**7.2 Binární logistická regrese**

Závisle proměnná je binární, nabývá jen dvou hodnot 0 a 1. Do tohoto tvaru je potřeba závisle proměnnou upravit, Stata totiž bere nulovou hodnotu jako nulu a jakoukoliv nenulovou hodnotu jako jedničku. Následně se používá stejně jako lineární regrese, jen místo příkazu REGRESS použijeme LOGIT. Samotné koeficienty se interpretují příliš složitě, lepší je interpretovat poměry šancí (odds ratio). Ty dostaneme pomocí parametru OR nebo použití příkazu LOGISTIC namísto příkazu LOGIT. Poměr šancí pak slouží jako multiplikativní konstanta (kolikrát vyšší šanci máme, že nastane vysvětlovaný jev, tedy že závisle proměnná nabývá hodnoty 1).

Výsledné poměry šancí pak interpretujeme buď jako zvýšení/snížení šance, nebo jako procentní nárůst. Např. poměr šancí s hodnotou 1,02 znamená, že pokud se nezávisle proměnná zvýší o jednotku (např. jeden rok), zvýší se pravděpodobnost, že jev nastane přesně 1,02×, neboli o 2 %.

*recode gearn (1 2 = 1) (3/5 = 0), gen (muz\_chlebodarce)*

*logit muz\_chlebodarce i.sex age i.educ*

*logit muz\_chlebodarce i.sex age i.educ, or*

*logistic muz\_chlebodarce i.sex age i.educ*

****

Obrázek Ukázka binární logistické regrese

**7.3 Ordinální logistická regrese**

Závisle proměnná je ordinální, nabývá několika kategorizovaných hodnot, které lze seřadit, a jsou od sebe stejně vzdáleny. Následně se používá stejně jako lineární regrese, jen místo příkazu REGRESS použijeme OLOGIT. Opět interpretujeme poměry šancí, které dostaneme pomocí parametru OR.

*ologit gearn i.sex age i.educ, or*

**

Obrázek Ukázka ordinální logistické regrese

**7.4 Multinomická logistická regrese**

Závisle proměnná je ordinální, nabývá několika kategorizovaných hodnot, které lze seřadit, ale – na rozdíl od ordinální logistické regrese – nejsou od sebe stejně vzdáleny. Následně se používá stejně jako lineární regrese, jen místo příkazu REGRESS použijeme MLOGIT. Opět interpretujeme poměry šancí (v tomto případě se ovšem statisticky korektně označují jako relative risk ratio), které dostaneme pomocí parametru RRR.

*mlogit gearn i.sex age, brr*

**

Obrázek Ukázka multinomické logistické regrese

Multinomická logistická regrese se interpretuje pro každou kategorii závisle proměnné zvlášť ve vztahu k referenční kategorii. Na výše uvedeném obrázku proto například platí, že pokud je respondent ženského pohlaví, je o 26,6 % nižší pravděpodobnost (1-0,734), že bude „ani souhlasit, ani nesouhlasit“, ve srovnání s tím, že bude „spíše nesouhlasit“. Interpretace multinomického logistického modelu je proto extrémně složitá, a pokud je to jen trochu možné, použijeme raději ordinální logistickou regresi.

Podmínkou je ale stejná vzdálenost mezi jednotlivými variantami závisle proměnné, kterou nejlépe ověří Brantův test. Postup je takový, že nejprve provedeme ordinální logistickou regresi (příkaz OLOGIT) a následně zadáme příkaz BRANT, DETAIL. Ten zjistí, jestli je použití ordinální logistické regrese vhodné.

*ologit gearn i.sex age*

*brant, detail*



Obrázek Použití Brantova testu

V našem případě bohužel Brantův test doporučuje použít multinomický logistický model.

**7.5 Kvalita regresních modelů**

Při práci s regresními modely postupujeme vždy od jednodušších modelů ke složitějším. Sledujeme přitom, jak se mění kvalita modelu. Cílem je dosáhnout stavu, kdy je model co nejjednodušší, ale přitom co nejkvalitnější (protichůdné požadavky). Než se podíváme na příklad budování modelů, probereme jednotlivé ukazatele kvality.

U lineární regrese je hlavním ukazatelem kvality R2, který vyjadřuje, do jaké míry model reprezentuje sociální realitu. Hodnota je sice udávána v desetinném čísle, ale po vynásobení stem dostáváme hodnotu v procentech (např. R2 = 0,1234 znamená, že model vysvětluje realitu z 12,34 %). Samozřejmě platí, že čím vyšší je R2, tím kvalitnější je model. U lineárních modelů je potřeba reportovat dvě hodnoty, kromě R2 také počet případů N.

U logistické regrese ukazatel R2 použít nemůžeme, i když ho Stata v podobě Pseudo R2 nabízí. Ukazateli kvality logistického modelu jsou hodnoty likelihood ratio, AIC a BIC. Alespoň jeden z nich je potřeba – spolu s počtem případů N – reportovat spolu s modelem. Ani jeden z uvedených ukazatelů přitom neříká nic o absolutní kvalitě modelu (oproti R2, které je použitelné i pro jeden samostatný model), slouží pouze ke srovnání kvality dvou modelů. Modely musí být do sebe vnořené (nested), to znamená, že jeden model musí obsahovat tytéž proměnné jako druhý model, případně nějaké navíc, a oba modely musí mít stejný počet případů N.

Poměr věrohodnosti (log-likelihood ratio, -2LL), obecně platí, že čím větší log-likelihood, tím lepší model. Akaikovo informační kriterium AIC podle některých zdrojů nevyžaduje, aby byly porovnávané modely do sebe vnořené, zohledňuje počet nezávislých proměnných. Platí, že čím menší AIC, tím lepší model. Bayesovo informační kritérium (BIC) penalizuje složité modely ještě silněji, lze ho ale teoreticky využít i pro modely s různým N. I v tomto případě volíme model, který má hodnotu BIC co nejnižší.

Charakteristiky modelu vypisují příkazy ESTAT IC nebo FITSTAT.



Obrázek Posuzování kvaliti regresního modelu

**7.6 Postupné budování modelu**

Při návrhu regresních modelů vycházíme, stejně jako při jiných sociologických analýzách, z teorie. Studiem literatury zjistíme, které nezávisle proměnné by měly ovlivňovat závisle proměnnou. Nevytváříme zvláštní regresní model pro každou hypotézu, ale jeden model celkový, z kterého následně rozhodneme o platnosti všech hypotéz. Všechny modely prezentujeme v jedné přehledné tabulce včetně potřebných ukazatelů kvality modelu.

Příklad: Souvislost vzdělání a výše příjmu

Literatura: Teorie lidského kapitálu říká, že výše příjmu je ovlivněna především délkou praxe (většinou nahrazeno věkem) a stupněm dosaženého vzdělání (lidským kapitálem). Mincerova rovnice empiricky ukazuje, že je potřeba porovnávat přirozený logaritmus hrubého příjmu s nezávisle proměnnými věk, druhá mocnina věku, stupeň vzdělání.

Nejprve si připravíme data:

*gen prijem=payn*

*replace prijem=payn/12 if paynam==1 //pokud je příjem roční, převedeme ho na měsíční*

*replace prijem=prijem/(hours\*4.5) //měsíční příjem převedeme na hodinový tak, že ho vydělíme počtem hodin odpracovaných týdně a počtem týdnů v měsíci*

replace prijem=ln(prijem) //vypočteme logaritmus hodinového příjmu

gen age2=age\*age //druhá mocnina věku

recode educ (0/4=1) (5/8=2) (9/11=3)

Nyní odhadneme první regresní model M1:

regress prijem i.educ age age2



Obrázek Regresní model M1

Vidíme, že regresní model vysvětluje realitu z 5,41 %. Dosažení maturitního vzdělání zvyšuje příjem o 32 %, dosažení vysokoškolského vzdělání o 61 % ve srovnání se vzděláním základním. Získání jednoho roku praxe zvýší příjem o 1,8 %. Předpokládáme ale, že na příjem má vliv i pohlaví respondenta, statistiky Eurostatu říkají, že Česko patří k zemím s nejvyšším genderovým rozdílem v příjmech. Přidáme proto do modelu M2 ještě pohlaví.

regress prijem i.educ age age2 i.sex

Kvalita modelu se mírně zvýšila, koeficienty se změnily (viz tabulka). Nyní nás zajímá, jak se liší návratnost vzdělání pro ženy v závislosti pro různé stupně vzdělání, tedy zda vysokoškolačky získají ze svého vzdělání stejně jako středoškolačky. Vytvoříme proto model M3 s přidanou interakcí mezi pohlavím a stupněm vzdělání.

regress prijem i.educ age age2 i.sex i.sex#i.educ

Všimněte si, že postupně jsme vytvořili čtyři hypotézy, ale výsledky budeme interpretovat z jediného regresního modelu:

H1: Čím více má člověk praxe, tím vyšší má příjem.

H2: Čím je člověk vzdělanější, tím vyšší má příjem.

H3: Ženy mají nižší příjem než muži bez ohledu na vzdělání i praxi.

H4: Absolvování vysoké školy dává ženám příležitost dohnat genderový rozdíl v příjmech.

Všechny modely shrneme do jediné tabulky, ze které pak výsledky slovně interpretujeme:

Tabulka Determinanty výše přijmu

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | M1 | M2 | M3 |
| Vzdělání | Vyučen | Ref. | Ref. | Ref. |
|  | Maturita | 0,319\*\*\* | 0,330\*\*\* | 0,387\*\*\* |
|  | VŠ | 0,613\*\*\* | 0,618\*\*\* | 0,683\*\*\* |
| Věk |  | 0,019\* | 0,024\* | 0,024\* |
| Věk\*Věk |  | -0,000 | -0,000+ | -0,000+ |
| Pohlaví | Muž |  | Ref. | Ref. |
|  | Žena |  | -0,211\*\*\* | -0,131\* |
| Interakce | Muž a Vyučen |  |  | Ref. |
|  | Žena \* Maturita |  |  | -0,112 |
|  | Žena \* VŠ |  |  | -0,128 |
| Konstanta |  | 3,570\*\*\* | 3,558\*\*\* | 3,517\*\*\* |
| N |  | 3 033 | 3 033 | 3 033 |
| R2 |  | 0,0541 | 0,0641 | 0,0648 |

Statistická signifikance: \*\*\* p<0.001 \*\* p<0.01 \* p<0.05 + p<0.1

Na základě modelu M3 nemůžeme vyvrátit hypotézu H1 (v našich datech skutečně s rostoucí praxí roste příjem) ani H2 (v našich datech rostoucí vzdělání skutečně zvyšuje příjem). Stejně tak hypotéza H3 musí zůstat podržena, v modelu M2 dokonce hodnota 21,1 % přibližně odpovídá rozdílu gender pay gap, který uvádí pro Českou republiku Eurostat (v modelu M3 bychom museli interpretovat vliv pohlaví spolu s vlivem interakce mezi pohlavím a vzděláním). Hypotézu H4 musíme zamítnout, VŠ žena má příjem nižší o (0,131 + 0,128) 25,9 %, zatímco žena s maturitou má příjem nižší jen o (0,131 + 0,112) 24,3 % nižší.

Nejčastější chyby:

Vytváření zvláštních modelů pro každou hypotézu – Je potřeba vytvořit „jeden velký model“, který kontroluje vliv ostatních proměnných.

Interpretace jen jednoho modelu – Je potřeba dokázat, že použitý model je ten nejlepší možný, proto je vhodné ukázat více modelů.

Chybějící charakteristiky modelu – U každého modelu musí být informace o jeho kvalitě. U lineární regrese N a R2, u logistické regrese N a buď AIC nebo BIC.

Chybějící referenční hodnoty – U kategorizovaných proměnných je potřeba ukázat čtenáři, která hodnota je referenční.

Vyřazení nesignifikantních proměnných – Nesignifikance nezávisle proměnných ztěžuje jejich interpretaci, na druhou stranu i nesignifikantní proměnná slouží jako kontrolní proměnná, která očišťuje vliv ostatních proměnných. Pokud je její přítomnost v modelu zdůvodnitelná (z literatury nebo z logické úvahy), je lepší ji tam ponechat.