poměr rozebíral ve svém článku Ray C. Fair (článek má název „A theory of extramartial affair“ a byl publikován v roce 1978 v časopise *Journal of Political Economy*). Tento článek využil výběrových šetření provedených populárními časopisy a poskytl vcelku zajímavý pohled na tuto problematiku. V rámci tohoto příkladu je využita jen část původních dat, konkrétně se jedná o $N = 601$ pozorování následujících proměnných (jsou obsahem souboru `affair.gdt`):

- $AFFAIR = 1$ pokud měl jednotlivec tento druh poměru ($= 0$ jinak);
- $MALE = 1$ pokud je jednotlivec mužem ($= 0$ jinak);
- $YEARS$ je počet let manželství daného jednotlivce;
- $KIDS = 1$ pokud má jednotlivec děti z manželství ($= 0$ jinak);
- $RELIG = 1$ pokud se jednotlivec pokládá za nábožensky založeného;
- $EDUC$ je počet ukončených let vzdělání;
- $HAPPY = 1$ pokud se jednotlivec cítí v manželství šťastný ($= 0$ jinak).

Tabulka 6.1: Logit – mimomanželské poměry.

	Logit			*Podíl šancí*	*Logit (robust)*	
		p-hodn.	95% int.			p-hodn.
Proměnná	Koef.	$\beta_j = 0$	spol.	Koef.	Koef.	$\beta_j = 0$
Konstanta	-1.29	0.07	[-2.71;0.13]	—	-1.29	0.09
$MALE$	0.25	0.26	[-0.18;0.67]	1.28	0.25	0.27
$YEARS$	0.05	0.03	[0.01;0.09]	1.05	0.05	0.03
$KIDS$	0.44	0.12	[-0.12;1.00]	1.55	0.44	0.13
$RELIG$	-0.89	0.00	[-1.32;-0.47]	0.41	-0.89	0.00
$EDUC$	0.01	0.75	[-0.07;0.10]	1.01	0.01	0.75
$HAPPY$	-0.87	0.00	[-1.28;-0.46]	0.42	-0.87	0.09

$AFFAIR$ je závisle proměnná. Tabulka 6.1 ukazuje výsledky odhadu koeficientů logit modelu (ML odhad) spolu s odhady mezních vlivů, vztažených v tomto případě k podílu šancí (tzn. $\exp(\beta_j)$), což je vliv jednotkové změny j-té vysvětlující proměnné na podíl šancí, pokud se ostatní proměnné nemění. Z pohledu na p-hodnoty vyplývá, že $YEARS$, $RELIG$ a $HAPPY$ jsou statisticky významné. Odhadnuté koeficienty neměří přímo mezní vlivy jednotlivých proměnných, můžeme však interpretovat znaménka těchto koeficientů. Koeficienty u $RELIG$ a $HAPPY$ jsou záporné, což znamená, že nábožensky založení nebo v manželství šťastní jednotlivci budou mít relativně menší pravděpodobnost mimomanželské aféry. Kladný koeficient $YEARS$ znamená, že pro jednotlivce, kteří jsou v manželství delší dobu, bude tato pravděpodobnost větší. (pokračování v příkladu 6.2)

Příklad 6.2. *Proč mít či nemít mimomanželský poměr? (pokračování příkladu 6.1)* Interpretace koeficientů není sama o sobě úplně zřejmá. Tabulka 6.1 tak obsahuje rovněž výsledky (ve sloupci „podíl šancí“) pro mezní vliv na podíl šancí. Tyto hodnoty je možno interpretovat mnohem snadněji. Pracujme například s odhadem mezního vlivu proměnné $HAPPY$. Hodnota je 0.42. Jak toto číslo interpretovat? $HAPPY$ je umělá proměnná, tudíž „jednotková změna“ znamená změnu z nešťastného manželství na šťastné. Můžeme tak říct: „Jestliže se manželství jednotlivce změní z nešťastného na šťastné (a ostatní proměnné se nezmění), potom bude podíl šancí ve prospěch mimomanželské aféry odpovídat 42% toho, jaký byl před touto změnou“. Konkrétně předpokládejme jednotlivce, který má podíl šancí 4. To znamená, že $\Pr(Y_i = 1) = \frac{4}{5}$ a $\Pr(Y_i = 0) = \frac{1}{5}$ a je zde tedy 80% šance že daný jednotlivec bude mít mimomanželskou aféru. Pokud se jehomanželství stane šťastným, bude tento podíl šancí tvořit 42% původní hodnoty. Bude tedy $4 \times 0.42 = 1.68$. To znamená, že zde nyní bude 63% šance, že jednotlivec bude mít mimomanželskýpoměr. Dochází tedy k poklesu pravděpodobnosti aféry o 17% (to odpovídá samozřejmě hodnotě mezního vlivu s příslušného sloupce tabulky 6.2 a i pro logit model lze tuto hodnotu získat přímo). Poslední dva sloupce tabulky 6.1 (označené „Logit robust“) ilustrují další možnosti výstupu ekonometrických programů. V kapitole 5 jsme diskutovali problém heteroskedasticity a heteroskedasticitě konzistentní estimátor. Ten se využívá pro výpočet korektních rozptylů OLS estimátoru (které dále vstupují do postupů testování hypotéz a konstrukce intervalů spolehlivosti), pokud máme problém s heteroskedasticitou. V případě logit a probit modelu lze rovněž využít *robustní estimátory rozptylů*. Odhady koeficientů jsou samozřejmě stejné jako v případě, kdy jsme robustní odhad nepoužili. Liší se však výsledné p-hodnoty pro $H_0 : \beta_j = 0$. Tyto p-hodnoty mají tendenci být o něco větší. Nicméně, v tomto případě se nezdá, že by byla heteroskedasticita problémem, protože výsledky s robustními i „nerobustními“ odhady rozptylů jsou fakticky stjné (to platí i pro probit model). Samozřejmě iv tomto případě lze použít testy heteroskedasticity, které však mají svou speciální podobu a nemusí být dostupné ve všech ekonometrických balíčcích. (dokončení v příkladu 6.3)

Příklad 6.3. *Proč mít či nemít mimomanželský poměr? (dokončení příkladu 6.2)*

Tabulka 6.2: Probit – mimomanželské poměry.

	Probit			*Mezní ef.*	*Probit (robust)*	
Proměnná	Koef.	p-hodn. $\beta_j = 0$	95% int. spol.	Koef.	Koef.	p-hodn. $\beta_j = 0$
Konstanta	-0.74	0.08	[-1.56;0.09]	—	-0.74	0.11
$MALE$	0.15	0.23	[-0.10;0.40]	0.05	0.15	0.24
$YEARS$	0.03	0.03	[0.00;0.05]	0.01	0.03	0.02
$KIDS$	0.25	0.12	[-0.07;0.57]	0.07	0.25	0.13
$RELIG$	-0.51	0.00	[-0.75;-0.27]	-0.15	-0.51	0.00
$EDUC$	0.01	0.81	[-0.04;0.06]	0.00	0.01	0.81
$HAPPY$	-0.51	0.00	[-0.76;-0.27]	-0.17	-0.51	0.09

Tabulka 6.2 ukazuje podobné výsledky jako tabulka 6.1, v tomto případě se jedná o probit model. Věcná interpretace výsledků se nezměnila: např. šťastné manželství nebo náboženské založení významně snižuje šance jednotlivce, že se zaplete do nějaké mimomanželské aféry. Podobně jako u logit modelu je problém přímé interpretace odhadů koeficientů. Většina programů počítá

$$\phi\left(\widehat{\alpha} + \widehat{\beta}_1 \overline{X}_1 + \ldots + \widehat{\beta}_k \overline{X}_k\right) \beta_j.$$

Jak bylo již dříve zmíněno, jedná se o hodnoty mezního vlivu vysvětlující proměnné j na pravděpodobnost volby 1 (při neměnnosti ostatních vysvětlujících proměnných). Podíváme-li se tedy do sloupce „Mezní ef.“ vidíme, že např. mezní vliv $KIDS$ je 0.07. Tuto hodnotu můžeme interpretovat tak, že „pokud manželství přinese narození dítěte, zvýší se pravděpodobnost mimomanželské aféry o 0.07 (pokud zůstávají ostatní vysvětlující proměnné neměnné a na svých průměrných hodnotách)“. Nicméně, koeficient u proměnné $KIDS$ je statisticky nevýznamný. S využitím probit a logit modelů můžeme ukázat řadu dalších zajímavých empirických závěrů. Můžeme např. spočítat predikovanou pravděpodobnost mimomanželské aféry hypotetického jednotlivce. Můžeme např. zodpovědět otázku: „Jaká je pravděpodobnost, že bude mít mimomanželský poměr nábožensky založený muž s 16 roky vzdělání, který je dva roky ženatý a má již z tohoto manželství děti?“. Alternativně lze spočítat predikovanou pravděpodobnost mimomanželské aféry pro každého z $i = 1, \ldots, N$ jednotlivců. Tyto pravděpodobnosti lze využít k hodnocení toho, jak kvalitně model „vyrovnává“ data. Pokud jsou predikované pravděpodobnosti vysoké pro jednotlivce, kteří skutečně mimomanželský poměr zažili (a jsou nízké pro věrné partnery), vyrovnává model data dobře. Existuje řada formálních měřítek kvality tohoto vyrovnání (podobných koeficientu determinace R^2) založených právě na predikovaných pravděpodobnostech. Pro bližší informace je vhodné sáhnout po některé z pokročilých učebnic ekonometrie.

6.2.2 Modely multinomiální volby

Multinomiální probit a multinomiální logit

Předchozí diskuzi nad logit a probit modelem lze rozšířit do podoby volby mezi několika alternativami. Z tohoto důvodu budeme předpokládat, že Y_i může nabývat hodnot $0, 1, \ldots, J$. Existuje tedy $j+1$ alternatit a k jejich označení budeme používat dolní indexy $\{j = 0, \ldots, J\}$. Jednotlivec si vybere tu volbu, která mu přináší největší užitek. Stejně jako v předchozí části si pomocí U_{ji} označíme užitek i-tého jednotlivce z volby alternativy j (pro $i = 1, \ldots, N$ a $j = 0, \ldots, J$). V modelu binomiální volby jsme hovožili o pravděpodobnosti volby 1 (kdy pravděpodobnost volby 0 byla jedna mínus pravděpodobnost volby 1). O volbě 1 jsme hovořili tehdy, pokud $U_1 \geq U_0$ a tatoúvaha byla ekvivalentní volbě v případě, kdy $U_1 - U_0 \geq 0$ (důležitý byl tedy rozdíl v užitcích z jednotlivých alternativ). V případě modelů multinomiální volby provedeme něco podobného s tím, že si jednu z alternativ zvolíme jako srovnávací měřítko (benchmark) či základní alternativu. Každá z ostatních alternativ je pak porovnávána s touto alternativou z pohledu užitků, které jednotlivcům jednotlivé alternativy přinesou. V praxi není důležité co zvolíme jako základní alternativu. Například pro problém z oblasti dopravy může každý jednotlivec volit mezi dopravou do práce prostřednitvím věřejné dopravy, automobilu nebo kola. Nehraje roli, jestli jako srovávací alternativu zvolíme hromadnou dopravu, auto nebo kolo. Pro naše potřeby budeme v případě modelu multinomiální volby jako benchmark chápat alternativu 0.

V modelech binomiální volby jsme vycházeli z rozdílu užitků mezi alternativami 1 a 0. V případě více alternativ rovněž tento postup zvolíme, přičemž kdy budeme uvažovat rozdíly v užitcích mezi každou z alternativ a základní alternativou. Budeme tedy mít

$$Y_{ji}^* = U_{ji} - U_{0i}$$

pro $j = 1, \ldots, J$. Proměnná Y_{ji}^* je nepozorovatelná, nicméně pozorujeme volbu každého z jednotlivců. Konkrétně tak pozorujeme $Y_i = j$ pokud jednotlivec i provede volbu j. Formálně máme následující vztah mezi nepozorovanými diferencemi v užitcích a skutečně provedenou volbou: pokud $Y_{ji}^* < 0$ pro $j = 1, \ldots, J$, potom jednotlivec i zvolí základní alternativu a $Y_i = 0$. V jiném případě i-tý jednotlivec zvolí alternativu, jterá mu přinese nejvyšší hodnotu užitku pro Y_{ji}^*.

Stejně jako v případě modelů probit a logit, mohou tyto rozdíly v užitcích záviset na vysvětlujících proměnných (např. doba strávená cestou do práce pro každou alternativu může pomoci vysvětlit proč preferujeme veřejnou dopravu před automobilem). POužijeme tedy regresní model:

$$Y_{ji}^* = \alpha_j + \beta_{j1} X_{1i} + \beta_{j2} X_{2i} + \ldots + \beta_{jk} X_{ki} + \epsilon_{ji}.$$

Příslušným dolním indexům je potřeba věnovat velkou pozornost. Oproti modelu binomiální volby nemáme jedinou regresi, ale J různých regresí (každou pro porovnání všech alternativ $j = 1, \ldots, J$ s alternativou 0). V každé z regresí máme různé koeficienty (tzn. α_j je úrovňová konstanta v regresi zahrnující diference v užitcích mezi alternativou j a alternativou 0, β_{j1} je koeficient první vysvětlující proměnné v této regresi, atd.).

Všimněme si ale, že jsme nedali dolní indexy j vysvětlujícím proměnným. Ačkoli tedy každá z regresí může mít různé vysvětlující proměnné (tzn., že některé koeficienty u vysvětlujících proměnných mohou být nulové, z čehož vyplývá, že příslušné vysvětlující proměnné z regrese vypouštíme), předpokládáme, že hodnota každé z vysvětlujících proměnných je stejná v každé z rovnic. To dává smysl v případě, že vysvětlující proměnné jsou charakteristiky daného jednotlivce. Například u našeho dopravního problému může výběr dopravního prostředku závisetna důchodu daného jednotlivce. Chtěli bychom tedy zahrnout jako vysvětlující proměnnou X_{1i} důchod i-tého jednotlivce. Hodnota této vysvětlující proměnné bude stejná pro každou regresi. Na druhé straně je však rozumné předpokládat, že volba dopravního prostředku může záviset na čase, který je potřeba na dopravení se alternativními prostředky do práce. Tento čas bude různý mezi jednotlivými alternativami (tzn. časová délka dopravy se bude ličit mezi autem, veřejnou dopravou a kolem). Naše specifikace modelu nám však napovídá, že to asi nebude možné. Nicméně, snadným trikem se budeme schopni vypořádat s vysvětlujícími proměnnými, které se budou lišit mezi alternativami. Místo toho, že budeme mít jedinou vysvětlující proměnnou zvanou např. „časová délka dopravy“, můžeme pracovat s $J + 1$ různými vysvětlujícími proměnnými, tzn. jedna vysvětlující proměnná je časová délka dopravyautem, druhá je čas dopravy prostředkem hromadné dopravy a poslední bude čas potřebný k dopravě do práce pomocí kola. Multinomiální probit a logit modely tak dovolují pracovat s vysvětlujícími proměnnými, jejichž hodnoty se liší mezi jednotlivci a/nebo mezi alternativami. Jak uvidíme později (v diskuzi nad podmíněným logit modelem), rozdíl mezi typy vysvětlujících proměnných může být v některých případech důležitý.

Multinomiální probit a multinomiální logit modely jsou založeny na předchozí množině J regresních rovnic, nicméně se liší v předpokladech o rozdělení náhodných složek. Multinomiální probit model předpokládá náhodné složky normálně rozdělené, oproti tomu multinomiální logit model přepdokládá jiný typ rozdělení náhodných složek. V našem případě se nemusíme zabývat přesnou podobou posledního ze jmenovaných rozdělení. Důležité je to, že v případě multinomiálního logit modelu je pravděpodobnost, že i-tý jednotlivec provede volbu j, dána jako

$$\Pr(Y_i = j) = \frac{\exp(\beta_j X_i)}{1 + \sum_{s=1}^{J} \exp(\beta_s X_i)},$$

kde pro jednoduchost výrazu uvažujeme jednoduchou regresní závislost s jedinou vysvětlující proměnnou.

V případě multinomiálního probit modelu si můžeme vyjádřit $\Pr(Y_i = j)$ s využitím vlastností normálního rozdělení. V tomto případě však musíme využít *vícerozměrné normální rozdělení*. Jedná se o rozšíření normálního rozdělení, nicméně jeho přesnou pdobu si zde není potřeba uvádět. Místo toho se zaměříme na trochu vysvětlení a intuice. Uvědomme si, že v případě multinomiálního probit modelu máme různé náhodné složky v každé z regresí zahrnujících všechny rozdíly v užitcích (tj. ϵ_{ij} pro $j = 1, \ldots, J$). Tyto náhodné složky mohou být vzájemně korelovány. Například pro případ tří alternativ (tzn. dvou regresí zahrnujících rozdíly v užitcích, Y_{1i}^* a Y_{2i}^*) můžeme v klidu počítat s $corr(\epsilon_{1i}, \epsilon_{2i}) \neq 0$. Vícerozměrné normální rozdělení tuto korelaci umožňuje. Jedná se o rozšíření normálního rozdělení, které nám říká, že každá

z náhodných složek (samostatně) je normálně rozdělená, ale v případě nutnosti korelovaná s ostatními (pro daného i-tého jednotlivce!). Počet možných korelací se nám zvyšuje s tím, jak narůstajímožnosti alternativ. Pokud tak máme například možnost volby mezi čtyřmi alternativami, mohli bychom mít tři různé korelace $corr(\epsilon_{1i}, \epsilon_{2i})$, $corr(\epsilon_{1i}, \epsilon_{3i})$a $corr(\epsilon_{2i}, \epsilon_{3i})$. Obecně s $J + 1$ alternativami máme $\frac{J(J+1)}{2}$ korelací. My ovšem nevíme jaké tyto korelace jsou a musíme je tudíž odhadnout. Multinomiální model sám o sobě má dost koeficientů k odhadu (tzn. musíme odhadnout všechny regresní koeficienty v každé z J rovnic). Multinomiálníprobit model však musí odhadnou i všechny korelace mezi náhodnými složkami. Pokud nemáme k dispozici velké množství dat pro relativně málo alternativ, je obtížné dostat přesné odhady parametrůmultinomiálního probit modelu. Z tohoto důvodu je multinomiální probitmodel typicky využíván jen v případech, kdy je počet alternativ relativně malý. Další komplikací s multinomiálním probitem je to, že výpočet pravděpodobností v rámci vícerozměrného normálního rozdělení může zabrat dost výpočetního času. Pro standardní normální rozdělení existují statistické tabulky (či jejich počítačový ekvivalent) a výpočet je tak snadný. V případě vícerozměrného rozdělení je potřeba nasadit výpočetně náročné simulační metody.

V případě hodně alternativ je tak populární multinomiální logit model. Tentomodel předpokládá, že náhodné chyby v různých rovnicích jsouvzájemně nekorelované. V tomto případě si tak člověk nemusí dělat starosti s odhadem všech korelací mezi náhodnými složkami, které jsou přítomné v multinomiálním probit modelu. (samozřejmě je nutný odhad regresních koeficientů v každé z J rovnic). Tento aspekt multinomiálního logit modelu má jednu důležitou implikaci, která v některých empirických aplikacích nemusí být žádoucí. Týká se to toho, že pravděpodobnosti volby implikované multinomiálním logitem musí splňovat vlastnost *nezávislosti irelevantních alternativ (independence of irrelevant alternatives – IIA)*.

Co je tím myšleno je nejlépe ilustrovatelné na příkladu našeho dopravního problému. Předpokládejme, že ve výchozí situaci má typický dojíždějící do práce volbu mezi automobilem ($Y = 0$) nebo veřejnou dopravou ($Y = 1$). IIA vlastnost se týká podílu šancí mezi těmito volbami, tedy podílu pravděpodobnosti těchto voleb (což je výraz $\frac{\Pr(Y=0)}{\Pr(Y=1)}$). Předpoklad IIA říká, že tento podíl šancí bude neměnný bez ohledu na jiné alternativy. Předpokládejme, že ve výchozí situaci je $\frac{\Pr(Y=0)}{\Pr(Y=1)} = 1$, tedy typický dojíždějící si se stejnou pravděpodobností vybere buď auto nebo veřejnou dopravu. Předpokládejme, že se postaví nová cesta pro cyklisty a dojíždějící má možnost cestovat dopráce na kole (nová alternativa je nyní $Y = 2$). Vlastnost IIA říká, že přidání této nové alternativy neovlivní skutečnost, že $\frac{\Pr(Y=0)}{\Pr(Y=1)} = 1$. V tomto případě může být IIA vlastnost rozumná. Původně jsme předpokládali, že $\Pr(Y = 0) = \Pr(Y = 1) = \frac{1}{2}$. Předpokládejme, že jakmile se postaví cyklistická stezka, existuje 20% šance, že náš dojíždějící bude jezdit do práce na kole. To je konzistentní s $\Pr(Y = 0) = \Pr(Y = 1) = 0.40$, což stále znamená, že $\frac{\Pr(Y=0)}{\Pr(Y=1)} = 1$.

Pro ilustraci příkladu, kdy vlasnost IIA není rozumná si můžeme představit to, co ekonometrové nazývají tzv. problémem „červených a modrých autobusů (red bus-blue bus problem)“. Budeme předpokládát, že dojíždějící má ve výchozí situaci mezi automobilem ($Y = 0$) a červeným autobusem ($Y = 1$). Nyní předpokládejme, že auto-

busová společnost natře půlku svých autobusů na modro a půlku na červeno a my to pojmeme jako novou alternativu ($Y = 2$). V takovémto případě asinebude rozumné předpokládat splnění podmínky IIA, což samozřejmě povede k nemožnosti práce s multinomiálním logit modelem. Předpokládejme, že původně platilo $\Pr(Y = 0) = \Pr(Y = 1) = \frac{1}{2}$ a tedy $\frac{\Pr(Y=0)}{\Pr(Y=1)} = 1$. Protože je modrý autobus virtuálně identický sčerveným autobusem, zavedení této alternativy by pravděpodobně vedlo k tomu, že bude volit auto se stejnou pravděpodobností, tedy $\Pr(Y = 0) = 0.50$ a $\Pr(Y = 1) = \Pr(Y = 2) = 0.25$. Zavedení nové alternativy implikuje $\Pr(Y = 0) = \Pr(Y = 1) = 2$. Takováto změna však porušuje vlastnost IIA a v případě použití multinomiálního modelu není dovolena. Samozřejmě je nutné zdůraznit, že otázka rozumnosti nebo nerozumnosti splnění IIA předpokladu závisí na datech, se kterými pracujeme. Někdy může být její splnění rozumné, někdy ne.

Pro opuštění omezující IIA vlastnosti pro použití multinomiálního logit modelu byla vyvinuta řada variant logit modelů. Jednou z těch populárních je *vnořený logit model (nested logit model)*, který předpokládá vnořenou strukturu rozhodovacího procesu. Například v případě našeho problému červených a modrých autobusů by ekonometr použil nejdříve logit model pro dojíždějícího, který se rozhoduje mezi automobilem a veřejnou dopravou. Pokud je zvolena druhá z alternativ, je použit druhý logit model pro volbu dojíždějícího mezi červený autobus a modrý autobus. Jeden logit model je tak vnořen do druhého logit modelu. Podrobněji tento typ modelůdiskutovat v tomto textu nebudeme, nicméně některé ekonometrické balíčky (LIMDEP a Stata) jejich odhad umožňují.

Nebudeme si popisovat do detailu odhad a testování modelů multinomiální volby. Bude nám stačit vědět, že typickými metodami odhadu jsou metody maximální věrohodnosti a jedná se zde o rozšíření těch metod popsaných v rámci diskuze nad probit a logit modely. Relevantní ekonometrické programy nám dávají bodové odhady koeficientů, p-hodnoty pro rozhodování o statistické významnosti parametrů a různé druhy predikčních pravděpodobnsotí a mezních vlivů. Testy věrohodnostního poměru lze využít pro testování dalších hypotéz. Ilustrace empirických odhadů multinomiálního logit model je obsahem příkladu 6.4. Odhadu multinomiálního modelu zde pozornost věnována není, neboť v komerčních programech jako LIMDEP a Stata je odhad dostupnýjenpro omezený počet alternativ. V případě gretlu je multinomiální model možno odhadnout v propojení s prostředím a balíčky R projektu, což ale vyžaduje lehce vyšší schopnost práce s tímto nástrojem.

Řada programů umožňuje pro případ více alternativ odhady tzv. *uspořádaného probit modelu (ordered probit model)*. Nebudeme se mu nijak blíže věnovat, nicméně jej lze snadno použít v případě, kdy jendotlivé alternativy lze seřadit ordinálním způsobem. Například v rámci dotazníků je spotřebitel dotazován na to, jestli je produkt vynikající, dobrý, uspokojivý, slabý nebo velmi slabý. Musí si tak vybrat mezi těmito pěti alternativami, které jsou vcelku přirozeně seřazeny od nejlepší po nejhorší (vynikající je lepší než dobrá, dobrá je lepší než uspokojivá atd.).

Než se dostaneme k ilustrativnímu příkladu, je dobré shrnout si hlavní charakteristiky multinomiálních logit a probit modelů. Oba dva zahrnují odhad několika rovnic a koeficienty v každé z rovnic se týkají rozdílu diferencí mezi konkrétní alternativou a základní alternativou (benchmarkem). Implikací je tak skutečnsot, že oba modely mají

Příklad 6.4. *Poptávka po crackerech*
Pro ilustraci multinomiálního logit modelu použijeme data z oblasti marketingu. Jedná se o data požitá v článku autorů Paap a Franses s názvem „A dynamic multinomial probit model forbrand choice with different long-run a short-run of marketing mix variables“, publikovaným v *Journal of Applied Econometrics* v roce 2000. Data jsou obsažena v souboru `cracker.gdt`, nicméně je lze stáhnout ze stránek Journal of Applied Econometrics Data Archive. Data jsou dostupná pro $N = 136$ domácností v Rome (Georgia) a týkají se nákupu čtyř druhů crackerů: Sunshine, Keebler, Nabisco a Private label. Data pocházejí z pokladních čteček čárových kódů místních supermarketů. Data odpovídají čtyřem alternativám. Jako základní alternativa je zvolena „Private label“ (nicméně benchmarku nám z věcného hlediska výsledky neovlivní). Pro každou z alternativ použijeme jako vysvětlující proměnné úrovňovou konstantu a cenu všech čtyř značek crackerů v daném obchodě a v daný čas nákupu. Náš multinomiální logit model tak zahrnuje tři různé regrese: první je založena na rozdílu užitků mezi crackery Sunshine a Private label, druhá je založena na diferencích užitků mezi crackery Keebler a Private label a poslední regrese vychází z rozdílu užitků mezi crackery Nabisco a Private label. Každá z regresí zahrnuje úrovňovou konstantu a čtyři vysvětlující proměnné odpovídající cenám těchto čtyř značek crackerů. Odhadem multinomiálního logit modelu metodou maximální věrohodnosti získáme výsledky obsažené v tabulce 6.3. Stejně jako v případě logit a probit modelů bychom mohli použít heteroskedasticitě konzistentní estimátory označované jako robustní estimátory rozptylu.
Jak tedy interpretovat koeficienty z tabulky 6.3? Nejprve si uvědomme, že zde máe tři oddělené výsledky regresí (označených v tabulce názvy crackerů, tedy Sunshine, Keebler a Nabisco). Jako v případě regresních koeficientů v modelu kvalitativní volby je obtížné interpretovat samotnou velikost koeficientů. Nicméně alespoň znaménka koeficientů nám mohou poskytnout užitečné informace. Tyto koeficienty lze interpretovat jako mezní vliv vztahující se k rozdílu užitků, Y_{ji}^*, a rozdíl v užitcích je vždy vztažen vzhledem k základní alternativě (což jsou v tomto případě crackery Private label). Pokud se tento rozdíl užitků zvyšuje, zvyšuje se pravděpodobnost, že si jednotlivec koupí alternativu j oproti benchmarkovému výběru. Tyto úvahy lze využít k interpretaci znamének všech koeficientů. Například odhad β_{11} je záporný. Tento koeficient je z regresního modelu pro crackery Sunshine (tj. z regrese, ve které je závisle proměnná rozdíl užitku mezi crackery Sunshine a Private label) a vztahuje se k vysvětlující proměnné „cena crackerů Sunshine“. Záporná hodnota koeficientu nám tak naznačuje následující interpretaci mezního vlivu: „Pokud se cena crackerů Sunshinezváší (a ostatní ceny se nemění), potom pravděpodobnost výběru Sunshine crackerů vzhledem ke crackerům Private label poklesne.“ To je celkem rozumný výsledek, protože s růstem ceny zboží se snižuje pravděpodobnost, že si tento výrobek spotřebitel koupí. (pokračování v příkladu 6.5)

dost koeficientů, které je nutné odhadnout, zejména v případě více alternativ. S malou datovou množinou tak může být odhad tolika koeficientů velmi nepřesný. V případě

Příklad 6.5. *Poptávka po crackerech (pokračování příkladu 6.4)*

Tabulka 6.3: Multinomiální logit pro data o crackerech

	Stř. hodnota	p-hodnota pro $\beta_j = 0$	95% int. spol.
Sunshine			
α_1	-10.06	0.15	[-23.59;3.46]
β_{11}	-7.98	0.01	[0.77;24.02]
β_{12}	12.39	0.04	[0.77;24.02]
β_{13}	0.37	0.91	[-5.83;6.57]
β_{14}	4.83	0.36	[-5.54;15.20]
Keebler			
α_2	-2.53	0.73	[-16.90;11.85]
β_{21}	-3.10	0.30	[-9.01;2.81]
β_{22}	-0.60	0.92	[-12.99;2.81]
β_{23}	1.15	0.70	[-4.67;6.97]
β_{24}	5.33	0.25	[-3.66;14.32]
Nabisco			
α_3	-7.01	0.09	[-15.09;1.07]
β_{31}	-1.38	0.48	[-5.23;2.48]
β_{32}	5.57	0.12	[-1.37;12.50]
β_{33}	0.86	0.65	[-2.84;4.56]
β_{34}	4.72	0.06	[-0.23;9.67]

Ostatní koeficienty z tabulky 6.3 můžeme interpretovat obdobným způsobem. Relevantní počítačové programy vracejí i p-hodnoty testu nulovosti každého z jednotlivých koeficientů. V našem případě jsou tyto výsledky ve sloupci „p-hodnota pro $\beta_{ji} = 0$.“ Až na několik výjimek se zdají být vysvětlující proměnné statisticky nevýznamné. To může být způsobeno malou velikostí vzorku. Stejně tak zde může existovat i řada opomenutých vysvětlujících proměnných (např. individuální charakteristiky domácností, jako je její příjem), které mohou být důležitým faktorem ve vysvětlení volby mezi jednotlivými crackery. (dokončení v příkladu 6.6)

multinomiálního probit modelu tento problém zhoršuje to ,že je nutné odhadnout i korelace mezi náhodnými složkami. Z těchto důvodů, pokud nenípočet alternativ malý, zde existuje tendence jednoznačně upřednostnit multinomiální logit model. Hlavním problémem multinomiálního logit modelu je to, že předpokládá splnění vlasnosti IIA. Jedná se o předpoklad, který lze testovat jednou z variant Hausmanova testu. Je nad rámec našich potřeb blíže si tento test rozebírat, nicméně bývá součástí mnohých ekonometrických programů a není ho tak těžké v případě nutnosti použít.

Příklad 6.6. *Poptávka po crackerech (dokončení příkladu 6.5)*
Stejně jako v případě logit a probit modelů, existují i jiné způsoby prezentace výsledků odhadů, které nesou ilustrativnější informaci než je prostý pohled naregresní koeficienty v tabulce 6.3. Například v naší diskuzi nad logit modelem byl zaveden koncept podílu šancí. I v případě multinomiálního modelu můžeme definovat podíly šancí pro každou z alternativ vzhledem k základní alternativě:

$$\frac{\Pr(Y_i = j)}{\Pr(Y_i = 0)},$$

pro $j = 1, 2, 3$. Vliv jednotkové změny vysvětlující proměnné na každý z podílu šancí je v některých počítačových programech rovněž počítán. Alternativně lze spočítat predikované pravděpodobnosti výběru každé z alernativ pro jednotlivce s danou množinou charakteristik.

Tabulka 6.4: Predikované pravděpodobnosti – data o crackerech

Pravděpodobnost nákupu	Stř. hodnota	Sm. odch.	Min.	Max.
Sunshine	0.08	0.11	0.01	0.64
Keebler	0.07	0.03	0.02	0.16
Nabisco	0.60	0.10	0.31	0.80
Private label	0.25	0.11	0.02	0.49

V tabulce 6.4 je ilustrován jiný typ informace, který lze snadno vytvořit v jakémkoli relevantním počítačovém programu. Vzorec pro výpočet $\Pr(Y_i = 1)$ byl prezentován v textu této části kapitoly. Tento vzorec byl dán pro případ jednoduché regrese, nicméně rozšíření pro případě vícenásobné regrese je zřejmé. Můžeme tak spočítat $\Pr(Y_i = j)$ pro $i = 1, \ldots, N$ a $j = 0, \ldots, J$. Jinými slovy, můžeme si spočítat odhad pravděpodobností výběru každé z alternativ pro všechny jednotlivce. V našem příkladu máme $N = 136$ jednotlivců a čtyři alternativy, na tomto základě vypočteme 544 pravděpodobností. Prezentace každé z nich by byla matoucí. Můžeme se však v některých případech zaměřit jen na vybraného jednotlivce (např. pokud pracujeme Nabisco a z nějakého důvodu nás zajímá jednotlivec č. 60, můžeme ukázat odhady pravděpodobností výběru crackerů Nabisco právě pro tohoto jednotlivce resp. domácnosti). Nebo můžeme vzít všechny pravděpodobnosti každé z alternativ pro $N = 136$ jednotlivců a spočítat jejich popisné statistiky. To je obsahem tabulky 6.4. Všimněme si například řádku pro crackery Nabisco. Průměrná pravděpodobnost volby crackerů Nabisco je 0.60, minimum je 0.31 a maximum je 0.80. Hodnota minima říká, že každý jednotlivec ve vzorku má minimálně 30% šanci výběru crackerů Nabisco. Hodnota maxima nám říká, že nejméně jeden člověk si s 80% pravděpodobností zvolí tyt crackery. Skutečnost, že tato hodnota není vyšší (např. nemáme osobu resp. domácnost s 99% šancí volby této značky crackerů), zohledňuje skutečnost, že většinavysvětlujících proměnných není významných (tzn., že náš model nemá mnoho vysvětlujících proměnných se silnou vysvětlující silou).

Podmíněný logit (conditional logit)

Podmíněný logit model je dalším populárním modelem pro případ více alternativ. Budeme uvažovat stejné značení jako v případě multinomiálního logit a probit modelu. V rámci těchto modelů bylo zmíněno, že mamé vždy jeden regresní model porovnávající každou alternativu se základní alternativou. Pro $j = 1, \ldots, J$ jsme tak měli

$$Y_{ji}^* = \alpha_j + \beta_{j1} X_{1i} + \beta_{j2} X_{2i} + \ldots + \beta_{jk} X_{ki} + \epsilon_{ji}.$$

S podmíněnýmlogit modelem je těchto J regresních modelů nahrazeno jediným regresním modelem. Podmíněný logit model je tak více kompaktní a zahrnuje odhad menšího počtu koeficientů. NIcméně, podmíněný logit model je vhodný jen pro určitý typ dat.

Pro přesnou definici podmíněného logit modelu si připomeňme, že pro multinomiální logit model s jedinou vysvětlující proměnnou

$$\Pr(Y_i = j) = \frac{\exp(\beta_j X_i)}{1 + \sum_{s=1}^{J} \exp(\beta_s X_i)}$$

pro $j = 1, \ldots, J$. V případě podmíněného logit modelu je tento výraz nahrazen vztahem

$$\Pr(Y_i = j) = \frac{\exp(\beta X_{ji})}{1 + \sum_{s=1}^{J} \exp(\beta X_{si})}.$$

Pokud se dobře podíváme na použité dolní indexy, můžeme vidět odlišnosti tohoto typu modelu od multinomiálního logit modelu. V případě podmíněného logitu máme pouze jede koeficient β (pro případ více vysvětlujících proměnných dojdeme k zobecnění na $\alpha, \beta_1, \ldots, \beta_k$). Pro multinomiální logit má parametr β dolní index j a máme tedy celkem J koeficientů (v příapadě úrovňové konstanty a k vysvětlujících proměnných máme celkem $(k + 1) \times J$ koeficientů). V multinomiálním logitu má vysvětlující proměnná dolní index i (tzn. vysvětlující proměnná může být např. důchod, který se liší mezi jednotlivci), přičemž podmíněný logit model má dodatečný dolní index j (tj. vysvětlující proměnná odpovídá proměnné časové vzdálenosti cesty do práce pro daného jednotlivce, která se liší v závislosti na použité alternativě dopravního prostředku).

Klíčový rozdíl mezi multinomiálním a podmíněným logitem je ten, že multinomiální logit je vhodný v případě,kdy se vysvětlující proměnné liší napříč jednotlivci, přičemž podmíněný logit se uplatní tam,kde se vysvěltující proměnné liší mezi alternativami. S podmíněným logitem nemůžeme mít vysvěltující proměnné, které by se lišily pouze mezi jednotlivci. Abychom si ukázali, proč tomu tak je, předpokládejme jednu vysvěltující proměnnou, která se liší mezi alternativami (X_j) a druhou, která se liší v rámci jednotlivců (Z_i). Podmíněný logit model tak rozšíříme do podoby

$$\Pr(Y_i = j) = \frac{\exp(\beta_1 X_{ji} + \beta_2 Z_i)}{\sum_{s=1}^{J} \exp(\beta_1 X_{si} + \beta_2 Z_i)}.$$

Vlastnosti operátoru exponenciály lze využít k ukázce toho, že nejsme nikdy schopni

odhadnout β_2. To znamená,

$$\begin{aligned}\Pr(Y_i = j) &= \frac{\exp(\beta_1 X_{ji} + \beta_2 Z_i)}{\sum_{s=1}^{J} \exp(\beta_1 X_{si} + \beta_2 Z_i)} \\ &= \frac{\exp(\beta_1 X_{ji}) \exp(\beta_2 Z_i)}{\sum_{s=1}^{J} \exp(\beta_1 X_{si}) \exp(\beta_2 Z_i)} \\ &= \frac{\exp(\beta_1 X_{ji}) \exp(\beta_2 Z_i)}{\exp(\beta_2 Z_i) \sum_{s=1}^{J} \exp(\beta_1 X_{si})} \\ &= \frac{\exp(\beta_1 X_{ji})}{\sum_{s=1}^{J} \exp(\beta_1 X_{si})}.\end{aligned}$$

Člen zahrnující vysvětlující proměnnou lišící se jen napříč jednotlivci lze z podmíněného logit modelu vykrátit.

Jako v případě všech modelů kvalitativní volby je obtížn¨á přímá interpretace koeficientů podmíněného logitu. Jak však bylo ukázáno v případě jiných modelů, můžeme odhadnout různé predikované pravděpodobnosti či různé mezní fekty s ohledem na tyto pravděpodobnosti. Například tak lze spočítat

$$\frac{\partial \Pr(Y_i = j)}{\partial X_{ji}},$$

i když stejně jako v případě logitu a probitu tento mezní vliv závisí na X_j, a tedy ne pouze na odhadu regresního koeficientu.

Bylo řečeno, že podmíněný logit by mohl být použit v případě, kdy se vysvětlující proměnné liší mezi alternativami (ale ne mezi jednotlivci). Jako poznámku k empirické praxi je třeba zmínit to, že podmíněný logit model lze upravit tak, že bude obsahovat vysvětlující proměnné lišící se mezi jednotlivci, a to vytvřoením odpovídajících vysvětlujících proměnných. Pro ilustrayi předpokládejme vysvětlující proměnnou Z_i, která se neliší v rámci alternativ. Můžeme vytvořit umělé proměnné pro alternativy $j = 1, \ldots, J$. Tyto umělé proměnnéoznačme jako D_{ji} (např. $D_{1i} = 1$ označuje první alternativu, $D_{2i} = 1$ označuje druhou alternativu apod.). Tyto umělé proměnné můžeme použít v podmíněném logit modelu (protože nabývají různých hodnot pro různé alternativy). Můžeme dokonce propojit tyto umělé proměnné s vysvětlujícími proměnnými jako Z_i. Jsme tedy schopni vytvořit $Z_i \times D_{ji}$ a použít je jako novou množinu vysvětlujících proměnných, které se již liší mezi alternativami. Tímto způsobem tak jsme schopni odhadnout podmíněný logitmodel i v případě vysvětlujících proměnných, které se neliší mezi alternativami. Samozřejmě tímto „trikem“ přidáváme do modelu velkou spoustu vysvětlujících proměnných (tzn. že máme vytvořeno J umělých proměnných a pokud je zkombinujeme s Z_i, dodáváme do modelu J vysvěltujícíchproměnných). Ve skutečnosti lze snadno ukázat, že při využití tohoto postupu se podmíněný logit model ukazuje jako ekvivalentní varianta multinomiálního logit modelu. Odpovídající definicí koeficientů a vysvětlujících proměnných pakoba modely mohou být naprosto totžné. Z tohoto důvodu tak nebylo nutné striktně rozdělit část této podkapitoly na pasáž věnovanou multinomiálnímu logit modelu a podmíněnému logit modelu. Oba pojmy jsou ale bežně používány, a tudíž je dobré analyzovat je odděleně.

Vzhledem k úzkému vztahu mezi podmínenými a multinomiálními logit modely nebudeme detailně probírat odhady a testování tohoto „nového“ typu logit modelů. Postačující je pro nás skutečnost, že řada ekonometrických programů je schopna provést odhad metodou maximální věrohodnosti, provést testy hypotéz na základě věrohodnostního poměru a implementovat i jiné estimátory (např. heteroskedasticitě konzistentní estimátory).

6.3 Modely omezených vysvětlovaných proměnných

V této podkapitole se budeme zabývat modely regresního typu, kdy závisle proměnná není spojitá, ale je určitým způsobem omezená. Začneme diskuzí nad *tobit* modelem, což je příklad modelu, kdy je vysvětlovaná proměnná cenzorována. V další části se budeme věnovat případu, kdy závisle proměnná vyjadřuje počet, může tedy nabývat hodnot $0, 1, 2, 3, \ldots$.

6.3.1 Tobit

V mnoha případech se můžeme setkat se situací, kdy námi využívaná data jsou nejakým způsobem cenzorována. Například v rámci dotazníkových šetření mohou být repsondenti dotazováni na svůj příjem. Může nastat situace, že v případě, pokud jejich příjem přesáhne 100000 dolarů, je tento jejich příjem zaznamenán jako „100000 dolarů a více“. V rámci ilustrace probit a logit modelů jsme využili data pojednávající o tom, jestli daný jednotlivec měl či neměl mimomanželskou aféru, což odpovídalo závisle umělé proměnné nabývající hodnot 1 nebo 0. V rámci výzkumu však naše odotazy mohly být směřovány na počet těchto afér. Repsondenti tak mohli poskytnout odpověď 0, 1, 2, atd. kdy poslední kategorie mohla být „deset a více“. Jednotlivcům s 11, 12 a více aférami by tak byla přiřazena odpověď „deset a více“. To jsou některé z příkladů, kdy se můžeme setkat s omezenými daty. Pro jednotlivce, kterých se takovéto omezení týká takneznáme skutečnou hodnotu závisle proměnné, víme jen, že je více či méně než konkrétní hodnota.

Existuje řada typů jakým mohou být data cenzorována, kdy pro jednotlivé typy omezení jsou vyvinuty odpovídající ekonometrické modely. V této části se zaměříme na jeden z nich: tobit model. Ostatní typy modelů mezených dat jsou podobné, a pokud pochopíme princip tobit modelu, neměli bychom mít problém představit si i další typy těchto modelů.

Tobit model má závisle proměnnou cenzorovanou na hodnotě nula. Máme tedy model vícenásobné regrese

$$Y_i^* = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i.$$

Závisle proměnnou, Y_i^* však přímo nepozorujeme. Místo toho pozorujeme Y_i, kde

$$\begin{array}{rcccl} Y_i & = & Y_i^* & \text{pokud} & Y_i^* > 0, \\ Y_i & = & 0 & \text{pokud} & Y_i^* \leq 0. \end{array}$$

Jako příklad uplatnění tobit modelu uvažujme ekonomický model, který analyzuje úroveň zamýšlených investic firem v závislsoti na jejich charakteristikách. V odpoví-

dajícím regresním modelu bychom měli zamýšlené investice jako závisle proměnnou a charkateristiky firemy jako vysvětlující proměnné. V praxi jsouvšak data o zamýšlených investických zřídkakdy dostupná. Místo toho pozorujeme skutečné (realizovan) investice. Pokud není možno provádět negativní investice, potom se skutečné investice budou rovnat zamýšleným investicím v případě, pokud je úroveň zamýšlených investic kladná. Negativní hodnoty zamýšlených investic odpovídají nulovým hodnotám skutečných investic.

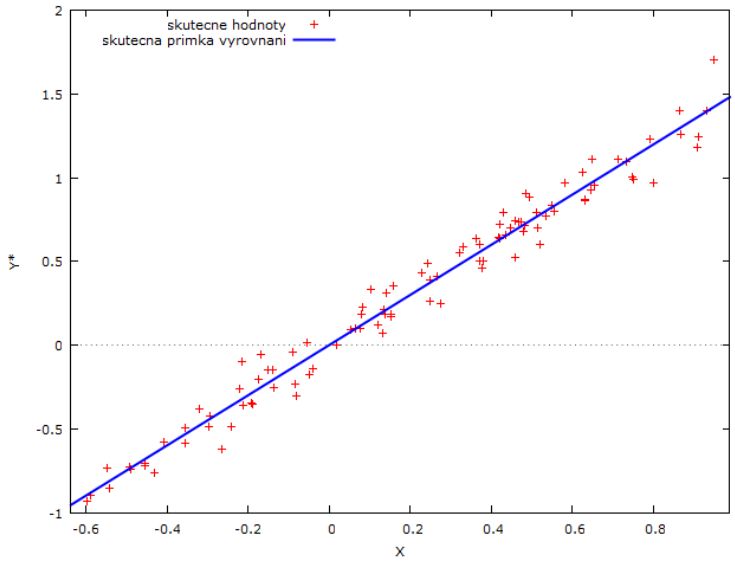
Než se dostaneme kanalýze tobit modelu, mělibychom si vysvětlit, proč je v tomto případě nevhodné použití metody nejmenších čtverců na nepozorovaná data. Mohli bychom tedy uvažovat o regresi

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i.$$

Je zřejmé, že estimátor metody nejmenších čtverců bude vychýlená. Nebudeme si tuto skutečnost dokazovat, spíše si ji ukážeme na obrázcích. Obrázek 6.1 je bodový graf obsahující uměle simulovaná data pro jednoduchý regresní model

$$Y_i^* = \alpha + \beta X_i + \epsilon_i.$$

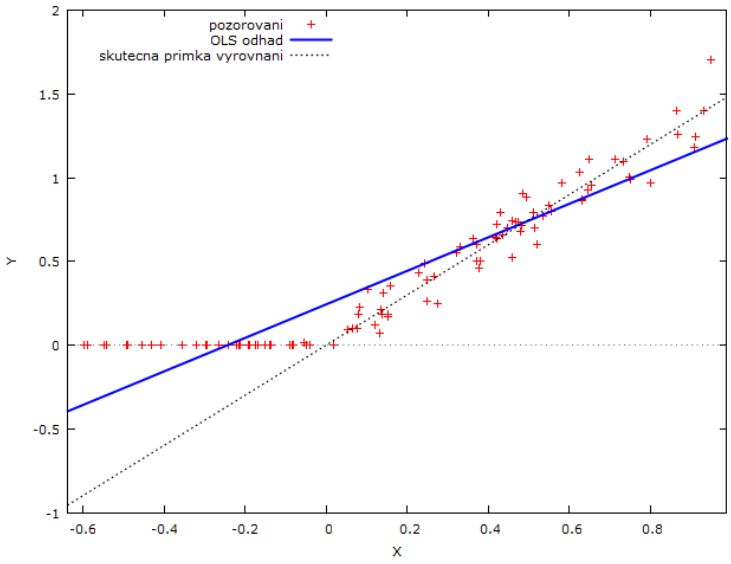
Na obrázku 6.1 je ukázána i skutečná regresní přímka na jejímž základě byla tato data vytvořena (soubor `tobit.gdt`).



Obrázek 6.1: Bodový graf původní časové řady a skutečná regresní přímka.

Předpokládejme nyní, že data z obrázku 6.1 jsoucenzorována. Konkrétně vytvoříme $Y_i = Y_i^*$ pro pozitivní hodnoty Y_i^*. Jinak platí $Y_i = 0$. Tato data jsou vykreslena na obrázku 6.2 (podívejme se zejména na data odpovídající Y=0). Na obrázku 6.2 je vykreslena skutečná regresní přímka (stejná jako na obrázku 6.1) a rovněž i přímka

vyrovnání získanána základě regrese Y na X. Vidíme, že regresní přímka na základě OLS estimátoru je poněkud vzdálena od té skutečné. OLS estimátor je tak skuečně vychýlený. Aby metoda OLS vyrovnala data co nejlépe (minimalizovala součet čtverců reziduí), natáčí se regresní přímka směrem ke všem omezeným pozorování. Obrázek 6.2 nám rovněž napovídá, že čím více takovýchto omezených pozorování budeme mít, tím horší bude vychýlení.



Obrázek 6.2: Bodový graf omezené datové množiny, skutečná a OLS regresní přímka.

Předchozí obrázky a diskuze zdůrazňují skutečnost, že pokud pracujeme s cenzorovanými pozorováními, neměli bychom používat OLS metodu. Místo toho bychom měli použít odpovídající estimátor, který rozpozná podstatu omezení závisle proměnné. Tímto estimátorem je v našem případě tobit estimátor. Nebudeme si ho do detailu popisovat. Jedná se o estimátor metody maximální věrohodnosti, který korektně bere v úvahu cenzorování dat. Tobit estimátor je dostupný ve většině ekonometrických programů. Můžeme tak automaticky využít tobit odhady parametrů $\alpha, \beta_1, \ldots, \beta_k$ spolu s obvyklými statistickými informacemi (např. p.hodnota testu statistické významnosti koeficientů, konfidenční intervaly parametr, apod.). Rovněž je vcelku snadné provést testy hypotéz založené na věrohodnostním poměru. Jako v případě předchozích regresních modelů můžeme narazit na problém heteroskedasticity, kdy některé počítačové programy obsahují heteroskedasticitě konzistentní estimátory pro tobit model.

U interpretace výsledků tobit odhadů není třeba se zdržovat, protože jejich interpretace je obdobná té pro jakýkoli jiný regresní model. Například β_j má svou tradiční interpretaci mezního vlivu vzhledem na Y^*. Parametr β_j je tedy mezní vliv proměnné X_j na Y^*, pokud seostatní vysvětlující proměnné nemění. Pokud tedy máme dobrý ekonometrický software, je odhad a testování tobit modelu zřejmé a interpretace výsledků je snadná.

6.3.2 Práce s daty vyjadřujícími počet

Existuje řada aplikací, ve kterých závisle proměnná může nabývat pouze celých čísel $(0, 1, 2, 3, \ldots)$. Tato data budeme označovat jako *data vyjadřující počet (count data)*. Empirický příklad uváděný dále je z odborné literatury věnované ekonomii zdravotnictví. Závisle proměnná je počet návštěv u lékaře v daném roce. Ekonomové zdravotnictví se zajímají o analýzu toho, jestli může být tato vysvětlovaná proměnná vysvětlena proměnnými jako je příjem rodiny, jestli má dotyčný soukromé zdravotní pojištění a dalšími individuálními charakteristikami. To nám naznačuje, že bychom mohli použít regresní model v podobě:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i.$$

Jediným problémem, který nám poněkud brání provést regresi s využitím OLS technik je to, že není moc rozumné předpokládat, že závisle proměnná a náhodné složky mají normální rozdělení. Normální rozdělení je vhodné pro spojité náhodné veličiny. Pro datavyjadřující počet není normální rozdělení vhodné, protože vysvtlovaná veličina nemůže nabývat jakýchkoli hodnot.

Připomeňme si, že veškerá teoretická odvození pro regresní model z kapitoly 3 byla prováděna za předpokladu splnění klasických předpokladů, tedy i předpokladu normality závisle proměnné resp. náhodných složek. Samozřejmě, že mnohá z odvození kapitoly 3 nevyžadovala splnění všech těchto klasických předpokladů. Důkaz nestrannosti OLS estimátoru splnění předpokladu normality nevyžadoval. Odvození rozptylu OLS estimátoru rovněž nemuselo zahrnovat přepdoklad normality. To nám napovídá, že i v případě dat vyjadřujících počet by OLS estimátor mohl být akceptovatelny. V řadě situací tomu tak skutečně je. Nicméně existují mnohem lepší estimátory, které lze využít právě pro případ vysvětlovaných proměnných vyjadřujících počet. V další části se zaměříme na jeden z nejvýznamnějších, který se týká *Poissonova regresního modelu*, a budeme se tak věnovat i odpovídajícímu estimátoru.

Poissonův regresní model

Poissonův regresní model je v podstatě regresní model splňující klasické předpoklady, až na jedinou výjimku. Touto výjimkou je to, že závisle proměnná předpokládá Poissonovo rozdělení. Nebudeme si zde formálně definovat tento typ rozdělení. Klíčovou vlastností je však to, že se jedná o velmi běžné rozdělení pro náhodné veličiny nabývající hodnot $0, 1, 2, 3, \ldots$. Za předpokladu Poissonova rozdělení máme definovanou věrohodnostní funkci a můžeme tak vyvinou estimátor metody maximální věrohodnosti. Tento estimátor je součástí standardních ekonometrických programů. V rámci Poissonova modelu tak můžeme automaticky získat odhady $\alpha, \beta_1, \ldots, beta_k$ spolu s obvyklými statistickými informacemi jako jsou p-hodnoty testů nulovosti každého z regresních koeficientů, konfidenční intervaly každého z parametrů, apod. Postupy testů hypotéz založené na věrohodnostním poměru lze stejně tak snadno provést. Jako v případě většiny modelů z této kapitoly umožňují mnohé programy robustní odha rozptylů koeficientů (můžeme tedy provést heteroskedasticitě konzistentní odhady).

Pokud máme vhodný software, lze odhad a testování Poissonova regresního modelu snadno provést. Může nás tedy zajímat samotná interpretace odhadů. Poissonovo

rozdělení (stejně jako normální rozdělení a řada dalších rozdělení) má svou střední hodnotu. Tato střední hodnota je označovaná jako λ. Pokud má závisle proměnná, Y_i, Poissonovo rozdělení, říkáme, že

$$E(Y_i) = \lambda_i.$$

Při splnění klasických předpokladů jsme pro případ jednoduché regrese měli situaci $E(Y_i) = \beta X_i$. Poissonův regresní model může pracovat i v tomto kontextu a platí

$$E(Y_i) = \lambda_i = \beta X_i.$$

Protože Y_i nemůže být záporné, je v praxi běžnější pracovat s výrazem

$$E(Y_i) = \lambda_i = \exp(\beta X_i),$$

a právě tato volba je obsažena ve většině ekonometrických programů (dobré je ale vždy nahlédnout do příslušné dokumentace). I v následujícím textu budeme uvažovat tuto podobu Poissonova regresního modelu.

Koeficient Poissonova regresního modelu lze vztáhnout k mezním vlivům na základě faktu, že

$$\frac{\mathrm{d}E(Y_i)}{\mathrm{d}X_i} = \beta \exp(\beta X_i).$$

Rozšíření na vícenásobnou regresi je velmi jednoduché a zřejmé. Pokud pracujeme s modelem, kde

$$E(Y_i) = \lambda_i = \exp(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}),$$

potom $\beta_j \exp(\beta_j X_j i)$ měří to, jak moc očekáváme, že se závisle proměnná změní, pokud se X_j změní o jednotku a předpokládáme-li neměnnost ostatních vysvětlujících proměnných.

Některé počítačové programy dávají výstup v podobě tzv. *podílu relativní incidence (incidence rate ratio)*. Tento podíl je založen na tom, že se nejprve spočítá $E(Y)$ pro konkrétní hodnoty $X_1, X_2, \ldots, X_k$ a potom se změní jedna z vysvětlujících proměnných o jednotku. Spočítá se nová střední hodnota, $E(Y)$, a vezme se podíl těchto dvou středních hodnot. Odvození s využitím exponenciálních funkcí nám ukazují, že tento poměr je

$$\frac{\exp(\alpha + \beta_1 X_1 + \ldots + \beta_j (X_j + 1) + \ldots + \beta_k X_k)}{\exp(\alpha + \beta_1 X_1 + \ldots + \beta_j X_j + \ldots + \beta_k X_k)} = \exp(\beta_j).$$

Podíl relativní incidence není úplně přesně to, co je mezní vliv dané proměnné diskutovaný v rámci modelu vícenásobné regrese. Je to však způsob, jak měřit vliv jednotkové změny vysvětlující proměnné na závisle proměnnou (při neměnnosti ostatních vysvětlujících proměnných). V modelu dat vyjadřujících počet je to užitečné měřítko, protože závisí jen na daném koeficientu, nikoli na vysvětlujících proměnných.

Poznamejme ještě, že pokud je vysvětlující proměnná vyjádřená v logaritmu (což je v mnohých aplikacích obvyklé), interpretace odhadů koeficientů se zjednodušuje. Pokud tedy

$$\begin{aligned} E(Y_i) = \lambda_i &= \exp(\beta \ln(X_i)) \\ &= X_i \exp(\beta), \end{aligned}$$

potom $\exp(\beta)$ je obvyklé měřítko mezního vlivu. Poslení obecnou poznámkou je to, že otázky týkající se toho, jak interpretovat mezní vlivy pro data vyjadřující počet jsou podobné jako ty diskutované v kapitole 4. V této kapitole jsme tedy diskutovali otázku toho, jak interpretovat regresní koeficienty v případě, kdy jsou závisle proměnná a/nebo vysvětlující proměnné vyjádřené v logaritmech.

Test přerozptýlenosti v Poissonově regresním modelu

Poissonův regresní model má jeden omezující předpoklad, který nenastává v případě regresního modelu s normálně rozdělenými náhodnými složkami. Vlastností Poissonova rozdělení je to, že střední hodnota a rozptyl jsou totožené. V případě normálního rozdělení jsou střední hodnota (často označovaná jako μ) a rozptyl (označován obvykle jako σ^2) zcela rozdílné veličiny. V rámci Poissonova regresního modelu jsme si řekli, že

$$E(Y_i) = \lambda_i,$$

ovšem nyní již víme, že nám to implikuje

$$var(Y_i) = \lambda_i.$$

Jedná se tedy o dobrý předpoklad? Odpověď na tuto otázkuzávisí na konkrétních datech, která používáme. V některých situacích je tento předpoklad rozumný, v některých rozumný být nemusí. Nícméně nás to vede k důležitému testu hypotézy

$$H_0 : E(Y_i) = var(Y_i).$$

Existuje řada testů této hypotézy (a opět, relevantní programy by měly tento test dokázat provést). Jeden test, který nazveme podle tvůrců *Cameronův-Trivediho test*, je prakticky snadno zvládnutelný. Takové testy jsou označovány jako *testy přerozptýlenosti (tests of overdispersion)*. Proč takovéto označení? Poznamenejm si že rozptyl je měřítkem „disperze“ a pokud je H_0 je chybná, rozptyl by měl být nad („over“) hodnotou, kterou nám dává Poissonův regresní model.

Cameronův-Trivediho test zahrnuje nejdříve odhad Poissonova regresního modelu (využitím metody maximální věrohodnosti) a získání vyrovnaných hodnot závisle proměnné. Označme si tyti vyrovnané hodnoty $\widehat{\lambda}_i$ pro $i = 1, \ldots, N$. Data a vyrovnané hodnoty lzepoužít pro konstrukci nové proměnné:

$$Z_i = \frac{\left(Y_i - \widehat{\lambda}_i\right)^2 - Y_i}{\widehat{\lambda}_i \sqrt{(2)}}.$$

Z důvodů, které si zde nebudeme uvádět má tato proměnná při platnosti nulové hypotézy střední hodnotu nulovou. Jednoduchým způsobem testu této hypotézy je provedení regrese Z na úrovňovou konstantu a zjištění, jestli je úrovňová konstanta (estimátor je aritmetický průměr Z) statisticky významná (pomocí nám známého t-testu). Pokud je úrovňová konstanta statisticky nevýznamná, nezamítáme nulovou hypotézu, H_0, a POissonův regresní model je vhodným modelem pro použití. Pokud však shledáme

úrovňovou konstantu jako statisticky významnou, neměli bychom Poissonův regresní model použít.

Co dělat v případě, když použijeme test přerozptýlenosti a bude nám indikovat nevhodnost použití Poissonova modelu? Existuje řada alternativ. Nejpopulárnější je tzv. *negativní binomiální regresní model*. Nemá smysl pouštět se do podrobného vysvětlování, vězme jen, tento model umožňuje odlišnost $E(Y_i)$ a $var(Y_i)$ a některé ekonometrické programy ho dokáží odhadnout. Interpretace regresních koeficientů je stejná jako u Poissonova regresního modelu. Tento model je diskutován v pokročilejších učebnicích ekonometrie (spolu s dalšími modely). Knížka C. Camerona a P. Trivediho, „Regression Analysis of Count Data," je věnovaná pouze modelům s daty vyjadřujícími počet.

Příklad 6.7. *Poptávka po zdravotní péči*
Předpokládejme, že nás zajímá vysvětlení faktorů, které ovlivňují poptávku po zdravotní péči mezi senitory. Máme data o $N = 4406$ Američanů ve věku 66 a více let, získaných na základě výběrového šetření. Data byla využita v článku „Demand for medical care by the elderly: a finite mixture approach" autorů P. Deb a P. Trivedi, publikovaného v *Journal of Applied Econometrics* v roce 1987. Tento článek poskytuje bližší informace ohledně dat. Data jsou obsahem souboru `deb_trivedi.gdt`. Závislá a vysvětlující proměnné jsou:

- $DRVISIT$ = počet návštěv u lékaře v minulém roce;
- $FAMINC$ = rodinný příjem (v desítkách tisíc dolarů);
- $MALE = 1$ pokud je jednotlivec muž ($= 0$ jinak);
- $EXCHLTH = 1$ pokud osoba cítí, že má výborné zdraví ($= 0$ jinak);
- $POORHLTH = 1$ pokud osoba cítí, že má chatrné zdraví ($= 0$ jinak);
- AGE věk respondenta (v letech dělený stovkou);
- $MARRIED = 1$ pokud je osoba ženatá nebo vdaná ($= 0$ jinak);
- $PRIVINS = 1$ pokud má osoba soukromé zdravotní pojištění ($= 0$ jinak).

Empirické výsledky s využitím Poissonova modelu jsou obsahem tabulky 6.5. S výjimkou proměnné $FAMINC$, která je významná na 10% hladině významnosti (ale ne na 5%), jsou všechny vysvětlující proměnné silně statisticky významné. To je obvyklé pro velké datové soubory (jako je ten náš s $N = 4406$).
(dokončení v příkladu 6.8

6.3.3 Rozšíření

V rámci této kapitoly jsme pracovali s průřezovými daty. Je dobré poznamenta , že všechny modely z této kapitoly lze rozšířit i pro panelová data. Nebudeme se jim detailně věnovat, nicméně mnohé modely z kapitoly 9 (např. model náhodných vlivů) lze přizpůsobit pro práci s kvalitativními vysvětlovanými proměnnými a máme tak k dis-

Příklad 6.8. *Poptávka po zdravotní péči (dokončení příkladu 6.7)*

Tabulka 6.5: Poissonův model pro data poptávky po zdravotní péči

Proměnná	Koef.	p-hodnota pro $\beta_j = 0$	95% int. spol.	IRR*
Konstanta	1.78	0.00	[1.62;1.94]	—
$FAMINC$	0.004	0.08	[-0.001;0.008]	1.004
$MALE$	-0.09	0.00	[-0.11;-0.06]	0.92
$EXCHLTH$	-0.49	0.00	[-0.54;-0.43]	0.62
$POORHLTH$	0.53	0.00	[0.49;0.56]	1.69
AGE	-0.03	0.00	[-0.05;-0.01]	0.97
$MARRIED$	-0.06	0.00	[-0.03;-0.09]	0.94
$PRIVINS$	0.29	0.00	[0.26;0.32]	1.33

* *Incidence rate ratio* – podíl relativních incidencí.

Jako příklad interpretace výsledků si všimněme, že odhad koeficientu u proměnné $EXHLTH$ je negativní. To zcela nepřekvapivě znamená, že osoba, která se cítí zdravotně skvěle bude navštěvovat lékaře s menší pravděpodobností. podíl relativní incidence je v tomto případě 0.62, což nám implikuje, že osoba s vynikajícím zdravím navštěvuje svého doktora z 62 % četnosti návštěv osoby, která vynikající zdraví nemá. Jiným příkladem je podíl relativních incidencí pro $PRIVINS$, který je 1.33. To znamená, že jednotlivec se soukromým zdravotním pojištěním navštěvuje svého doktora o 33% častěji než osoba bez tohoto pojištění (při němnnosti ostatních parametrů).
Test přerozptýlenosti by však poskytoval silný důkaz pro to, že Poissonův regresní model není v případě této datové množiny vhodný. Pro vážnější práci s těmito daty by tak byl lepší negativní binomiální regresní model.

pozici např. probit modely náhodných vlivů. Řada ekonometrických balíčků umí pracovat i s panelovými variantami modelů kvalitativních a omezených vysvětlovaných proměnných.

Existuje řada modelů použitelných v situacích, kdy je závisle proměnné omezená specifickým způsobem. Není myslitelné pokrýt zde všechny možné modely kvalitativní volby a omezených vysvětlovaných proměnných. Existují však například *modely trvání (durations models)*, které se používají v případě, kdy závisle proměnná vyjadřuje dobu trvání (např. v aplikacích z ekonomie práce taková proměnná může být počet měsíců, po které jsou zkoumaní jednotlivci nezaměstnaná). Důležité je rovněž vědět, že v rámci obecné třídy logit a probit modlů existují jejich speciální varianty, jako je např. dříve krátce zmíněný uspořádaný probit model nebo vnořený logit model, ale nejen tyto.

Jiný typ modelů, který lze interpretovat jako model omezené vysvětlované proměnné je tzv. *treatment effects model (model efektů léčby)*. Pro jehomotivace si všimněme, že v lékařské statistice existuje řada případů, kdy se náš zájem zaměřuje na efekt nějaké léčby na konečný výsledek v podobě zdraví pacientů (léčených). V této souvislosti byly vyvinuty modely k odhadu těchto *efektů léčby (treatment effects)*. V

ekonomii je také řada případů, kdy ná zajímá efekt „léčení“ (což může být např. účast v programu rekvalifikací) na „výsledek“ (např. dosažený plat). V řadě ekonomických aaplikací a aplikací z oblasti medicíny totiž statistickou analýzu komplikuje skutečnost, že jednotlivci nejsou do „programu léčby“ náhodně přiřazeni. S tím spojený problém je problém nespolupráce, kdy po předepsání léčby se tento jednotlivec příslušnéléčbě nepodřizuje. Toto je tedy oblast, kde ekonomická i medicínská literatura mají mnoho společného. V ideálním světě, kde bychom mohli sledovat výsledke pro každého jednotlivce ve stavu léčby či neléčby, by bylo velmi snadné vidět příslušný efekt léčby. Předpokládejme například, že nychom mohli nalézt plat jednotlivce, který by se účastnil programu rekvalifikace (či školení) a stejně tak bychom byli schopni zjistit mzdu,kterou by daný jednotlivec dostal, kdyby v tomto programu účasten nebyl. Rozdíl mezi těmito platy by samozřejmě bylo měřítko zisku z programu rekvalifikace. V praxi bychom však nebyli chopni oba tyto výsledky současně sledovat. V tomto smyslu jetak závisle proměnné omezena a spada do obecné kategorie modelů diskutovaných v této kapitole. Treatment effects modely se stávají čímdál více populární v rámci výzkumu odhadu zisků či benefitů z různých typů programů a politik.

Tyto modely jsou zde zmiňovány proto, abychom měli povědomí o tom, že existují pro případ, že bychom narazili na empirický problém, který by si jejich použití vyžadoval. Minimálně znalost názvů těchto modelů by nám měla pomoci v tom, že budeme vědět do kterých kapitol knížek pokročilé ekonometrie je potřeba se podívat.

6.4 Shrnutí

Tato kapitola popisoval modely, které je vhodné využít v případě, kdy je závisle proměnná omezená v rámci omezené množiny hodnot. Na základě této kapitoly tedy můžeme vyslovit následující závěry:

- Všechny modely diskutované v této kapitole jsou podobné modelu vícenásobné regrese v tom ohledu, že závisle proměnná závisí na vysvětlujících proměnných. Omezenost vysvětlované proměnné však naznačuje, že použití standardních OLS či GLS metod není vhodné.

- Modely kvalitativní (diskrétní) volby se tákají dat, kdy jednotlivci volí mezi různými alternativami. Modely binární volby zahrnují dvě alternativy a závisle proměnná je takumělá proměnná. Modely multinomiální volby zahrnují více než dvě alternativy.

- Probit a logit jsou dvanejpopulárnější modely binární volby. Odhad koeficientů těchto modelů lze provést metodami maximálnívěrohodnsoti. Ekonometrické programy nám poskytují standardní statistické informace (např. p-hodnoty pro testování toho, jestli je každý z koeficientů roven nule, intervaly spolehlivosti každého z koeficientu, atd.).

- Interpretace koeficientů v logit a probit modelech není na první pohled zřejmá tak, jako v případě modelů vícenásobné regrese. Koeficienty nelze interpretovat jendoduše jako mezní vlivy. V této kapitole byly ukázány některé další způsoby

prezentace empirických výsledků (např. predikované pravděpodobnosti volby, podíly šancí, atd.)-

- Multinomiální logit a multinomiální probit jsoudva nejpopulárnější modely multinomiální volby. Odhad koeficientů v těchto modelech lze provést s využitím metod maximální věrohodnosti. Počítačové programy nám opět poskytují standardní statistickouinformaci (např. p-hodnoty pro testování toho, jestli je každý z koeficientů roven nule, intervaly spolehlivosti každého z koeficientů, apod.).

- Multinomiální logit a probit modely zahrnují více regresních rovnic, kdy každá z rovnic je vztažena k volbě mezi každou jednou z alternativ a základní (benchmarkovou) alternativou.

- Koeficienty v těchto modelech nelze interpretovat přímo jako marginálnívlivy, nicméně v této kapitole byly diskutovány některé z alternativních způsobů prezentace empirických výsledků.

- Multinomiální logit model zahrnuje předpoklad nezávislosti irelevantních alternativ, který však nemusí být v některých aplikacích splněn. Multinomiálníprobit modeltento předpoklad neobsahuje, nicméně je velmi obtížné odhadnout ho v případě, kdy je počet alternativ velký.

- Podmíněný logit model lze využít pokud máme více než dvě alternativy, ale naše vysvětlující proměnné jsou vztaženy k těmto alternativám (tzn. liší se v rámci jednotlivých alternativ a nejsou výhradními charakteristikami jednotlivců provádějících danou volbu).

- Existují další typy modelů, v rámci kterých je závisle proměnné omezena (a na konci kapitoly je uveden jejich stručný přehled). V této kapitole byly diskutovány dva důležité modely: tobit model a Poissonův regresní model.

- Tobit model je příkladem cenzorovaného regresního modelu. Má vysvětlovanou proměnnou která je cenzorována k hodnotě nula.

- Odhad koeficientů tobit modelu lze provést metodami maximální věrohodnosti. Počítačové programy námposkytují standardní statistické informace (např. p-hodnoty protest statistické nevýznamnosti každého z koeficientů, konfidenční intervaly každého z koeficient, apod.). Koeficienty lze v tomto případě interpretovat standardně jako meznívlivy vysvětlující proměnné na vysvětlovanou proměnnou.

- Poissonův regresní model je obvykle používán v případě, kdy závisle proměnná vyjadřuje počet.

- Odhad koeficientů Poissonova regresního modelu lze provést metodami maximální věrohodnsoti. Vhodný počítačový software nám poskytně standardní statistiky (např. p-hodnota testů hypotéz nevýznamnosti každého z parametrů, konfidenční intervaly parametrů,atd.).

- Interpretace koeficientů v Poissonově regresního modelu je velmi podobná interpretaci koeficientů v rámci vícenásobné regrese. V této kapitole však byly diskutovány některé interpretační problémy, které nám vznikají a byl zaveden koncept podílu relativních incidencí.
- Poissonův regresní model máomezující předpoklad, že střední hodnota a rozptyl závisle proměnné jsoustejné (pro dané hodnoty vysvětlujících proměnných). Testy přerozptýlenosti lze snadno provést pro ověření tohoto předpokladu. Pokud je tento předpoklad porušen, existuje jiný typ modelu zvaný negativní binomiální regresní model, který je dorbé použít.

Měli bychom tak již znát a umět vysvětlit obsah následujících klíčových pojmů:

- Kvalitativní vysvětlovaná proměnná
- Omezená vysvětlovaná proměnná
- Modely binární volby
- Pravděpodobnost volby
- Probit model
- Logit model
- Podíl šancí
- Mezní vliv v probit modelu
- Multinomiální probit model
- Multinomiální logit model
- Základní alternativa (benchmark)
- Nezávislost irelevatních alternativ (IIA)
- Prediktivní pravděpodobnosti volby
- Podmíněný logit model
- Proměnná vyjadřující počet
- Tobit model
- Poissonův regresní model
- Podíl relativní incidence
- Test přerozptýlenosti
- Camronův-Trivediho test
- Treatment effects modely
- Modely trvání

Kapitola 7

Analýza jednorozměrných časových řad

V této kapitole se dozvíme:

☞

7.1 Úvod

Řada aplikací v ekonomii, zejména v oblasti makroekonomie a financí, je zaměřena na analýzu časových řad. Většina příkladů, na které jsme se v předchozích kapitolách soutředili, byla zaměřena na práci s průřezovými daty. To nám umožnilo relativně snadno si vybudovat základní představu o problematice regrese, zahrnující takové koncepty, jako je testování hypotéz a intervaly spolehlivosti. Při práci s časovými řadami je znalos těchto základů naprosto klíčová. V případě práce s časovými řadami nám však vyvstávají některé nové otázky a problémy. Tato kapitola a kapitola následující je zaměřena právě na tyto problémy a otázky. V této kapitole se změříme na *analýzu jednorozměrných časových řad*, tedy časových řad jediné proměnné. Většina empirických studií pracuje s regresním modelem či modelem podobného typu. Zaměřuje se na práci s modelem, kde vystupuje několik proměnných (tj. jedna závisle proměnná a několik vysvětlujících proměnných). Analýza jednorozměrných časových řad není našim konečným cílem. Tím je regresní modelování s časovými řadami, což je náplní následující kapitoly. Z ohledem na toto tak může být překvapující, proč věnovat celou kapitolu analáze jednorozměrných časových řad. Můžeme pro to mít dva důvod. Prvním z nich je to, že práce s jedinou proměnnou nám umožňuje velmi snadné zavedení základních myšlenek a značení vztažených k časovým řadám. Druhým důvodem je to, jak uvidíme, že při práci s časovými řadami je důležité pochopení vlastností každé z jednotlivých proměnných před tím, než se pustíme do regresního modelu zahrnujícího více proměnných. V této kapitole si tak ukážeme základní nástroje spojené s prací s časovými řadami.

Abychom si alespoň intuitivně ukázali to, jak se ekonometrické metody pro časové řady liší od těch využívajících data průřezová, připomeňme si, že našim cílem je regrese závisle proměnné na nějaké vysvětlující proměnné. Analýza časových řad však bude čelit dvěma problémům, na které při práci s průřezovými daty nenarazíme:

1. Časová řada jedné proměnné může ovlivňovat jinou proměnnou s časovým zpožděním.

2. Pokud jsou proměnné *nestacionární*, může nám vzniknout problém *zdánlivé regrese (spurious regression)*.

První problém lze intuitivně pochopit na jednoduchých příkladech. Poud odhadujeme regresní model, zajímá nás zjištění míry vlivu jedné nebo více vysvětlujících proměnných na závisle proměnnou. V případě časových řad musíme být velmi obezřetní při výběru vysvětlujících proměnných, protože jejich vliv na závisle proměnnou se může projevit až s odstupem času. Například, pokud je centrální banka znepokojena rostoucí inflací, zvýší pravděpodobně úrokové sazby. Dopad takovéto změny úrokových sazeb se v ekonomice bezpochyby projeví až se zpožděním, obnášejícím třeba i více než rok. Těmito projevy máme na mysli zejména dopad na inflaci a jíné významné makroekonomické veličiny (např. míru nezaměstnanosti). Obecně se dopady většiny nástrojů monetární a fiskální politiky projeví až v budoucích obdobích (obvykle čtvrtletích či letech). Tento problém je tedy problémem v oblasti makroekonomie, nicméně i v oblasti mikroekonomie se s obdobnou problematikou můžeme setkat. Příkladem může být rozhodnutí firmy o provedení nové investice (např. nákup nových počítačů), které neovlivníokamžitě její produkci. Nějaký čas zabere samotný nákup, jejich instalace a zaškolení pracovníků, kteří je budou využívat. Investice tak ovlivní produkci až za nějaký čas. V kontextu regrese to znaměna, že hodnota závisle proměnné v daném časovém okamžiku může záviset nejen na hodnotách vysvětlující proměnné (či proměnných) ve stejném čase, alei i na hodnotách těchto vysvětlujících proměnných v časech minulých.

Na tomto místě nám asi nemusí být zcela zřejmý druhý problém zmíňovaný výše. Pojmy nestacionarity, stacionarity a zdánlivé regrese budou podrobněji diskutován v následujícím textu. Nicméně, můžeme si již teď zapamatovat obecené pravidlo (diskutované ještě později), že pokud máme nestacionární časové řady proměnných, neměli bychom je jen tak v regresi používat. Než se pustíme do regrese, je vhodným postupem transformace proměnných do podoby stacionárních řad. Jediná vyjímka z tohoto obecného pravidla nastává v případě proměnných, které jsou tzv. *kointegrované*. Co je tím myšleno se dozvíme v následující kapitole. Pokud se nám zdají některé z doposud zavedených pojmů bez formálních definic, stačí si uchovat v paměti to, že zkrátka v případě práce s časovými řadami nám vznikajé nové problémy, které v případě průřezových dat nenastanou. Tyto problémy mají za následek to, že naivní použití vícenásobné regrese v duchu předchozích kapitole může být s ohledem na platnost výsledků empirické analýzy riskantní. Cílem této a další kapitoly je ukázat korektní, modifikovanéregresní techniky pro práci sčasovými řadami.

V této kapitole se zaměříme na tzv. *autoregresní modely*. Tyto modely odpovídají tomu, s čím jsme se setkali v kapitole 5, části 5.4, v případě autokorelace náhodných složek. Autoregresní modely jsou nejpopulárnějšími modely jednorozměrných

časových řad, neboť je lze zapsat jako regresní modely. Můžeme tak využít řadu závěrů z předchozích kapitol. Autoregresní model nám umožňuje diskutovat důležitou problematiku nestacionarity. Konkrétně si zde uvedeme motivaci a definici pro koncept *jednotkového kořene (unit root)*. Pokud má časové řada jendotkový kořen, je nestacionární. Existuje celá řada testů jednotkového kořene. Nejoblíbenějším z nich je *Dickeyho-Fullerův test*, který si v této kapitole probereme. Existuje řada jiných modelu jednorozměrných časových řad. V příloze této kapitoly se zaměříme na jeden z nich, kterým je *model klouzavých součtů (moving average model)*. V této kapitole je diskutována rovněž problematika volatility, což je koncept významný pro ekonomii financí.

7.2 Značení v rámci časových řad

V části 5.4, kapitole 5, jsme si zavedli regresní model s autokorelovanými náhodnými chybami. Protoe autokorelované náhodné složky jsou velmi úzce spjaty s daty v podobě časových řad, nabídla nám tato podkapitolaurčitou diskuzi nad relevantností tématu časových řad. Není od věci si trohcu osvěžit koncepty řásti předchozí kapitoly, neboť tato podkapitola pojednává o podobné problematice. V rámci práce s časovými řadami jsme používali značení pro jednotlivá pozorování $t = 1, \ldots, T$ (místo $i = 1, \ldots, N$). Protože se zabýváme analýzou jednorozměrných časových řadm budeme si naši proměnnou označovat jako Y_t.

Významným konceptem při práci s časovými řadami je koncept zpožděné proměnné. Tento koncept si popíšeme vcelku podrobně, aby bylo jasné, co znaměná a jak konstruovat odpovídající proměnné. Co jsou to zpožděné proměnné člověk nejlépe vidí právě z postupu jejich konstrukce. Mějme tedy časovou řadu pro $t = 1, \ldots, T$ období proměnné Y. Předpokládejme, že si vytvoříme novou proměnnou W, která má pozorování $W_t = Y_t$ pro $t = 2, \ldots, T$ a novou proměnnou Z, která obsahuje pozorování $Z_t = Y_{t-1}$ pro $t = 2, \ldots, T$. Proč používáme zápis $t = 2, \ldots, T$ místo původního $t = 1, \ldots, T$? Pokud bychom psali $t = 1 \ldots, T$, potom by první pozorování proměnné Z, Z_1, bylo rovno Y_0. Ale my nevíme, jakou hodnotu má Y_0, protože proměnná Y je pozorovaná od $t = 1, \ldots, T$. Jinými slovy, W a Z obsahují pouze $T - 1$ pozorování. Poznamenejme rovněž, že pokud bychom vytvořili další novou proměnnou $X_t = Y_{t-2}$, potom by tato proměnní X měla pozorování or $t = 3, \ldots, T$ a obsahovala by tak pouze $T - 2$ pozorování.

Nové proměnné W a Z mají $T - 1$ pozorování. Pokud si představíme W a Z jako dva sloupce obsahující každý $T - 1$ čísel, vidíme, že první prvek proměnné W je Y_2 a první prvek Z je Y_1. Druhý prvek W a Z bude Y_3 resp. Y_2, atd. Jinými slovy tím říkáme, že Z obsahuje Y *zpožděné (lagged) o jedno období*.. Obecně můžeme vytvořit „Y zpožděnou o j období“. Můžeme tak uvažovat o proměnných „Y“, „Y zpožděné o jedno období,“ „Y zpožděné od dvě období,“ atd., jako o různých proměnných, podobně jako v příkladu s cenami domů jsme uvažovali různé proměnné „cena domu“, „velikost domu“ nebo „počet ložnic“. Velká část analýzy jendorozměrných časových řad může být chápána skrze tvrobu regrese s využitím Y jako závisle proměnné a zpožděných proměnných Y jako vysvětlujících proměnných.

Poznamenejme ovšem, že pokud chceme zahrnout několik vysvětlujících proměn-

ných do modelu vícenásobné regrese, musí mít všechny tyto proměnné stejný počet pozorování. Co to pro nás znamená v souvislosti s použitím zpožděných proměnných? Protože zpožděné proměnné obsahují méně pozorování, než původní časové řada, jakákoli regrese založená na zpožděných proměnných musí tuto skutečnost respektovat. Pokud je například Y závisle proměnná a vysvětlující proměnné jsou „Y zpožděné o jedno obodbí“ a „Y zpožděné o dvě období,“ potom posledně zmiňovaná proměnná má pouze $T-2$ pozorování. Musíme se ujistit, že všechny proměnné budou obsahovat tento počet pozorování, a to tak, že jednoduše příslušná první pozorování odřízneme. Níže uvedený příklad nám tento postup ilustruje. Obecně, každá proměnná v regresi s použitím časových řad, musí obsahovat počet pozorovánírovenT mínus maximální počet zpoždění u použitých proměnných. Otázka diskutovaná v tomto odstavci se vztahuje k tzv. *počátečním podmínkám (initial conditions)*. V rámci tohoto textu budeme chápat počáteční podmínky doposud popsaným způsobem, ovšem je třeba zdůraznit, že existuj i mnohem sofistikovanější metody pro zacházení s počátečními podmínkami, které v sobě nezahrnují odříznutí počátečních pozorování.

Jako příklad uvažujme situaci, že máme k dispozici celkem deset pozorování proměnné Y (tzn. $Y_1, Y_2, \ldots, Y_{10}$) a chceme provést regresi zahrnující Y, zpožděné Y, Y zpožděné o dvě období a Y zpožděné o tři období. Tabulka 7.1 ukazuje, jak budou příslušné proměnné vypadat.

Tabulka 7.1: Tvorba zpožděných proměnných.

Číslo řádku	Y	Zpožděné Y	Zpožděné Y o dvě období	Zpožděné Y o tři období
1	Y_4	Y_3	Y_2	Y_1
2	Y_5	Y_4	Y_3	Y_2
3	Y_6	Y_5	Y_4	Y_3
4	Y_7	Y_6	Y_5	Y_4
5	Y_8	Y_7	Y_6	Y_5
6	Y_9	Y_8	Y_7	Y_6
7	Y_{10}	Y_9	Y_8	Y_7

Všimněme si, že každá z proměnných obsahuje sedm pozorování, což je T mínus maximálnířád zpoždění (tj. $10-3=7$). Podíváme-li se na některý zřádků (např. řádek 4), vidíme, že: (a) Y obsahuje data v konkrétním čase (např. řadek čtyři obsahuje Y v čase $t=7$); (b) zpožděné Y obsahuje pozorování o jedno období dřívější /např. řádek 4 obsahuje Y v čase $t=6$); (c) Y zpožděné o dvě období obsahuje pozorování o dvě období starší (např. řádek 4 obsahuje Y v čase $t=5$); a (d) Y zpožděné o tři období obsahuje pozorování z před třemi periodami (např. řádek čtyři obsahuje Y v čase $t=4$). Většina ekonometrických programů dokáže vytvářet zpožděné proměnné automaticky pomocí jednoduchého příkazu či jediným kliknutím.

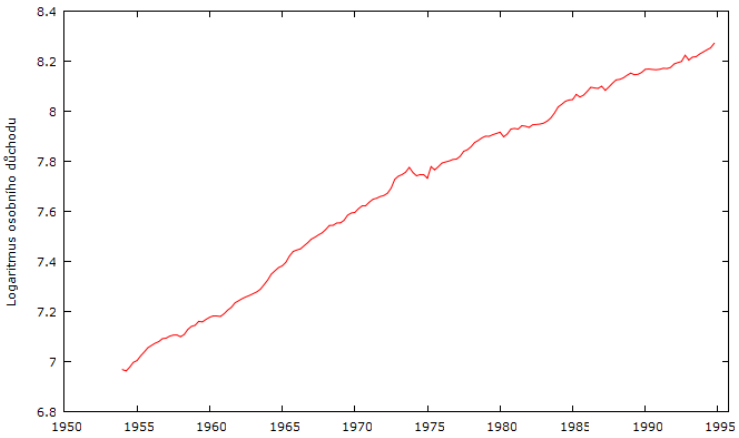
Důležité je si rovněž ujasnit další obvyklé značení. V případě každé proměnné příslušný dolní index označuje odkaz na odpovídající pozorování (např. Y_{10} je desáté pozorování proměnné Y). V takovémto případě je Y_{t-1} $(t-1)$-té pozorování proměnné

Y. Někdy však použijeme Y_{t-1} ve významu celé proměnné „Y zpožděné o jedno období“ (nebo Y_{t-j} pro označení celé proměnné „Y zpožděné o j období“). Naše značení tak bude s pohledu rozlišení proměnné a konkrétního pozorování poněkud volnější. To bývá v řadě textů (nejen učebních) obvyklé, protože z kontextu většinou jasně vyplývá, co zrovna máme na mysli, a není tak nutné zaplevelovat značení v rovnicích dalšími nadbytečnými indexy.

7.3 Trendy v časových řadách

Na začátku této kapitoly byla uvedena motivace pro analázu jednorozměrných časových řad spočívající v tom, že před tím, než se pustíme do výstavby regresního modelu, je důležité porozumět vlastnostem jednotlivých proměnných (v podobě časových řad). Bez dalšího vysvětlení bylo zmíněno, že v případě nestacionárních proměnných můžeme narazit na problém zdánlivé regrese. V praxi je otázka stacionarity a nestacionarity časových řad úzce spojena s konceptem trendu. V této části kapitoly se budeme věnovat diskuzi a analýze konceptu trendu, což nám poskytně dostatek intuice pro pochopení klíčové problematiky nestacionarity.

Relevantní otázky si můžeme ilustrovat s využitím důležité makroekonomické veličiny: osobního důchodu (data jsou obsahem souboru `USincome.gdt` dostupného v *gretlu* na záložce `Koop` položka `income.gdt`). Obrázek 7.1 je grafem časové řady logaritmu osobního důchodu v USA od prvního čtvrtletí roku 1954 do posledního čtvrtletí roku 1994. Jinými slovy, Y_t je logaritmus osobního důchodu pro $t = 1954Q1, \ldots, 1994Q4$. Původní časová řada osobního důchodu (nelogaritmovaná) je měřena v miliónech dolarů,



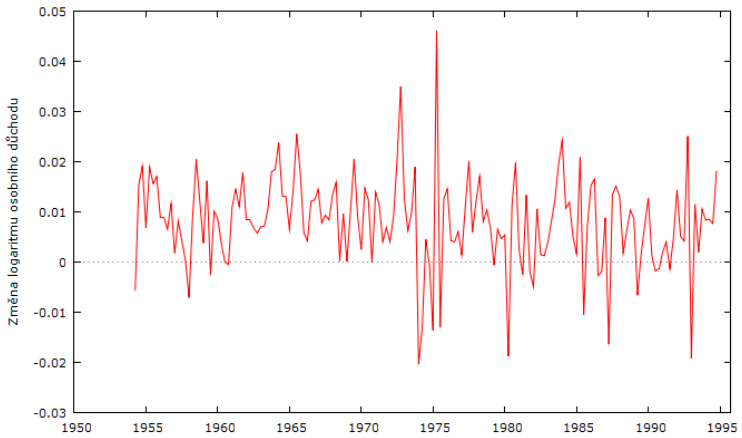
Obrázek 7.1: Graf časové řady logaritmu osobního důchodu v USA.

Všimněme si, že logaritmus osobního důchodu se zdá být v čase rostoucí,a to přibližně konstantním tempem růstu. Můžeme zde vidět i určitou variabilitu (např. krátký propad v osobním důchodu odpovídající recesi poloviny 70. let a počátku let osmde-

sátých), nicméně celkově odpovídá graf časové řady přibližně přímce s kladným sklonem. Tento vytrvalý vývoj (v tomto případě směrem nahoru) jeoznačován jako *trend*. Řada makroekonomických a finančních proměnných (např. HDP, cneová hladina, průmyslová produkce, spotřeba, vládní výdaje, indexy akciových trhů, atd.) vykazují trend podobného typu.

Na tomto místě je obvyklé zavést si pojem *diferencování*. Formálně,pokud je Y_t časová řada, potom $\Delta Y_t = Y_t - Y_{t-1}$ je první diference Y_t. Proměnná ΔY_t měří změnu či růst proměnné v čase. Pokud je Y_t proměnná v logaritmovanáa, potom nám vlasntnosti operátoru logaritmu říkají, že $100Y_t$ bude vyjadřovat procentní změnu původní proměnné (nelogaritmované) mezi obdobím $t-1$ a t. Tento výsledek je dalším důvodem pro práci s logaritmy proměnných. Ve své podstatě je práce s logaritmy makroekonomických veličin natolik obvyklá, že hovoří-li se o „osobním důchodu“, mívá se tím na mysli „logaritmus osobního důchodu“. Výraz ΔY_t je označován jako „delta Y“ nebo “změna v proměnné Y“.

Obrázek 7.2 vyresluje změnu v logaritmu osobního důchodu s využitím dat z obrázku 7.1. Hodnoty na ose y lze interpretovatjako procentní změnu v tom smyslu, že 0.02 odpovídá 2 %, apod. Všimněme si, že obrázek 7.2 vypadá zcela odlišně od obrázku 7.1. Trendové chování z obrázku 7.1 nám zcela zmizelo (k tomuto faktu se ještě vrátíme). Obrázek 7.2 nám ukazuje, že osobní důchod má tendenci růst tempem přibližně 1 % za čtvrtletí, přestože se toto tempo růstu vyznačuje významnou variabilitou v čase. V mnohých obdobích recese osobní důchod klesal a v některých obdobích expanze napoak rostl tempem 3 % nebo 4 % za čtvrtletí.



Obrázek 7.2: Graf časové řady uměny logaritmu osobního důchodu v USA.

Další vlastností časových řad, kterou v případě průřezových dat obvykle nenalezneme, je existence korelace mezi pozorováními. Například osobní důchod dnes, bývá vysoce korelován s osobnímdůchodem zpředchozího čtvrtletí (tedy „včera“). Pokud bychom ve skutečnosti spočítali korelaci mezi osobním důchodem a zpožděným osobním důchodem, získali bychom hodnotu 0.999716, což je hodnota velmi blízko jedné, která označuje perfektní korelaci. Pokud však spočítáme korelaci mezi změnou osob-

ního důchodu a změnou osobního důchodu zpožděnou o jedno období, získáme hodnotu $= 0.00235$, která je velmi blízko nule, tedy hdnotě nekorelovanosti. Tato zjištění mají svůj intuitivní smysl. Makroekonomické časové řady, jako důchod, HDP, spotřeba, atd., se v čase vyvíjejí velmi pozvolna. I v obdobích hluboké recese jen zřídka kdy spadnou o více než $1\ \%$ nebo $2\ \%$ za čtvrtletí. Důsledkem je to, že čtvrtletní důchod v jednom období bude mít tendenci být velmi podobný důchodu minulého odbobí, což se projeví ve vysokém korelačním koeficientu. Oproti tomu jsou změny v makroekonomických veličinách velmi nevyzpytatelné. Změny v důchodu toto čtvrtletí a minulé čtvrtletí mohou být vzájemně velmi odlišné, což se nám projevuje v téměř nulovém korelačním koeficientu mezi těmito veličinami.

Obrázky 7.1 a 7.2 a výsledky korelace diskutované v předchozím odstavci byly založeny na osobním důchodu ve Spojených státech. Jiné makroekonomické časové řady ve spoustě dalších zemí mají podobný charakter chování. Proměnná Y tak, jinými slovy, má tendenci vykazövat trendové chování spojené s vysokými hodnotami korelace v čase, přičemž proměnná ΔY má tendenci zcela opačného chování, tedy žádný trend a korelaci v čase. Tyto vlastnosti jsou vcelku důležité v rámci regresního modelování s využitím časových řad, neboť se dotýkají problému nestacionarity. V souladu s tím se budeme ve zbytku této kapitoly zabývat formálními nástorji a modely pro uchopení tohoto problému.

7.4 Autokorelační funkce

Korelace diskutované v předchozí části jsoujednoduchými příklady *autokorelace*, tedy korelace mezi jednou proměnnou a stejnou proměnnou zpožděnou o jedno či více období. Obvyklým nástrojem využívaným v praxi k pochopení vlastností časových řad je *autokorelační funkce*. Obecně nás může zajímat korelace mezi Y a Y zpožděným o p období. Například, data o osobním důchodu jsou pozorována každé čtvrtletí, tedy korelace mezi Y a Y zpožděným o $p = 4$ období je korelace mezi důchodem dnes a důchodem před jedním rokem (tzn. že rok mý čtyři čtvrtletí). Takovou korelaci budeme značit jako r_p, tedy

$$r_p = corr(Y_t, Y_{t-p}).$$

Autokorelační funkce chápe r_p jako funkci p, tedy r_p je počítáno pro $p = 1, \ldots, P$. Předpokládáme, že P je maximální uvažovaná délka zpoždění a typicky se voli jako relativně vysoká hodnota (např. $P = 12$ pro měsíční data).

Z praktického hlediska si všimněme, že r_p je korelace mezi proměnnou (řekněme, Y) a Y zpožděným o p období. Připomeňme si, že Y zpožděné o p období obsahuje $T - p$ pozorování. Při výpočtu r_p tak implicitně „vyhazujeme“ prvních p pozorování. POkud bychom předpokládali extrémně dlouhá zpoždění, dostali bychom se do problému výpočtu autokorelaci s velmi malým počtem pozorování. V krajním případě, kdybychom zvolili $p = T$, neměli bychom žádné pozorování k použití. To nás opravňuje k tomu, abychom nedávali p příliš vysoké. Všimněme si, že autokorelační funkce v sobě zahrnuje možnost použití různě zpožděných proměnných. Teoreticky můžeme použít data pro $t = 2, \ldots, T$ k výpočtu r_1, data pro $t = 3, \ldots, T$ k výpočtu r_2, atd., až konečně data pro $t = P + 1, \ldots, T$ k výpočtu r_p. To ale znamená, že každá z

autokorelací byla spočítána s různým počtem data, a nejsou tak tyto korelační koeficienty vzájemně porovnatelné. Z tohoto důvodu je standarní praxí výběr maximálního zpoždění, P, a použití dat pro $t = P + 1, \ldots, T$ pro výpočet všech autokorelací.

Tabulka 7.2: Autokorelační funkce.

Délka zpoždění, p	Osobní důchod, Y	Změna v osobním důchodu, ΔY
1	0.9997	-0.0100
2	0.9993	0.0121
3	0.9990	0.1341
4	0.9986	0.0082
5	0.9983	-0.1562
6	0.9980	0.0611
7	0.9978	-0.0350
8	0.9975	-0.0655
9	0.9974	0.0745
10	0.9972	0.1488
11	0.9969	0.0330
12	0.9966	0.0363

Tabulka 7.2 ukazuje autokorelační funkce pro Y = osobní důchod v USA a pro ΔY = změna v osobním důchodě, a to s využitím maximální délky zpoždění 12 (tj. $P = 12$). V této tabulce je nápadné to, že autokorelace pro osobní důchod jseou téměř rovny jedné, a to i pro případ vysokých délek zpoždění. Oproti tomu, autokorelace pro změny v osobním důchodu jsou velmi malé a vykazují více či méně náhodné chování. Jinými slovy, autokorelace jsou v tomto případě fakticky nulové. Toto chování je obvyklé u většiny makroekonomických časových řad: časová řada má autokorelace blízké jedničce, nicméně diference této řady mají autokorelace výrazně nižší, mnohdy téměř nulové.

O hodnotách těchto korelací můžeme uvažovat různým způsobem:

- Y je vysoce korelované v čase. Osobní důchod před třemi lety (tj. $p = 12$, pro čtvrtletní data) je vysoce korelován s důchodem dnes. Proměnná ΔY však tuto vlastnost nemá. Tempo růstu osobního důchodu v aktuálním čtvrtletí je v zásadě nekorelováno s růstem v minulém čtvrtletí.

- Pokud bychom znali minulé hodnoty osobního důchodu, dokázali bychom velmi dobře odhadnout, jak by mohl vypada osobní důchod dnes, tedy v aktuálním čtvrtletí. POkud bychom však znali jen tempa růstu v minulosti, pak tyto hodnoty by nám pramálo pomohli s predikcí změn v osobním důchodu v aktuálním čtvrtletí.

- Neformálně řečeno, proměnná Y „si pamatuje minulost“ (tzn. je vysoce korelovaná se svými minulými hodnotami). To je příklad chování s *dlouhou pamětí (long memory)*. Změna Y, ΔY tuto vlastnost nemá.

- Y je nestacionární řada, avšak ΔY stacionární řada je. Doposud jsme si formálně nedefinovali výrazy „nestacionarita“ a „stacionarita“, nicméně oba tyto pojmy mají důležitou roli v ekonometrii časových řad. Řeč bude o nich ještě později. Na tomto místě však mějme na paměti, že vlastnosti autokorelační funkce pro Y jsou charakteristické pro nestacionární řady.

7.5 Autoregresní model

Autokorelační funkce je užitečným nástrojem pro shrnutí informace o časové řadě.[15] Jak však již bylo v úvdoních kapitolách zmíněno, ukazatelé korelací mají svá omezení a v řadě případů je tak upřednosťnovaným nástrojem regrese. To platí i v případě analýzy časových řad. Autokorelace je korelace, a z tohoto důvodu je žádoucí vyvinout sofistikovanější modely pro analýzu závislostí mezi proměnnou a jejími zpožděnými hodnotami. Řada takových modelů byla rizvinuta v rámci statistické literatury věnované jednorozměrným časovým řadám, obvyklejší model, interpretovatelný jako regresní model, je však tzv. *autoregresní model*. Jak jméno napovídá, jedná se o regresní model, ve kterém jsou vysvětlující proměnné zpožděné hodnoty závisle proměnné (autoregrese je tak regrese proměnné sama na sebe, resp. své zpožděné hodnoty). Slovo „autoregresní“ je obvykle zkracováno na „AR“. Některé aspekty AR modelu byly diskutovány v kapitole 5. V této kapitole jsme předpokládali, že náhodné chyby regrese mají AR strukturu. V této kapitole budeme předpokládat AR strukturu samotné vysvětlované proměnné.

7.5.1 $AR(1)$ model

Naši analýzu autoregresních modelů začneme jednoduchým případem, kdy je vysvětlující proměnná závisle proměnná zpožděná o jedno období. Jedná se p případ $AR(1)$ modelu:

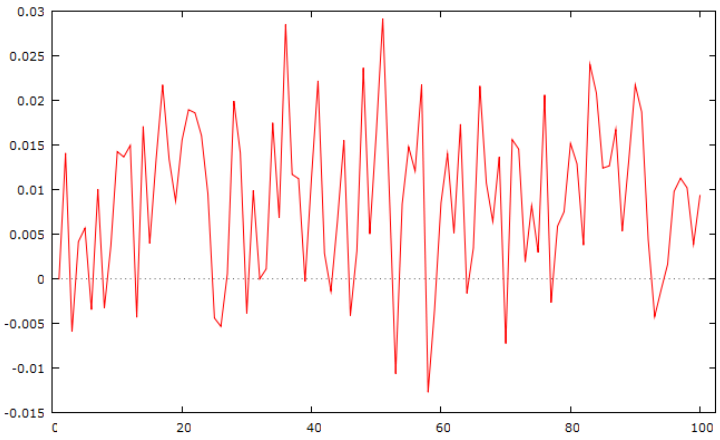
$$Y_t = \alpha + \rho Y_{t-1} + \epsilon_t$$

pro $t = 2, \ldots, T$. Tento model vypadá jako jednoduchý regresní model diskutovaný v předchozích kapitolách, s výjimkou toho, že vysvětlující proměnná je zpožděná závisle proměnná, Y_{t-1}. HOdnota ρ v $AR(1)$ modelu je úzce vztažena k chování autokorelační funkce a konceptu nestacionarity. Pokud jde o další terminologii, pokud je Y_t popsaáno AR modelem, říkáme, že se jedná o *autoregresní proces*.

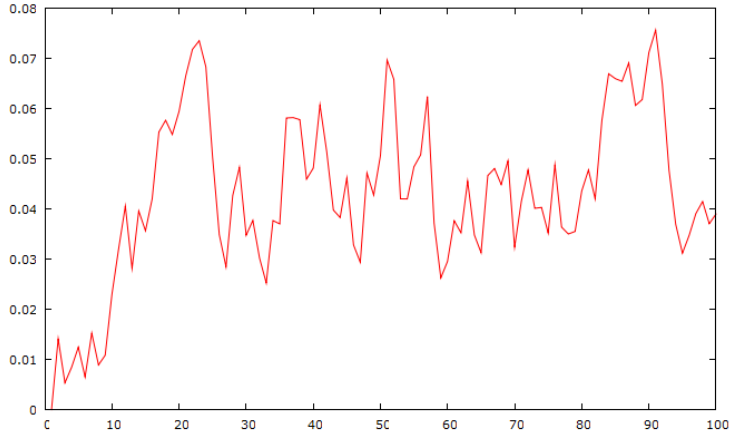
V kapitole 5 jsme v rámci diskuze nad autokorelovanými náhodnými složkami použili značení ρ jako koeficient u zpožděné náhodné složky v AR modelu chyb regrese. KOeficient ρ v této kapitole hraje podobnou roli. Podíváme-li se zpátky do kapitoly 5, části 5.4.1 týkající se „vlastnostem autokorelovaných chyb“, vidíme, že ρ determinuje vlastnosti náhodných složek. Analogickým odvození jsme schopni získat vlastnosti časové řady odpovdíající $AR(1)$ modelu. Tato odvození si zde nebudeme pro zřejmou podobnost uvádět. Podívejme se však na praktickou ilustraci vlastností $AR(1)$ modelu.

[15] Pokud hovožíme o proměnné Y, potom vyjádření „Y je proměnná v podobě časové řady“, „Y je časová řada“ nebo prostě „Y je řada“ budeme chápat jako ekvivalentní.

Abychom pochopili chování $AR(1)$ proceu, můžeme si uměle nasimulovat tři různé časové řady s různýou volbou pro ρ: $\rho = 0$, $\rho = 0.8$ a $\rho = 1$, Všechny tři řady mají stejnou hodnotu α (konkrétně $\alpha = 0.01$) a stejné náhodné složky. Obrázky 7.3, 7.4 a 7.5 ukazují grafy časové řady pro takto generovaná data.

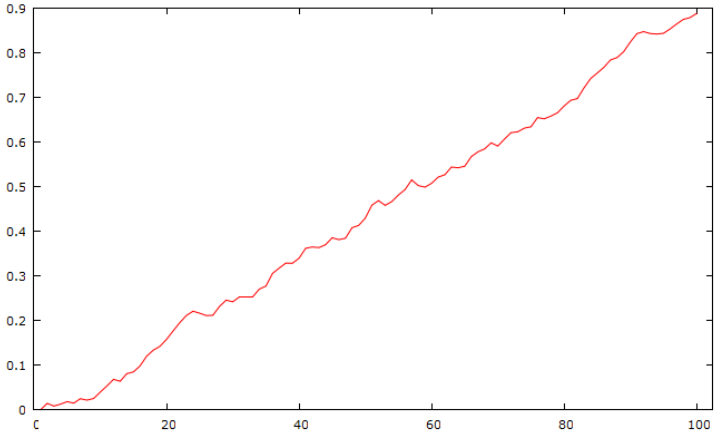


Obrázek 7.3: $AR(1)$ časová řada s $\rho = 0$.



Obrázek 7.4: $AR(1)$ časová řada s $\rho = 0.8$.

Všimněme si, že 7.3 (s $\rho = 0$) vykazuje náhodné fluktuace kolem průměru odpovídající hodnotě 0.01 (což je hodnota parametru α). Ve skutečnosti je tento obrázek velmi podobný obrázku 7.2, který obsahuje graf časové řady změny v osobním důchodu. Obrázek 7.5 (s $\rho = 1$) vykazuje trendové chování a je velmi podobný obrázku 7.1, tedy grafu vývoje osobního důchodu. Obrázek 7.4 ($\rho = 0.8$) vykazuje chování, odpovídající něčemu mezi náhodnými fluktuacemi obrázku 7.3 a silným trendovým chováním



Obrázek 7.5: $AR(1)$ časová řada s $\rho = 1$.

obrázku 7.5.

Obrázky 7.3 až 7.5 ilustrují typy chování, které je schopen $AR(1)$ model podchytit a vzhledem k podobnostem s vývojem prezentovaným na obrázcích 7.1 a 7.2 tak naznačují, proč je tento typ modelů obvykle využíván v makroekonomii a financích. Pro různé hodnoty ρ jsou tyto modely schopny vykazovat chování v podobě jak náhodných fluktuací, což je typické pro tempa růst řady makroekonomických a finančních časových řad, tak i trendové chování typické pro makroekonomické i finanční časové řady. Samozřejmostí je i chování ležící mezi těmito dvěma extrémy.

Hodnota $\rho = 1$ nám implikovala typ chování, které jsme nazývali jako nestacionární, Hodnoty jiné (menší než jedna) pak implikují stacionární chování. Tento fakt nám poskytuje formální definici konceptu stacionarity a nestacionarity, přinejmenším pro $AR(1)$ model: pro $AR(1)$ model můžeme říct, že Y je stacionární, pokud $|\rho| < 1$ a nestacionární v případě, kdy $\rho = 1$. Další možnost, $|\rho| > 1$, je málokdy v ekonomii uvažována. Tato možnost totiž implikuje explozivní chování, které je pozorovatelné jen v neobvyklých případech (např. hyperinflace), pro naše potřeby má tak menší empirickou relevanci a nebudeme se tímto případem zabývat.

Máme tedy zavedeny pojmy „nestacionární“ a „stacionární“ bez formální definice (kromě případu $AR(1)$ modelu). Jak uvidíme, rozdíl mezi stacionárními a nestacionárními časovými řadami je extrémně důležitý. Formální definici si ukážeme zanedlouho, nicméně, ještě před tím si uvedeme obecnou intuici stojící v pozadí.

Formálně „nestacionární“ znamená „něco co není stacionární“. V ekonomii se obvykle zaměřujeme na speciální typ nestacionarity, který bývá přítomen v řadě makroekonmických časových řad: nestacionarita jednotkového kořene. Tento koncept si zobecníme později, nicméně na tomto místě je užitečné chápat *jednotkový kořen (unit root)* jako skutečnost implikující $\rho = 1$ v $AR(1)$ modelu. O tom, jestli je časová řada, Y, stacionární nebo má jednotkový kořen, můžeme uvažovat různým způsobem:

- V $AR(1)$ modeleu, jestliže $\rho = 1$, potom Y má jednotkový kořen. Pokud $|\rho| < 1$, potome je Y stacionární.

- Jestliže má Y jednotkový křen, potom jsou její autokorelace blízko jedné a příliš neklesají s tím, jak se zvyšuje délka zpoždění.
- Pokud má Y jednotkový kořem. potom má dlouhou paměť. Stacionární časové řady nemají dlouhou paměť.
- Jestliže má Y jednotkový kořen, potom řada vykazuje trendové chování, zvláště, pokud je α nenulové.
- Pokud má Y jednotkový kořen, potom ΔY bude stacionární. Z tohoto důvodu jsou řady s jednotkovým kořenem označovány jako *diferenčně stacionární*, resp. *stacionární diferencích*.

Poslední zmíněný bod lze vidět nejlépe tak, že odečteme Y_{t-1} od obou stran rovnice v $AR(1)$ modelu, čímž dostaneme

$$Y_t - Y_{t-1} = \alpha + \rho Y_{t-1} - Y_{t-1} + \epsilon_t$$

nebo

$$\Delta Y_t = \alpha + \phi Y_{t-1} + \epsilon_t,$$

kde $\phi = \rho - 1$. Všimněme si, že pro $\rho = 1$ je $\phi = 0$ a předchozí rovnice tak implikuje, že ΔY_t náhodně fluktuuje kole α. Všimněme si rovněž (s ohledem na další diskuzi), že můžeme testovat $\phi = 0$ pro ověření toho, jestli má časová řada jednotkový kořen. Dále jsme si řekli, že časová řada bude stacionární, poku $-1 < \rho < 1$, což je ekvivalentní $-2 < \phi < 0$. Toto tvrzení budeme označovat jako *podmínku stacionarity*.

Pro zavedení dalších pojmů v souvislosti s $AR(1)$ modelem předpokládejme, že $\rho = 1$ (či ekvivalentně $\phi = 0$). V tomto případě můžeme $AR(1)$ model zapsat jako

$$Y_t = \alpha + Y_{t-1} + \epsilon_t.$$

Tento model je nazýván modelem *náhodné procházky (random walk)*. Přesněji řečeno, model náhodné procházky nemá úrovňovou konstantu (tj. $\alpha = 0$). Případ s úrovňovou konstantou je nazýván *modelem náhodné procházky s driftem (random walk wit drift)*. Existence úrovňové konstanty nám umožňuje nenulové průměrné tempu růstu časové řady. Protože $\rho = 1$, proměnná Y má jednotkový kořen a je nestacionární. Model náhodné procházky je obvykle využíván k modelování vývoje cen akcií. Cena akcie dnes je dána cenou akcie včera a (nepredikovatelné) náhodnou složkou. Model náhodné procházky s driftem je vhodnější pro makreokonomické veličiny jako HDP, které jsou v průměru rostoucí v čase. To nám opět napovídá, že nestacionarita exsituje v řadě časových řad z oblasti makreokonomie a financí.

Model $AR(1)$ je regresní model. Díky tomu můžeme použít OLS k regresi proměnné Y na úrovňovou konstantu a zpožděné hodnoty Y. Pokud tak učiníme pro případ dat osobního důchodu ve Spojených státech (použitých k vytvoření obrázku 7.1), získáme $\widehat{\alpha} = 0.039$ a $\widehat{\rho} = 0.996$. Protože OLS odhad, $\widehat{\rho}$, a skutečná hodnota $AR(1)$ koeficientu, ρ, budou zřídkakdy identické, je možné, že $\rho = 1$, protože OLS odhad je velmi blízko této hodnotě. Všimněme si, že pokud provedeme regresi ΔY_t na Y_{t-1}, získáme OLS odhad $\widehat{\phi} = -0.004$, což je $1 - \widehat{\rho}$, přesně tak jak bychom asi očekávali.

7.5.2 Rozšíření $AR(1)$ modelu

V předchozích odstavcích bylo argumentováno, že $AR(1)$ model lze interpretovat jako jednoduchý regresní model, kde Y je závisle proměnná a zpožděné Y je vysvětlující proměnná. Je ale rovněž možné, že bychom měli jako vysvětlující proměnné dodat i další zpoždění Y. To lze udělat rozšířením $AR(1)$ modelu do podoby autoregresního modelu řadu p, $AR(p)$ modelu:

$$Y_t = \alpha + \rho_1 Y_{t-1} + \ldots + \rho_p Y_{t-p} + \epsilon_t$$

pro $t = p+1, \ldots, T$. Nebudeme řešit problematiku vlastností tohoto modelu. Stačí nám vedět, že se jedná o podobný model jako je $AR(1)$ model, nicméně je obecnější povahy. Tento model dokáže generovat jake trendové chování typické pro makroekonmické časové řady a tak i náhodné chování typické pro tempa růstu. Samozřejmostí je i modelování chování ležícího mezi těmito dvěma „extrémními“ případy.

Stejně jako u $AR(1)$ modelu je v rámci diskuze nad jednotkovým kořenem této proměnné obvyklé odečtení Y_{t-1} z obou stran rovnice a drobná úprava $AR(p)$ modelu. Pokud to učiníme, vidímě, že $AR(p)$ model lze zapsat jako

$$\Delta Y_t = \alpha + \phi Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \ldots + \gamma_{p-1} \Delta Y_{t-p+1} + \epsilon_t,$$

kde koeficienty sklonu v této regresi, $\phi, \gamma_1, \ldots, \gamma_{p-1}$, jsou jednoduchými funkcemi $\rho_1, \ldots, \rho_p$ původního $AR(p)$ modelu. Například, $\phi = \rho_1 + \ldots + \rho_p - 1$. Je třeba zdůraznit, že se jedná stále o jeden a tentýž $AR(p)$ model, jen zapsaný v jiné podobě. Proto se na oba modely budeme odkazovat jako na $AR(p)$ modely. Pokud nás překvapuje, kam se poděl člen Y_{t-p} z první rovnice, potom vězme, že je uveden v rámci členu ΔY_{t-p+1} (tzn. $\Delta Y_{t-p+1} = Y_{t-p+1} - Y_{t-p}$). Obě dvě varianty modelu mají stejný počet koeficientů, $p+1$ (první varianta má $\alpha, \rho_1, \ldots, \rho_p$, druhá pak $\alpha, \phi, \gamma_1, \ldots, \gamma_{p-1}$).

Klíčové body hodné zřetel jsou následující: výše uvedená rovnice má stále podobu regresního modelu; $\phi = 0$ implikuje, že $AR(p)$ model časové řady Y obsahuje jednotkový kořen; pokud je $-2 < \phi < 0$, potom je řada stacionární. POdíváme-li se na předchozí rovnici s $\phi = 0$, vídíme, že důležitou souvislost týkající se řad s jednotkovým kořenem, který byl již dříve zmíněna. Pokud totiž časový řada obsahuje jednotkový kořen, potom regresní model zahrnující proměnnou ΔY a její zpožděné hodnoty je vhodným modelem (tzn. pro $\phi = 0$ člen Y_{t-1} vypadává a v regresi se objevují jen členy obsahující ΔY a její zpožděné varianty). Obvykle se tak říká, že „pokud je v řadě přítomen jednotkový kořen, lze řadu diferencováním převést na stacionární řadu“.

V regresních modelech nechceme zahrnout řady mající jednotkový kořen. Výjimkou je jen případ zvaný kointegrace, který bude diskutovn v následující kapitole. To je dost dobrou motivací pro analýzu toho, jestli má Y jednotkový kořen. V předchozích podkapitolách bylo zmiňováno, že časové řady s jednotkovým kořenem vykazují trendové chování. Znamená to ovšem to, že nám pro zjištění existence jednotkového kořene stačí podívat se na graf vývoje proměnné Y, jeslti má trendový vývoj? Odpověď je záporná. K vysvětlení toho, proč tomu tak je, se podívejme na jiný model.

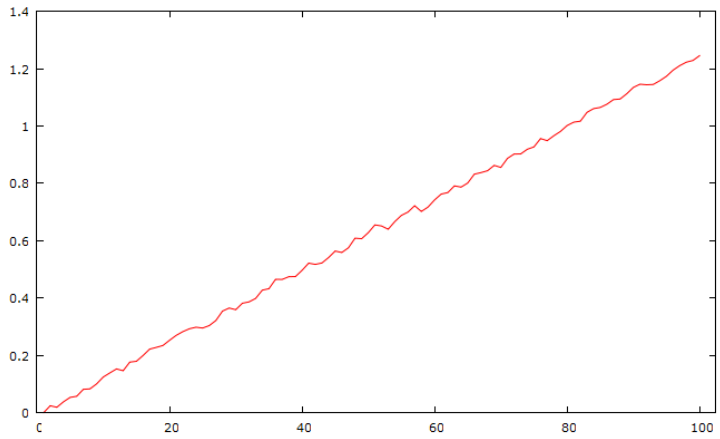
V předchozím textu jsme si ukázali, že mnohé časové řady obsahují trendy a AR modely s jednotkovými kořeny implikují trendové chování. Představme si, že obrázek 7.1 (nebo obrázek 7.5) je bodový XY graf, kde osa x ozančuje čas, a my bychom

rádi vytvořili regresní mode s využitím těchto dat. Chceme tedy odhadnout následující regresní přímku:

$$Y_t = \alpha + \delta t + \epsilon_t,$$

kde vysvětlující proměnná je čas (a pro označení koeficientu této vysvěltující proměnné jsme použili δ). Předchozí regresi můžeme interpretovat jako regresi zahrnující Y a další proměnnou s pozorováními $1, 2, 3, 4, \ldots, T$. Jedná se tak o další regresní model, který má trendové chování. Člen δt je označován jako *deterministický trend* protože se jedná o exaktní (deterministickou) funkci času. Oproti tomu řada s jednotkovým kořenem obsahuje tzv. *stochastický trend*.

Tento model deterministického trendu můžeme zkombinovat s $AR(1)$ modelem, čímž dostaneme $Y_t = \alpha + \rho Y_{t-1} + \delta t + \epsilon_t$. Abychom si ilustrovali některé vlastnosti tohoto modelu, podívejme se na obrázek 7.6, což je graf časové řady umělých dat generovaných předchozím modelem s $\alpha = 0$, $\rho = 0.2$ a $\delta = 0.01$. Poznamenejme, že tato řada je stacionární, protože $|\rho| < 1$. Obrázek 7.6 vypadá podobně jako obrázek 7.5 popř. obrázek 7.1. Stacionární modely s deterministickým trendem mohou vykazovat graficky podobný průběh, jako modely nestacionární, které obsahují stochastický trend. Důležité je zapamatovat si, že pohled na samotný graf časové řady není dostatečný k tvrzení o tom, jestli řada má nebo nemá jednotkový kořen.



Obrázek 7.6: Trendově stacionární $AR(1)$ časová řada.

Předchozí odstavce nám tedy motivovaly použití standardních pojmů, které budeme používat a které si uvedeme v následujícím přehledu:

- Nestacionární časové řady, na které se zaměřujeme, jsou ty, které obsahují jednotkový kořen. Tyto řady obsahují stochastický trend. POkud však provedeme diferenci těchto řad, bude výsledná časová řada stacionární. Z tohoto důvodu se nazývají *diferenčně stacionární (difference stationary)*.

- Stacionární časové řady, na které upínáme naši pozornost, mají tu vlastnost, že $-2 < \phi < 0$ v rámci $AR(p)$ modelu. Tyto řady však mohou vykazovat tren-

dové chování skrze zavedený deterministický trend. V takové případě hovoříme o *trendově stacionárních* řadách.

Pokud dodáme deterministický trend do $AR(p)$ modelu, získáme velmi obecný model, který je obvykle využíván v rámci analýzy jednorozměrných časových řad:

$$\Delta Y_t = \alpha + \phi Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \ldots + \gamma_{p-1} Y_{t-p+1} + \epsilon_t.$$

Výše uvedená rovnice se nazývá $AR(p)$ *model s deterministickým trendem*. Může být pro nás překvapivé to, proč nevyužijeme původní $AR(p)$ specifikaci zavedenou v úvodu této podkapitoly (tzn. specifikaci, kde jako vysvětlující proměnné vystupují $Y_{t-1}, \ldots, Y_{t-p}$), Existují pro to dva důvody. Prvním důvodem je to, že brzy přikročíme k testu jednotkového kořene. V rámci naší nové specifikace je velmi snadné testování $\phi = 0$. Testování toho, jestli jsou regresní koeficienty nulové, patří do problematiky, která nám je důvěrně známá (viz problém t-statistik v kapitole 2 a kapitole 3). V rámci původního modelu $AR(p)$ je testování jednotkového kořene komplikovanější. Druhý důvod je ten, že $Y_{t-1}, \ldots, Y_{t-p}$ jsou obvykle silně vzájemně korelovány (viz autokorelační funkce v tabulce 7.2). POkud bychom tyto vysvětlující proměnné použili v naši regresi, museli bychom obvykle čelit problému multikolinearity (viz kapitoly 2 a 4). V současném modelu však používáme jako vysvětlující proměnné $Y_{t-1}, \Delta Y_{t-1}, \ldots, \Delta Y_{t-p+1}$, které již nemají tendenci být vzájemně korelovány, a problém multikolinearity tak obvykle odpadá.

Naše diskuze nad jednorozměrnými časovými řadami se zatím soustředila zejména na modely (tj. autoregresní modely) a méně pak na metody (tj. estimátory nebo testové statistiky). K testování hypotéz se zanedlouho dostaneme. Pokud jde o techniku odhadu, existuje velké množství zůsobů pro odhad AR modelů a většina počítačových programů nám umožňuje volbu mezi různými odhadovými postupy, jakým je např. odhad metodou maximální věrohodnosti. Velmi často je však používán i OLS estimátor. Pravdou je, že s jeho použitím vznikají drobné statistické problémy, zejména v případě, kdy je model nestacionární případně se nestacionárnímu přibližuje (tj. ϕ je blízké nule). Konkrétně to znamená to, že pokud máme malou velikost vzorku, může být OLS estimátor vychýlen směrem dolů. Lze ukázat, že OLS estimátor odpovídá estimátoru metodou maximální věrohodnosti, pokud budeme chápat prvních p pozorování jako pevně daná (což jsme doposud činili), abychom se tak vyhnuli problému, že $Y_0, \ldots, Y_{1-p}$ nepozorujeme.

Příklad 7.1. *Osobní důchod ve Spojených státech*
Tabulka 7.3 obsahuje výsledky OLS regrese ΔY_t na úrovňovou konstantu, Y_{t-1}, ΔY_{t-1}, ΔY_{t-2}, ΔY_{t-3} a deterministický trend, a to s využitím dat o osobním důchodu ve Spojených státech. Jinými slovy tak máme výsledky regrese pro $AR(4)$ model s deterministickým trendem. Můžeme mít podezření, že tato časová řada může obsahovat jednotkový kořen, což je podezření, které nám tabulka jistmý způsobem potvrzuje. Konkrétně, jednotkový kořen je v řadě přítomen tehdy, pokud je parametr ϕ (koeficient u Y_{t-1}) nulový. Jak můžeme vidět, odhad ϕ je velmi malý (tj. $\widehat{\phi} = -0.018$). Je třeba upozornit, že tradiční t-test není možno pro případ tohoto koeficientu použít. Tomu bude ale věnovaná samostatná pozornost později.

Tabulka 7.3: $AR(4)$ model s deterministickým trendem.

Proměnná	OLS odhad	t-statistika	p-hodnota
Konstanta	0.138	1.279	0.203
Y_{t-1}	-0.018	-1.190	0.236
ΔY_{t-1}	-0.017	-0.217	0.829
ΔY_{t-2}	0.014	0.172	0.863
ΔY_{t-3}	0.130	1.627	0.106
t	0.0001	0.955	0.341

7.5.3 Testování $AR(p)$ modelu s deterministickým trendem

V kapitolách 2, 3 a 4 byly popsány postupy testování různých hypotéz v rámci regresního modelu. Konkrétně bylo ppsáno jak testovat hypotézu o nulovosti regresních koeficientů s využitím t-statistiky a jak testovat hypotézy zahrnující více lineárních restrikcí pomocí F-statistiky. Tyto techniky lze pužít i v $AR(p)$ modelu s deterministickým trendem (tzn. pokud chceme vynechat vysvětlující proměnné, jejichž koeficienty nejdou statisticky významně odlišné od nuly). Tento druh testování je užitečný v případě, kdy chceme zvoit vhodnou délku zpoždění, p, a v případě, kdy cheme testovat jednotkový kořen. Obvykle nejprve testujeme výběr vhodné délky zpoždění a po té testujeme přítomnost jednotkového kořene.

V případě $AR(p)$ modelu se však objevuje důležitá komplikace, se kterou jsme se doposud nesetkali. Abychom ji viděli, rozdělíme si parametry modelu do dovu skupin: (1) $\alpha, \gamma_1, \ldots, \gamma_{p-1}, \delta$ a (2) ϕ. Jinými slovy, budeme uvažovat testy hypotéz zahrnujících ϕ nezávisle od těch zahrnujících ostatní koeficienty.

Testy zahrnující $\alpha, \gamma_1, \ldots, \gamma_{p-1}$ a δ

Pro volbu odpovídající délky zpoždění v $AR(p)$ modelu existuje řada sofistikovaných statistických kritérií a testovacích metod. Níže budou diskutovány tzv. *informační kritéria*, která lze k k volbě řádu zpoždění využít. Nicméně, i prostý pohled na t-statistiku nebo F-statistiky může být dostatečně informativní. Pokud se například podíváme na tabulku 7.3, vidíme, že p-hodnoty u jednotlivých koeficientů členů zpožděné proměnné, ΔY, jsou nevýznamné a měly by být z regrese odstraněny (tj. jejich p-hodnoty jsou větší než 0.05). Alternativním způsobem je volba maximální délky zpoždění, $pmax$, a následný postupné vyhazování poslední délky zpoždění, pokud je u této proměnné nevýznamný parametr.

V $AR(p)$ modelu s deterministickým trendem můžeme řešit rovněž test toho, jestli je $\delta = 0$. To lze provést standardním způsobem, např. pohledem na p-hodnotu a její srovnání s hladinou významnosti (např. 0.05). Tento test lze provést kdykoli, ale obvyklé je jeho provední až po stanovení optimální délky zpoždění p.

Strategie testování výše uvedených hypotéz může být shrnuta do následujících bodů:

Krok 1. Volba rozumné maximální délky zpoždění, $pmax$.

Krok 2. Odhad (s využitím OLS) modelu $AR(pmax)$ s deterministickým trendem. Pokud je p-hodnota testu $\gamma_{pmax-1} = 0$ menší než zvolená hladina významnosti (např. 0.05), potom pokračujeme krokem 5, a to s využtím $pmax$ jako délky zpoždění ($AR(p)$ model s deterministickým trendem má dle našeho značení parametr u maximální délky zpoždění právě γ_{p-1}). Jinak přistupujeme k dalšímu kroku.

Krok 3. Odhad $AR(pmax - 1)$ modelu. Pokud je p-hodnota testu $\gamma_{pmax-2} = 0$ menší hladina významnosti, potom přejdeme k bodu 5 a využijeme hodnotu $pmax - 1$ jako optimální délku zpoždění. V opačném případě pokračujeme v následujícím kroku.

Krok 4. Opakovaný odhad AR modelů nižšího řádu, dokud nenalezneme $AR(p)$ model, pro který je γ_{p-1} statisticky významné (nebo nám „dojdou“ zpoždění).

Krok 5. Test zanedbání deterministického trendu. Testujeme tedy, jestli je p-hodnota testu $\delta = 0$ větší než zvolená hladina významnosti. V takovém případě vyhazujeme proměnou deterministického trendu.

Příklad 7.2. *Osobní důchod ve Spojených státech (pokračování)* Pokud aplikujeme strategii volby optimální délky zpoždění na data o osobním důchodu ve Spojených státech a začneme s $pmax = 4$, dojdeme k modelu

$$\Delta Y_t = \alpha + \phi Y_{t-1} + \epsilon_t.$$

Nejprve tedy odhadneme $AR(4)$ models deterministickým trendem (viz tabulka 7.3) a zjistíme, že je koeficient u ΔY_{t-3} statisticky nevýznamný. V souladu s tímto závěrem odhadneme $AR(3)$ model s deterministickým trendem, kdy opět zjišťujeme, že parametr u ΔY_{t-2} je nevýznamný. Opět tedy tuto proměnnou vypouštíme a zaměříme se na $AR(2)$, atd. Po té, co zjistíme i statistickou nevýznamnost deterministického trendu, zůstaneme u $AR(1)$ modelu. OLS odhady tohoto modelu jsou obsahem tabulky 7.4.

Tabulka 7.4: $AR(1)$ model.

Proměnná	OLS odhad	t-statistika	p-hodnota
Konstanta	0.039	2.682	0.008
Y_{t-1}	-0.004	-2.130	0.035

Výběr modelu versus průměrování modelů

Předchzí postup je dobrým příkladem *postupu sekvenčního testování (sequential testing procedure)*. To znamená, že je prováděna sekvence testů hypotéz s cílem nalezení

jednoho nejlepšího modelu. Většina ekonometrů využívá právě této strategii *výběru modelu (model selection)*. V našem příkladě jsme tak odhadli pět modelů: $AR(4)$ s deterministickým trendem, $AR(3)$ s deterministickým trendem, $AR(2)$ s deterministickým trendem, $AR(1)$ s deterministickým trendem, a $AR(1)$ model.

Postupy sekvenčního testování jsou mnohými kritizovány. Nebudeme se zabývat formálními důkazy, na kterých jsou tyto kritiky založeny, místo toho si jen intuitivně naznačíme východiska, na kterých jsou postaveny. Prvním z nich je to, že vždy, když je proveden test hypotézy, existuje možnost, že se dopouštíme chyby, spočívající v tom, že zamítneme „lepší mode“ oproti „ne tak dobrému“. Možnost existence této chyby se rychle znásobuje v rámci provádění celé sekvence testů hypotéz. Například tvrzení, že p-hodnota 0.05 regresního koeficientu (např. β_j) znamená, že $H_0 : \beta_j = 0$ je zamítána na 5% hladině významnosti může být potenciálně zavádějící, pokud je regrese vybrána na základě předchozích testů hypotéz. Druhým problémem, a to dokonce i v situaci, kdy nás sekvenční testování hypotéz dovedlo k výběru „nejlepšího modelu“, je to, že je zřídkakdy žádoucí prezentovat výsledky pouze tohoto modelu a ignorovat výsledky a závěry z „ne tak dobrých modelů“. Tyto obavy spojené s výběrem modelu se promítají do obvyklé empirické poučky, že pokud „težíme“ informaci z dostatečně dlouhého datového vzorku, vždy „něco“ najdeme, ale nesmíme těmto nálezům zcela bezvýhradně věřit.

Z těchto důvodů byla strategie výběru jednoho modelu kritizována zejména bayesiánskými ekonometry. Bayesiánsi (a někteří nebayesiánci) mají tendenci preferovat strategii *průměrování modelů (model averaging)*. Místo jediného modelu jsou empirické výsledky prezentovány na základě váženého průměrování všech využitých modelů. Samozřejmě ne všechny modely jsou rovnocenné. Některé modely popisují data lépe než jiné modely. Váhy jsou z tohoto důvodu počítány tak, že modelům lépe popisujícím data přiřazují vyšší váhu. Tento stručný popis nám rozhodně nestačí pro praktickou aplikace průměrování modelů. Nicméně pro intuitivní chápání problematiky je to dostačující. Vhodnou literaturou k této problematice je např. kapitolá 11 z Koop [16].

Využití informačních kritérií k výběru modelu

Alternativní přístup k výběru modelu je použití *informačních kritérií*. Intuitivně se jedná o měřítko toho, jak je model dobrý. Použití je snadné, stačí spočítat informační kritéria pro každý z modelů a vybrat model, který dává nejvyšší hodnotu. Většina ekonometrických programů tyto hodnoty spočítá automaticky, tudíž je můžeme v praxi lehce využít. Informační kritéria lze použít pro jakýkoliv typ modelů. Pokud tak máme například dva regresní modely se stejnou závisle proměnnou, ale různými vysvětlujícími proměnnými, lze použít informační kritérium pro rozhodnutí o tom, který z nich použít. Ve skutečnosti lze i nám známý korigovaný koeficient determinace, $\overline{R}^2$, diskutovaný v kapitole 4, interpretovat podobným způsobem jako informační kritérium. Informační kritéria jsou však nejběžněji používána v rámci modelů časových řad (např. pro volbu optimální délky zpoždění).

Formální derivace a motivace kteréhokoli z různých informačních kritérií využívaných ekonometry vyžaduje technické detaily jdoucí nad rámec obsahu tohoto textu. Zaměříme se tedy jen na obecnou intuici a zadefinujeme si několi nejpopulárnějších kritérií. V rámci různých regresních modelů jsme používali doposud i různá značení.

Například v jednoduchém regresním mdelu jsme používali β k ozačení regresního koeficientu, v $AR(1)$ modelu jsme jako koeficient zpožděné závisle proměnné používali ρ, a v dalších variantách $AR(1)$ modelu se nám objevoval koeficient ϕ. Informační kritéria lze využít ve všech těchto modelech, a proto použijeme obecné značení, θ, označující „všechny koeficienty v modelu“. Věrohodnostní funkce je tedy $L(\theta)$. Věrohodnostní funkci můžeme volně intepretovat jako měřítko toho, jak dobře model popisuje chování dat. Je tedy logické, že toto měřítko by mělo vstupovat i do informačního kritéria. Informační kritéria mají obvykle podobu

$$IC(\theta) = 2\ln[L(\theta)] - g(p)$$

kde p je počet koeficientů modelu a $g(p)$ je rostoucí funkce p. Tradiční použití informačního kritéria zahrnuje vyhodnocení $IC(\theta)$ v konkrétních hodnotách parametrů (např. OLS nebo ML odhadech vektoru parametrů θ), a to pro každý z modelů, který uvažujeme. Na tomto zákaldě pak zvolíme model s nejvyšší hodnotou informačního kritéria. Informační kritéria se liší v závislosti na použité funkci $g(p)$. Jedná se o funkci, která zohledňuje šetrnost pokud jde o počet parametrů. Modely s přílišným počtem koeficientů tak jsou výrazně penalizovány.

V naší diskuzi nad koeficientem determinace R^2 a $\overline{R}^2$ v kapitole 4 bylo řečeno, že R^2 by neměl být používán k rozhodování o upřednostňovaném modelu, protože s přidáním nevých vysvětlujících proměnných nikdy neklesne (a to i v případě nevýznamných proměnných). Přidáním nových vysvětlujících proměnných neexistuje způsob, jak by mohlo být vyrovnání dat horší. V rámci regrese by koeficienty u všech nových proměnných mohly být nulové, čímž bychom získali stejné vyrovnání dat, jako před jejich přidáním. Totéž se děje i s věrohodnostními funkcemi: přidáním nové vysvětlující proměnné se hodnota věrohodnostní funkce (vyhodnocená v maximálně věrohodném odhadu) vždy zvětší, a to i v případě nevýznamné proměnné. To je motivace pro přidání členu $g(p)$, v rámci informačního kritéria $IC(\theta)$, který nám zaručí, že dodání dodatečné vysvětlující proměnné je penalizováno. Informační kritérium tak například naznačí vhodnost přidání dalšího zpoždění v $AR(p)$ modelu, pokud zisk z přidání této proměnné (vyjádřený růstem věrohodnostní funkce) převýší náklady (vyjádřené skrze penalizační funkci $g(p)$).

Pravděpodobně nejběžnějším informačním kritérim je *Bayesovské informační kritérium (Bayesian BIC)*:

$$BIC(\theta) = 2\ln[L(\theta)] - p\ln(T).$$

Někdy je toto kritérium nazýváno *Schwarzovo kritérium*. Dvě další populární informační kritéria jsou *Akaikeho informační kritérium (AIC)*, dané jako

$$AIC(\theta) = 2\ln[L(\theta) - 2p]$$

a *Hannanovo-Quinnovo kritérium (HQ)*, dané výrazem

$$HQ(\theta) = 2\ln[L(\theta)] - pc_{HQ}\ln[\ln(N)],$$

kde c_{HQ} je konstanta, kdy se doporučují různé volby této konstanty. HQ je konzistentní kritérium volby modelu pro $c_{HQ} > 2$. Konzistentní kritérium volby modelu je takové, které s pravděpodobností jedna vybere korektní model pro délku vzorku jdoucí k nekonečnu.

Testy zahrnující ϕ: testy jednotkového kořene

K dokončení naši analýzy testování hypotéz v $AR(p)$ modelu s deterministickým trendem si musíme položit ještě jednu velmi důležitou otázku: má Y jednotkový kořen? Připomeňme si, že pokud je $\phi = 0$, potom Y obsahuje jednotkový kořen. V tomto případě musí být řada v regresním modelu diferencovány (tzn. řada je diferenčně stacionární). Na první pohled by se mohlo zdát, že bude stačit jednoduchý test $\phi = 0$, a to stejným způsobem, jako jsme testovali statistickou významnost ostatních koeficientů. Např. v tabulce 7.4 je t-statistika koeficientu ϕ rovna -2.13. POdíváme-li se do tabulek Studentova t-rozdělení (pro odpovídající velikost vzorku), vidímě, že 5% kritická hodnta je 1.97. Protože je absolutní hodnota testové statistiky větší než kritická hodnota, mohlo by nás to svádět k závěru, že ϕ je nenulové a že Y není řada s jednotkovým kořenem. Tento závěr je však nekorektní! V rámci testování hypotéz je parametr ϕ odlišné od ostatních koeficientů a musíme s ním tak pracovat odděleně.

Abychom plně pochopily důvod, proč nemůžem provádět test $\phi = 0$ stejným způsobem jako pro případ testování nulovosti koeficientů u ostatních proměnných, musely bychom mít statistické zákaldy převyšujicí doposavadní požadavky textu. Dostatečné pro nás může být to, že pokud zvolíme v rámc využívání ekonometrických programů OLS odhad, implicitně se předpokládá stacionarita všech proměnných v modelu při počítání p-hodnot. Pokud je vysvětlující proměnná Y_{t-1} nestacionární, její p-hodnota bude nekorektní.

Alternativní způsob vysvětlení vychází z toho, že, jak jsme si uvedli v kapitole 3, t-statistika má Studentovo t-rozdělení. Kritické hodnoty t-testu tak musíme brát z tabulek právě Studentova t-rozdělení. POkud testujeme to, jestli je $\phi = 0$, můžeme spočítat t-statistiku. Ovšem tato statistika již nemá Studentovo t-rozdělení. Důkaz by vyžadoval trochu více obtížnější matematiky, nicméně alespoň intuitivní vysvětlení lze ukázat v následujícím odstavci.

V kapitole 3 jsme pracovali s jednoduchým regresním modelem, kdy jsme si odvodili, že při splnění klasických předpokladů platí

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum X_i^2}\right)$$

a tento fakt jsme využili pro odvození rozdělení t-statistiky. V rámci autoregresního modelu však klasické předpoklady neplatí. Zejména nemůžeme předpokládat, že vysvětlující proměnné jsou pevně daná čísla (tj. nenáhodné veličiny). Vysvětlující proměnné jsou totiž zpožděné hodnoty závisle proměnné. Pokud je závisle proměnná náhodná, potom by i vysvětlující proměnné měly být náhodné. V modelu jednoduché regrese s náhodnými vysvětlujícími proměnnými lze předchozí výraz nahradit výrazem

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{TE(X^2)}\right).$$

V $AR(1)$ modelu (kde ϕ je koeficient u zpožděné závisle proměnné) bychom mohli uvažovat, že tento výraz bude mít podobu

$$\widehat{\phi} \sim N\left(\phi, \frac{\sigma^2}{TE(Y_{t-1}^2)}\right).$$

Pravdou však je, že $E(Y_{t-1}^2) = var(Y_{t-1}) + [E(Y_{t-1})]^2$ a připomeňme si, že Y (a tedy i zpožděné Y, což je Y_{t-1}) má jednotkový kořen v případě, že platí $\phi = 0$. Rozptyl proměnné s jednotkovým kořenem je však nekonečno. Postup odvození z kapitoly 3 tak nebude fungovat. Odvození skutečného rozdělení pro $\widehat{\phi}$ je poněkud komplikované a vyžaduje vcelku pokročilé nástroje. Tento odstavec měl však sloužit jen jako určitá intuice toho, proč nelze pracovat i v případě parametru ϕ se standardními odvozeními jako v případě modelu jednoduché regrese.

Korektní způsob testování jednotkového kořene byl vyvinut dvěma statistiky jménem Dickey a Fuller a je znám jako *Dickeyho-Fullerův test*. Jako testovou statistiku pro test $\phi = 0$ využívá obvyklou t-statistiku. Oba autoři však ukázali, že při platnosti nulové hypotézy není rozdělení této statistiky Studentovo t-rozdělení, ale jiné rozdělení zvané *Dickeyho-Fullerovo rozdělení*. Aby se věc zkomplikovala ještě více, rozdělení této testové statistiky se liší v závislosti na tom, jestli AR model obsahuje nebo neobsahuje deterministický trend. Nebudeme si ukazovat, proč tomu tak je, místo toho si v tabulce 7.5 ukážeme kritické hodnoty nutné pro praktické provádění Dickeyho-Fullerova testu.

Tabulka 7.5: Kritické hodnoty Dickeyho-Fullerova testu.

	$T = 25$	$T = 50$	$T = 100$	$t = \infty$
AR model bez deterministického trendu				
1% kritická hodnota	-3.75	-3.59	-3.50	-3.42
5% kritická hodnota	-2.99	-2.93	-2.90	-2.80
AR model s deterministickým trendem				
1% kritická hodnota	-4.38	-4.15	-4.04	-3.96
5% kritická hodnota	-3.60	-3.50	-3.45	-3.41

Pokud jde ještě o terminologii spojenou s tímto testem, někteří autoři používají termín „Dickeyho-Fullerův test“ pro test $\phi = 0$ v $AR(1)$ modelu, a termín „augmented (obohacený) Dickeyho-Fullerův test“ pro případ $AR(p)$ modelu. V tomto případě je základní test jednotkového kořene „obohacen (augmented)“ extra zpožděními. Protože se jedná v zásadě o stejný test, odpovídající stejným statistickým tabulkám, budeme v obou případech hovořit obecně o „Dickeyho-Fullerovu testu“.

Praktické provedení testu jednotkového kořene začíná tím, že odhadneme $AR(p)$ model s deterministickým trendem

$$\Delta Y_t = \alpha + \phi Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \ldots + \gamma_{p-1} \Delta Y_{t-p+1} + \epsilon_t$$

a využijeme postupy testování hypotéz popsané dříve nebo informační kritéria k výběru optimální délky zpoždění a rozodneme i o tom, jestli zahrnout deterministický trend. Po výběru preferované specifikace si zaznamenáme t-statistiku odpovídající koeficientu ϕ a porovnáme ji s příslušnou Dickeyho-Fullerovou kritickou hodnotou z tabulky 7.5.

Tabulka 7.5 poskytuje 1% a 5% kritické hodnoty Dickeyho-Fullerova testu. Nezapomeňme, že $\phi = 0$ znamená, že časová řada Y generovaná $AR(p)$ procesem obsahuje jednotkový kořen. Pokud platí $-2 < \phi < 0$, potom je řada stacionární. Stacionární

řada je konzistentní s $\widehat{\phi}$ a proto by příslušná t-statistika měla být záporná, a to více než kritická hodnota z tabulky 7.5. Pokud je $\widehat{\phi}$ kladná, napovídá to tomu, že Y vykazuje expanzivní chování (což je silně nestacionární případ). Pravdou je, že t-statistika pro ϕ ma rozdělení, které závisí na velikosti vzorku a na tom, jestli model obsahuje nebo neobsahuje deterministický trend. Tabulka 7.5 obsahuje kritické hodnoty pro celou řadu velikostí vzorku. Kritické hodnoty pro velikost vzorků mezi tabelovými hodnotami odpovídají hodnotám ležícím mezi příslušnými tabulovanými hodnotami. Pokud máme např. $T = 78$ a pracujeme s AR modelem bez deterministického trendu, můžeme použít tabulku 7.5 k tvrzení, že 5% kritická hodnota leží mezi -2.93 a -2.90 (to jsou hodnoty pro $T = 50$ a $T = 100$).Pro většinu aplikací se jedná o dostatečnou informaci pro praktickou implementaci Dickeyho-Fullerova testu (podrobnější tabulky lze nalézt např. skrze Google).

Dickeyho-Fullerův test je nejpopulárnější test jednotkového kořene, ale existují i jiné testy. Nebudeme si je zde popisovat, nicméně řada ekonometrických programů je standardně obsahuje.

Na závěr je třeba zmínit i několik varování spojených s testy jednotkového kořene. Dickeyho-Fullerův test vykazuje tu vlastnost, která je statistiky označována jako „malou sílu (low power)“. Jinými slovy, test může vést k chybnému nalezení jednotkového kořene, i přesto, že jednotkový kořen není v řadě přítomen. Intuitivně, trendově stacionární řady mohou vypadat jako řady s jednotkovým kořene (stačí srovnat obrázky 7.5 a 7.6) a může být mnohdy těžké je oddělit. Jiné typy modelů časových řad mohou vykazovat jednotkový kořen, i když jej ve skutečnosti nemají. Zřejmým příkladem je model časové řady charakterizovaný náhlými změnami či zlomy. Tyto strukturální zlomy lze najít v makroekonomických časových řadách a mohou být způsobeny událostmi jako jsou války nebo kizová období (např. ropné embargo zemí OPEC). Ceny akcií mohou vykazovat strukturální zlomy v důsledků krachů na trhu, a ceny komodit mohou bt rovněž charakterizovány strukturálními propady v důsledku sucha či přírodních katastrof. Strukturální zlomy jsou potenciální součástí řady typů časových řad a je tak třeba dávat pozor při interpretacu výsledků Dickeyho-Fullerova testu.

Příklad 7.3. *Osobní důchod v USA (pokračování)*
V předchozím příkladu byla využita data o osobním důchodu. Zvolili jsme $AR(1)$ model, který neobsahoval deterministický trend. Máme $T = 163$ (z původních $T = 164$, díky jednomu zpoždění v modelu). Tabulka 7.5 nám říká, že kritická hodnota na hladině významnosti 5 % je mezi -2.90 a -2.80. Z tabulky 7.4 vidíme, že t-statistika pro ϕ je -2.13, což není „zápornější“ než kritická hodnota. Na tomto základě nemůžeme zamítnou hypotézu, že osobní důchod obsahuje jednotkový kořen (na hladině významnosti 5 %.

7.6 Definice stacionarity

Koncept stacionarity a nestacionarity jsme si již motivovali. Týká se absence či přítomnost trendu. V kontextu konkrétní třídy modelů ($AR(p)$) jsme si definovali kon-

krétní typ nestacionarity: nestacionaritu jednotkového kořene. Doposud jsme si však neukázali formální definici těchto konceptů. Pro empirickou praxi to asi není úplně to nejpodstatnější. Samozřejme je pravda, že autoregresivní modely (a jejich rozšíření diskutované v další kapitole) dominují v empirických pracích a nestacionarita v podobě existence jednotkového kořene má pro empirickou praxi důležité implikace. Pro úplnost si však formální definici uvedeme:

Časová řada Y je stacionární, pokud:

1. $E(Y_t)$ je stejné pro všechna t.
2. $var(Y_t)$ je konečný a stejný pro všechny hodnoty t.
3. $cov(Y_t, Y_{t-1})$ závisí pouze na s, nikoli na t.

V tomto případě se jedná o definici tzv. slabé resp. kovarianční stacionarity. Definice silné stacionarity má dodatečnou implikaci v tom, že všechna Y_t mají stejné rozdělení.

Intuitivně tato definice pokrývá myšlenku, že základní statistické vlasntosti modelu (tzn. střední hodnota, rozptyl a kovariance) se v čase nemění. Jak jsme viděli, nestacionarita znamená „něco co není stacionární“.

Pro $AR(1)$ model

$$Y_t = \alpha + \rho Y_{t-1} + \epsilon_t,$$

kde náhodné složky splňují klasické předpoklady, lze ukázat, že stacionarita nastává v případě, pokud $|\rho| < 1$, a že žádná jiná hodnota ρ nevede k stacionaritě. Pokud platí, že $|\rho| < 1$, můžeme psát (odvození jsou podobná těm z kapitoly 5, části 5.4):

$$\begin{aligned} E(Y_t) &= \frac{\alpha}{1-\rho}, \\ var(Y_t) &= \frac{\sigma^2}{1-\rho^2}, \\ cov(Y_t, Y_{t-s}) &= \frac{\rho^s \sigma^2}{1-\rho^2}, \end{aligned}$$

kde $\sigma^2 = var(\epsilon_t)$. Střední hodnota a rozptyl jsou tak konstantní v čase a kovariance mezi proměnnými vzdálenými od sebe s období závisí jen na s (tzn. ne na t). Podmínky stacionarity jsou splněny pro $|\rho| < 1$. Tyto podmínky nejsou splněny pokud má řada jednotkový kořen, tedy $|\rho| = 1$.

Pro model s deterministickým trende jako je například

$$Y_t = \alpha + \delta t + \epsilon_t,$$

lze ukázat, že

$$\begin{aligned} E(Y_t) &= \alpha + \delta t, \\ var(Y_t) &= \sigma^2, \\ cov(Y_t, Y_{t-s}) &= 0. \end{aligned}$$

Protože $E(Y_t)$ závisí na t, je tento model nestacionární. Tento model však bude stacionární, pokud odstraníme deterministický trend. To nám napomáhá motivovat značení zavedené pro tyto modely (řady), které označujeme jako trendově stacionární.

7.7 Modelování volatility

7.7.1 Volatilita v cenách aktiv: úvod

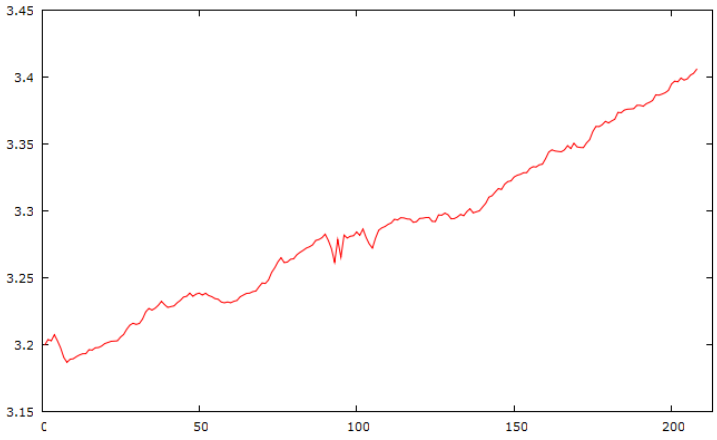
7.7.2 ARCH modely – autoregresní modely s podmíněnou heteroskedasticitou

Příklad 7.4. *Volatilita v cenách akcií*
(dokončení v příkladu 7.5)

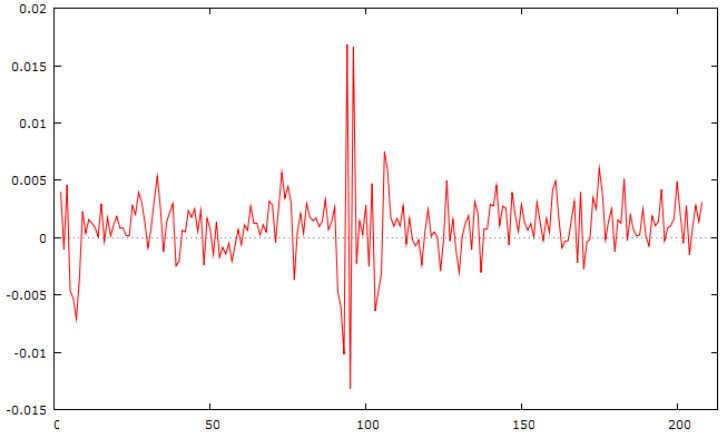
Příklad 7.5. *Volatilita v cenách akcií (dokončení z příkladu 7.4)*

Tabulka 7.6: $AR(1)$ model pro Δy_t^2.

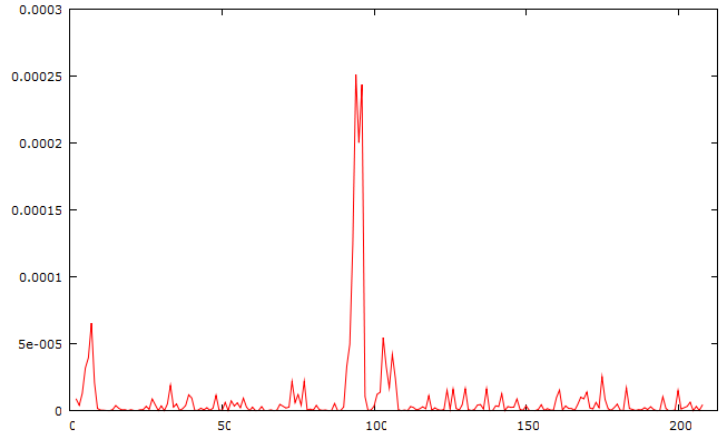
Proměnná	OLS odhad	t-statistika	p-hodnota
Konstanta	0.024	1.624	0.106
Δy_{t-1}^2	0.737	15.552	0.000



Obrázek 7.7: Graf časové řady logaritmu ceny akcie.



Obrázek 7.8: Graf časové řady změny logaritmu ceny akcie.



Obrázek 7.9: Graf časové řady volatility akcie.

Příklad 7.6. *Volatilita v cenách akcií (pokračování)*

Tabulka 7.7: ARCH(1) model pro data výnosnosti akcií.

Proměnná	Odhad koeficientu	p-hodnota	95% int. spolehlivosti
Regresení rovnice s vysvětlovanou proměnnou ΔY			
Konstanta	0.105	0.000	[0.081;0.129]
ARCH rovnice			
Konstanta	0.024	0.000	[0.016;0.032]
$\Delta\varepsilon_{t-1}^2$	0.660	0.000	[0.302;1.018]

Tabulka 7.8: ARCH(2) model pro data výnosnosti akcií.

Proměnná	Odhad koeficientu	p-hodnota	95% int. spolehlivosti
Regresení rovnice s vysvětlovanou proměnnou ΔY			
Konstanta	0.109	0.000	[0.087;0.131]
ARCH rovnice			
Konstanta	0.025	0.000	[0.016;0.033]

Příklad 7.7. *Volatilita v cenách akcií (pokračování)*

Tabulka 7.9: GARCH(1,1) model pro data výnosnosti akcií.

Proměnná	Odhad koeficientu	p-hodnota	95% int. spolehlivosti
Regresení rovnice s vysvětlovanou proměnnou ΔY			
Konstanta	0.109	0.000	[0.087;0.131]
ARCH rovnice			
Konstanta	0.026	0.000	[0.015;0.038]
$\Delta\varepsilon_{t-1}^2$	0.714	0.000	[0.327;1.101]
σ_{t-1}^2	-0.063	0.457	[-0.231;0.104]

7.8 Shrnutí

Na základě této kapitoly tedy již víme:

✌

Měli bychom tak již znát a umět vysvětlit obsah následujících klíčových pojmů:

- ❀ Ekonomický model
- ❀ Ekonometrický model
- ❀ Průřezová data
- ❀ Časové řady
- ❀ Panelová data
- ❀ Střední hodnota
- ❀ Rozptyl
- ❀ Kovariance
- ❀ Korelace a korelační koeficient
- ❀ Korelační matice

Příloha: Modely MA a ARMA

Kapitola 8

Regrese s časovými řadami

V této kapitole se dozvíme:

☞

8.1 Úvod

8.2 Regrese stacionárních časových řad

Příklad 8.1. *Efekt nákupů výpočetní techniky na tržby*

Tabulka 8.1: ADL(2,2) model s deterministickým trendem.

Proměnná	OLS odhad	t-statistika	p-hodnota
Konstanta	-0.028	-0.685	0.495
Y_{t-1}	-0.120	-9.460	0.000
ΔY_{t-1}	0.794	25.628	0.000
X_t	0.125	2.605	0.011
ΔX_t	0.838	19.111	0.000
ΔX_{t-1}	0.002	0.103	0.918
t	0.0001	0.984	0.328

8.3 Regrese časových řad s jednotkovým kořenem

8.3.1 Zdánlivá regrese

8.3.2 Kointegrace

8.3.3 Odhad a testování s kointegrovanými proměnnými

Tabulka 8.2: Kritické hodnoty Engle-Grangerova testu.

	$T = 25$	$T = 50$	$T = 100$	$t = \infty$
1% kritická hodnota	-4.37	-4.12	-4.01	-3.90
5% kritická hodnota	-3.59	-3.46	-3.39	-3.33

8.3.4 Regrese kointegrovaných časových řad – model korekce chyb

Příklad 8.2. *Kointegrace mezi cenami dvou zboží (pokračování)*

Tabulka 8.3: Dvoukrokový odhad jednoduchého modelu korekce chyb.

Proměnná	OLS odhad	t-statistika	p-hodnota
Konstanta	-0.023	-0.068	0.946
$\hat{\epsilon}_{t-1}$	-1.085	-14.458	0.000
ΔX_t	1.044	5.737	0.000

8.4 Regrese nekointegrovaných časových řad s jednotkovým kořenem

8.5 Grangerova kauzalita

8.5.1 Grangerova kauzalita v ADL modelu

Příklad 8.3. *Grangerovská kauzalita mzdové inflace na inflaci cenou?*

Tabulka 8.4: ADL s cenovou inflací jako vysvětlovanou proměnnou.

Proměnná	OLS odhad	t-statistika	p-hodnota
Konstanta	-0.751	-1.058	0.292
ΔP_{t-1}	0.822	4.850	0.000
ΔP_{t-2}	-0.041	-0.222	0.825
ΔP_{t-3}	0.142	0.762	0.448
ΔP_{t-4}	-0.181	-1.035	0.303
ΔW_{t-1}	-0.016	-0.114	0.909
ΔW_{t-2}	-0.118	-0.823	0.412
ΔW_{t-3}	-0.042	-0.292	0.771
ΔW_{t-4}	0.038	0.266	0.791
t	0.030	2.669	0.009

Příklad 8.4. *Grangerovská kauzalita cenové inflace na inflaci mzdovou?*

Tabulka 8.5: ADL se mzdovou inflací jako vysvětlovanou proměnnou.

Proměnná	OLS odhad	t-statistika	p-hodnota
Konstanta	-0.609	-0.730	0.467
ΔW_{t-1}	0.053	0.312	0.755
ΔW_{t-2}	-0.040	-0.235	0.814
ΔW_{t-3}	-0.058	-0.348	0.728
ΔW_{t-4}	0.036	0.215	0.830
ΔP_{t-1}	0.854	4.280	0.000
ΔP_{t-2}	-0.217	-0.993	0.323
ΔP_{t-3}	0.234	1.067	0.288
ΔP_{t-4}	-0.272	-1.323	0.188
t	0.046	3.514	0.001

8.5.2 Grangerova kauzalita s kointegrovanými proměnnými

8.6 Vektorová autoregrese

Příklad 8.5. *VAR(1) s proměnnými RMPY*

Tabulka 8.6: RMPY VAR(1) se závisle proměnnými ΔR, ΔM, ΔP, a ΔY.

	Závisle proměnná							
Nezávisle	ΔR		ΔM		ΔP		ΔY	
proměnná	Koef.	p-hodn.	Koef.	p-hodn.	Koef.	p-hodn.	Koef.	p-hodn.
Konstanta	-3.631	0.162	0.335	0.001	0.161	0.138	0.495	0.005
ΔR_{t-1}	0.222	0.003	-0.013	0.000	0.010	0.002	0.000	0.940
ΔM_{t-1}	3.391	0.007	0.749	0.000	0.121	0.021	0.283	0.001
ΔP_{t-1}	1.779	0.228	0.061	0.303	0.519	0.000	-0.117	0.242
ΔY_{t-1}	3.224	0.004	-0.032	0.480	-0.039	0.407	0.309	0.000
t	-0.056	0.011	0.000	0.695	0.002	0.048	-0.003	0.035

8.6.1 Prognózování s VAR modely

Tabulka 8.7: RMPY VAR(2) se závisle proměnnými ΔR, ΔM, ΔP, a ΔY.

	Závisle proměnná							
Nezávisle	ΔR		ΔM		ΔP		ΔY	
proměnná	Koef.	p-hodn.	Koef.	p-hodn.	Koef.	p-hodn.	Koef.	p-hodn.
Konstanta	-4.000	0.103	0.261	0.017	0.113	0.311	0.513	0.006
ΔR_{t-1}	0.315	0.000	-0.017	0.000	0.009	0.004	0.002	0.670
ΔM_{t-1}	2.824	0.106	0.655	0.000	0.086	0.280	0.310	0.019
ΔP_{t-1}	3.049	0.061	-0.020	0.785	0.366	0.000	0.074	0.545
ΔY_{t-1}	3.696	0.000	-0.051	0.270	-0.010	0.835	0.270	0.001
ΔR_{t-2}	-0.346	0.000	0.003	0.298	-0.001	0.795	-0.010	0.085
ΔM_{t-2}	-2.201	0.213	0.157	0.045	0.025	0.755	-0.094	0.480
ΔP_{t-2}	1.164	0.457	0.095	0.170	0.282	0.000	-0.233	0.049
ΔY_{t-2}	1.085	0.303	0.036	0.445	-0.046	0.334	0.153	0.054
t	-0.045	0.029	0.000	0.798	0.001	0.209	-0.003	0.104

Tabulka 8.8: Prognóza inflace a růstu HDP.

	ΔP		ΔY	
	Prognóza	Skutečnost	Prognóza	Skutečnost
$1992Q1$	0.626	0.929	-0.019	0.865
$1992Q2$	0.731	0.689	0.220	0.698
$1992Q3$	0.862	0.289	0.275	0.838
$1992Q4$	0.940	0.813	0.271	1.393

8.6.2 Vektorová autoregrese s kointegrovanými proměnnými

Příklad 8.6. *Spotřeba, agregátní bohatství a očekávané výnosy akcií*

Tabulka 8.9: Johansenův test kointegrace – CAY data.

Řád	Trace statistika	5% kritická hodnota
0	37.27	29.68
1	6.93	15.41
2	0.95	3.79

8.6.3 Využití VAR modelů – impulzní odezvy a varianční dekompozice

8.7 Shrnutí

Na základě této kapitoly tedy již víme:

✌

Měli bychom tak již znát a umět vysvětlit obsah následujících klíčových pojmů:

- ❀ Ekonomický model
- ❀ Ekonometrický model
- ❀ Průřezová data
- ❀ Časové řady
- ❀ Panelová data
- ❀ Střední hodnota
- ❀ Rozptyl
- ❀ Kovariance
- ❀ Korelace a korelační koeficient
- ❀ Korelační matice

Příloha: Teorie prognózování

Kapitola 9

Modely panelových dat

V této kapitole se dozvíme:

☞

9.1 Úvod

9.2 Souhrnný model

9.3 Modely individuálních vlivů

9.3.1 Model fixních vlivů

9.3.2 Model náhodných vlivů

9.3.3 Rozšíření modelů individuálních vlivů

9.4 Shrnutí

Na základě této kapitoly tedy již víme:

✌

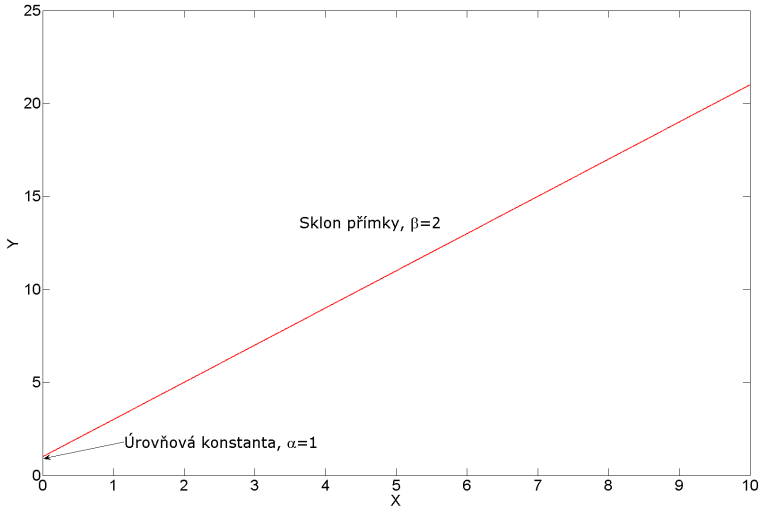
Měli bychom tak již znát a umět vysvětlit obsah následujících klíčových pojmů:

- ❀ Ekonomický model
- ❀ Ekonometrický model
- ❀ Průřezová data
- ❀ Časové řady
- ❀ Panelová data
- ❀ Střední hodnota
- ❀ Rozptyl
- ❀ Kovariance
- ❀ Korelace a korelační koeficient
- ❀ Korelační matice

Příloha A

Základy matematiky

Funkce a rovnice přímky



Obrázek A.1: Přímka.

Logaritmy

Sumační a multiplikační operátor

Příloha B

Základy pravděpodobnosti a statistiky

B.1 Základy pravděpodobnosti

Definice B.1.1 (Experiment a náhodné jevy)**.** Experiment je proces, jehož výsledek není předem znám. Možné výsledky experimentu se nazývají náhodné jevy. Množina všech možných výsledků je výběrový prostor.

Definice B.1.2 (Diskrétní a spojité proměnné)**.** Proměnnou (či veličinu) nazýváme diskrétní, jestliže existuje konečný či spočítatelný počet hodnot, jichž může nabývat. Proměnná je spojitá, pokud může nabývat jakékoliv hodnoty na přímce reálných hodnot nebo na určitém intervalu reálných hodnot.

Definice B.1.3 (Náhodné proměnné a pravděpodobnost)**.** Obvykle bývají otázky vztahující se k pravděpodobnosti, experimentu a náhodným jevům reprezentovány proměnnou (veličinou), ať již diskrétní nebo spojitou. Jelikož není výsledek experimentu předem znám, je tato proměnná nazývána náhodnou veličinou. Pravděpodobnost lze intuitivně chápat jako reflexi věrohodnosti toho, že každý z náhodných jevů nastane. Pravděpodobnost realizace jevu A je označována $Pr(A)$. Je důležité rozlišovat mezi náhodnou veličinou X označovanou velkými písmeny a její realizací x obvykle značenou písmenem malým. Jako příklad lze uvažovat experiment házení kostkou. Výběrový prostor je v tomto případě $[1, 2, 3, 4, 5, 6]$ a diskrétní náhodná veličina X nabývá hodnot $1, 2, 3, 4, 5, 6$ s pravděpodobnostmi danými jako $Pr(X = 1) = Pr(X = 2) = \ldots = Pr(X = 6) = \frac{1}{6}$. Alternativně je náhodná veličina X funkcí definovanou v bodech $1, 2, 3, 4, 5, 6$. Funkce je implicitně definována pravděpodobnostmi $Pr(X = 1) = Pr(X = 2) = \ldots = Pr(X = 6) = \frac{1}{6}$.

Definice B.1.4 (Nezávislost)**.** Dva jevy A a B jsou nezávislé, jestliže platí $Pr(A, B) = Pr(A)Pr(B)$, kde $Pr(A, B)$ je pravděpodobnost současné realizace jevů A a B.

Definice B.1.5 (Podmíněná pravděpodobnost)**.** Podmíněná pravděpodobnost jevu A jevem B, označována $Pr(A|B)$, je pravděpodobnost realizace jevu A za podmínky realizace B.

Teorém B.1.1 (Pravidla podmíněné pravděpodobnosti)**.** *Nechť A a B označují dva jevy, potom platí:*

$$Pr(A|B) = \frac{Pr(A,B)}{Pr(B)},$$

$$Pr(B|A) = \frac{Pr(A,B)}{Pr(A)},$$

a kombinací pak získáme Bayesův teorém

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}.$$

Definice B.1.6 (Pravděpodobnost a distribuční funkce)**.** Nechť diskrétní náhodná veličina X je definovaná na výběrovém prostoru $x_1, \ldots, x_n$. Pravděpodobnostní funkce této veličiny se označuje $p(x)$, přičemž platí:

$$p(x) = \begin{cases} Pr(X = x_i) & \text{pro } x = x_i, \\ 0 & \text{jinak,} \end{cases}$$

pro $i = 1, 2, \ldots, N$. Distribuční funkce (DF) označována $P(x)$ je definována jako:

$$P(x) = Pr(X \leq x) = \sum_{j \in J} Pr(x_j),$$

kde J je množina indexů j pro které platí $x_j \leq x$.

Pravděpodobnostní a distribuční funkce mají následující vlastnosti:

$$p(x_i) > 0 \quad i = 1, 2, \ldots, N$$

$$\sum_{i=1}^{N} p(x_i) = P(x_N) = 1$$

Definice B.1.7 (Hustota pravděpodobnosti a DF)**.** DF příslušná spojité náhodné veličině X je $P(x) = Pr(X \leq x) = \int_{-\infty}^{x} p(t)\mathrm{d}t$, kde $p(\cdot)$ je hustota pravděpodobnosti (probability density function - p.d.f.). Pro tyto funkce platí:

$$p(x) \geq 0 \quad \forall x,$$

$$\int_{-\infty}^{\infty} p(t)\mathrm{d}t = P(\infty) = 1,$$

$$p(x) = \frac{\mathrm{d}P(x)}{\mathrm{d}x},$$

$$Pr(a \leq x \leq b) = P(b) - P(a) = \int_a^b p(x)\mathrm{d}x.$$

Definice B.1.8 (Očekávaná hodnota)**.** Nechť $g(\cdot)$ je funkce, pak očekávaná hodnota $g(X)$, označována jako $E[g(X)]$ je definována jako:

$$E[g(X)] = \sum_{i=1}^{N} g(x_i)p(x_i),$$

pokud X je diskrétní náhodná veličina na výběrovém prostoru $x_1, \ldots, x_N$, a

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)p(x)\mathrm{d}x,$$

pokud X je spojitá náhodná veličina (platí-li $E[g(X)] < \infty$). Speciálními případy této obecné definice zahrnují:

- střední hodnota $\mu \equiv E(X)$,
- rozptyl $\sigma^2 \equiv var(X) = E[X - \mu]^2 = E(X^2) - \mu^2$,
- r-tý moment $E(X^r)$,
- r-tý centrovaný moment $E(X - \mu)^r$.

Třetí a čtvrtý centrovaný moment jsou obvyklé míry šikmosti a špičatosti náhodné veličiny, což reflektuje tloušťku krajů (chvostů) funkce hustoty pravděpodobnosti.

Teorém B.1.2 (Vlastnosti očekávané hodnoty)**.** *Nechť jsou dány náhodné veličiny X a Y, funkce $g(\cdot)$ a $h(\cdot)$ a konstanty a a b, potom platí:*

- $E[ag(X) + bh(Y)] = aE[g(X)] + bE[h(Y)]$,
- $var[ag(X) + bh(Y)] = a^2 var[g(X)] + b^2 var[h(Y)]$, *pokud X a Y jsou nezávislé.*

B.2 Základy asymptotické teorie

Konvergence v pravděpodobnosti

Základní zákon velkých čísel

Další zákon velkých čísel

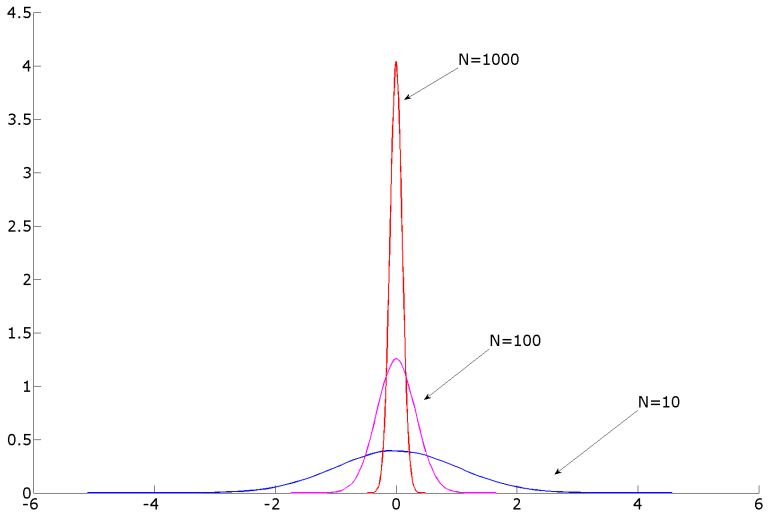
Konvergence v distribuci

Základní centrální limitní věta

Další užitečné teorémy

Slutského teorém

Cramérovy teorémy



Obrázek B.1: Rozdělení výběrového průměru pro různé velikosti vzorku.

Příloha C

Jak zpracovat empirický projekt?

V této části se zaměříme na užitečné poznatky týkající se práce na projektu, který vychází z vlastního empirického výzkumu či analýzy. Poznatky v této příloze jsou uplatnitelné nejen pro semestrální projekty ve škole, ale určitě jsou využitelné i pro práci na větších projektech či výzkumných úkolech, jejichž výsledkem může být článek či série článku v renomovaných odborných časopisech nebo nějaký druh závěrečné zprávy pro potřeby manažerského rozhodování v rámci nějakého projektu na podnikové úrovni. Budeme-li tedy hovořit o odborné práci, může tím být myšlen článek, working paper, výzkumnou zprávu nebo jakýkoli další dokument typu odborné studie, pojednávající o výsledcích (a postupu) nějaké vědecké práce s empirickým obsahem o nějž se chceme podělit.

Předchozí kapitoly se zabývaly (mimo jiné) problematikou formulace ekonometrického modelu, jeho odhadem adekvátními technikami, interpretací výsledků, testováním ekonomických hypotéz a dalšími aspekty, které jsou s tím vším spojeny. Specifikace modelu, výběr odhadové metody a získání vhodných dat je součástí jakéhokoliv ekonometrického výzkumného projektu. Zaměříme se tedy na problematiku výběru vhodného tématu výzkumného projektu, na základní součástí výzkumné zprávy a v příloze D pak nakousneme téma zdrojů ekonomických dat.

Ekonomický výzkum může být velkým dobrodružstvím. Nikdy dopředu nevíme, jaké výsledky nám přinese ekonometrická analýza dat. Výsledky nám mohou potvrdit naše očekávání, stejně tak nám mohou ale přinést překvapivé výsledky, které standardní teorie nepředpokládá a může se nám tak naskytnout jedinečná příležitost pokusit se vysvětlit analyzovaný problém zcela nově a neotřele. Ekonomický výzkum je tak rovněž zdrojem zábavy, neboť koho by nepotěšilo, objevit vlastními silami něco nového a zajímavého.

Výzkumný projekt znamená příležitost prozkoumat nějaké důležité ttéma či problém, který nás zajímá. Před tím, než se však člověk pustí to vlastního výzkumu a psaní příslušné zprávy či článku, není od věci věnovat trochu času dobrému promyšlení výběru samotného tématu a problému. Pokud nás tedy nějaká myšlenka napadne,

je dobré napsat si abstrakt projektu, ve kterém si obvykle člověk utřídí poznatky o tom, co o problematice ví a v co doufá, že mu daný projekt přinese.

C.1 Výběr tématu

Výběr dobrého tématu je zásadní pro úspěšné zvládnutí projektu. V úvodu je třeba položit si otázku: „Co mě zajímá?“. Zájem o konkrétní téma, které hodláme řešit, je hodně důležitý, protože dělat něco, co nás nebaví nebo nezajímá, určitě k dobré výkonnosti nepřispívá. Začneme-li tedy pracovat na nějaké zajímavé otázce či problému, obvykle se postupně začnou vynořovat další otázky a problémy, které s ní souvisí. Tyto mohou přinést jiný pohled na původní téma, nebo pro nás mohou znamenat nové směry, které se můžeme vydat prozkoumávat, a které se nám mohou ukázat jako ještě zajímavější.

Již v rámci několika semestrů studia člověk obvykle získá nějakou představu o tom, jaké oblasti ekonomie jej zajímají více a jaké méně. Pro každého z nás mají specializované oblasti ekonomie různou přitažlovost, ať již jde o oblasti monetární ekonomie, ekonomického růstu, marketingu, veřejných financí, finančních trhů, ekonomie práce, enviromentální ekonomie, mezinárodního obchodu apod. Pokud nás některá z takových oblastí zajímá, ovšem nemáme specifickou představu o tom, kde začít při výběru konkrétního tématu, není od věcí probrat to s osobou, která se v dané oblasti výzkumu pohybuje. On či ona budou určitě schopni poskytnout nějakou inspiraci, které nás posune dále a určitě doporučí vhodnou literaturu (monografie, články, working papers), kterou nebude od věci si pročíst. Mohou nám stejně tak doporučit vhodné odborné časopisy, který publikuje články aplikovaného výzkumu v dané oblasti. Pokud máme vybránu oblast problémů, které nás zajímají, můžeme si prostřednictvím některé z databází ekonomické literatury (např. EBSCOhost, EconLit, ProQuest, JSTOR) projít seznam článků k danému tématu. Každý článek bývá doprovázen tzv. JEL klasifikací (podle *Journal od Economic Literature*), což je klasifikační schéma, podle něhož jsou vcelku přehledným způsobem od sebe „odděleny“ články různých oblastí ekonomie.

Jakmile máme alespoň rámcově nalezený problém, na kterém chceme pracovat, je třeba vyřešit otázku dat. Při zpracování semestrálního projektu (v rámci jednosemestrálního kurzu) asi ne každý má dostatek času pro sběr vlastních dat, které by v projektu využil. V takovémto případě je potřeba hledat vhodná a dostupná data vztahující se k řešenému problému. I v tomto případě je možné využít rady a pomoci zkušenějších matadorů z řad akademických pracovníků. V praxi je obvyklé, že při zpracování problémových otázek získáváme data prostřednictvím vlastních průzkumů, což může zabrat velkou část času vymezeného (byť rámcově) na jeho zpracování.

Tímto jsme tedy popsali dva aspekty kvalitního námětu výzkumu: problém by měl být pro nás zajímavý a relevantní data by měla být dobře dostupná. Třetí aspekt dobrého projektu je opět pragmatického rázu: měli bychom být schopni úkol dokončit ve vymezeném čase, což v případě semestrálního projektu odpovídá zbytku semestru. Tento časový aspekt je úzce svázán nejen s dostupností dat. Vyžaduje totiž, abychom byli dostatečně obeznámeni s vhodnými ekonometrickými postupy pro analýzu dat a abychom je byli schopni výpočetně implementovat s využitím vhodného software, případně abychom si byli schopni v rozumném časovém horizontu osvojit nutné do-

vednosti jeho využití. Tato pragmatická pravidla platí analogicky i pro další typy výzkumných úkolů, liší se jen s ohledem na dostupný čas, případně s ohledem na další nároky pro jejich zpracování. Některé úkoly s sebou přinášejí nutnost naučit se využívat specifické softwarové nástroje, vyvinuté pro řešení právě typu úkolu, který nás zajímá. Většinou se vyplatí věnovat svůj čas jejich studiu (obvykle bývají uživatelsky přívětivé a doprovázené kvalitní dokumentací), neboť příprava vlastních technických nástrojů nemusí být v našich silách.

C.2 Abstrakt

Pokud tedy máme zvolenou problematiku, kterou chceme řešit, je dobrým nápadem vytvoření si stručného abstraktu. To nám pomůže uspořádat si myšlenky a zaměřit se tak na to skutečně podstatné, co chceme dělat, případně můžeme naše nápady konzultovat s vyučujícími (v případě semestrálního projektu) nebo s jinými odborně zdatnými osobami v případě jiného typu projektů a úkolů. Abstrakt by měl být krátký a neměl by přesáhnout 500 slov. Měl by zahrnovat následující body:

1. stručné popsání problému;
2. komentář k dostupným informacím (dosavadním přístupům) doprovázený jednou nebo dvěma klíčovými odkazy na literaturu;
3. popis metodiky výzkumu, která zahrnuje
   (a) ekonomický model,
   (b) metody ekonometrického odhadu a analýzy,
   (c) zdroje dat,
   (d) techniky odhadu, testování hypotéz, případně predikce;
4. potenciální přínos výzkumu.

Je však třeba třeba zdůraznit, že abstrakt jako takový se rozsahem může lišit např. v případě odborného článku. Zde by neměl přesáhnout 100 slov (což se liší dle požadavků vydavatele časopisu) a zaměřuje se na zdůraznění toho, co v našem příspěvku řešíme, v čem jsou náš přístup a výsledky zajímavé, přínosné, a proč vůbec tedy stojí za to si náš článek přečíst.

C.3 Struktura odborné studie

Odborné zprávy či studie v oblasti ekonomie mají svůj standardní formát, v rámci kterého je diskutován průběh zpracování projektu a interpretovány jeho výsledky. Samotná struktura hodně závisí na účelu, k jakému je daná zpráva či studie zpracovávána (seminární práce na vysoké škole, článek do odborného časopisu, podkladová zpráva pro rozhodování vedoucích pracovníků ve státní správě nebo centrálních bankách, analytická zpráva profesionálních ekonomů v soukromém sektoru bank, průmyslu.). Je

tak samozřejmě logické, že ne všechny níže uváděné komponenty musí obsahovat- např. školní semestrální projekt. Kolik prostoru budeme věnovat tomu či onomu bodu závisí jen a jen na nás. Sami musíme rozhodnout, co v našem projektu pokládáme za zásadní a nejzajímavější a čemu tedy budeme věnovat největší prostor. Jednotlivé části jsou tedy následující:

1. *Úvodní představení problému:* Většina zpráv a příspěvků začíná stručným představením toho, jaké otázky jsou řešeny, co nás k jejich řešení vedlo a shrnuje dosažená empirická zjištění. Úvod by měl být většinou psán jednoduchým, „netechnickým"' jazykem s minimem odborných ekonomickcýh a statistických výrazů. I laický čtenář by tak měl obecně pochopit problém a získané závěry, které jsou v projektu či příspěvku řešeny. Pokud se jedná např. o odborný článek, tak právě zde je nejlepší příležitost k tomu, vyzdvihnout originalitu a přínos našeho článku, tedy čím je zajímavý a proč má cenu ho vůbec číst. Není od věci přehledně představit i obsah jednotlivých částí zprávy.

2. *Přehled literatury:* Tato část představuje stručné a výstižné zhrnutí relevantní literatury v oblasti výzkumu, kterou jsme si zvolili. Součástí je popis toho, co bylo v dané oblasti již vyzkoumánoi a vysvětlení toho, jak naše práce přispívá k dosavadnímu stavu poznání. Velmi žádoucí je zde citovat práce jiných, které byly motivací pro náš výzkum. To vše opět ve stručné podobě. Není však třeba představovat přehled veškeré literatury, která se k danému tématu vztahuje.

3. *Ekonomický model:* Pokud se jedná o akademický příspěvek s nějakým formálním toeretickým modelem, pak je jeho popis obsažen v této části. Pro zprávy typu "policy reports" je v této části spíše prostor pro detailnější popsání ekonomických a institucionálních otázek řešených v této práci. Tato část je obvykle techničtěji založená a využívá se zde více jazyk ekonomie a matematiky. V této části se je možné zaměřit na posluchače, kteří jsou experty v daném oboru. V této části zprávy je věnován prostor specifikaci používaného ekonomického modelu a definování ekonomických proměnných. Na tomto místě je třeba deklarovat předpoklady modelu a identifikovat hypotézy, které chceme testovat. Ekonomický model může být mnohdy rozsáhlý a komplikovaný. Našim úkolem je vysvětlit model zcela jasně, a to co nejstručnějším a nejjednodušším způsobem. Není třeba používat ryze technického žargonu. Kde je to možné, snažme se používat jednoduchých a výstižných pojmů a obratů namísto zbytečně komplikovaných výrazů. Našim cílem je ukázat kvalitu našich myšlenek, nikoli šíři a rozsah naši slovní zásoby.

4. *Data:* Nesmíme zapomínat na popis dat, které budeme využívat, odkud je čerpáme (tedy jejich zdroj) a případná omezení pokud jde o jejich dostupnost. Data by měla být pokud možno bez větších problémů volně dostupná, aby kdokoliv měl možnost replikovat v případě zájmu naše výsledky a postupy.

5. *Ekonometrický model:* Tato část by se měla věnovat diskuzi nad tím, jak chceme využít data k analýze ekonomického problému popisovaného v třetí části. Tato část se bude lišit v závislosti na řešeném tématu a v závislosti na tom, komu je váš příspěvek určen. Například zde bude nutné argumentovat, že nás v rámci

studie zajímá určitá regrese a že konkrétní proměnná bude vysvětlovaná proměnná a další proměnné budou vysvětlující. Podobně, pokud se budete zajímat o analýzu makroekonomických časových řad, měli bychom zde představit implikace ekonomické teorie, že určité proměnné jsou kointegrovány a že z tohoto důvodu je třeba provést test kointegrace. Stručně řečeno, v této části je třeba popsat a zejména obhájit využívané postupy a techniky a zdůvodnit jejich volbu. Neměli bychom opomenout vysvětlit postupy testování hypotéz a jejich praktické použití. Kromě diskuze nad zahrnutím těch či oněch proměnných je dobré zdůvodnit rovněž funkční podobu modelu, předpoklady pokud jde o náhodnou složku a další přepdoklady, které uvažujeme. Značení by mělo být co možná nejjednodušší a není dobré zaneřádit stránky naší studie sáhodlouhými důkazy a definicemi (samozřejmě za předpokladu, že tyto důkazy netvoří jádro naší práce a nejsou našim hlavním přínosem). Důkazy a definice je dobré dát do technických příloh na závěr dokumentu. Vždy je ale třeba zvážit jejich relevantnost.

6. *Empirické výsledky a příslušné závěry:* Jádrem zprávy či projektu je právě tato část. Na tomto místě je žádoucí popsat dosažená empirická zjištění a vysvětlit jejich vztah k otázce či otázkám, které řešíte. Měla by zde být obsažena jak ekonomická, tak i statistická interpretace výsledků. Ekonomickou interpretací jsou zde myšleny např. odhady parametrů nebo závěry o kointegraci proměnných, a jaký vztah mají tato zjištění k ekonomické teorii. Statistická intepretace zahrnuje výsledky testování hypotéz, které ukazují statistickou významnost parametrů nebo potvrzení zvolené délky zpoždění, vysvětlení pro odstranění některé z vysvětlujících proměnných, diskuze nad kvalitou shody modelu s daty (koeficient determinace), testy heteroskedasticity atd. Většina těchto informací je prezentována v podobě tabulek a grafů. Není neobvyklé, když články začínají jednoduchými grafy (např. vykreslení časových řad dat) a za nimi následuje tabulka příslušných popisných statistik (střední hodnota, směrodatná odchylka, maximální a minimální hodnota, korelační matice). Další tabulky pak mohou obsahovat výsledky více formálnější analýzy, jako např. odhad parametrů modelu metodou nejměnších čtverců spolu s příslušnými t-statistikami (či p-hodnotami), koeficienty determinace R^2 a F-statistikami testujícími významnost regrese jako celku. Kromě představení našich vlastních odhadů parametrů (jejich interpretace a příslušných testů) je žádoucí komentovat vztah našich výsledků k výsledkům (odhadům) jiných autorů či odhadům našich předchozích studií, pokud možno s příslušnými ekonomickými implikacemi.

7. *Možná rozšíření a omezení studie a závěr*: Výzkum s sebou obvykle přináší řadu otázek spojených s ekonomickým modelem, daty a odhadovými technikami. Není od věci zamyslet nad dalším výzkumem s ohledem na dosažené výsledky a jak se s ním vypořádat. V závěru by měly být stručně shrnuty problémy, kterými se náš příspěvek zabýval a určitě by zde neměla chybět hlavní empirická zjištění.

8. *Poděkování:* Velmi vhodné je v samostatné sekci uvést jména těch osob (vědecké kolegy, spolupracovníky, přátele apod.), které významně přispěli k našemu výzkumu svými radami či komentáři.

9. *Reference:* Před případnými přílohami se nesmí zapomenout na seznam literatury, kterou ve své studii citujeme, a odkaz na datové zdroje, které využíváme.

Je dobré zdůraznit ještě několik věcí. První z nich je ta, že nejsou „dobré“ a „špatné“ empirické výsledky. Empirické výsledky jsou takové jaké jsou a člověk by tak neměl být zklamaný, pokud neukazují to, v co doufal, že by ukazovat měly. V ideálním světě přichází výzkumník s novou teorií a provede empirickou práci, která podpoří tuto novou teorii statisticky významným způsobem. Reálný svět je ale poněkud jiný. V reálném světě jsou mnohdy ty proměnné, kde bychom očekávali jejich statistickou významnost, statisticky nevýznamné. Proměnné, které by podle nás měly být kointegrovány, často kointegrovány nejsou. Koeficienty, které by měly být kladné, mohou být někdy záporné. Takové výsledky jsou dosahovány i v rámci velmi sofistikovaných studií. Je třeba mít na paměti, že zjištění, že nějaká teorie nepopisuje dobře realitu, má stejnou váhu jako zjištění, že teorie tuto realitu popisuje velmi dobře. Proto není třeba zoufat, pokud nějaké takové zdánlivě špatné výsledky dostaneme. Vždy si ale zkontrolujme (a to i v případě výsledků potvrzujících nějakou teorii), jestli jsme správným způsobem použili adekvátní metody. Rozhodně není dobrý přístup, nějakým způsobem „šolichat“ s daty či metodami, aby to „takříkajíc vyšlo“ (a takovéto pochybné postupy nějak zatajovat a tvářit se, že jsme postupovali korektně).

Empirické výsledky mohou být mnohdy nejasné a matoucí. Jeden statistický test může indikovat jednu věc, druhý naopak věc zcela opačnou. Vysvětlující proměnná v jedné regresi významná, může být ve druhé nevýznamná. V takovém případě toho opět mnoho nenaděláme, kromě toho, že tyto výsledky ve vší počestnosti zvěřejníme (tedy nevybereme si jen jeden, který se nám hodí) a pokusíme se v rámci možností porozumět tomu, proč takovýto konflikt či nejasnost vzniká a jaké může být jeho logické vysvětlení. Jen ve výjimečných případech dochází k úplnému falšování výsledků. Velmi časté je však pokušení používat nečestných postupů, aby člověk mohl ukázat výsledky, které lze ekonomicky rozumně očekávat. O tom již byla řeč výše. Je samozřejmě obvyklé, že výzkumník provede velký počet regresí s různými vysvětlujícími proměnnými. To je vcelku rozumné, protože člověk tak detailněji analyzuje data z více úhlů pohledu. Ovšem, pokud je pak prezentován výsledek, který podporuje požadovanou teorii a naopak jsou zamlčeny výsledky, které hovoří proti ní, jedná se o úmyslné matení čtenáře. Tohoto pokušení je třeba se vždy vyvarovat. Obecně je tak dobré prezentovat spíše výsledky více regresí než vybrat jen jeden model a prezentovat pouze tyto jeho výsledky.

Pokud jde o samotnou prezentaci výsledků, důležitou roli hraje jasnost a stručnost. Nikdo nemá zájem číst dlouhé, špatně strukturované a rozvláčné zprávy. Dokázat rozhodnout, co je potřeba prezentovat (např. které testové statistiky z různých regresí, které jsme provedli) a co naopak ne, je hodně významná dovednost. Do zprávy vyberme ty nejdůležitější informace a prezentujme dosažené výsledky čestně a otevřeně. Není od věci zkontrolovat si i gramatickou a stylistickou stránku naší zprávy, protože tento druh chyb může zbytečně srážet kvalitu našeho příspěvku. Pravdou je, že člověk, který na nějakém úkolu dlouho pracuje, je do něj pohroužen natolik, že mu tento druh chyb při pročítání uniká. Ale od toho tu máme řadu ochotných přátel, kteří si naši práci rádi přečtou a upozorní na chyby a překlepy, které nám unikly, případně dodají nejednu moudrou radu a komentář, díky čemuž se jejich jména nepochybně objeví na čestném místě sekce „Poděkování“.

Příloha D

Zdroje dat

K empirické analýze samozřejmě potřebujeme adekvátní data. Otázkou tedy je, kde je získat? Mnohá užitečná data, jsou obsahem příloh či doprovodných webových stránek ekonometrických učebnis. Díky nim si člověk může sám provést příslušnou ekonometrickou analýzu problému, porovnat ji s řešením, které nabízí dané učebnice a osvojit si tak potřebné používání ekonometrických technik. Většina těchto dat je přístupná skrze datovou bázi příkladových dat v gretlu [1], případně je snadno doinstalovatelná pomoc souborů dostupných z přísušných stránek gretlu. Gretl nabízí snadné stažení i dalších datových zdrojů z centrálních bank a jiných ekonomickcýh institucí, ať už jde o data vybraných evropských zemí či Spojených států. Tato data jsou aktuální s větším či menším zpožděním (v řádu měsíců či let) a snadno se můžeme podívat, co všechno nabízejí. Výhodou je zde to, že si hledanou řadu či řady můžeme rovnou stáhnout a okamžitě s ní v programu pracovat.

Ekonomická data jsou dostupná z různých zdrojů, a asi je těžké nabídnout nějaký obecný komentář k jejich získávání. Opomeneme-li tvorbu vlastních dat ve smyslu vlastních měření či dotazníkových šetření, je nejlepším zdrojem dat internet, který nabízí řadu relevantních webových stránek, ze kterých je možno potřebná data získat. Protože je webová síť oblastí velmi dynamickou, mohou být dále uváděné informace již zastralé, tudíž je musíme brát zejména z pohledu námětů toho, co a kde hledat.

Řada dat je dostupná volně, mnohá data jsou však zpoplatněna. Samotné univerzity či univerzitní knihovny obvykle poskytují svým studentům a zaměstnancům volným přístup k široké paletě placených databází.

Jako první zdroj dat bych uvedl stránky statistických úřadů jednotlivých zemí. V našem případě to je samozřejme Český statistický úřad. Ten nabízí odkazy i na další mezinárodní zdroje dat, kterými může být Eurostat, a v rámci nichž se bezpochyby snadno proklikáme k hledaným údajům. Jako příklad zahraničního statistického úřadu uveďme Tatauranga Aotearoa, což je maorské označení statistického úřadu Nového Zélandu popřípadě statistiky Spojených národů. Odkazy na volně dostupné oficiální statistiky nabízí například stránka knihovny University of Auckland.

Bohaté zdroje čistě ekonomických dat můžeme najít na stránkách národních centrálních bank a dalčích mezinárodních institucí. Příkladem může být Česká národní banka, Evropská centrální banka, americký FED a novozélandská RBNZ. Bohaté, ale

ne zcela volně dostupné, jsou databáze OECD popř. Světové banky. Zajímavá data nabízí i Statistická divize Spojených národů.

Samostatnou kapitolu tvoří různorodé rozcestníky k datům. Užitečná je tak americká stránka Resources for Economists on the Internet, která nabízí ohromné množství materiálů k celé řadě ekonomických témat a poskytuje i odkazy k užitečným datovým zdrojům. Na této stránce můžete nalézt i odkaz na datový archiv některých časopisů. Řada časopisů totiž požaduje od autorů, aby svá data v rámci možností dávala veřejnosti k dispozici, tím pádem se nabízí skvělá možnost získat potřebná data k zajímavým článkům a zkoušet si replikovat výsledky. Skvělým příkladem je Journal of Applied Econometrics Data Archive. Příslušné podkladové články snadno získáme skrze JSTOR či EBSCOhost (pokud k nim má univerzita samozřejmě předplacen přístup).

Odkazy na volně dostupná ekonomická data se hemží stránka Economics Network. Vybraná historická data je možno získat přes stránky Portálu historické statistiky. Odkazy na spoustu časových řad nabízí Rob Hyndman.

Další stránkou s užitečnými odkazy je National Bureau of Economic Research. Skrze tyto stránky se můžeme dostat na tzv. Penn World Table (opět jsou dostupné i skrze gretl), které nabízejí makroekonomická data více než stovky zemí v průběhu několika desetiletí. Rovněž tak hodně jednotlivých zemí má své stránky s velkým množstvím panelových dat, kdy probíhá každým rokem průzkum mezi respondenty. Ve Spojených státech je to Panel Study of Income Dynamics, pro Spojené království je srovnatelnou datovou bází British Household Panel Survey. Pro českou republiku je bohužel získání přístupu k tomuto typu dat (které pravidelně provádí Český statistický úřad) z nepochopitelných důvodů velmi obtížné a fakticky nemožné.

Pokud jde o „čistě“ finanční data, je situace složitější. Existují vynikající databáze cen akcií a informací o účetnictví firem pro velkou řadu společnsotí za řadu let, nicméně tyto databáze jsou dosti nákladné a ne každá univerzita či fakulta k nim má předplacen přístup. Pro Českou republiku přichází v úvahu např. server Patria plus. Finanční data nabízí i americký FED. Financial Data Finder je vyhledávač dat poskytovaný Fisher College of Business (Ohio State University). Řada akademiků poskytuje svá využívaná data veřejnosti na svých webových stránkách. Příkladem může být Robert Shiller z Yale Univerzity, který na své webové stránce poskytuje odkazy k různým zajímavým finančním datovým zdrojům. Obecně však můžeme říct, že věnovat nějaký čas prohledáváním webové sítě může být velmí užitečné.

Literatura

[1] Gnu regression, econometrics and time-series library (gretl). Version 1.8.3.

[2] Matlab. The MathWorks, version 2008b.

[3] BALTAGI, B. H. *Econometric Analysis of Panel Data*, 4 ed. John Wiley & Sons, 2008.

[4] BENEŠ, J. Iris toolbox.

[5] BROOKS, C. *Introductory Econometrics for Finance*, 2 ed. Cambridge University Press, 2008.

[6] CIPRA, T. *Finanční ekonometrie*, 1 ed. Ekopress, 2008.

[7] DAVIDSON, R., AND MACKINNON, J. *Econometric Theory and Methods*. Oxford University Press, 2004.

[8] DOUGHERTY, C. *Introduction to Econometrics*, 3 ed. Oxford University Press, 2007.

[9] ENDERS, W. *Applied Econometric Time Series*, 2 ed. Wiley Series in Probability and Statistics. John Wiley & Sons, 2005.

[10] GREENE, W. H. *Econometric Analysis*, 6 ed. Prentice Hall, 2008.

[11] GUJARATI, D. N., AND PORTER, D. C. *Basic Econometrics*, 5 ed. McGraw-Hill, 2009.

[12] HAYASHI, F. *Econometrics*. Princeton University Press, 2001.

[13] HILL, R. C., GRIFFITHS, W. E., AND LIM, G. C. *Principles of Econometrics*, 3 ed. John Wiley & Sons, 2008.

[14] JUILLARD, M. Dynare toolbox for matlab. ver. 4.

[15] KENNEDY, P. *A Guide to Econometrics*, 6 ed. Wiley-Blackwell, 2008.

[16] KOOP, G. *Bayesian Econometrics*. Wiley, 2003.

[17] KOOP, G. *Introduction to Econometrics*. John Wiley & Sons, 2008.

[18] KOOP, G. *Analysis of Economic Data*, 3 ed. John Wiley & Sons, 2009.

[19] LESAGE, J. P. Econometrics toolbox. version 7.

[20] STOCK, J. H., AND WATSON, M. W. *Introduction to Econometrics*, 2 ed. Pearson/Addison Wesley, 2007.

[21] STOCK, J. H., AND WATSON, M. W. *Introduction to Econometrics*, brief ed. Pearson/Addison Wesley, 2008.

[22] VERBEEK, M. *A Guide to Modern Econometrics*, 3 ed. Wiley, 2008.

[23] WOOLDRIDGE, J. M. *Introductory Econometrics (A Modern Approach)*, 4 ed. South-Western College, 2009.