10 Mnohonásobná lineární regrese

V kapitole 7 jsme se zabývali situací, v níž bylo zapotřebí analyzovat vztah mezi jednou závisle a jednou nezávisle proměnnou. V této kapitole rozšíříme tento přístup, abychom mohli analyzovat komplexnější data s více nezávislými proměnnými. Popisné prostředky, které jsme poznali – bodové dvojrozměrné grafy, grafy reziduálních hodnot, korelace a regresní koeficienty, metoda nejmenších čtverců – se používají i v tomto případě. Jsou předpokladem pro získání hodnověrného modelu dat a oprávněnou aplikaci statistických testů významnosti a intervalů spolehlivosti.

Příklady otázek, které mohou být zodpovězeny pomocí mnohonásobné regrese:

- Jak závisí hodnota výsledku znalostního testu na demografických charakteristikách žáka?
- Jak závisí úspěšnost hokejového mužstva na průměrném BMI indexu, přesnosti přihrávek a úspěšnosti střelby z trestného nájezdu?
- Jak závisí celková osmolalita roztoku na koncentracích různých osmoticky aktivních látek?

Mnohonásobná regrese je prostředkem pro zkoumání statistické závislosti pomocí modelu, jenž zahrnuje jednu závislou proměnnou a několik nezávislých proměnných. Data získáme tak, že u prvků výběru zjistíme hodnoty všech uvažovaných proměnných. Rozlišujeme tři druhy úloh, pro jejichž řešení je vhodné aplikovat mnohonásobnou regresní analýzu:

1. Chceme poznat efekt, který má na cílovou proměnnou Y souhrn změn ovlivňujících parametrů $X_1, X_2, \ldots, X_k$.
2. Chceme predikovat hodnotu závisle proměnné Y pro budoucí hodnoty proměnných $X_1, X_2, \ldots, X_k$.
3. V rámci explorační statistické analýzy chceme vyhledat statistické vztahy mezi závisle proměnnou a několika nezávisle proměnnými.

Náš výklad bude mít charakter průvodce základními kroky při využívání regresního modelu s více nezávislými proměnnými při analýze dat. Také objasníme vztah tohoto modelu k analýze rozptylu a k analýze kovariance. Stejně jako v jednoduché regresi dvou proměnných, hraje i v mnohonásobné lineární regresi důležitou roli metoda nejmenších čtverců, pomocí níž odhadujeme parametry modelu. Na obrázku 10.1 zachycujeme prvky regresního modelu, jež v kapitole podrobněji popíšeme.

10.1 Mnohonásobná regrese a metoda nejmenších čtverců

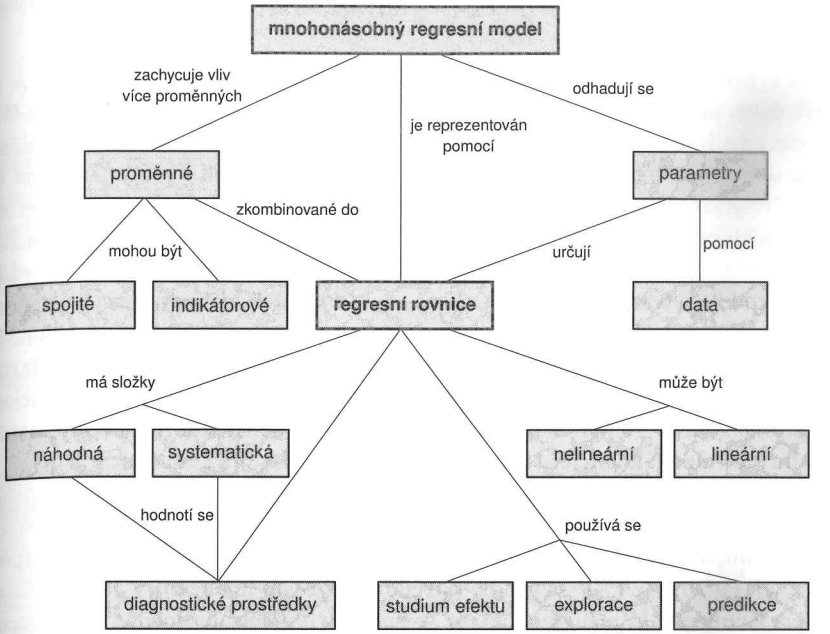
Regresní analýza označuje obecně širokou třídu statistických technik, které jsou navrženy pro zkoumání vztahu mezi závisle proměnnou Y a nezávislými proměnnými $X_1, X_2, \ldots, X_k$. Připomeňme si z kapitoly 7.3, že pro oba typy proměnných se používají v aplikacích různé názvy. Například pro nezávisle proměnnou používáme označení prediktor, regresor, predikující, vysvětlující, ovlivňující proměnná. Predikční funkce má obecný tvar

$$y = f(x_1, x_2, \ldots, x_k; b_0, b_1, \ldots, b_m),$$

kde $b_0, b_1, \ldots, b_m$ jsou hodnoty parametrů upřesňující tvar funkce. Za $f(.)$ volíme nejčastěji vážený součet prediktorů nebo součet jejich jednoduchých transformací. Funkci $f(.)$ tvoří v nejjednodušším případě prostá lineární funkce $y = b_0 + b_1 x_1$, jak jsme poznali v kapitole 7.3. Na druhé straně se může jednat o komplikovanější vážený součet několika prediktorů, jež jsou navíc transformovány umocněním nebo jinou transformací – např. funkcemi $\ln(x)$ nebo $\exp(x)$. Také uvažujeme násobení prediktorů mezi sebou (tedy např. $y = b_0 + b_1 \ln(x_1) + b_2 \ln(x_2) + b_3 x_1 x_2$). Všechny nelineární komponenty ve vztahu – v tomto příkladu $\ln(x_1)$ apod. – jsou ve výpočtech pokládány za nové nezávislé proměnné. Tímto přístupem (pro jednu nezávisle proměnnou jsme jej popsali v kap. 7.3.6) se daty mohou prokládat i nelineární funkce.

Ideálně by měla být splněna podmínka, že obecný tvar funkce je volen na základě nějakých teoretických úvah. Teorie by měla navrhnout, zda použijeme lineární, kvadratické nebo logaritmické funkce proměnných. V mnoha aplikacích však není teorie natolik rozpracovaná, aby určila typ závislosti. Často se proto postupuje pomocí metody pokusu a omylu. Uveďme, že kvalitu proložení dat regresní funkcí přezkušujeme podobně jako v případě jednoduché regrese (kap. 7.3.2). Tomuto problému se budeme věnovat také v kapitole 10.4.

Obr. 10.1 Schéma konceptu mnohonásobné regrese



Data zachycujeme tabulkou, kde pro každý objekt uvádíme hodnoty prediktorů a závisle proměnné. Například zjišťujeme u n žáků hodnoty k nezávisle proměnných $X_1, X_2, \ldots, X_k$ a závisle proměnnou Y. Matice měření X pak má tvar:

$$\begin{array}{ll} \text{žák 1:} & (x_{11}, x_{12}, \ldots, x_{1k}, y_1) \\ \text{žák 2:} & (x_{21}, x_{22}, \ldots, x_{2k}, y_2) \\ \vdots & \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ \text{žák } n\text{:} & (x_{n1}, x_{n2}, \ldots, x_{nk}, y_n) \end{array}$$

Poznamenejme, že řádku v matici říkáme **vektor měření**. Při hledání hodnot koeficientů $b_0, b_1, \ldots, b_m$ využíváme podobně jako v jednoduché regresi **metodu nejmenších čtverců**. Vycházíme ze součtu čtverců rozdílů změřených hodnot y_i a hodnot $\hat{y}_i$ vypočítaných pomocí regresní funkce, do níž dosadíme hodnoty

nezávisle proměnných z matice pozorování X:

$$\sum_{i=1}^{n} v_i(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} v_i(y_i - f(x_1, x_2, \ldots, x_k; b_0, b_1, \ldots, b_m))^2,$$

kde v_i jsou váhy, které přisuzujeme jednotlivým odchylkám. Běžně volíme $v_i = 1$. Hledáme taková $\{b_j\}$, jež minimalizují součet čtverců odchylek. Říkáme, že $\{b_j\}$ jsme zjistili metodou nejmenších čtverců. Jestliže se váhy v_i nerovnají jedničce, jde o váženou metodu nejmenších čtverců. Pro posouzení kvality proložení se používají kromě součtů čtverců reziduálních hodnot další kritéria, např. součet absolutních reziduálních hodnot, maximum z absolutních reziduálních hodnot atd.

Nejčastěji se uvažuje prostý lineární vztah mezi závisle proměnnou a prediktory, který má tvar $y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$. Predikční funkci jako vážený součet nezávislých proměnných někdy nazýváme lineární prediktor proměnné Y.

Metodou nejmenších čtverců dosáhneme toho, že predikční funkce bude mít pro soubor měření mezi všemi lineárními prediktory největší korelační koeficient s proměnnou Y. Tomuto největšímu korelačnímu koeficientu říkáme **mnohonásobný korelační koeficient** mezi Y a množinou prediktorů a značíme ho $r_{y.x_1x_2\ldots x_k}$. Ten se také někdy zkráceně označuje symbolem R. Jeho rozsah je $0 \leq r_{y.x_1\ldots x_k} \leq 1$. Hodnoty blízké k 0 indikují statistickou nezávislost Y na množině proměnných X_j, zatímco hodnota 1 označuje dokonalou lineární vazbu, která je dána regresní funkcí. Parametry $\{b_j\}$ nazýváme **parciální regresní koeficienty** a b_0 absolutní člen. Koeficient b_j je hodnota, o niž se změní v průměru proměnná Y při změně X_j o jednu jednotku (při ostatních proměnných zafixovaných). Jestliže všechny proměnné standardizujeme před regresní analýzou transformací

$$y' = \frac{(y - m_y)}{s_y}, \quad x'_j = \frac{(x_j - m_j)}{s_j},$$

potom vztah mezi těmito proměnnými bude vyjádřen rovnicí

$$y' = b'_0 + b'_1x'_1 + b'_2x'_2 + \cdots + b'_kx'_k,$$

kde koeficienty b'_j jsou **standardizované regresní koeficienty**. Podle jejich velikosti někdy posuzujeme relativní přínos prediktorů X_j k predikci proměnné Y.

Míru přesnosti odhadu proměnné Y regresním vztahem a rozptýlenost hodnot Y kolem regresní funkce posuzujeme velikostí **reziduální směrodatné odchylky při regresi** $s_{y.x_1x_2\ldots x_k}$. Hodnotu $s^2_{y.x_1x_2\ldots x_k}$ nazýváme reziduální (zbytkový) rozptyl kolem regresní funkce a vypočítáme ho pomocí vztahu

$$s^2_{y.x_1x_2\ldots x_k} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - k}.$$

Odpovídá průměrné kvadratické odchylce měření závisle proměnné od regresní funkce. Jestliže s_y je směrodatná odchylka proměnné Y, pak pro výše uvedené korelační koeficienty platí vztah

$$R^2 = r^2_{y.x_1x_2\ldots x_k} = \frac{s_y^2 - s^2_{y.x_1x_2\ldots x_k}}{s_y^2},$$

který říká, že čtverec mnohonásobného korelačního koeficientu $R^2 = r^2_{y.x_1x_2\ldots x_k}$ je roven části variability proměnné Y, která je vysvětlena prediktory $X_1, X_2, \ldots, X_k$. Koeficient R^2 se nazývá **koeficient determinace**:

$$\textit{koeficient determinace } R^2 = \frac{\textit{variabilita vysvětlená modelem}}{\textit{celková variabilita}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}.$$

Někdy se uvádí koeficient determinace v procentech. Pak má hodnotu $100\,R^2$.

PŘÍKLAD 10.1

Mnohonásobná lineární regrese

Výsledky získané mnohonásobnou regresní analýzou demonstrujeme na modelovém příkladu z psychologického výzkumu. Matice pozorování X (tab. 10.1) se získala tak, že šest žáků ($\hat{o}_i$) absolvovalo čtyři testy, které měří následující veličiny: X_1 = přírodovědné vědomosti; X_2 = literární vědomosti; X_3 = schopnost koncentrace; X_4 = logické myšlení. Testy se hodnotí na škále od 1 do 10 (1 = špatný výsledek, 10 = výborný výsledek). V našem případě je počet objektů $n = 6$ a počet proměnných $k = 4$. V praxi by měl být počet zkoumaných žáků mnohem větší.

Tab. 10.1 Matice měření X

Pokusná osoba	Test			
	X_1	X_2	X_3	X_4
$\hat{o}_1$	7	9	10	8
$\hat{o}_2$	9	8	8	10
$\hat{o}_3$	4	3	1	2
$\hat{o}_4$	2	3	2	2
$\hat{o}_5$	3	1	2	4
$\hat{o}_6$	1	1	1	4

Tab. 10.2 Korelační matice R

	X_1	X_2	X_3	X_4
X_1	1,00	0,91	0,87	0,87
X_2	0,91	1,00	0,96	0,82
X_3	0,87	0,96	1,00	0,89
X_4	0,87	0,82	0,89	1,00

Tab. 10.3 Porovnání skutečných hodnot proměnné Y a hodnot odhadnutých modelem

Modelem odhadnutá hodnota Y	9,3	8,6	1,4	2,1	0,9	1,6
Naměřená hodnota Y	10	8	1	2	2	1

Pro výpočty v lineární regresní analýze je zapotřebí zjistit korelace všech proměnných mezi sebou, které uspořádáváme do tzv. korelační matice (tab. 10.2). Dále počítáme průměry a směrodatné odchylky:

vektor průměrů $m = (4{,}33; 4{,}18; 4{,}00; 5{,}00)$;
vektor směrodatných odchylek $s = (2{,}80; 3{,}18; 3{,}60; 3{,}00)$.

Ptáme se, kolik bodů můžeme očekávat u testu koncentrační schopnosti žáka, jestliže známe výsledky jeho testů pro literární schopnosti, přírodovědné schopnosti a logické myšlení? V tomto problému je tedy třetí proměnná X_3 závislá ($= Y$) a ostatní jsou nezávislé. Hledat lineární prediktor pro koncentrační schopnost má cenu pouze tehdy, když je významná závislost mezi Y a ostatními proměnnými. Uveďme některé výsledky.

Predikční rovnice získaná metodou nejmenších čtverců má tvar:

$$y = -0{,}38x_1 + 0{,}98x_2 + 0{,}53x_4 - 1{,}09.$$

Mnohonásobný koeficient korelace $R = 0{,}96$.

Pomocí uvedené rovnice jsme v tabulce 10.3 vypočítali odhadnuté hodnoty proměnné $Y = X_3$ pro kontrolu správnosti regrese.

Jestliže budeme považovat získanou predikci za dostatečnou, lze v budoucnu použít získanou predikční rovnici pro odhad koncentrační schopnosti (pokud budou platit podmínky regresního modelu, viz dále) a nemusíme ji zvlášť testovat (měřit).

Vezměme např. výsledky predikčního vektoru 2. žáka v matici dat X, které jsou $x = (9; 8; 10)$. Odhad schopnosti koncentrace získáme následujícím výpočtem:

$$\hat{y}_2 = -0{,}38 \times 9 + 0{,}98 \times 8 + 0{,}537 \times 10 - 1{,}09 = 8{,}6.$$

Reziduální hodnota pro tento odhad je $e_2 = y_2 - \hat{y}_2 = -0{,}6$.

V regresní a korelační analýze počítáme také tzv. **parciální koeficienty korelace** mezi vybranými proměnnými. Tyto koeficienty nám pomáhají řešit *problém třetí proměnné*, tedy problém možného efektu rušivých proměnných, jejž jsme popsali v kapitole 7.2.4. Parciální koeficient korelace $r_{x_1x_2.x_3...x_k}$ je mírou asociace mezi proměnnými X_1 a X_2 po vyloučení „efektu" množiny proměnných $X_3, \ldots, X_k$ na obě zkoumané proměnné. Využívá se při posuzování zdánlivých asociací mezi proměnnými. Například známe korelační matici psychologických testů T_1, T_2 a T_3. Chceme zjistit, jaký efekt na jejich korelační strukturu má věk (V). Vypočteme parciální korelační koeficienty proměnných T_i po vyloučení vlivu věku. Ty opět tvoří korelační matici. Jestliže nová korelační matice je podobná matici původní, můžeme tvrdit, že původní korelační matice odráží korelační poměry testů ve všech uvažovaných věkových kategoriích.

Parciální korelační koeficient $r_{x_1x_2.x_3...x_k}$ se zjistí tak, že predikujeme zvlášť proměnné X_1 a X_2 pomocí lineárních prediktorů vytvořených z proměnných $X_3, \ldots, X_k$. Pro obě predikce vypočteme reziduální hodnoty, které považujeme za nové proměnné E_1 a E_2. Následně korelujeme tyto nové proměnné, jejichž hodnoty jsme dopočítali pro každý měřený objekt. Platí, že $r_{x_1x_2.x_3...x_k} = r_{E_1E_2}$.

Také se zjišťuje **semiparciální koeficient korelace**. Ten se používá podobně jako parciální koeficient korelace, jestliže chceme posoudit přínos proměnné X_2 k predikci proměnné X_1, kterou již predikujeme pomocí proměnných $X_3, \ldots, X_k$. Proměnná X_1 nyní představuje závisle proměnnou. Počítáme ho tak, že korelujeme proměnnou E_2, jež vznikla jako rozdíl mezi proměnnou X_2 a její predikcí pomocí proměnných $X_3, \ldots, X_k$, s proměnnou X_1. Čtverec tohoto koeficientu určuje hodnotu, o niž se zvětší původní koeficient determinace, který je daný mnohonásobným koeficientem korelace $r_{x_1x_3...x_k}$. Kvůli této výhodné interpretaci mu dáváme přednost.

Je zřejmé, že všechny uvedené korelační koeficienty jsou z intervalu $\langle -1; 1 \rangle$.

10.2 Lineární model, statistické testy a intervalové odhady

Předchozí odstavec se zabýval tím, jak postihnout vztahy mezi proměnnými pro data z výběru pomocí vhodně zvolených lineárních funkcí. Jestliže chceme provést zobecnění na populaci, musíme předpokládat odpovídající typ výběru a platnost určitého pravděpodobnostního modelu. Pokud má navržený model oprávnění, lze provést další zkoumaní dat pomocí statistických testů významnosti a intervalů spolehlivosti. Dále popíšeme model, který se používá při analýze dat pro svoji relativní jednoduchost nejčastěji. Metody ověření modelu popíšeme v příslušném odstavci. Platnost modelu je také základním předpokladem pro využívání získaných predikčních funkcí.

Obecně regresní model vyjadřujeme rovnicí, jež popisuje vztah mezi hodnotou závisle proměnné a hodnotami nezávislých proměnných

$$y = f(x_1, x_2, \ldots, x_k; \beta_0, \beta_1, \ldots, \beta_m) + e,$$

kde e označuje chybu predikce a $\beta_0, \beta_1, \ldots, \beta_m$ jsou parametry modelu.

Lineární model vyjadřujeme rovnicí

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e.$$

Předpokládáme, že chybová komponenta e má normální rozdělení s nulovou střední hodnotou a rozptylem $\sigma^2_{y.x}$, který je stejný pro všechny hodnoty prediktorů. Rozptyl $\sigma^2_{y.x}$ je dalším parametrem modelu. Vztahy mezi proměnnými musí splňovat předpoklad

$$E(Y \mid X_1 = x_1,\ X_2 = x_2,\ \ldots,\ X_k = x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

což znamená, že podmíněná střední hodnota proměnné Y je lineární funkcí hodnot $x_1, x_2, \ldots, x_k$.

Data pro regresní analýzu – matici měření X – lze získat obecně metodami výběru typu I a II, jež jsme pro jednoduchý případ analýzy závislosti vymezili v kapitole o jednoduché lineární regresní analýze (kap. 7.3.3). Ve výběru typu I nekontrolujeme žádnou nezávislou proměnnou – hodnoty měříme na objektech z prostého náhodného výběru. Ve výběru typu II fixně určujeme hodnoty některých nezávislých proměnných.

Odhady parametrů získané metodou nejmenších čtverců $\{b_j\}$ zajišťují optimální proložení dat závisle proměnné pro daný výběr. Nemusí to platit pro celý základní soubor. Získané hodnoty $\{b_j\}$ jsou odhady teoretických parametrů $\{\beta_j\}$ regresní funkce f. Hodnota reziduálního zbytkového rozptylu $s^2_{y.x_1x_2\ldots x_k}$ je odhadem parametru $\sigma^2_{y.x}$.

Podobně jako v jednoduché lineární regresi nulová hodnota mnohonásobného korelačního koeficientu implikuje nulovou hodnotu všech regresních koeficientů $\beta_1, \beta_2, \ldots, \beta_k$. Chceme-li testovat tuto hypotézu, použijeme testovací statistiku F, která má tvar

$$F = \frac{(n-k-1)r^2_{y.x_1x_2\ldots x_k}}{k(1-r^2_{y.x_1x_2\ldots x_k})}.$$

Jestliže F je větší než kritická mez F-rozdělení s k a $n-k-1$ stupni volnosti, lze se přiklonit k hypotéze, že aspoň jeden regresní koeficient je různý od nuly. Statistickou významnost jednotlivých regresních koeficientů testujeme statistikou b_i/SE_b, jež je běžně uvedena ve výstupu programu pro mnohonásobnou regresi.

Tab. 10.4 Tabulka analýzy rozptylu pro případ mnohonásobné lineární regrese

Zdroj variability	S	st.v.	MS	F
model	$S_M = \sum(\hat{y}_i - \bar{y})^2$	k	$MS_M = \frac{S_M}{k}$	$F = \frac{MS_M}{MS_e}$
reziduální	$S_e = \sum(y_i - \hat{y}_i)^2$	$n-k-1$	$MS_e = \frac{S_e}{n-k-1}$	
Celková variabilita	$S_T = \sum(y_i - \bar{y})^2$	$n-1$	$MS_T = \frac{S_T}{n-1}$	

Koeficient SE_b je směrodatná chyba odhadu parametru. Například posuzujeme-li pouze j-tý regresní koeficient, pak jestliže je hodnota b_j/SE_b v intervalu $(-2; +2)$, lze pochybovat, zda příslušná nezávisle proměnná přispívá k predikci závisle proměnné v rámci odhadnuté regresní rovnice. Lépe určíme intervaly spolehlivosti pro vybraný regresní koeficient β_j pomocí intervalu:

$$\left(b_j - tSE_b;\ b_j + tSE_b\right),$$

kde SE_b je směrodatná chyba odhadu koeficientu β_j a t je kritická hodnota t-rozdělení o $n-k-1$ stupních volnosti pro danou hladinu spolehlivosti.

Ve výstupech programů pro mnohonásobnou lineární regresi se objevuje tabulka analýzy rozptylu, která umožňuje sestavení testu globální hypotézy, že všechny regresní koeficienty se rovnají nule. Tento test je ekvivalentní statistickému testu nulovosti mnohonásobného korelačního koeficientu. Tabulku analýzy rozptylu jsme již poznali v kapitole o jednoduché a dvoufaktorové analýze rozptylu. Pro náš případ její tvar uvádí tabulka 10.4. Testovací statistika F z tabulky má za platnosti hypotézy nulovosti všech regresních koeficientů F-rozdělení s k a $n-k-1$ stupni volnosti. Platí vztah

$$F = \frac{(n-k-1)r^2_{y.x_1x_2\ldots x_k}}{k(1-r^2_{y.x_1x_2\ldots x_k})} = \frac{MS_M}{MS_e},$$

přičemž

$$R^2 = r^2_{y.x_1x_2\ldots x_k} = \frac{S_M}{S_T}.$$

Směrodatná chyba při regresi $s_{y.x_1x_2\ldots x_k}$ má následující pravděpodobnostní interpretaci: Jestliže $\hat{y}$ je hodnota vypočtená pomocí regresní rovnice a y je naměřená hodnota, pak při větších rozsazích výběrů nerovnost $|y - \hat{y}| \leq 2s_{y.x_1x_2\ldots x_k}$ platí přibližně v 95 % případů při větším počtu měření.

10.3 Hledání optimální množiny prediktorů

V mnoha situacích chce výzkumník zjistit **optimální podmnožinu prediktorů** závisle proměnné z větší množiny potenciálních prediktorů. K dosažení tohoto cíle se používají různé taktiky. Dnes nejvíce používaná taktika je založena na maximalizaci mnohonásobného koeficientu korelace, a provádí se tak, že se vypočtou regrese všech možných podskupin nezávislých proměnných na Y a pak se postupuje podle následujícího pravidla: Ze všech podskupin o j prediktorech se vybere jako nejlepší ta, která má s Y největší mnohonásobný koeficient korelace. To se provede pro všechny rozsahy podskupin $1 \leq j \leq k$. Z takto vybraných optimálních podskupin se pak vybere ta, jež má statisticky významně nejlepší predikci, neboli má tyto dvě vlastnosti:

- má největší mnohonásobný koeficient korelace ze všech podskupin prediktorů o rozsahu j;
- jestliže k ní přibereme další proměnné, nevede to k významnému zlepšení predikce.

V popsaném algoritmu se využívá statistický test nulového efektu dodatečné proměnné nebo nulového efektu skupiny dodatečných proměnných na závisle proměnnou. Opírá se o testovací F-statistiku, jež má tvar

$$F = \frac{(n-k-1)(r^2_{y.x_1x_2\ldots x_k} - r^2_{y.x_1x_2\ldots x_r})}{(k-r)(1-r^2_{y.x_1x_2\ldots x_k})},$$

kde dodatečná je množina prediktorů $X_{r+1}, \ldots, X_k$, kterou přidáváme k množině prediktorů $X_1, X_2, \ldots, X_r$. Uvedená statistika má za předpokladu nulového přídavného efektu F-rozdělení se stupni volnosti $k-r$ a $n-k-1$.

Algoritmus **postupné regrese** vybírá nejlepší podskupinu prediktorů následujícím způsobem:

a) V prvním kroku vybere jako nejlepší prediktor proměnnou s největším korelačním koeficientem s Y a zařadí ji do tvořené množiny prediktorů.
b) V následujícím kroku se přibere proměnná, která nejlépe zlepšuje predikční mohutnost těch proměnných, které již byly do predikce zařazeny (má největší parciální korelační koeficient s Y).
c) Z predikce je odstraněna ta proměnná, jejíž příspěvek pro predikci Y klesl pod určitou úroveň (její parciální korelační koeficient s Y klesl pod mez významnosti). Přejde se na krok b).

Proces přibírání prediktorů skončí, když již žádný další prediktor významně nezlepší predikci. Tento algoritmus však nevede nutně k nejlepší skupině prediktorů Tímto postupem se obvykle „podaří" seřadit prediktory podle velikosti jejich predikční schopnosti.

V jednodušším algoritmu **dopředná regrese** se vynechává krok c). **Zpětná regrese** znamená, že se postupně odebírají z větší množiny prediktorů ty, jejichž predikční hodnota je malá.

10.4 Předpoklady lineárního modelu

Na začátku je zapotřebí zjistit scházející údaje v matici dat a zkoumat přítomnost extrémních hodnot u jednotlivých proměnných. V průběhu tvorby a ověřování vhodnosti vytvořeného modelu je nutné ověřit pět specifických předpokladů:

1. Reziduální hodnoty $e_i = y_i - \hat{y}_i$ mají normální rozdělení s nulovou střední hodnotou.
2. Rozptyl reziduálních hodnot je stejný pro uvažované rozsahy nezávislých proměnných.
3. Hodnoty predikované proměnné jsou na sobě nezávislé.
4. Vztahy mezi prediktory a závisle proměnnou jsou lineární.
5. Neexistuje multikolinearita mezi prediktory (definice viz dále).

Většinu kontrol jsme popsali v souvislosti s jednoduchou regresní analýzou (kap. 7.3.4). Abychom ověřili tyto předpoklady, musíme specificky provést následující kontroly:

1. Zobrazíme reziduální hodnoty pomocí grafu stonku a listu nebo pomocí normálního grafu a zkontrolujeme normalitu jejich rozdělení.
2. Zobrazíme vztah mezi reziduálními hodnotami a prediktory a zkontrolujeme, zda rozptýlenost reziduálních hodnot je homogenní.
3. Někdy je závislost mezi měřeními závisle proměnné způsobena efektem pořadí, v němž byly objekty měřeny. Zobrazíme reziduální hodnoty proti pořadí měření a kontrolujeme přítomnost rozlišitelné konfigurace nebo cyklu.
4. Zobrazujeme bodové dvojrozměrné grafy závisle a nezávisle proměnné.
5. *Multikolinearita* znamená, že nezávisle proměnné nebo jejich podmnožina jsou vzájemně silně korelovány. Odhady regresních koeficientů jsou pak velice nestabilní – když změníme několik málo hodnot měření, odhady regresních koeficientů se mohou dramaticky změnit. Programy pro výpočet lineární regrese poskytují různé koeficienty pro detekci tohoto jevu (např. koeficient tolerance).

Také zjišťujeme pro tzv. vybočující pozorování a odlehlá pozorování při regresi (Reif, 2000, s. 102), zda mají charakter vlivných bodů. **Vlivné body** jsou takové,

jež podstatně ovlivňují odhady regresních koeficientů. **Vybočující pozorování** jsou nezvyklé konfigurace hodnot týkající se společného rozdělení nezávislých proměnných. **Odlehlé hodnoty** při regresi jsou nápadně velké reziduální hodnoty, upozorňující na špatnou predikci závisle proměnné.

10.5 Aplikační problémy v regresní analýze

Zmíníme některé problémy, které je nutné řešit v průběhu aplikace regresní analýzy. Při návrhu uspokojivého a validního regresního modelu si musíme klást následující otázky:

- Jaké jsou cíle při vytváření regresního modelu?
- Splňují analyzovaná data požadavky, jež vyplývají z těchto cílů? Přitom jde o velikost výběru; o pokrytí populace výběrem; o to, zda se změřily všechny relevantní proměnné a zda jsou rozmezí hodnot nezávislých proměnných dostatečně veliká.
- Které proměnné je potřeba zařadit do modelu? Jaký efekt má nezahrnutí dané proměnné do modelu?
- Jak se vybere ten nejlepší model? Jak se validizuje vytvořený model?

Stručně komentujeme uvedené body:

Cíle statistické analýzy pomocí regresního modelu jsme zmínili na začátku kapitoly. Jestliže nás zajímá jenom predikce, není tak důležité, jak vypadá skutečně správný model, ani se příliš nezajímáme o význam jednotlivých proměnných. Hlavním cílem je predikce s dostatečnou přesností. Při exploraci nezávisle proměnných posuzujeme, které z nich zlepšují predikci závisle proměnné. Tento krok je často předstupeň pro vytvoření validního modelu. Data obvykle vyhovují několika různým modelům a určení optimálního může být obtížným úkolem.

Důležitým krokem v analýze je posouzení **kvality dat**. Zajímáme se, zda rozsah výběru je dostatečný. V kapitole 11 o volbě rozsahu výběru naznačujeme v jejím závěru řešení i pro případ více nezávislých proměnných. Není však důležitý pouze rozsah výběru. Aby byl vytvořený model validní, musí výběr pokrývat celou populaci. Přitom musíme dbát, abychom dostatečně postihli možnou variaci nezávislých proměnných. Také musíme uvážit chyby měření nezávislých proměnných. Jestliže jsou značné, pak jsou vypočítané regresní koeficienty podezřelé, protože nepopisují dobře vztah mezi skutečnými hodnotami nezávisle proměnných a závisle proměnnou. Také je důležité zařadit do vytvářeného modelu všechny významné regresory. Jestliže některá z důležitých proměnných schází, dostáváme nesprávné odhady regresních koeficientů pro zařazené proměnné a výsledná regresní rovnice má slabou kvalitu predikce.