4 ANALÝZA ZÁVISLOSTÍ

V této kapitole se budeme věnovat vztahům mezi dvěma nebo více statistickými znaky. Budeme tedy pro každou statistickou jednotku sledovat více znaků, což nám umožní zabývat se například vztahem mezi vzděláním a výběrem témat filmů, vzděláním a příjmem nebo třeba cenou elektrické energie a její spotřebou. Budeme závislosti vyhledávat, popisovat a zkoumat. Metody, které se k tomu účelu používají, přitom závisejí na cílech analýzy a povaze dat, která zpracováváme. Hlubší zkoumání závislostí z hlediska charakteru popisované skutečnosti a dat, která máme k dispozici, může v některých případech vést k úsudkům o příčinných souvislostech, které nazýváme **kauzálním** vztahem.

Z hlediska metody zkoumání je vhodné rozlišení **pevných** (také **deterministických**) a **volných** (také **stochastických** nebo **statistických**) závislostí. Pevná závislost mezi dvěma veličinami představuje vztah mezi veličinami, podle jejich charakteru buď popisovaný matematickou funkcí, nebo projevující se jednoznačným přiřazením hodnot jedné veličiny hodnotám veličiny jiné. Setkáváme se například se závislostí ceny deseti výrobků na ceně jednoho, se závislostí doby trvání cesty na vzdálenosti a rychlosti pohybu apod. Volná závislost popisuje jakousi tendenci, nikoliv neměnný vztah. Pokud uvažujeme o souvislosti vzdělání a mzdy, můžeme říci, že s vyšším vzděláním roste mzda, tento vztah ale nelze jednoduše popsat (určit mzdu při znalosti vzdělání): musíme mít na paměti, že někteří lidé s vysokým vzděláním mohou pobírat nízkou mzdu, nebo naopak. Obdobně mezi cenou zájezdu a jeho délkou je přímá závislost (s délkou cena roste), nicméně další podrobnosti zájezdu, které cenu ovlivňují, znemožňují výpočet ceny jen na základě délky zájezdu (v takovém případě by závislost byla pevná). Snadno si lze představit, že existují drahé krátké zájezdy, stejně jako levné dlouhé; tendence je ale zřejmá. Všimněme si, že v obou uvedených případech je závislost přímá. Pokud bychom uvažovali závislost ceny ojetého automobilu na jeho stáří či počtu ujetých kilometrů, jednalo by se o závislost nepřímou v tom smyslu, že menší hodnota jedné proměnné přináší spíše větší hodnotu druhé.

V reálných empirických situacích se setkáváme prakticky výhradně s volnými závislostmi. Za obecnými tendencemi projevujícími se v souboru statistických údajů se však mohou skrývat hlubší zákonitosti vztahů mezi veličinami. K poznání a matematickému popisu statistických závislostí, jakož i k ověřování platnosti výzkumných teorií slouží metody **analýzy kontingenčních tabulek, analýzy rozptylu a regresní a korelační analýzy**, kterými se budeme zabývat v této kapitole.

Pro potřeby těchto metod bývá vhodné rozlišit jednostranné a vzájemné závislosti. Jednostrannými závislostmi se zabývá například **regresní analýza**. Jedná se o situaci, kdy proti sobě stojí vysvětlující (nezávisle) proměnná v úloze „příčin" a vysvětlovaná (závisle) proměnná v úloze „následků". V těchto případech bývá zvykem zkoumat obecné tendence ve změnách vysvětlovaných proměnných vzhledem ke změnám vy-

světlujících proměnných. Snahou je odpovědět na otázky, které se týkají formy změn například vysvětlované proměnné y při změnách vysvětlující proměnné x. Vzájemnými (většinou lineárními) závislostmi se zabývá **korelační analýza**. V korelační analýze se klade důraz více na intenzitu (sílu) vzájemného vztahu než na zkoumání závislosti veličin ve směru příčina – následek. Z výpočetních i interpretačních hledisek však dochází ke značnému prolínání obou přístupů, jak uvidíme dále.

V úvodních odstavcích knihy jsme objasnili, že údaje, které ze statistického šetření vyplynou, jsou různého typu a je velmi důležité zvolit vhodný nástroj jejich statistické analýzy. Hovořili jsme o tom, že možnosti jsou dány charakterem dat – počtem jejich hodnot (variant, kategorií) a především typem relací mezi nimi. To samozřejmě platí i pro volbu některé z výše uvedených metod.

Z hlediska počtu variant je nejjednodušším typem alternativní proměnná; obsahuje jen informaci o tom, u které jednotky jsme určitou sledovanou vlastnost zaznamenali a u které nikoliv. Na alternativní lze v případě potřeby převést každou proměnnou: jedna její vybraná varianta je v takovém případě pro analýzu označena jedničkou a všechny ostatní nulou.

Diskrétní kvantitativní proměnné (počet členů domácnosti) nabývají většinou malého počtu hodnot, jejichž výčet lze snadno pořídit. U spojitých kvantitativních proměnných je počet naměřených hodnot v souboru sice konečný, ale (v závislosti na rozsahu souboru) obvykle natolik vysoký, že pro účely třídění je pouhý seznam nevhodný. Jak již víme, hodnoty je v takovém případě nutné roztřídit do skupin – obnáší to jednoduché postupy nastavení stejně či různě širokých intervalů.

Pokud jsou hodnoty (zde spíše varianty či kategorie) proměnné vyjádřeny slovně (nebo číselným kódem) a nemají vlastnost objektivního uspořádání, jde o nominální proměnné; pro kvantitativní analýzu je v první řadě žádoucí, aby variant takové proměnné nebylo příliš mnoho.

Pořadové (ordinální) proměnné mohou vyjadřovat odstupňování určité vlastnosti nebo činnosti (velká nespokojenost, nespokojenost, spokojenost, velká spokojenost), souhlasu (od naprostého nesouhlasu po naprostý souhlas), změny v čase (služby se zhoršily, nezměnily, zlepšily se) preference jedné ze dvou alternativ (preference spíše značky A, žádná preference, preference spíše značky B) apod. Vzhledem k vžité praxi použití lineární stupnice přirozených (nebo celých) čísel pro vyjádření pořadí kategorií (např. od 1 = naprostý nesouhlas po 5 = naprostý souhlas) jsou pořadové proměnné všeobecně vnímány jako měřitelné a je s nimi často nakládáno tak, jako by jejich uspořádané kategorie byly od sebe stejně vzdáleny. Dobře lze ilustrovat riziko takového přístupu například u proměnné vzdělání (1 = základní, 2 = středoškolské bez maturity, 3 = středoškolské s maturitou, 4 = vysokoškolské) či u spojité číselné proměnné, jejíž nestejně široké intervaly nahradíme jejich pořadovými čísly.

Pro varianty hodnot nominálních a ordinálních proměnných se přirozeně používá termín kategorie; označíme je proto jako **kategoriální proměnné**.

4.1 Dvourozměrné třídění dat

Výsledkem **dvourozměrného** třídění (třídění souboru podle dvou proměnných), a sice podle r hodnot (či kategorií) proměnné x a s hodnot (či kategorií) proměnné y, je tabulka dvourozměrného rozdělení četností (viz tabulka 4.1). Obsahuje údaje o hodnotách (či kategoriích) obou proměnných a shrnuje výskyt kombinací hodnot (kategorií) obou proměnných u jednotek v souboru. Podle charakteru proměnných se pak tabulka označuje jako **korelační** (v případě kvantitativních proměnných) nebo **kontingenční** (v případě kategoriálních proměnných). Ta se používá častěji, budeme zde tedy dále používat terminologii týkající se kontingenční tabulky.

Tab. 4.1 Dvourozměrné rozdělení četností

x / y	y_1	y_2	…	y_s	n_{i+}
x_1	n_{11}	n_{12}	…	n_{1s}	n_{1+}
x_2	n_{21}	n_{22}	…	n_{2s}	n_{2+}
…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rs}	n_{r+}
n_{+j}	n_{+1}	n_{+2}	…	n_{+s}	n

Sdružené absolutní četnosti n_{ij} v kontingenční tabulce udávají, kolik je v souboru jednotek, u nichž proměnná x nabývá kategorie x_i, $i = 1, 2, \ldots, r$ a proměnná y kategorie y_j, $j = 1, 2, \ldots, s$. Součty těchto četností v řádcích, okrajové (marginální) četnosti n_{i+}, informují o výskytu jednotlivých kategorií proměnné x v souboru (bez ohledu na výskyt kategorií proměnné y). Popisují tak jednorozměrné rozdělení znaku x. Stejnou informaci z pohledu výskytu jednotlivých kategorií proměnné y (bez ohledu na výskyt kategorií proměnné x) obsahují okrajové četnosti n_{+j}, které jsou součty sloupcovými. Popisují tedy jednorozměrné rozdělení znaku y.

Součet všech sdružených absolutních četností v tabulce, stejně jako součet všech okrajových četností v řádku, a také součet všech okrajových četností v sloupci se rovná rozsahu souboru n, tedy

$$\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r} n_{i+} = \sum_{j=1}^{s} n_{+j} = n. \tag{4.1}$$

Relativní četnosti lze v případě dvourozměrného třídění stanovit několika způsoby:

a) Výsledkem porovnání sdružených absolutních četností s rozsahem souboru jsou **sdružené relativní četnosti** $p_{ij} = n_{ij}/n$. Charakterizují tedy dvourozměrnou strukturu souboru podle obou sledovaných proměnných a jejich součet se rovná jedné. Obvykle se interpretují v procentech, pak je jejich součet 100 %.

b) **Okrajové (marginální) relativní četnosti** $p_{i+} = n_{i+}/n$ (a $p_{+j} = n_{+j}/n$) vyjadřují podíl jednotlivých kategorií proměnné x (nebo proměnné y) v celém souboru a popisují tedy jednorozměrná rozdělení obou proměnných. Platí

$$\sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = \sum_{i=1}^{r} p_{i+} = \sum_{j=1}^{s} p_{+j} = 1 . \qquad (4.2)$$

c) Sdružené absolutní četnosti v kontingenční tabulce je ovšem možné také porovnat s jejich řádkovými, nebo s jejich sloupcovými součty (tedy s četnostmi marginálními) a zjistit tak **relativní četnosti podmíněné**.

- V prvním případě získáme strukturu souboru vzhledem k proměnné y, přičemž x nabývá vždy jedné vybrané kategorie (na i-tém řádku tabulky). Touto kategorií proměnné x je podmíněna struktura proměnné y. Takových podmíněných (řádkových) struktur je tedy celkem r.
- Ve druhém případě získáme strukturu souboru vzhledem k proměnné x, přičemž tato struktura je podmíněna jednou vybranou kategorií proměnné y (v j-tém sloupci tabulky). Takových podmíněných (sloupcových) struktur je celkem s.

Graficky lze výskyt kombinací kategorií dvou proměnných znázornit trojrozměrným grafem; názornější a lépe čitelné však jsou grafy konfrontující podmíněnou strukturu řádků nebo sloupců v tabulce (pruhové či mozaikové grafy).

Pravidla pro sestavení korelační tabulky jsou analogická a analogicky lze také interpretovat její obsah. Poznamenejme, že v případě kvantitativní proměnné o velkém počtu hodnot je nutno postupovat podobně jako u jednorozměrného třídění, tedy přejít na intervaly. Může se to týkat jenom jedné, ale i obou uvažovaných proměnných. Interval hodnot lze chápat jako kategorii ordinální proměnné a opět tedy platí vše, co jsme o konstrukci a obsahu dvourozměrné tabulky uvedli výše.

Příklad 4.1

Kontingenční tabulka 4.2 vznikla tříděním odpovědí osob při zjišťování jejich spokojenosti s členstvím ČR v Evropské unii. V roce 2006 bylo dotázáno 300 a v roce 2016 odpovídalo 200 osob (smyšlená data). Provedeme výpočet sdružených a okrajových relativních četností, a také popíšeme všechna podmíněná rozdělení obou proměnných.

Tab. 4.2 Data k příkladu 4.1

Spokojenost / Rok	−−	−	+	++	Celkem
2006	12	72	171	45	300
2016	28	76	82	14	200
Celkem	40	148	253	59	500

Obě sledované proměnné jsou pořadové a obě nabývají malého počtu hodnot ($r = 2$, $s = 4$). Jednotky, dotázané osoby, tedy byly roztříděny do osmi skupin. Například v roce 2006 bylo dotázáno 300 osob, z nichž 12 vyjádřilo naprostou nespokojenost (− −). V roce 2016 z 200 dotázaných bylo 14 osob naprosto spokojených (++).

Celkem bylo v obou šetřeních 148 nespokojených (−) a 253 spokojených (+), 40 zcela nespokojených a 59 naprosto spokojených atd.

Nejprve provedeme výpočet sdružených relativních četností, tedy všechny četnosti v tabulce budeme dělit rozsahem souboru 500. Výslednou strukturu souboru (v %) obsahuje tabulka 4.3.

Tab. 4.3 Sdružené relativní četnosti v procentech pro příklad 4.1

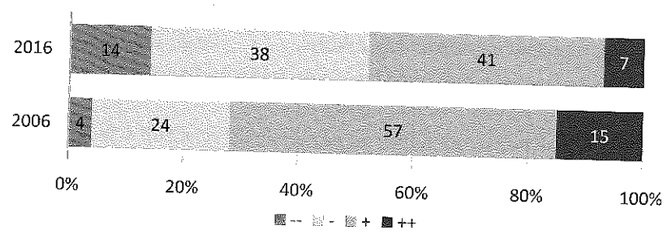
Spokojenost / Rok	−−	−	+	++	Celkem
2006	2,4	14,4	34,2	9,0	60
2016	5,6	15,2	16,4	2,8	40
Celkem	8	29,6	50,6	11,8	100

Sdružené absolutní (i relativní) četnosti ve druhém sloupci tabulek 4.2 a 4.3 jsou podobné, absolutní i relativní četnost v roce 2006 je poněkud nižší. Přesto nelze říci, že podíl nespokojených v obou letech zůstal zhruba stejný, že v roce 2016 jen mírně vzrostl. Liší se totiž počty dotázaných v obou letech, řádkové součty. Proveďme výpočet řádkových relativních četností (dělením součtem 300 v prvním řádku a 200 ve druhém řádku). Strukturu odpovědí v obou letech obsahuje tabulka 4.4 (v %). Z druhého sloupce tabulky je zřejmé, že se podíl nespokojených v roce 2016 významně zvýšil.

Tab. 4.4 Řádkové relativní četnosti v procentech pro příklad 4.1

Spokojenost / Rok	−−	−	+	++	Celkem
2006	4	24	57	15	100
2016	14	38	41	7	100
Celkem	8	29,6	50,6	11,8	100

V pruhovém grafu na obrázku 4.1 lze porovnat strukturu odpovědí v obou letech.



Obr. 4.1 Pruhový graf pro příklad 4.1

Odlišná struktura řádků tabulky podle kategorií sloupcové proměnné odpovídá různým kategoriím řádkové proměnné (a naopak). Zjevně se tak projevuje vztah **asociace obou proměnných**. V následujících odstavcích ukážeme, jak změřit sílu asociace a položíme si také otázku, jak ověřit existenci takového vztahu, pokud do dvourozměrné tabulky byla roztříděna data, pocházející z výběrového šetření.

■

4.1.1 Měření asociace dvou kategoriálních proměnných

Změnou struktury souboru z hlediska jedné proměnné při změně kategorie druhé proměnné v tabulce se projevuje souvislost (asociace) obou veličin. Kontingenční tabulky jsou tak východiskem při zkoumání této souvislosti.

Nezávislosti dvojice sledovaných proměnných se tedy projevuje tak, že podmíněná struktura v řádcích se nemění a shoduje se se strukturou řádku součtového. Totéž lze zároveň říci i o sloupcích tabulky, jejichž struktura se rovněž shoduje. Platí tedy (pro $i = 1, 2, \ldots, r$ a $j = 1, 2, \ldots, s$)

$$\frac{n_{ij}}{n_{i+}} = \frac{n_{+j}}{n}, \quad \frac{n_{ij}}{n_{+j}} = \frac{n_{i+}}{n}, \tag{4.3}$$

a odtud

$$n_{ij} = \frac{n_{i+} n_{+j}}{n}. \tag{4.4}$$

Skutečnost, že se při změně kategorie jedné proměnné nemění podmíněné rozdělení druhé proměnné, se označuje jako statistická nezávislost. Asociace kategoriálních proměnných je obecně považována za tím slabší, čím více se přibližuje statistické nezávislosti, a za tím silnější, čím více se blíží pevné závislosti, kdy každé kategorii jedné proměnné je jednoznačně přiřazena jediná kategorie druhé proměnné. Vzájemně jednoznačné přiřazení dvojic kategorií proměnných (symetrická asociace) ovšem předpokládá jejich stejný počet, jinak řečeno čtvercovou kontingenční tabulku. V obdélníkové tabulce, je-li např. počet řádků nižší než počet sloupců ($r < s$), mohou být sloupcové kategorie jednoznačně přiřazeny řádkovým (ve smyslu asymetrické asociace y na x), v opačném směru však nikoliv.

Pro měření síly asociace byla navržena řada statistik různého typu. V tomto textu se omezíme pouze na několik nejznámějších.

Ze vztahu (4.4) plyne, že odchylku od nezávislosti vyjadřují v jednotlivých políčkách kontingenční tabulky rozdíly $n_{ij} - n_{i+}n_{+j}/n$; v souhrnu za celou tabulku pak Pearsonova statistika G,

$$G = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n} = n\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}}. \tag{4.5}$$

Hodnotu této statistiky však ovlivňuje nejen rozsah souboru n, ale také velikost kontingenční tabulky, což znesnadňuje její interpretaci. Pro měření asociace se proto používají její jednoduché funkce, například **Pearsonův koeficient kontingence** C

$$C = \sqrt{\frac{G}{G+n}}, \tag{4.6}$$

nebo **Cramérův koeficient** V

$$V = \sqrt{\frac{G}{n(m-1)}}, \quad m = \min(r, s). \tag{4.7}$$

Oba tyto koeficienty odpovídají požadavku, aby v případě nezávislosti proměnných nabývaly hodnoty 0. V případě nenulové asociace se však jejich hodnoty liší:

- Pearsonův koeficient kontingence se v případě pevné závislosti blíží k jedné, přičemž stupeň tohoto přiblížení je tím větší, čím větší je kontingenční tabulka. Tak například pro nejmenší (čtyřpolní) tabulku je jeho maximálně možná hodnota 0,707, u tabulky 5×4 je toto maximum 0,866 atd. To ovšem poněkud komplikuje interpretaci hodnot tohoto koeficientu. Na druhou stranu je třeba říci, že u reálných dat se koeficient kontingence nejčastěji nachází v dolní polovině příslušného intervalu možných hodnot.
- Cramérův koeficient nabývá hodnoty v intervalu od 0 do 1. Maxima ovšem může nabýt i v případě obdélníkové kontingenční tabulky, kdy, jak už víme, se v jednom směru o jednoznačné přiřazení kategorií obou proměnných, a tedy o pevnou závislost, nejedná.

U čtyřpolní tabulky je Cramérův koeficient vždy vyšší než Pearsonův (a dává se mu přednost). U větších tabulek je tomu většinou naopak. Nejmenší kontingenční tabulka o čtyřech polích je výsledkem třídění souboru podle dvou alternativních proměnných. Statistika G (4.5) se v takovém případě zjednoduší na tvar

$$G = \frac{n(nn_{11} - n_{1+}n_{+1})^2}{n_{1+}n_{+1}n_{2+}n_{+2}} = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{+1}n_{2+}n_{+2}} \tag{4.8}$$

a Cramérovo V lze psát jako

$$V = \frac{(n_{11}n_{22} - n_{12}n_{21})}{\sqrt{n_{1+}n_{+1}n_{2+}n_{+2}}}. \tag{4.9}$$

Příklad 4.2

Kontingenční tabulka 4.5 vznikla tříděním odpovědí osob při zjišťování jejich zájmu o problémy životního prostředí (velký, průměrný, malý) a vzdělání (základní, středoškolské, vysokoškolské). Provedeme výpočet měr asociace zájmu o životní prostřední a vzdělání (4.6) a (4.7).

Tab. 4.5 Data k příkladu 4.2

Zájem o ŽP / Vzdělání	velký	průměrný	malý	Celkem
Základní	70	32	28	130
Středoškolské	130	64	24	218
Vysokoškolské	72	48	32	152
Součet	272	144	84	500

Nejprve určíme v jednotlivých polích tabulky četnosti očekávané v případě shody všech řádkových (a také sloupcových) podmíněných rozdělení. Tak například pro kategorie vzdělání = základní a zájem = velký bude očekávaná četnost počítána jako

$$\frac{130 \cdot 272}{500} = 70{,}72\,,$$

pro kategorie vzdělání = základní a zájem = průměrný bude očekávaná četnost počítána jako

$$\frac{130 \cdot 144}{500} = 37{,}44$$

atd. Povšimněme si, že okrajové četnosti musí odpovídat původní kontingenční tabulce (viz tabulky 4.5 a 4.6).

Tab. 4.6 Očekávané četnosti k příkladu 4.2

Zájem o ŽP / Vzdělání	velký	průměrný	malý	Celkem
Základní	70,72	37,44	21,84	130
Středoškolské	118,60	62,78	36,62	218
Vysokoškolské	82,68	43,78	25,54	152
Součet	272	144	84	500

Dosazením do sčítanců v (4.5) vypočteme jednotlivé složky statistiky G.

Tab. 4.7 Výpočet statistiky G v příkladu 4.2

Zájem o ŽP / Vzdělání	velký	průměrný	malý	Celkem
Základní	0,01	0,79	1,74	2,54
Středoškolské	1,10	0,02	4,35	5,47
Vysokoškolské	1,38	0,41	1,63	3,42
Součet	2,49	1,22	7,72	11,43

Statistika G nabývá tedy hodnoty 11,43. Pearsonův koeficient kontingence potom určíme jako

$$C = \sqrt{\frac{11{,}43}{11{,}43 + 500}} = 0{,}150$$

a Cramérovo V jako

$$V = \sqrt{\frac{11{,}34}{500 \cdot 2}} = 0{,}106\,.$$

Vzhledem k nízkým hodnotám obou koeficientů se jedná o závislost slabou.

■

4.1.2 Test nezávislosti v kontingenční tabulce

Předpokládejme nyní, že údaje v kontingenční tabulce jsme získali tříděním údajů o jednotkách, které tvoří vzorek rozsáhlé populace, pořízený prostým náhodným výběrem. Ověření existence souvislosti mezi dvojicemi kategoriálních proměnných v populaci je obvykle prvním krokem analýzy vztahů mezi nimi.

Vyslovme hypotézu, že veličiny X a Y asociovány nejsou. V tom případě se populační podmíněná rozdělení shodují. Uspořádání četností v kontingenční tabulce však mohou být od uspořádání odpovídající nezávislosti odchýlena, a to:

- jen náhodně v důsledku skutečnosti, že vztah mezi proměnnými posuzujeme pouze na základě vzorku,
- systematicky, neboť proměnné jsou asociovány.

Je proto nutné provést výpočet testové statistiky se známým rozdělením a na zvolené hladině významnosti rozhodnout, zda lze odchýlení uspořádání kontingenční tabulky od hypotézy nezávislosti ještě považovat za náhodné, či zda již hypotéze o nezávislosti neodpovídá, a tedy je třeba tuto hypotézu zamítnout.

Odchylky četností zjištěných tříděním vzorku a četností očekávaných při platnosti testované hypotézy o nezávislosti, zahrnuje statistika G z (4.5). Tato statistika má přibližně χ^2 rozdělení s $(r-1)(s-1)$ stupni volnosti, a je proto často používaným testovým kritériem. Postup se nazývá **chí-kvadrát test nezávislosti v kontingenční tabulce** a je zvláštním případem testu dobré shody, kterým jsme se zabývali v části 3.4.5. Zvolíme-li hladinu významnosti α, kritický obor má tvar

$$W_\alpha = \{g;\, g \geq \chi^2_{1-\alpha}\} = \langle \chi^2_{1-\alpha}, \infty \rangle,$$

viz tabulka 4.8 (symbol g zde značí hodnotu testového kritéria G).

Z třetí kapitoly již víme, že rozdělení chí-kvadrát testového kritéria je pouze přibližné. Pro zajištění přijatelné aproximace rozdělení této statistiky při určitém počtu polí v kontingenční tabulce se zpravidla vyžaduje takový rozsah výběru n, aby očekávané (teoretické) četnosti vesměs dosahovaly hodnoty alespoň 5. Vzhledem k častým praktickým obtížím byla tato podmínka mnohokrát ověřována, což vedlo ke zjištění, že ji lze sice poněkud oslabit, nicméně menších než 5 by mělo být maximálně 20 % z očekávaných četností (a každá v takovém případě musí být alespoň jednotková).

V případě nedodržení této podmínky nelze považovat aproximaci za vyhovující. Často doporučovaným řešením je zmenšení počtu kategorií proměnných jejich slučováním. Kromě věcných námitek je však třeba rovněž uvážit, že takový postup může ovlivnit samotný výsledek testu, neboť rozměry tabulky určují parametr rozdělení testového kritéria. Lépe je v případě nevyhovujícího rozsahu výběru zvážit použití jiného vhodného testu (např. Fisherova přesného testu).

Tab. 4.8 Test nezávislosti v kontingenční tabulce

H_0	H_1	Testové kritérium	Kritický obor
znaky jsou nezávislé	existuje závislost	$G=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n}$ $G\approx\chi^2\left((r-1)(s-1)\right)$	$W_\alpha=\{g;\, g\geq\chi^2_{1-\alpha}\}$

Příklad 4.3

V tabulce 4.5 s daty z příkladu 4.2 provedeme test nezávislosti obou znaků.

Již víme, že testové kritérium G nabývá hodnoty 11,43. Náhodný výběr je přitom dostatečně velký, neboť všechny očekávané četnosti v tabulce 4.6 jsou větší než 5. Rozdělení testového kritéria je tedy přibližně chí-kvadrát, počet stupňů volnosti je roven $(3-1)\cdot(3-1)=4$. Pro hladinu významnosti 0,05 je kritická hodnota rovna 9,49 ($\chi^2_{0,95}(4)=9,49$) a kritický obor $W_{0,05}$ má tvar $\langle 9,49;\infty)$. Hodnota testového kritéria je prvkem kritického oboru, a tedy hypotézu o nezávislosti znaků lze zamítnout. Usoudíme, že existuje závislost mezi vzděláním a zájmem o životní prostředí.

Chí-kvadrát test nezávislosti v kontingenční tabulce je běžnou součástí statistického softwaru. Pro účely jeho vyhodnocení v takovém případě poznamenejme, že p-hodnota uvedeného testu se stanoví (podle (3.74)) jako

$1-F(11,43)=1-0,978=0,022,$

kde F je distribuční funkce rozdělení $\chi^2(4)$. p-hodnota je menší než 0,05 a závěr pro zvolenou hladinu významnosti je samozřejmě týž, můžeme usoudit, že existuje statisticky významná asociace mezi vzděláním a zájmem o životní prostředí. V tomto případě také říkáme, že existence asociace byla potvrzena.

■

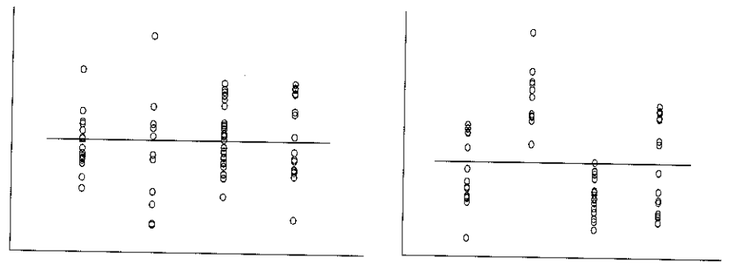
4.2 Použití analýzy rozptylu pro analýzu závislostí

V předchozí kapitole této knihy byla probrána jednofaktorová analýza rozptylu (ANOVA), která umožňuje testovat shodu středních hodnot na základě nezávislých náhodných výběrů z normálního rozdělení. Tento postup můžeme použít i pro zkoumání závislostí mezi jednou proměnnou s několika málo variantami hodnot nebo kategorií (budeme ji také někdy nazývat **faktorem**) a jednou spojitou náhodnou veličinou s normálním rozdělením. Spojitou proměnnou označíme Y a faktor označíme X, toto značení budeme dodržovat dále v celé této kapitole. Představme si například, že chceme posoudit závislost ceny bytu (Y v Kč) na počtu pokojů (X). V tomto případě je proměnná X číselná s diskrétním rozdělením. Pokud bychom (jako v příkladu 3.15) zkoumali závislost doby do doručení zásilky na způsobu dopravy, je faktorem kvalitativní proměnná se třemi kategoriemi (I, II, III).

Všimněme si, že uvažovaná závislost je jednostranná, ptáme se, zda střední hodnota spojité náhodné veličiny Y je stejná pro všechny hodnoty (kategorie) faktoru X. ANOVA (část 3.4.4) umožňuje posoudit existenci závislosti mezi sledovanými znaky. Nulovou hypotézu nezávislosti obou znaků si můžeme představit jako

H_0: střední hodnota veličiny Y nezávisí na hodnotě faktoru X. (4.10)

Alternativou je, že existuje závislost mezi oběma znaky, tedy že střední hodnoty spojité veličiny nejsou stejné pro všechny hodnoty faktoru a aspoň jedna je jiná než ostatní. V našem případě řekneme, že cena bytu závisí na počtu pokojů, a představíme si, že střední ceny nejsou stejné pro všechny možné počty pokojů. Pokud tedy zamítneme nulovou hypotézu (4.10), usoudíme, že existuje statisticky významná závislost mezi Y a X. Na obrázku 4.2 jsou hodnoty spojité náhodné veličiny (svislá osa) roztříděny podle faktoru se čtyřmi hodnotami (vodorovná osa). V levé části obrázku neexistuje závislost, v pravé části závislost existuje a můžeme ji formulovat tak, že hodnoty ve druhé skupině jsou spíše větší a ve třetí skupině spíše menší, než v první a čtvrté skupině.



Obr. 4.2 Volná závislost mezi jedním faktorem a jednou spojitou veličinou

Z třetí kapitoly již víme, že rozhodnutí o hypotéze nezávislosti mezi oběma proměnnými lze založit na rozkladu součtu čtverců spojité veličiny Y. Nejdříve zavedeme podrobnější značení

- k počet možných hodnot faktoru X,
- n rozsah výběru,
- n_j počet pozorování pro j-tou hodnotu faktoru ($j=1, 2, \ldots, k$),

y_{ji} hodnota i-tého pozorování proměnné Y ($i = 1, 2, \ldots, n_j$) pro j-tou hodnotu faktoru ($j = 1, 2, \ldots, k$),

$\bar{y}_j$ průměr v j-té skupině, $\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ji}$,

$\bar{y}$ průměr proměnné Y bez ohledu na hodnotu faktoru, $\bar{y} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ji}$.

Označíme S_y součet čtverců odchylek hodnot proměnné Y od průměru $\bar{Y}$. Jedná se tedy o vyjádření celkové variability proměnné Y v souboru n hodnot a platí

$$S_y = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(y_{ji} - \bar{y}\right)^2. \tag{4.11}$$

Tento součet rozdělíme na součet čtverců mezi skupinami popsanými hodnotami faktoru

$$S_{y.m} = \sum_{j=1}^{k} n_j \left(\bar{y}_j - \bar{y}\right)^2 \tag{4.12}$$

a součet čtverců uvnitř těchto skupin

$$S_{y.v} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(y_{ji} - \bar{y}_j\right)^2. \tag{4.13}$$

Všimněme si, že podobný problém byl zkoumán také v první kapitole této knihy v souvislosti s vlastnostmi rozptylu. Postup při provedení testu hypotézy (4.10) lze shrnout buď do tabulky analýzy rozptylu 4.9, která je nejčastěji používána v analýze rozptylu v softwaru, nebo do tabulky 4.10.

Tab. 4.9 Tabulka analýzy rozptylu

Zdroj variability	Stupně volnosti	Součet čtverců	Průměrný čtverec	F	p-hodnota
Mezi skupinami	$k-1$	$S_{y.m}$	$S_{y.m}/(k-1)$	$\frac{S_{y.m}/(k-1)}{S_{y.v}/(n-k)}$	
Uvnitř skupin	$n-k$	$S_{y.v}$	$S_{y.v}/(n-k)$		
Celkem	$n-1$	S_y			

Tabulka 4.9 obsahuje vše, co potřebujeme k provedení testu. Ve sloupci „Stupně volnosti" jsou uvedeny počty stupňů volnosti rozdělení testového kritéria F za platnosti nulové hypotézy. Další sloupec obsahuje součty čtverců podle (4.11)–(4.13) a je vidět, že skutečně platí

$$S_y = S_{y.m} + S_{y.v}.$$

Ve sloupci nadepsaném „F" je vypočtena hodnota testového kritéria, poslední sloupec obsahuje p-hodnotu, kterou lze nalézt jako doplňkovou pravděpodobnost k distribuční

funkci Fisherova-Snedecorova rozdělení se stupni volnosti $k-1$ a $n-k$ v bodě vypočítaného testového kritéria. Pokud chceme rozhodnout pomocí kritického oboru, použijeme podle tabulky 4.10 kritický obor

$$W_\alpha = \{F; F \geq F_{1-\alpha}\}.$$

Proti hypotéze nezávislosti tedy hovoří velké hodnoty testového kritéria F. Pokud zamítneme hypotézu nezávislosti, závislost mezi znaky označíme za statisticky významnou na zvolené hladině významnosti α. Kvantil $F_{1-\alpha}(k-1, n-k)$ je tedy kritickou hodnotou pro test hypotézy nezávislosti.

Tab. 4.10 Test nezávislosti

H_0	H_1	Testové kritérium	Kritický obor
$\mu_1 = \mu_2 = \ldots = \mu_k$	neplatí H_0	$F = \frac{S_{y.m}/(k-1)}{S_{y.v}/(n-k)}$ $F \sim F(k-1, n-k)$	$W_\alpha = \{F; F \geq F_{1-\alpha}\}$

V předchozím výkladu jsme uvedli, že nás zajímá také síla závislosti. Všimněme si, že jsme nepředpokládali nějaké uspořádání hodnot faktoru, proto není možné sledovat také směr závislosti proměnných X a Y. Intenzitu závislosti mezi oběma znaky popíšeme **poměrem determinace** P^2, který je definován jako poměr součtu čtverců mezi skupinami $S_{y.m}$ a celkového součtu čtverců S_y

$$P^2 = \frac{S_{y.m}}{S_y}. \tag{4.14}$$

Pracujeme-li s tabulkou analýzy rozptylu 4.9, hodnoty součtů čtverců jsou obsaženy v jejím druhém sloupci. Hodnoty koeficientu determinace leží v intervalu $\langle 0,1\rangle$, čím větší je hodnota poměru determinace, tím silnější je zkoumaná závislost, a tedy tím více se liší střední hodnoty spojité proměnné pro jednotlivé hodnoty faktoru. Hodnoty jedna koeficient determinace dosahuje v případě, že pro jednotlivé hodnoty faktoru jsou hodnoty spojité proměnné stejné (nulová variabilita proměnné uvnitř skupin) a tyto hodnoty jsou různé pro různé hodnoty faktoru. Naopak poměr determinace nabývá hodnoty nula tehdy, když aritmetické průměry jsou stejné pro všechny hodnoty faktoru. Taková situace naznačuje také rovnost (nebo aspoň blízkost) středních hodnot veličiny Y pro všechny hodnoty faktoru X.

Příklad 4.4

Majitel firmy zkouší tři postupy, jak doručit zboží zákazníkům. Náhodně vybral 35 zakázek, rozdělil je podle způsobů dopravy a zjistil přesnou dobu do jejich doručení. Budeme se zabývat otázkou, zda existuje závislost mezi způsobem dopravy a dobou potřebnou k doručení, tedy zda všechny způsoby dopravy jsou v průměru stejně rychlé, nebo některý z nich je rychlejší nebo pomalejší. Z pozorovaných dob byly nalezeny

skupinové charakteristiky uvedené v tabulce 4.11. Průměrné hodnoty jsou přibližně od 59 do 79 minut, analýza rozptylu umožňuje posoudit, zda pozorované hodnoty průměrů svědčí proti rovnosti skupinových středních hodnot, nebo zda se jedná jen o náhodný výkyv. Podíl nejvyšší a nejnižší směrodatné odchylky je menší než dva (13,16/6,70 = 1,964), proto je splněna přibližná podmínka použití testu. Pokud bychom použili Bartlettův test homogenity rozptylů (část 3.4.4), hypotézu o rovnosti rozptylů ve všech třech skupinách nezamítneme (p-hodnota je rovna 0,107).

Tab. 4.11 Popisné charakteristiky doby do doručení

Způsob dopravy	n	Průměr $\bar{y}_j$	Směrodatná odchylka s_j
I	12	58,91	9,73
II	9	65,11	13,16
III	14	78,97	6,70

Hypotéza nezávislosti obou znaků tedy má tvar

H_0: střední doba do doručení zásilky nezávisí na způsobu dopravy,
neboli

H_0: $\mu_1 = \mu_2 = \mu_3$,

kde $\mu_j, j = 1, 2, 3$ jsou střední doby do doručení zásilky při způsobech doručení I, II a III. Průběh testu a dosažené výsledky jsou shrnuty v tabulce analýzy rozptylu 4.12 (podle obecné tabulky 4.9). Na základě p-hodnoty, která je podle tabulky menší než 0,000 1, zamítáme nulovou hypotézu, že neexistuje závislost mezi způsobem dopravy a dobou doručení. Hodnotu poměru determinace určíme jako

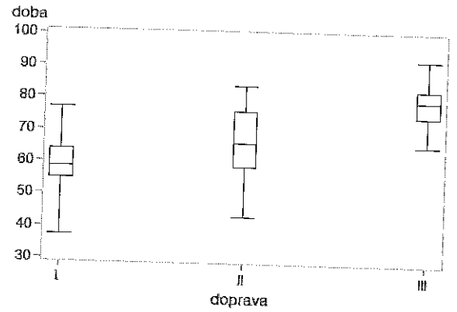
$$P^2 = \frac{2741,12}{5749,61} = 0,477,$$

tedy závislost je (statisticky) významná, její intenzita je středně silná.

Tab. 4.12 Tabulka analýzy rozptylu pro příklad 4.4

Zdroj variability	Stupně volnosti	Součet čtverců	Průměrný čtverec	F	p-hodnota
Mezi skupinami	2	2 741,12	1 370,55	14,58	<0,000 1
Uvnitř skupin	32	3 008,49	94,02		
Celkem	34	5 749,61			

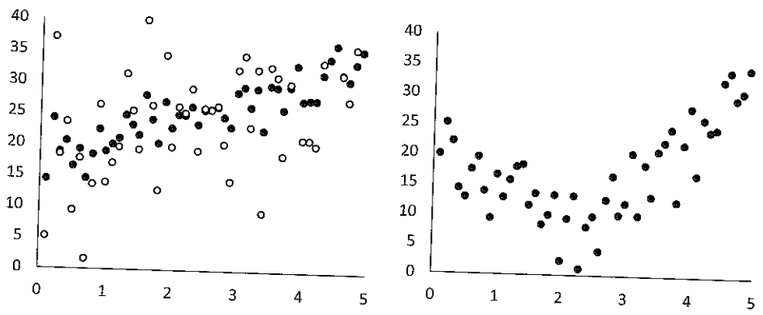
O povaze závislosti vypovídá obrázek 4.3, který obsahuje krabičkové grafy pro jednotlivé způsoby dopravy. Podle LSD testu (část 3.4.4) jsou nerozlišitelné způsoby doručení 1 a 2, třetí způsob je zřejmě pomalejší.



Obr. 4.3 Krabičkové grafy pro data z příkladu 4.4

4.3 Regresní a korelační analýza

V této části se budeme zabývat závislostmi mezi dvěma a více kvantitativními proměnnými, v části 4.3.7 se stručně zmíníme také o zahrnutí kvalitativních proměnných. Na rozdíl od předchozí kapitoly se tyto závislosti budeme snažit popsat vhodnými matematickými funkcemi. Na obrázku 4.4 jsou znázorněny tři formy závislosti mezi dvěma znaky. Závislosti v levé části obrázku mají zřejmě přibližně lineární průběh, liší se však ve velikosti kolísání. Závislost, znázorněná v pravé části obrázku, je nelineární. Pokud nás bude zajímat intenzita závislosti, závislost mezi hodnotami znázorněnými plnými kolečky v levé části obrázku je zřejmě silnější, než závislost znázorněná prázdnými kolečky.



Obr. 4.4 Příklady volné závislosti mezi dvěma znaky

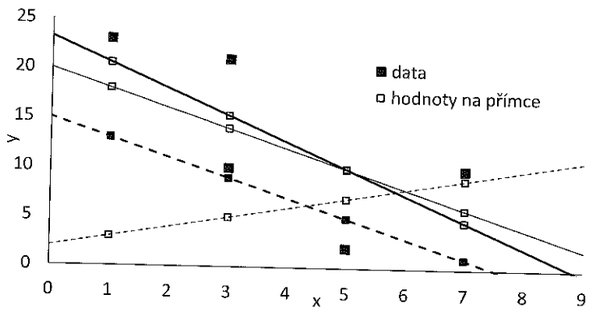
Předpokládejme nyní, že jsme pozorovali n dvojic hodnot (x_i, y_i) dvou znaků x a y. Hodnoty zobrazíme v grafu a pokusíme se vystihnout závislost mezi znaky znázorněnou pozorovanými body nějakou funkcí, kterou budeme nazývat **regresní funkce**. Při konstrukci obrázku jsme se museli rozhodnout, kterou proměnnou umístíme na vodorovnou osu a kterou na svislou osu, závislost popsaná regresní funkcí je tedy **jednostranná**. Proměnnou x, jejíž hodnoty nanášíme na vodorovnou osu, budeme nazývat vysvětlující proměnná (nebo také nezávisle proměnná). Proměnnou y budeme nazývat vysvětlovaná proměnná (nebo také závisle proměnná).

4.3.1 Přímková regrese

Nejjednodušším a nejčastěji používaným typem regresní funkce je přímka. Regresní funkci $\eta(x)$, popisující regresní přímku, zapíšeme ve tvaru

$$\eta(x; \beta_0, \beta_1) = \beta_0 + \beta_1 x, \tag{4.15}$$

kde jsme ještě vyznačili parametry, na kterých funkce závisí a které je třeba zvolit nebo nalézt. Parametr β_0 je **posunutí přímky** a je roven hodnotě na přímce v bodě $x = 0$, tedy průsečíku přímky se svislou osou. Parametr β_1 se nazývá **směrnice přímky** a popisuje její sklon vzhledem k vodorovné ose.



Obr. 4.5 Znázornění prokládání dat regresní přímkou

Na obrázku 4.5 je situace znázorněna pro $n = 5$ bodů. Ze čtyř zobrazených regresních přímek jsou dvě jasně nevhodné (čárkované čáry). Rozhodnout, která ze dvou zbývajících přímek lépe vystihuje pozorované body, je ale těžké. Proto potřebujeme nějaké kritérium, jenž nám umožní vybrat přímku, která bude závislost mezi x a y vystihovat co nejlépe.

Pro popis závislosti zvolíme takovou přímku, pro kterou součet Q čtverců odchylek pozorování y_i od odpovídajících hodnot na přímce $\beta_0 + \beta_1 x_i$ nabývá minimální hodnoty, tj. hledáme minimum výrazu

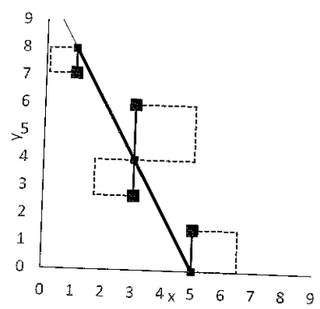
$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{4.16}$$

Metodu nalezení neznámých parametrů přímky nazýváme **metodou nejmenších čtverců** a volíme takové parametry β_0 a β_1, pro které nabývá výraz minimální hodnoty. V češtině používáme pro tuto metodu zkratku MNČ, někdy také zkratku z angličtiny OLS (**O**rdinary **L**east **S**quares).

Uvažujme například přímku s rovnicí $y = 10 - 2x$ a čtyři body uvedené v tabulce 4.13. Veličina Q je definována jako součet ploch čtverců znázorněných na obrázku 4.6. Stranou čtverce jsou vždy odchylky mezi pozorováními x_i a body na regresní přímce $\eta(x)$. Postupným výpočtem podle definice (4.16) dostaneme hodnotu Q jako součet hodnot v posledním sloupci v tabulce 4.13, tedy $Q = 8{,}75$.

Tab. 4.13 Výpočet součtu čtverců odchylek Q

x_i	y_i	$10 - 2x_i$	$y_i - 10 + 2x_i$	$(y_i - 10 + 2x_i)^2$
1	7,1	8	0,9	0,81
3	6,0	4	2,0	4,00
3	2,7	4	1,3	1,69
5	1,5	0	1,5	2,25
Součet				8,75



Obr. 4.6 Výpočet součtu čtverců odchylek Q

Pro přímky znázorněné na úvodním obrázku 4.5 jsou hodnoty kritéria Q po řadě 707 (rostoucí čárkovaná), 335 (klesající čárkovaná), 170 (plná čára slabá) a 159,25 (plná čára silná). Nejvhodnější z nich je tedy přímka nakreslená tlustou čárou.

Nevíme ale, zda neexistuje ještě nějaké lepší řešení. Proto odvodíme obecný vzorec pro výpočet hodnot parametrů b_0 a b_1 takových, že Q je pro tyto hodnoty minimální, a tedy přímka $b_0 + b_1 x$ bude minimalizovat součet čtverců (4.16). K tomu účelu nejprve najdeme parciální derivace kritéria Q podle obou parametrů

$$\frac{\partial Q}{\partial \beta_0}(\beta_0,\beta_1) = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-1),$$

$$\frac{\partial Q}{\partial \beta_1}(\beta_0,\beta_1) = 2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)(-x_i).$$

Pokud obě derivace položíme rovné nule a použijeme pro parametry symboly b_0 a b_1 namísto parametrů β_0 a β_1, dostaneme dvě **normální rovnice**, jejichž řešením získáme obě hledané hodnoty, které jsou parametry představujícími nejlepší volbu směrnice a posunutí z hlediska zvoleného kritéria Q.

Je třeba řešit dvě rovnice pro dvě neznámé b_0 a b_1

$$\frac{\partial Q}{\partial \beta_0}(b_0,b_1) = 2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)(-1) = 0,$$

$$\frac{\partial Q}{\partial \beta_1}(b_0,b_1) = 2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)(-x_i) = 0.$$

Po algebraických úpravách dostaneme rovnice (dále v nich už nebudeme uvádět meze sčítání, sčítat budeme vždy přes všechna pozorování, tedy $i = 1, 2, \ldots, n$)

$$\sum y_i = nb_0 + b_1 \sum x_i, \tag{4.17}$$

$$\sum y_i x_i = b_0 \sum x_i + b_1 \sum x_i^2, \tag{4.18}$$

jejichž řešením je

$$b_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum y_i x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}, \tag{4.19}$$

$$b_1 = \frac{n\sum y_i x_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}. \tag{4.20}$$

V praktických úlohách používáme k výpočtu optimální hodnoty posunutí a směrnice statistické programy. Užitečným a univerzálním postupem je také maticové vyjádření řešení, který zavedeme v dalším textu.

Hodnotu na regresní přímce označíme Y, tedy

$$Y = \eta(x; b_0, b_1) = b_0 + b_1 x, \tag{4.21}$$

pro pozorované body (menší čtverečky na obrázku 4.5) pak

$$Y_i = \eta(x_i; b_0, b_1) = b_0 + b_1 x_i.$$

Rozdíl mezi hodnotou y_i a hodnotou na regresní funkci Y_i nazveme **reziduum** a označíme ho e_i. Je tedy

$$e_i = y_i - Y_i,\ i = 1, 2, \ldots, n.$$

Vzorce pro výpočet směrnice a posunutí regresní přímky (4.19) lze s výhodou přepsat do jednoduššího tvaru obsahujícího průměry $\bar{x}$ a $\bar{y}$, kovarianci s_{xy} mezi x a y, danou vztahem

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{n}{n-1}\left(\overline{xy} - \bar{x}\,\bar{y}\right), \tag{4.22}$$

a výběrový rozptyl s_x^2, tj.

$$b_1 = \frac{s_{xy}}{s_x^2}, \tag{4.23}$$

$$b_0 = \bar{y} - b_1 \bar{x}. \tag{4.24}$$

Výsledná regresní přímka (4.21) má po úpravě tvar

$$Y = \bar{y} + b_1(x - \bar{x}). \tag{4.25}$$

Pro každé $i = 1, 2, \ldots, n$ tedy máme pozorovanou hodnotu y_i, hodnotu na regresní funkci Y_i a reziduum e_i. Při tomto značení můžeme minimální hodnotu součtu čtverců Q zapsat jako

$$Q(b_0, b_1) = \sum_{i=1}^{n}(y_i - Y_i)^2 = \sum_{i=1}^{n} e_i^2. \tag{4.26}$$

Tab. 4.14 Výpočet kritéria Q pro body na obrázku 4.5

i	x	y	xy	x^2	Y	e_i	e_i^2
1	1	23	23	1	20,5769	2,4230	5,8713
2	3	21	63	9	15,3076	5,6923	32,4023
3	3	10	30	9	15,3076	−5,3076	28,1716
4	5	2	10	25	10,0384	−8,0384	64,6168
5	7	10	70	49	4,7692	5,2307	27,3609
Součet	19	66	196	93	66	0	158,4231
Průměr	3,8	13,2	39,2	18,6	13,2	0	31,6846

Po dosazení souřadnic bodů z obrázku 4.5 do (4.20) a (4.24) dostaneme (potřebné hodnoty jsou vypočteny v tabulce 4.14)

$$b_1 = \frac{n\sum y_i x_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{5\cdot 196 - 19\cdot 66}{5\cdot 93 - 19^2} = -2{,}635,$$

$$b_0 = \bar{y} - b_1\bar{x} = 13{,}2 - (-2{,}635)\cdot 3{,}8 = 23{,}212.$$

Přímka minimalizující součet čtverců Q má tvar

$$23{,}212 - 2{,}635x \tag{4.27}$$

a součet čtverců odchylek je roven 158,42. Například pro $x_1 = 1$ a $y_1 = 23$ je postupně

$Y_1 = 23{,}212 - 2{,}635 \cdot 1 = 20{,}577,$

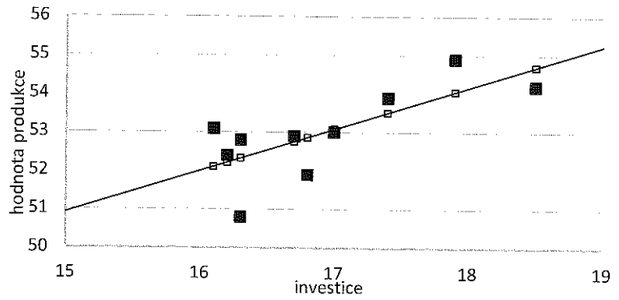
$e_1 = y_1 - Y_1 = 23 - 20{,}577 = 2{,}423,$

$e_1^2 = (y_1 - Y_1)^2 = 2{,}423^2 = 5{,}871.$

Nejlepší z přímek na obrázku 4.5 s rovnicí $23 - 2{,}5x$ a hodnotou $Q = 159{,}25$ je blízká přímce, která je podle zvoleného kritéria nejlepší. Tvrzení plyne z porovnání hodnot parametrů, stejně jako ze srovnání hodnot kritéria Q.

Příklad 4.5

V tabulce 4.15 jsou uvedeny údaje o hodnotě produkce v 100 000 Kč (vysvětlovaná proměnná y) a o výši investic v 10 000 Kč (vysvětlující proměnná x) v roce 2005 v souboru 10 firem. Data jsou znázorněna na obrázku 4.7, závislost je zřejmě možné dobře popsat regresní přímkou.



Obr. 4.7 Lineární závislost hodnoty produkce na investicích

Tab. 4.15 Výpočet parametrů regresní přímky pro příklad 4.5

Firma	y_i	x_i	$y_i x_i$	x_i^2	Y_i	e_i	e_i^2
1	52,8	16,3	860,64	265,69	52,32	0,481	0,231
2	51,9	16,8	871,92	282,24	52,86	−0,960	0,922
3	54,2	18,5	1 002,70	342,25	54,70	−0,499	0,249
4	50,8	16,3	828,04	265,69	52,32	−1,519	2,309
5	54,9	17,9	982,71	320,41	54,05	0,850	0,723
6	53,9	17,4	937,86	302,76	53,51	0,391	0,153
7	53,1	16,1	854,91	259,21	52,10	0,997	0,994
8	52,4	16,2	848,88	262,44	52,21	0,189	0,036
9	53,0	17,0	901,00	289,00	53,08	−0,077	0,006
10	52,9	16,7	883,43	278,89	52,75	0,148	0,022
Součet	529,9	169,2	8 972,09	2 868,58	529,9	0,000	5,643

Ze vzorců (4.19) a (4.20) plyne, že pro výpočet koeficientů regresní přímky je třeba z pozorovaných hodnot x_i, y_i, $i = 1, 2, \ldots, 10$ určit součty $\sum x_i$, $\sum y_i$ a $\sum y_i x_i$ (součty jsou uvedeny v posledním řádku tabulky 4.15).

Dosadíme-li údaje z posledního řádku tabulky do (4.19) a (4.20), dostaneme

$$b_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum y_i x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{529{,}9 \cdot 2868{,}58}{10 \cdot 2868{,}58 - 169{,}2^2} = 34{,}691,$$

$$b_1 = \frac{n\sum y_i x_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{10 \cdot 8972{,}09 - 169{,}2 \cdot 529{,}9}{10 \cdot 2868{,}52 - 169{,}2^2} = 1{,}082.$$

Regresní přímka má tedy tvar

$Y = 34{,}691 + 1{,}082x,$

nebo také podle (4.25)

$Y = 52{,}99 + 1{,}082(x - 16{,}92),$

neboť je

$$\bar{y} = \frac{1}{10} 529{,}9 = 52{,}99 \text{ a } \bar{x} = \frac{1}{10} 169{,}2 = 16{,}92.$$

Hodnoty Y ležící na regresní přímce a popisující závislost mezi hodnotou produkce a výší investic jsou uvedeny ve sloupci tabulky označeném Y_i a byly určeny dosazením hodnot vysvětlující proměnné x_i do regresní funkce podle (4.21). Například tedy

$Y_1 = b_0 + b_1 x_1 = 34{,}691 + 1{,}082 \cdot 16{,}3 = 52{,}32.$

V posledním sloupci tabulky je dále vypočtena hodnota kritéria Q a platí $Q = 5{,}643$.

■

Sdružené regresní přímky

V předchozí části jsme zkoumali závislost vysvětlované proměnné na vysvětlující proměnné, závislost tedy byla jednostranná. Pokud bychom ale například uvažovali výdaje domácností na potraviny a na bydlení, mohli bychom stejně dobře za vysvětlovanou proměnnou považovat obě veličiny. Proto se můžeme zabývat také sdruženou regresní přímkou, pro kterou budeme místo bodů (x_i, y_i), $i = 1, 2, \ldots, n$, uvažovat body (y_i, x_i). K přímce

$$y = \beta_0 + \beta_1 x \tag{4.28}$$

je **sdružená regresní přímka** definována jako

$$x = \alpha_0 + \alpha_1 y. \tag{4.29}$$

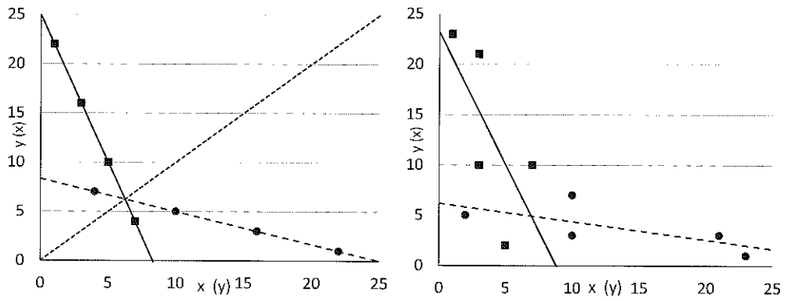
Z matematiky víme, že pokud by všechny body ležely na jedné přímce, obě funkce by byly inverzní (obrázek 4.8 vlevo) a mohli bychom z rovnice (4.28) vypočítat x ve tvaru

$$x = \frac{1}{\beta_1} y - \frac{\beta_0}{\beta_1}.$$

Porovnáním s (4.29) snadno v takovém případě dostaneme

$$\alpha_0 = -\frac{\beta_0}{\beta_1} \text{ a } \alpha_1 = \frac{1}{\beta_1}. \tag{4.30}$$

Obyčejně ale body (x_i, y_i) na jedné přímce neleží, proto nelze parametry regresních přímek takto snadno přepočítat a je třeba parametry sdružené přímky odhadnout pomocí (4.19) a (4.20) z (4.29). V pravé části obrázku 4.8 jsou čtverečky znázorněna data z obrázku 4.5 a již dříve nalezená regresní přímka $Y = 23{,}212 - 2{,}635x$ (plná čára). Kolečky jsou znázorněny body (y_i, x_i), sdružená regresní přímka (na obrázku je vyznačena čárkovaně) má rovnici $X = 6{,}189 - 0{,}181y$.



Obr. 4.8 Sdružené přímky, vpravo data z obrázku 4.5

4.3.2 Párový korelační koeficient

Intenzitu lineární závislosti mezi dvěma kvantitativními veličinami vyjadřujeme pomocí **korelačního koeficientu** (někdy se nazývá **Pearsonův korelační koeficient**).

Výběrovou hodnotu korelačního koeficientu počítáme z dvojic hodnot proměnných (x_i, y_i), $i = 1, 2, \ldots, n$, výběrová kovariance s_{xy} je definována v (4.22)

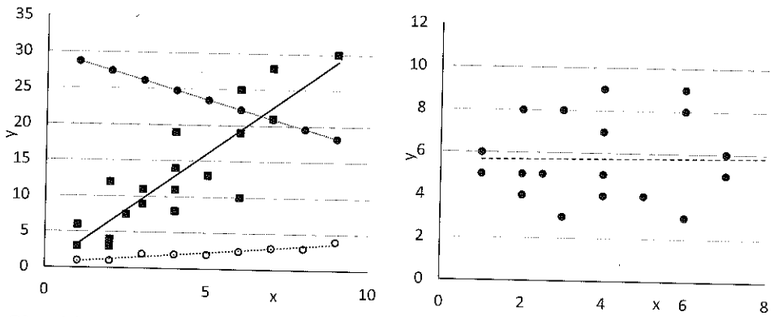
$$r_{yx} = \frac{s_{xy}}{s_x s_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\left(\overline{x^2} - \bar{x}^2\right)\left(\overline{y^2} - \bar{y}^2\right)}}. \tag{4.31}$$

Všimněme si, že hodnota koeficientu nezávisí na pořadí proměnných, a tedy platí $r_{yx} = r_{yx}$. Jedná se vlastně o normovanou kovarianci s_{xy}, která může nabývat libovolných reálných hodnot. Takové koeficienty se těžko interpretují. Proto ji při výpočtu korelačního koeficientu normujeme součinem směrodatných odchylek proměnných x (s_x) a y (s_y) tak, aby hodnoty výsledného koeficientu byly v intervalu $\langle -1; 1 \rangle$. Hodnoty -1 nabývá koeficient v případě, že všechny body leží na jedné klesající přímce, která má zápornou směrnici, a hodnoty 1, pokud je přímka rostoucí a má tedy kladnou směrnici. Pokud je korelační koeficient (nebo obdobně kovariance) roven 0, považujeme x a y za **lineárně nezávislé.** V takovém případě je regresní přímka rovnoběžná s vodorovnou osou.

Z tabulky 4.15 lze korelační koeficient mezi pozorovanými investicemi a produkcí v příkladu 4.5 určit jako (bylo třeba ještě dopočítat $\sum y_i^2 = 28\,091{,}73$)

$$r_{yx} = \frac{\frac{8972{,}09}{10} - \frac{169{,}2}{10}\cdot\frac{529{,}9}{10}}{\sqrt{\left[\frac{2868{,}58}{9} - \left(\frac{169{,}2}{9}\right)^2\right]\left[\frac{28091{,}73}{10} - \left(\frac{529{,}9}{10}\right)^2\right]}} = 0{,}736\,4.$$

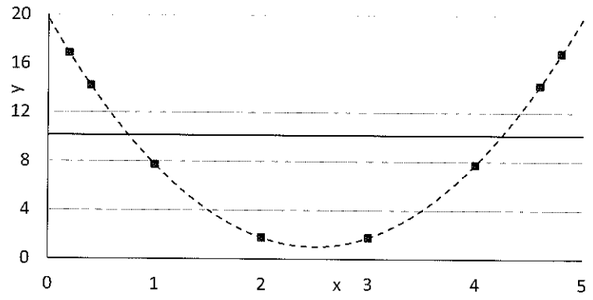
Závislost je tedy přímá (s růstem jedné z proměnných spíše roste i druhá proměnná) a intenzita závislosti je spíše silnější.



Obr. 4.9 Lineární závislosti pro různé množiny bodů

Na obrázku 4.9 jsou znázorněny lineární závislosti mezi různými dvojicemi proměnných x a y. Body jsou vždy proloženy přímkami charakterizujícími lineární závislost; korelační koeficienty jsou rovny $r_{xy} = -1$ (plné kroužky v levé části obrázku), 0,87 (čtverečky v levé části obrázku) a 0,96 (prázdné kroužky v levé části obrázku), a $r_{xy} = 0{,}02$ pro pravou část obrázku.

Je třeba zdůraznit, že korelační koeficient se vztahuje pouze k lineární závislosti, přibližně nulová hodnota koeficientu může znamenat, že veličiny jsou silně závislé, ale závislost není lineární. Takovou závislost nazveme nelineární a některými takovými závislostmi se budeme zabývat dále v části věnované regresním funkcím nelineárním v parametrech. Na obrázku 4.10 jsou znázorněny body, které leží na parabole, a jsou tedy deterministicky závislé, nicméně regresní přímka je rovnoběžná s vodorovnou osou a korelační koeficient je roven nule.



Obr. 4.10 Body ležící na parabole, $r_{xy} = 0$

Pokud vyjde hodnota korelačního koeficientu v absolutní hodnotě blízká jedné, neříká nám to mnoho o příčinné závislosti. Jednak korelační koeficient, na rozdíl od regresní funkce, popisuje oboustrannou závislost a také vysoká korelace může být způsobena vlivem nějaké další skryté proměnné. Takovou korelaci nazýváme pouze zdánlivou a neměli bychom s ní pracovat jako se skutečnou závislostí. S tímto problémem se setkáváme v různých aplikacích, v analýze časových řad je zvláště častým jevem. Možný postup si ukážeme na příkladu pouze jedné dodatečné proměnné (označíme ji z a budeme uvažovat trojice (x_i, y_i, z_i), $i = 1, 2, \ldots, n$), která způsobuje pouze zdánlivou závislost x a y. Označme r_{xy}, r_{xz} a r_{yz} párové korelační koeficienty mezi všemi dvojicemi z proměnných x, y a z určené podle (4.31) a definujeme **dílčí korelační koeficient** $r_{xy.z}$ popisující lineární závislost mezi x a y s vyloučením vlivu proměnné z jako

$$r_{xy.z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}. \tag{4.32}$$

Tento koeficient je stále korelačním koeficientem, proto jeho interpretace je stejná jako interpretace prostého párového koeficientu výše, jen dodáváme, že všechny úvahy jsou pro závislost s vyloučením vlivu proměnné z.

Spearmanův koeficient pořadové korelace

Jinou možností, jak posoudit závislost mezi dvěma znaky, je Spearmanův korelační koeficient. Jeho výhodou je, že je použitelný nejen na kvantitativní proměnné, ale také pro pořadové veličiny, u kterých známe pouze pořadí.

V případě kvantitativních proměnných x_i a y_i je nejprve třeba hodnoty uspořádat podle velikosti od minima do maxima a každé hodnotě přiřadit pořadí i_x v uspořádaném výběru pro proměnnou x a i_y pro proměnnou y. V případě pořadových veličin jsou obyčejně dána přímo pořadí i_x a i_y.

Těsnost závislosti se pak popisuje **Spearmanovým korelačním koeficientem** (také **Spearmanovým koeficientem pořadové korelace**) r_S, který je definován jako

$$r_S = 1 - \frac{6\sum (i_x - i_y)^2}{n(n^2 - 1)}. \tag{4.33}$$

Tento koeficient, obdobně jako párový korelační koeficient r_{xy}, může nabývat hodnoty od –1 do +1. Hodnoty 1 nabývá v případě, že pořadí hodnot je stejné v obou proměnných, hodnoty –1 v případě, že pořadí je právě opačné.

Příklad 4.6

Předpokládejme, že se o funkci uchází 8 zájemců, kteří byli podrobeni dvěma úkolům. Výsledky jsou uvedeny v tabulce 4.16, uchazeči byli seřazeni od nejlepšího (pořadí je 1) do nejhoršího (pořadí je 8).

Tab. 4.16 Výpočet Spearmanova koeficientu pořadové korelace

Uchazeč	první úkol	druhý úkol	$i_x - i_y$	$(i_x - i_y)^2$
A	1	2	−1	1
B	2	3	−1	1
C	3	1	2	4
D	4	6	−2	4
E	5	8	−3	9
F	6	4	2	4
G	7	5	2	4
H	8	7	1	1
Součet	36	36	0	28

V posledním řádku tabulky jsou součty. Pro první i druhý úkol platí, že součet je roven $1 + 2 + 3 + \ldots + 8 = 36$ a součet rozdílů je vždy roven nule. Dosadíme-li do vzorce (4.33), dostaneme

$$r_S = 1 - \frac{6 \cdot 28}{8(8^2 - 1)} = 0{,}667.$$

Koeficient naznačuje středně silnou shodu pořadí. Znamená to, že uchazeči, kteří uspěli v prvním úkolu, uspějí zřejmě dobře i v druhém.

■

4.3.3 Polynomická regrese

Polynomická regrese poskytuje možnost, jak vystihnout vztah, pro jehož popis se nehodí přímka, pomocí funkce lineární v parametrech. Pokud závislost proměnných x a y není lineární, jako například na obrázcích 4.4 v pravé části, 4.10 nebo 4.11, můžeme

zkusit použít nějaký polynom vyššího stupně, například parabolu. Regresní parabola je popsána třemi parametry β_0, β_1 a β_2 ve tvaru

$$\eta(x; \beta_0, \beta_1, \beta_2) = \eta(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2. \tag{4.34}$$

Pro snazší zápis jsme parametry umístili do sloupcového vektoru

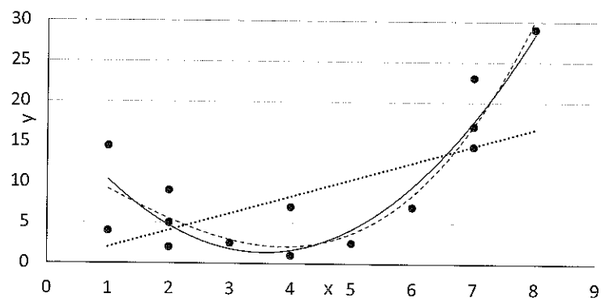
$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Pokud budeme vektor $\boldsymbol{\beta}$ psát do řádku, je třeba použít transpozici $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$.

Na obrázku 4.11 jsou čtrnácti body proloženy přímka, parabola a kubická parabola popsaná čtyřmi parametry β_0, β_1, β_2 a β_3 a rovnicí

$$\eta(x; \beta_0, \beta_1, \beta_2, \beta_3) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

Přímka není pro popis závislosti vhodná, mezi oběma parabolami není na základě obrázku snadné rozhodnout, která je lepší.



Obr. 4.11 Polynomická regrese přímka, parabola (plná čára) a kubická parabola (čárkovaně)

Polynomická regresní funkce stupně k má $p = k + 1$ neznámých parametrů a lze ji zapsat jako

$$\eta(x; \boldsymbol{\beta}) = \eta(x; \beta_0, \beta_1, \ldots, \beta_k) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k. \tag{4.35}$$

Všimněme si, že funkce jsou nelineární vzhledem k proměnné x, ale lineární vzhledem k neznámým koeficientům polynomů. Z hlediska matematické statistiky tedy stále hovoříme o lineární regresní funkci.

Nejvhodnější parametry budeme opět hledat metodou nejmenších čtverců. Například pro parabolu (4.34) to znamená minimalizovat funkci

$$Q(\beta_0, \beta_1, \beta_2) = Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \right]^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

vzhledem k hledaným parametrům. K tomu účelu opět najdeme parciálních derivace kritéria Q podle parametrů, v tomto případě budou derivace tři:

$$\frac{\partial}{\partial \beta_0} Q(\beta_0, \beta_1, \beta_2) = -2 \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \right],$$

$$\frac{\partial}{\partial \beta_1} Q(\beta_0, \beta_1, \beta_2) = -2 \sum_{i=1}^{n} x_i \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \right],$$

$$\frac{\partial}{\partial \beta_2} Q(\beta_0, \beta_1, \beta_2) = -2 \sum_{i=1}^{n} x_i^2 \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2) \right].$$

Po algebraických úpravách dostaneme, že hodnoty hledaných parametrů b_0, b_1 a b_2 by bylo možné nalézt řešením tří normálních rovnic

$$\sum_{i=1}^{n} y_i - n b_0 - b_1 \sum_{i=1}^{n} x_i - b_2 \sum_{i=1}^{n} x_i^2 = 0,$$

$$\sum_{i=1}^{n} y_i x_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 - b_2 \sum_{i=1}^{n} x_i^3 = 0,$$

$$\sum_{i=1}^{n} y_i x_i^2 - b_0 \sum_{i=1}^{n} x_i^2 - b_1 \sum_{i=1}^{n} x_i^3 - b_2 \sum_{i=1}^{n} x_i^4 = 0.$$

Je snadné si představit, že tento postup je velmi technicky i časově náročný, zvláště s rostoucím stupněm regresního polynomu. Proto nyní ukážeme jinou možnost, jak nalézt hodnoty regresních parametrů metodou nejmenších čtverců. Řešení problému lze snadno najít v maticovém tvaru. Pozorované hodnoty proměnné y, tj. $y_1, y_2, \ldots, y_n$, umístíme do n-rozměrného vektoru $\mathbf{y}$. Dále vytvoříme matici $\mathbf{X}$, kterou budeme nazývat **regresní matice**. Tato matice má n řádků, do prvního sloupce umístíme jedničky (bude odpovídat absolutnímu členu polynomu β_0 v (4.35)) a do dalších sloupců hodnoty $x_i, x_i^2, \ldots, x_i^k$. Potom místo (4.35) můžeme n rovnic pro n pozorovaných bodů zapsat v maticovém tvaru jako

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}. \tag{4.36}$$

V případě pozorovaných n dvojic bodů (x_i, y_i) a regresní přímky, paraboly, polynomu k-tého stupně (4.35) dostaneme regresní matice $\mathbf{X}$ s n řádky a dvěma, třemi a $p = k + 1$ sloupci ve tvaru

$$\mathbf{X}_{přímka} = \begin{pmatrix} 1 & x_1 \\ \ldots & \ldots \\ 1 & x_n \end{pmatrix}_{n \times 2}, \quad \mathbf{X}_{parabola} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 \end{pmatrix}_{n \times 3} \tag{4.37}$$

a $\mathbf{X}_{kubická\ regresní\ funkce} = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ \ldots & \ldots & \ldots & \\ 1 & x_n & x_n^2 & x_n^3 \end{pmatrix}_{n \times 4}$, $\mathbf{X}_{polynom} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^k \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^k \end{pmatrix}_{n \times (k+1)}$. (4.38)

Potom lze kritérium Q v maticovém tvaru zapsat obecně jako

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{4.39}$$

a vektor $\mathbf{b}$, pro který je dosaženo minima, může být vyjádřen explicitním vzorcem

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{4.40}$$

Pokud ve vzorci píšeme $(\mathbf{X}'\mathbf{X})^{-1}$, máme na mysli inverzní matici k matici $\mathbf{X}'\mathbf{X}$. Všimněme si, že v tomto případě není třeba konstruovat a řešit normální rovnice, které jsme odvodili pro regresní přímku a regresní parabolu, ale stačí násobit matice a najít inverzní matici.

Ukážeme, že pro regresní přímku jsou opravdu oba způsoby nalezení optimálních parametrů (podle (4.23) a (4.24) a pomocí (4.40)) shodné. Představme si, že v rovnici $\mathbf{y} = \mathbf{X}\mathbf{b}$ obě strany vynásobíme zleva maticí $\mathbf{X}$ a dostaneme rovnici

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}.$$

V případě přímky můžeme obě strany rovnice rozepsat jako

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix},$$

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}.$$

Dostali jsme tedy rovnice

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i, \qquad \sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2.$$

Tyto rovnice jsou shodné s rovnicemi (4.17) a (4.18), které byly odvozeny pomocí soustavy normálních rovnic. Proto pro přímkovou regresi jsou také stejná řešení. Obdobně by bylo možné ukázat, že totéž platí i pro regresní parabolu nebo další regresní funkce lineární v parametrech.

4.3.4 Další typy regresních funkcí

Lineární regresní funkce je nejjednodušším nástrojem pro popis závislosti mezi veličinami x a y, kterému dáváme přednost právě pro snadné výpočty a zřejmou interpretovatelnost koeficientů (regresních parametrů). Na druhé straně je zřejmé, že při modelování vztahů ekonomických veličin jen s lineární závislostí nevystačíme. V předchozí části byla probrána možnost použití polynomické regrese, nyní ukážeme další možnosti – použití přímkové regrese na transformované hodnoty vysvětlované nebo vysvětlující proměnné nebo regresní funkci nelineární v parametrech. Často je konkrétní

regresní funkce dána teoretickou podstatou řešeného problému a všeobecně akceptovaným teoretickým vztahem. Tuto situaci vidíme například v případě Cobbovy-Douglasovy produkční funkce.

Hyperbolická regresní funkce

Užitečnou regresní funkcí pro popis závislostí mezi veličinami v ekonomii je hyperbolická regresní funkce

$$\eta(x; \beta_0, \beta_1) = \beta_0 + \frac{\beta_1}{x}. \tag{4.41}$$

Všimněme si, že jde o lineární funkci vzhledem k regresním parametrům, vzhledem k vysvětlující proměnné y je lineární v transformované vysvětlující proměnné $1/x$. Znamená to, že přímku prokládáme body $(1/x_i, y_i)$, $i = 1, 2, \dots, n$; regresní matice pro výpočet optimálních parametrů v rovnici (4.41) jsou uvedeny dále ve vzorci (4.43).

Logaritmická regresní funkce

V některých situacích místo hyperbolické funkce používáme lineární funkci závislou na logaritmu proměnné x, tedy regresní funkci

$$\eta(x; \beta_0, \beta_1) = \beta_0 + \beta_1 \ln x. \tag{4.42}$$

Logaritmické regresní funkce jsou vhodné k modelování závislostí parabolického typu, které však nemají maximum a u nichž při vyšších hodnotách vysvětlující proměnné x vzrůstají hodnoty závisle proměnné y pouze velmi pozvolna, eventuálně se prakticky nemění (prodlužují regresní křivku v horizontálním směru). V případě regresní hyperboly a regresní logaritmické funkce neprovádíme žádnou transformaci proměnné y, proto vektory a matice, vstupující do výpočtu regresních parametrů (4.40), mají tvar

$$y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}_{n\times 1}, \quad \mathbf{X}_{\text{hyperbola}} = \begin{pmatrix} 1 & 1/x_1 \\ 1 & \dots \\ 1 & 1/x_n \end{pmatrix}_{n\times 2}, \quad \mathbf{X}_{\text{logaritmus}} = \begin{pmatrix} 1 & \ln x_1 \\ 1 & \dots \\ 1 & \ln x_n \end{pmatrix}_{n\times 2}. \tag{4.43}$$

Regresní funkce nelineární v parametrech

Dále ukážeme funkce, které nebudou lineární v neznámých parametrech. Takové funkce se nazývají nelineární a jejich zkoumání potom nelineární regresní analýza. Ukážeme pouze dvě funkce z velkého množství nelineárních funkcí, které se pro popis závislostí mezi proměnnými používají.

Pro výpočet parametrů regresních funkcí, které nejsou lineární z hlediska parametrů, nemůžeme přímo použít rovnice (4.40) ani nalézt explicitní řešení pomocí normálních rovnic. V některých případech je možné použít **linearizující transformaci**, tedy nalézt transformace proměnných, které by umožňovaly převést regresní funkci na funkci lineární v parametrech. Pokud taková transformace existuje, hledáme parame-

try v transformovaném modelu a zpětnými transformacemi se vrátíme k původní regresní funkci, popisující závislost mezi proměnnými x a y. Tento postup by ovšem dával stejný výsledek jako při přímém hledání pouze v případě, kdy by všechny pozorované body ležely přesně na uvažované regresní funkci. V opačném případě můžeme řešení pomocí linearizace považovat pouze za aproximativní a přibližné, jak ukážeme v příkladech v dalším textu. Při hledání nejlepších hodnot parametrů nelineárních regresních funkcí metodou nejmenších čtverců se postupuje tak, že na základě vhodného počátečního přiblížení pro parametry (například hodnot získaných pomocí linearizace) použijeme nějaký postup numerické matematiky, který umožňuje iterativním postupem zlepšovat počáteční řešení tak dlouho, až je dosažena požadovaná přesnost řešení. Touto problematikou se více v tomto textu zabývat nebudeme, odkazujeme čtenáře na práci Hebák, Hustopecký, Malá (2005) nebo na další rozsáhlou literaturu, zabývající se nelineárními regresními modely. V současné době můžeme s výhodou využít implementaci takových postupů ve statistickém softwaru.

Exponenciální regresní funkce

Nejznámější a také nejčastěji používanou regresní funkcí, která je nelineární v parametrech, je exponenciální regresní funkce

$$\eta(x; \beta_0, \beta_1) = \beta_0 \beta_1^x. \tag{4.44}$$

Tuto funkci můžeme použít pro velmi rychle rostoucí proměnnou y (například pro tržby v závislosti na počtu zaměstnanců rychle rostoucí firmy). Vzhledem k velmi rychlému růstu je ovšem třeba tuto funkci používat opatrně. Rovnici

$$y = \beta_0 \beta_1^x$$

můžeme zlogaritmovat a dostaneme

$$\ln y = \ln \beta_0 + \ln(\beta_1) x.$$

Je tedy možné hodnotami $\ln y_i$ a x_i proložit lineární funkci

$$\eta(x; \alpha_0, \alpha_1) = \alpha_0 + \alpha_1 x. \tag{4.45}$$

Pro výpočet parametrů a_0 a a_1 přímky v linearizovaném tvaru použijeme vzorce (4.19) a (4.20) pro dvojice $(\ln y_i, x_i)$, $i = 1, 2, \ldots, n$. Druhou možností zápisu je použít maticový výraz (4.40), kde zvolíme vektor závisle proměnné a regresní matici $\mathbf{X}$ ve tvaru

$$\ln \mathbf{y} = \begin{pmatrix} \ln y_1 \\ \ldots \\ \ln y_n \end{pmatrix}_{n\times 1}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & \ldots \\ 1 & x_n \end{pmatrix}_{n\times 2}.$$

Porovnáme-li parametry obou přímek, dostaneme při znalosti hodnot a_0 a a_1, parametry funkce b_0 a b_1 z rovnic

$$\ln b_0 = a_0 \quad \text{a} \quad \ln b_1 = a_1$$

(porovnáním v (4.45)) jako $b_0 = e^{a_0}$ a $b_1 = e^{a_1}$.

Jak jsme již uvedli, při popisu vztahu nelineárními funkcemi, jako je také exponenciální funkce, dáváme přednost nalezení parametrů přímým hledáním minimální hodnoty součtu čtverců odchylek

$$Q = \sum_{i=1}^{n} \left(y_i - \beta_0 \beta_1^{x_i} \right)^2.$$

V podobných nelineárních problémech již neexistuje explicitní vzorec pro b_0 a b_1, numerické metody běžně implementované v softwaru ale umožňují rychlé a spolehlivé řešení problému. Nemusí jít o specializovaný statistický software, doplněk Řešitel v programu MS Excel také umožní realizovat potřebné výpočty. V příkladu 4.7 ukážeme, že obě řešení (na základě linearizace a přímé numerické optimalizace) nejen nejsou identická, mohou se od sebe podstatně lišit. Hodnoty parametrů získané oběma postupy porovnáváme hodnotou součtu čtverců Q, stejně jako grafickou formou.

Příklad 4.7

Hotelová společnost vlastnící v rámci svého rezervačního řetězce 12 hotelů zkoumá vztah mezi celkovými měsíčními tržbami těchto hotelů (vysvětlovaná proměnná y, mil. Kč) a tržbami vyprodukovanými stravovacími úseky v nich (vysvětlující proměnná x, rovněž mil. Kč); data jsou uvedena v tabulce 4.17.

Řešení

Vztah mezi proměnnými x a y popíšeme exponenciální funkcí podle vzorce (4.44) a exponenciální funkcí s konstantou

$$\eta(x; \beta_0, \beta_1, \beta_2) = \beta_2 + \beta_0 \beta_1^x.$$

Tab. 4.17 Data pro exponenciální regresi z příkladu 4.7

Hotel	1	2	3	4	5	6	7	8	9	10	11	12
x_i	2	1,2	14,8	8	8	3	4,8	15,6	16	12	14	14
y_i	12	8	76,4	17	21,3	10	12,5	97,3	88	25	39	47,3

Uvažujme nejprve numerický postup, který umožní nalézt parametry, pro které je dosaženo minima kritéria nejmenších čtverců

$$Q = \sum_{i=1}^{n} \left(y_i - \beta_0 \beta_1^{x_i} \right)^2.$$

Najít řešení umožňuje například doplněk Řešitel v Excelu nebo jakýkoliv statistický program. Nalezené řešení $b_0 = 1{,}893\ 1$ a $b_1 = 1{,}272\ 8$ popisuje exponenciální regresní funkci ve tvaru

$$1{,}893\ 1 \cdot 1{,}272\ 8^x.$$

Hodnota kritéria Q pro tyto hodnoty parametrů je rovna 1072.

Pokud bychom chtěli použít linearizaci, je třeba nejprve pracovat s lineární funkcí popisující vztah mezi body $(x_i, \ln y_i)$ podle (4.45). Z odhadnuté lineární funkce

$a_0 + a_1 x = 1{,}854\,0 + 0{,}150\,7x$

dostaneme po zpětné transformaci

$b_0 = e^{1{,}8540} = 6{,}385\,2$ a $b_1 = e^{0{,}1507} = 1{,}162\,6$.

Získaná exponenciální funkce má tvar

$6{,}385\,2 \cdot 1{,}162\,6^x$.

Hodnota součtu čtverců v řešení je rovna 1895.

Všimněme si, že obě vypočtené funkce jsou odlišné, jak je vidět z nalezených hodnot parametrů, hodnot kritéria Q i obrázku 4.12. Funkce nalezená nelineární optimalizací lépe vystihuje pozorované body a hodnota součtu čtverců je také výrazně menší. Pro nelineární postup byly použity hodnoty parametrů získané linearizací jako počáteční aproximace.

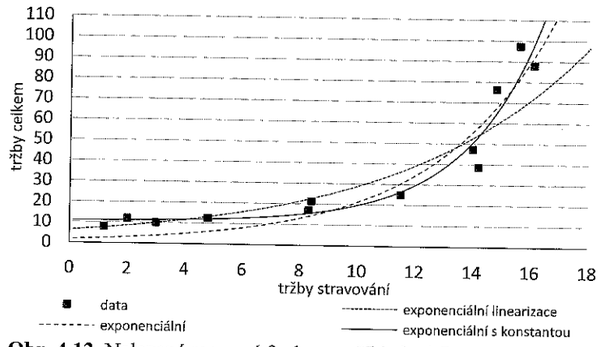
Z grafu je patrné, že by bylo vhodné do modelu zahrnout konstantu. Proto byly odhadnuty parametry regresní funkce

$$\eta(x; \beta_0, \beta_1, \beta_2) = \beta_2 + \beta_0 \beta_1^x.$$

Všimněme si, že tuto funkci již není možné linearizovat. Minimum funkce

$$Q(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} \left[y_i - (\beta_2 + \beta_0 \beta_1^{x_i}) \right]^2$$

bylo dosaženo pro $b_2 = 10{,}774\,8$, $b_0 = 0{,}208\,1$ a $b_1 = 1{,}455\,3$ (opět je třeba použít vhodný software). Hodnota kritéria Q v minimu je 729. Nalezená funkce je znázorněna na obrázku 4.12 plnou čarou. Je patrné, že velmi dobře vystihuje body, kterými jsme regresní funkce prokládali.



Obr. 4.12 Nalezené regresní funkce v příkladu 4.7

Mocninná regresní funkce

Další běžně používanou nelineární funkcí je mocninná funkce, definovaná jako

$$\eta(x; \beta_0, \beta_1) = \beta_0 x^{\beta_1}. \tag{4.46}$$

Tato funkce patří mezi Cobbovy-Douglasovy jednofaktorové produkční funkce, model (4.46) slouží například k odhadu objemu produkce y na základě změn známého produkčního faktoru x (produktivity práce a podobně). I v tomto případě si při výpočtu parametrů můžeme pomoci logaritmickou transformací. Po logaritmování vztahu

$$y = \beta_0 x^{\beta_1}$$

dostaneme

$$\ln y = \ln \beta_0 + \beta_1 \ln x.$$

Můžeme tedy uvažovat lineární regresní funkci

$$\eta(\ln x; \alpha_0, \alpha_1) = \alpha_0 + \alpha_1 \ln x;$$

hodnoty a_0 a a_1 minimalizující kritérium metody nejmenších čtverců můžeme dostat z (4.19) a (4.20), použijeme-li dvojice bodů $(\ln x_i, \ln y_i)$, $i = 1, 2, \ldots, n$. Jinou možností pro nalezení optimálních parametrů je použít maticové vyjádření (4.40), ze kterého dostaneme $\mathbf{a} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ln\mathbf{y}$, kde je

$$\ln \mathbf{y} = \begin{pmatrix} \ln y_1 \\ \ldots \\ \ln y_n \end{pmatrix}_{n\times 1}, \quad \mathbf{X} = \begin{pmatrix} 1 & \ln x_1 \\ \ldots & \ldots \\ 1 & \ln x_n \end{pmatrix}_{n\times 2}.$$

Najdeme-li hodnoty a_0 a a_1, použijeme zpětnou transformaci a z rovnic

$$\ln b_0 = a_0 \quad \text{a} \quad b_1 = a_1$$

dostaneme

$$b_0 = e^{a_0} \quad \text{a} \quad a_1 = b_1.$$

I v tomto případě je lépe nalézt optimální parametry b_0 a b_1 pomocí numerické minimalizace součtu čtverců

$$Q = \sum_{i=1}^{n} \left(y_i - \beta_0 x^{\beta_1} \right)^2. \tag{4.47}$$

Příklad 4.8

Uvažujme opět hodnoty uvedené v tabulce 4.17, tentokrát proložíme mocninnou funkci (4.46) a mocninnou funkci s posunutím

$$f(x; \beta_0, \beta_1, \beta_2) = \beta_2 + \beta_0 x^{\beta_1}.$$

Nejprve najdeme parametry mocninné funkce. Numerická optimalizace funkce (4.47) nalezne hodnoty parametrů a součtu čtverců Q

$b_0 = 0{,}003\,9$, $b_1 = 3{,}620\,8$, $Q(0{,}003\,9, 3{,}620\,8) = 1449$,

odhadnutá regresní funkce má tedy tvar

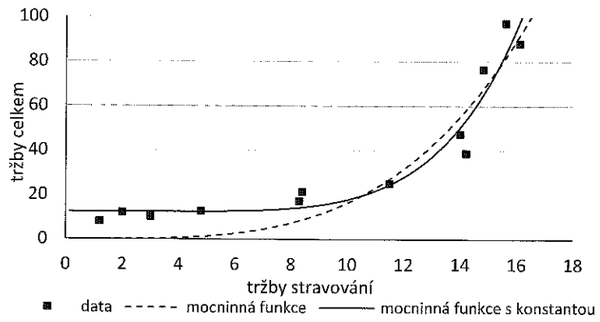
$0{,}003\,9 x^{3{,}620\,8}$.

Tato funkce je na obrázku 4.13 znázorněna čárkovaně. Je patrné, že by opět bylo vhodné přidat jeden parametr, v tomto případě posunutí. Odhadnutá funkce pomocí přímé minimalizace součtu čtverců Q má tvar

$$12{,}358\,7 + 0{,}000\,008\,86x^{5{,}788}.$$

Hodnota funkce Q je v tomto případě rovna 737. Funkce je v grafu znázorněna plnou čarou.

Hodnoty Q pro nelineární funkce, nalezené ze stejných dat uvedených v tabulce 4.17 v příkladech 4.7 a 4.8, lze porovnat. Například pro exponenciální a mocninnou funkci s konstantou, které podle obrázku velmi dobře vystihly data, je pro exponenciální funkci $Q = 737$ a pro mocninnou funkci $Q = 729$. Menší součet čtverců odchylek byl dosažen pro mocninnou funkci s konstantou, i když rozdíl není příliš velký.



Obr. 4.13 Nalezené mocninné funkce k příkladu 4.7

4.3.5 Klasický lineární regresní model

Všechny předchozí úvahy této části bylo možné provést pro n libovolných dvojic (trojic) čísel. Postupy připomínaly úvahy z první části věnované popisné statistice a neobsahovaly žádné induktivní úsudky probírané ve třetí části. Nyní budeme předpokládat, že hodnoty nezávisle proměnné $x_1, x_2, \ldots, x_n$ jsou pevné a budeme klást předpoklady na pravděpodobnostní rozdělení závisle proměnné y. Na základě těchto předpokladů budeme zkoumat vlastnosti závislosti mezi vysvětlující proměnnou x a vysvětlovanou proměnnou y nejen v pozorovaných bodech, ale budeme výsledky zobecňovat na celou populaci. Budeme tedy, stejně jako ve třetí kapitole, předpokládat, že v celé populaci existuje vztah mezi zkoumanými proměnnými, popsaný **populační regresní funkcí**. Například mezi těmi, kteří v určitém období pobírají mzdu (zkoumaná **populace**), existuje vztah mezi mzdou a věkem, délkou praxe nebo třeba počtem let, které zaměstnanec strávil ve škole. Proměnná *počet let ve škole* se používá jako kvantitativní (číselná) proměnná, která může nahradit ordinální kategoriální proměnnou vzdělání. Závislost je volná (stochastická), neboť neexistuje pevný, neboli deterministický vztah mezi mzdou a počtem let praxe. Pokud by existoval, bylo by možné ze znalosti počtu let praxe přesně určit mzdu. V regresní analýze se snažíme závislosti popsat regresními funkcemi; na základě pozorovaných hodnot tedy odhadujeme populační regresní funkci. V induktivní povaze úsudků spočívá rozdíl mezi předchozími částmi a následujícím textem, který se zabývá klasickým lineárním regresním modelem.

Zatímco v minulé kapitole jsme hodnoty parametrů regresních funkcí minimalizující součty čtverců počítali nebo hledali, nyní je budeme odhadovat ve smyslu postupů, které byly probrány ve třetí kapitole, věnované induktivním úsudkům. V takovém případě budeme moci posuzovat statistické vlastnosti bodových odhadů, jako jsou nezkreslenost nebo konzistence. Obdobně bude možné testovat statistické hypotézy týkající se **regresního modelu**, se kterým budeme pracovat.

Cílem regresní analýzy je hlubší vniknutí do obecných rysů zkoumaných závislostí a nalezení takových matematických předpisů, které by umožňovaly obecné úvahy o podstatě sledovaných vztahů. Odhad konkrétní regresní funkce na základě daných (výběrových) pozorování není ještě zárukou správně zvolené cesty k poznání příčinných souvislostí mezi veličinami a pochopitelně nedovoluje ztotožnit vypočítanou (výběrovou) regresní funkci s hypotetickou regresní funkcí základního souboru. Potřebujeme najít odpověď na celou řadu závažných otázek, souvisejících s posouzením regresní funkce jako nástroje k analýze vnitřních souvislostí mezi veličinami a k odhadům vysvětlované proměnné při volbě libovolných kombinací vysvětlujících proměnných. Uveďme některé z nich:

- Byl zvolen vhodný typ regresní funkce?
- Byl proveden správný výběr vysvětlujících proměnných?
- Jak lze hodnotit význam jednotlivých vysvětlujících proměnných zařazených do regresní funkce?
- Do jaké míry se ukázaly oprávněné předpoklady, za kterých je vhodné použít metodu nejmenších čtverců v popsané podobě, nebo by bylo vhodné hledat jinou metodu odhadu parametrů?
- Jak velké jsou výběrové chyby odhadů parametrů a v jakých mezích se pohybují hodnoty neznámých parametrů a hodnoty závisle proměnné při zvolených kombinacích hodnot vysvětlujících proměnných?

V dalším textu se budeme věnovat výše zmíněným problémům, nalezení přesných a obecných odpovědí na všechny uvedené otázky ale překračuje možnosti této kapitoly, věnované problematice regresní a korelační analýzy. Čtenáře odkazujeme na obsáhlou literaturu, věnovanou regresním modelům a jejich použití.

Nejprve budeme zkoumat případ, ve kterém jsou pouze dvě proměnné (závisle proměnná y a nezávisle proměnná x) a závislost budeme popisovat regresní přímkou. Potom se budeme věnovat nelineárnímu regresnímu modelu s mocninnou a expo-

nenciální regresní funkcí a podrobně se budeme zabývat modelem s polynomickou regresní funkcí (podle 4.3.3). Posledním uvažovaným regresním problémem bude vícenásobná regrese, v níž používáme jednu vysvětlovanou proměnnou a více vysvětlujících proměnných.

Předpokládejme tedy, že v celé populaci je závislost mezi proměnnými x a y popsána populační regresní přímkou $y = \beta_0 + \beta_1 x$ a parametry přímky β_0 a β_1 jsou neznámé. Máme k dispozici pouze n dvojic (x_i, y_i), $i = 1, 2, \ldots, n$, kde x_i je pevná hodnota vysvětlující proměnné, y_i je náhodná veličina a předpokládáme, že platí

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{4.48}$$

Náhodná veličina ε_i se nazývá **náhodná složka** a můžeme ji interpretovat jako výsledek působení všech v regresním modelu neuvažovaných vlivů, které ovlivňují změny vysvětlované proměnné y. Model (4.48) nazýváme **(stochastickým) regresním modelem**. O náhodné složce ε budeme předpokládat, že má nulovou střední hodnotu, konstantní rozptyl a normální rozdělení. Pokud v regresním modelu (4.48) platí

$$\varepsilon_i \sim N(0, \sigma^2),\ \varepsilon_i \text{ jsou nezávislé}, \ i = 1, 2, \ldots, n, \tag{4.49}$$

hovoříme o **klasickém lineárním regresním modelu**. O rozptylu náhodné složky σ^2 předpokládáme, že je neznámý a v průběhu práce s regresním modelem ho bude třeba také odhadnout. Vzhledem k tomu, že jsme stanovili předpoklady na náhodnou složku, můžeme při práci s regresním modelem použít postupy teorie pravděpodobnosti z druhé kapitoly a matematické statistiky z třetí kapitoly. Budeme tedy počítat střední hodnoty a rozptyly a mluvit o odhadech neznámých charakteristik (například parametrů modelu) a statistických vlastnostech těchto odhadů nebo testovat statistické hypotézy.

Pro každou hodnotu vysvětlující proměnné x_i můžeme najít střední hodnotu $E(y_i)$ vysvětlované proměnné

$$E(y_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i) = \beta_0 + \beta_1 x_i$$

a rozptyl $D(y_i)$

$$D(y_i) = D(\beta_0 + \beta_1 x_i + \varepsilon_i) = D(\varepsilon_i) = \sigma^2.$$

Použijeme-li dále vlastnosti normálního rozdělení, můžeme říci, že y_i jsou náhodné veličiny s normálním rozdělením $N(\beta_0 + \beta_1 x_i, \sigma^2)$ a můžeme tedy psát

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

Pro odhad parametrů regresního modelu (4.48) použijeme opět metodu nejmenších čtverců a pomocí metody normálních rovnic nebo maticového zápisu najdeme odhady $\hat{\beta}_0$ a $\hat{\beta}_1$ neznámých parametrů β_0 a β_1 jako

$$\hat{\beta}_0 = b_0 \quad \text{a} \quad \hat{\beta}_1 = b_1,$$

kde b_0 a b_1 jsou dány v (4.19) a (4.20). Vzhledem k předpokladům (4.49), které jsme učinili o náhodné složce, můžeme o odhadech parametrů regresní přímky říci (Hebák, Hustopecký, Malá, 2005; Hebák a další, 2013), že

- jsou konzistentní,
- jsou nezkreslené,
- mají nejmenší možný rozptyl,
- jsou lineární funkcí pozorovaných hodnot vysvětlující proměnné,
- mají normální rozdělení.

Z druhé až čtvrté vlastnosti plyne, že odhady jsou typu BLUE (z anglického **B**est **L**inear **U**nbiased **E**stimator). Nezkreslenost odhadů v případě regresní přímky znamená, že platí

$$E(\hat{\beta}_0) = \beta_0 \quad \text{a} \quad E(\hat{\beta}_1) = \beta_1.$$

Pro všechna x_i označíme hodnoty na odhadnuté regresní přímce

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \tag{4.50}$$

stejné značení zvolíme také pro libovolnou hodnotu vysvětlující proměnné x ve tvaru

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Tento postup umožní na základě odhadnutého regresního modelu odhadovat hodnoty vysvětlované proměnné pro všechny možné hodnoty vysvětlující proměnné x. Volíme-li hodnotu nezávisle proměnné x_i $(i = 1, 2, \ldots, n)$, pak rozdíl mezi pozorovanou hodnotou y_i a hodnotou na regresní funkci $\hat{y}_i$ nazveme reziduum a označíme jej $\hat{\varepsilon}_i$. Je tedy

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i). \tag{4.51}$$

Tato hodnota je realizací náhodné složky ε_i v bodě x_i. Pokud pozorovaný bod leží přímo na regresní přímce, je reziduum nulové. Pokud leží pozorování nad regresní funkcí, je reziduum kladné, v opačném případě je záporné. Pokud rezidua sečteme, z konstrukce odhadů parametrů plyne, že součet je roven nule.

Příklad 4.9

Vraťme se k údajům z příkladu 4.5. Již jsme našli odhady regresních parametrů $\hat{\beta}_0 = 34{,}691$, $\hat{\beta}_1 = 1{,}082$, a je tedy možné nalézt také hodnoty reziduí. Budeme předpokládat, že platí regresní model (4.48). Potom má odhadnutá regresní funkce tvar

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 34{,}691 + 1{,}082x.$$

Hodnota odhadnuté směrnice regresní přímky $\hat{\beta}_1$ udává, že pokud výše investic vzroste o 10 000 Kč, vzroste hodnota produkce v průměru o

$1{,}082 \cdot 100\,000 = 108\,200$ Kč.

V tabulce 4.18 jsou, pro všechny pozorované hodnoty x_i, $i = 1, 2, \ldots, n$, nalezeny odhadnuté hodnoty na regresní přímce $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, hodnoty reziduí (určené podle (4.51)) a jejich druhé mocniny. V prvním řádku je tedy například

$\hat{y}_1 = 34{,}691 + 1{,}082 \cdot 16{,}3 = 52{,}32$ (·10 000 Kč),

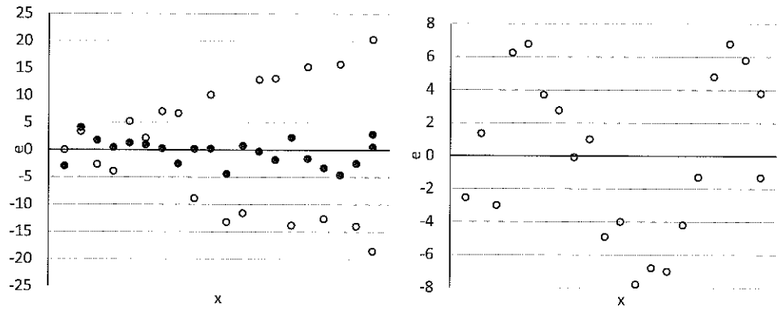
$\hat{\varepsilon}_1 = 52{,}8 - 52{,}319 = 0{,}481$ (·10 000 Kč).

Tab. 4.18 Výpočet reziduí v příkladu 4.9

Firma	y_i	x_i	$\hat{y}_i$	$\hat{\varepsilon}_i = y_i - \hat{y}_i$	$\hat{\varepsilon}_i^2$
1	52,8	16,3	52,319	0,481	0,231
2	51,9	16,8	52,860	–0,960	0,922
3	54,2	18,5	54,699	–0,499	0,249
4	50,8	16,3	52,319	–1,519	2,309
5	54,9	17,9	54,050	0,850	0,723
6	53,9	17,4	53,509	0,391	0,153
7	53,1	16,1	52,103	0,997	0,994
8	52,4	16,2	52,211	0,189	0,036
9	53,0	17,0	53,077	–0,077	0,006
10	52,9	16,7	52,752	0,148	0,022
Součet	529,9	169,2	529,900	0,000	5,643

■

Na hodnotách reziduí jsou založeny mnohé diagnostické postupy, které umožňují posoudit, zda uvažovaný regresní model splňuje předpoklady, které jsme na náhodnou složku kladli. Různé diagnostické postupy jsou popsány například v Hebák a další, (2013), jsou také implementovány ve statistických programech. V tomto textu se omezíme pouze na dva grafy, které nám umožní přibližně posoudit, zda rezidua odporují nebo neodporují předpokladům, které na náhodnou složku klademe. Nejedná se tedy o žádný test, pouze o grafickou metodu, která může pomoci při regresní diagnostice, kdy posuzujeme splnění předpokladů kladených na regresní model. Pokud je v modelu pouze jedna vysvětlující proměnná, sestrojíme graf reziduí tak, že na vodorovnou osu umístíme hodnoty x_i a na svislou osu hodnoty reziduí $\hat{\varepsilon}_i$, $i = 1, 2, \ldots, n$, obrázek tedy obsahuje n bodů. Obrázek doplníme vodorovnou přímkou $\varepsilon = 0$, znázorňující nulové reziduum (obrázek 4.14). Platí-li předpoklady kladené na náhodnou složku, rezidua by měla náhodně kolísat kolem nulové hodnoty, jak plyne z předpokladu o nulové střední hodnotě náhodné složky. Vzhledem k předpokladu konstantního rozptylu rezidua kolísají uvnitř nějakého pásu. Tato situace je znázorněna na obrázku 4.14 v levé části plnými body. V levé části obrázku je dále (prázdnými kolečky) znázorněn průběh reziduí, kde je porušen předpoklad stejných rozptylů a zdá se, že rozptyl roste s hodnotou nezávisle proměnné x. V takové situaci se nabízí hledat nějakou transformaci proměnných stabilizující rozptyl, nejběžnější volbou je logaritmus nebo odmocnina (podrobnější informace můžeme nalézt například v Hebák a další, 2013). Pokud rezidua vypadají tak, jako v pravé části obrázku, naznačuje to, že zvolená lineární regresní funkce není vhodná pro popis závislosti mezi proměnnými a pro tento popis je třeba hledat vhodnou transformaci proměnných, polynomickou regresní funkci či nelineární regresní funkci.



Obr. 4.14 Grafická diagnostika reziduí

Nyní označíme průměr hodnot vysvětlované proměnné $\bar{y} = \sum_{i=1}^{n} y_i / n$. Definujme (v souladu s první kapitolou knihy věnovanou popisné statistice a analýze rozptylu v části 3.4.4 a 4.2):

celkový součet čtverců odchylek (bez ohledu na zvolený regresní model) jako

$$S_y = \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{4.52}$$

součet čtverců odchylek vysvětlených modelem S_T

$$S_T = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

a **reziduální součet čtverců odchylek** S_R

$$S_R = Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

Lze ukázat, že platí

$$S_y = S_T + S_R. \tag{4.53}$$

Na tomto rozkladu celkového součtu čtverců můžeme založit posouzení kvality regresního modelu. Hodnota S_y je pro daná pozorování $y_1, y_2, \ldots, y_n$ vysvětlované proměnné konstantní a podle velikosti S_T a S_R můžeme posuzovat kvalitu regresní funkce jako modelu pro závislost mezi proměnnými x a y. Všimněme si, že čím regresní model vystihuje lépe zkoumanou závislost, tím je větší hodnota S_T, a tedy je menší hodnota S_R. V případě deterministické lineární závislosti, kdy všechny pozorované body leží na jedné přímce, je $S_y = S_T$ a $S_R = 0$. Naopak v případě, kdy lineární závislost mezi zkoumanými veličinami neexistuje, je $S_y = S_R$ a $S_T = 0$. Populační regresní přímka je pak rovnoběžná s vodorovnou osou, odhadnutá regresní přímka je přibližně rovnoběžná s touto osou.

Posouzení intenzity regresní závislosti je jedním z úkolů regresní a korelační analýzy. Posuzovaný vztah je tím silnější a regresní funkce tím lepším popisem, čím více

jsou empirické hodnoty vysvětlované proměnné soustředěné kolem odhadnuté regresní funkce. A naopak – vztah je tím slabší, čím více jsou empirické hodnoty vzdáleny hodnotám vyrovnaným regresní funkcí, tedy hodnotám na odhadnuté regresní funkci. Je vidět, že nalezení míry intenzity závislosti souvisí s kvalitou regresního odhadu.

Kvalitu regresního modelu můžeme posoudit **indexem determinace** I^2, který je definován jako

$$I^2 = \frac{S_T}{S_y} (\cdot 100\,\%) \tag{4.54}$$

a můžeme ho interpretovat jako procento variability vysvětlované proměnné vysvětlené regresním modelem přímky. Z výše uvedeného plyne, že pro lineárně nezávislé veličiny je $I^2 = 0$. V případě deterministické lineární závislosti je $I^2 = 1$, a regresním modelem je tedy vysvětleno 100 % variability vysvětlované proměnné. V takovém případě v regresním modelu (4.48) opravdu žádná variabilita není, $\varepsilon = 0$ a $\sigma^2 = 0$. Index determinace charakterizuje pouze intenzitu závislosti popsané zvolenou regresní funkcí, nikoliv její směr. Proto neodpovídá na otázku, zda je závislost přímá (rostou-li hodnoty jedné proměnné, spíše rostou také hodnoty druhé proměnné) nebo nepřímá (roste-li jedna proměnná, spíše klesá druhá). V prvním případě je odhadnutá směrnice regresní přímky kladná a odhadnutá přímka je rostoucí, ve druhém případě je směrnice regresní přímky záporná a přímka je klesající. Lze ukázat, že pokud koeficient determinace I^2 odmocníme (charakterizuje sílu lineární závislosti) a přidáme znaménko směrnice regresní přímky (charakterizuje směr lineární závislosti), dostaneme korelační koeficient r_{xy}, definovaný v (4.31). Vlastnostmi (výběrového) korelačního r_{xy} a možnostmi statistických úsudků založených na tomto koeficientu se budeme dále zabývat v části 4.3.8, která je věnována korelační analýze.

Jiné běžné označení pro koeficient I^2 je R^2, toto značení se používá ve výstupech v softwaru a vychází ze vztahu mezi korelačním koeficientem a indexem determinace, kdy $r_{xy}^2 = I^2$.

K dokončení popisu regresního modelu ještě zbývá nalézt odhad reziduálního rozptylu σ^2, který potřebujeme pro induktivní úsudky o parametrech regresní funkce. Odhad rozptylu náhodné složky nazveme **reziduální rozptyl**, označíme ho s_R^2 a definujeme jako

$$s_R^2 = \frac{Q(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2. \tag{4.55}$$

Velikost reziduálního rozptylu může sloužit jako charakteristika kvality regresního modelu, neboť pokud je jeho hodnota malá, jsou pozorované body blízké regresní funkci, a tedy regresní funkce dobře popisuje zkoumanou závislost. Takže můžeme hledat regresní funkce, které mají malý reziduální rozptyl. Odmocnina z reziduálního rozptylu $s_R = \sqrt{s_R^2}$ se nazývá **reziduální směrodatná odchylka**.

Pro příklad 4.9 můžeme nyní určit reziduální rozptyl ve tvaru

$$s_R^2 = \frac{5,643}{8} = 0,705$$

a reziduální směrodatnou odchylku jako

$$s_R = \sqrt{0,705} = 0,840.$$

Nyní bodové odhady regresních koeficientů (směrnice a posunutí regresní přímky) doplníme intervaly spolehlivosti. Vzorec (4.40), umožňující nalézt odhad $\hat{\boldsymbol{\beta}}$ pomocí maticového zápisu, lze doplnit o kovarianční matici vektoru odhadů ve tvaru

$$\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}. \tag{4.56}$$

Odhadneme-li neznámý rozptyl σ^2 náhodné složky reziduálním rozptylem s_R^2 podle (4.55), lze odhady směrodatné chyby odhadů regresních parametrů $\hat{\beta}_0$ a $\hat{\beta}_1$ najít jako

$$s_{\hat{\beta}_0} = s_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}, \tag{4.57}$$

$$s_{\hat{\beta}_1} = s_R \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}. \tag{4.58}$$

Hledané oboustranné $100(1 - \alpha)\%$ intervaly spolehlivosti pro parametry β_0 a β_1 populační regresní přímky pak mají tvar (při všech následujících úvahách používáme pro Studentovo rozdělení $n - 2$ stupni volnosti)

$$\left(\hat{\beta}_0 - t_{1-\alpha/2} s_{\hat{\beta}_0},\ \hat{\beta}_0 + t_{1-\alpha/2} s_{\hat{\beta}_0} \right), \tag{4.59}$$

$$\left(\hat{\beta}_1 - t_{1-\alpha/2} s_{\hat{\beta}_1},\ \hat{\beta}_1 + t_{1-\alpha/2} s_{\hat{\beta}_1} \right). \tag{4.60}$$

Dále budeme testovat hypotézy o regresních parametrech a regresním modelu. V regresní analýze nás nejčastěji zajímá, jestli je některý z parametrů roven nule; takový test nazýváme **individuálním testem o regresním parametru**. V takovém případě testujeme každý parametr zvlášť. V případě testu hypotézy o parametrech regresní přímky (pro j rovné 0 a 1) vyjádříme nulovou a alternativní hypotézu ve tvaru

$$H_0: \beta_j = 0 \quad H_1: \beta_j \neq 0. \tag{4.61}$$

Test můžeme založit na testovém kritériu

$$T = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}, \tag{4.62}$$

které má za platnosti nulové hypotézy Studentovo rozdělení s $n - 2$ stupni volnosti. Pokud platí nulová hypotéza, je hodnota zlomku blízká nule (platí $\beta_j = 0$, a proto také

jeho odhad $\hat{\beta}_j$ je blízký nule) a proti nulové hypotéze tedy hovoří příliš nízké nebo příliš vysoké hodnoty testového kritéria T; kritický obor má potom tvar

$$W_\alpha = \{t; |t| \geq t_{1-\alpha/2}\}. \tag{4.63}$$

Kritická hodnota je popsána kvantilem Studentova rozdělení, stupně volnosti jsou rovny $n-2$. Průběh testu je shrnut v tabulce 4.23, zvolíme-li počet parametrů modelu $p = 2$ a $\beta_{0j} = 0, j = 0, 1$.

Poznamenejme, že test o směrnici β_1 je pro regresní přímku obdobný **testu o regresním modelu**. V testu o regresním modelu je v nulové hypotéze lineární nezávislost proměnných X a Y. V takovém případě není model přímky vhodný pro popis závislosti mezi zkoumanými proměnnými. Alternativní hypotéza říká, že závislost popsaná modelem existuje, že je statisticky významná. Nulová hypotéza a alternativa mají v takovém případě tvar

$$H_0: \beta_0 = c, \beta_1 = 0 \quad H_1: \beta_1 \neq 0.$$

Test je založen na rozkladu celkového součtu čtverců S_y podle (4.53). Testovým kritériem je statistika

$$F = \frac{\frac{S_T}{S_R}}{n-2}, \tag{4.64}$$

která má za platnosti nulové hypotézy Fisherovo-Snedecorovo rozdělení F s parametry 1 a $n-2$. Kritický obor má tvar

$$W_\alpha = \{F; F \geq F_{1-\alpha}\}, \tag{4.65}$$

kde stupně volnosti pro kritickou hodnotu jsou rovny parametrům F rozdělení za platnosti hypotézy. Test je speciálním případem testu o regresním modelu, který je shrnut v tabulce 4.24, kde opět volíme $p = 2$.

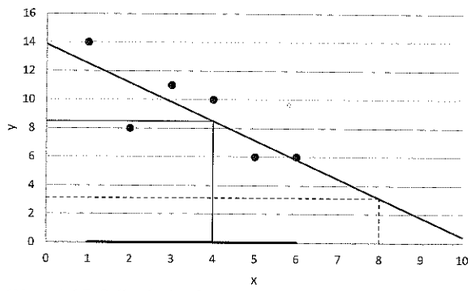
Regresní model můžeme použít i pro odhady středních hodnot vysvětlované proměnné Y pro zvolenou hodnotu x nebo individuální hodnoty y v tomto bodě. Takové postupy nazýváme také **předpovědi** nebo **predikce**. Popíšeme-li přímkou vztah mezi mzdou a délkou praxe, můžeme se zajímat o střední mzdu pro zaměstnance s 10 lety praxe (odhad střední hodnoty Y) nebo o mzdu jednoho konkrétního zaměstnance s 10 lety praxe. V takovém případě se snažíme odhadnout jednu konkrétní hodnotu mzdy.

Z hlediska hodnoty vysvětlující proměnné x rozlišujeme u těchto předpovědí **interpolace** a **extrapolace**. O interpolaci se jedná v případě, kdy hodnota nezávisle proměnné je obsažena v intervalu mezi nejmenší a největší hodnotou pozorované nezávisle proměnné, ve druhém případě leží mimo tento interval. Jestliže jsme regresní přímku odhadovali z hodnot v intervalu (x_{min}, x_{max}), pak se jedná o interpolaci, pokud by nás zajímala předpověď pro hodnotu vysvětlující proměnné x takovou, že $x_{min} < x < x_{max}$, a o extrapolaci pro $x < x_{min}$ nebo $x > x_{min}$. V případě extrapolace používáme odhadnutou regresní přímku mimo interval, ve kterém jsme ji odhadovali. V takovém případě musíme být v úsudcích opatrnější než v případě, kdy usuzujeme uvnitř

intervalu pozorování vysvětlující proměnné. Extrapolace využijeme například při analýze časové řady v následující kapitole pro predikce hodnot časové řady.

Zvolme tedy jednu konkrétní hodnotu vysvětlující proměnné x. Bodovým odhadem pro střední hodnotu i individuální hodnotu vysvětlované proměnné je hodnota na regresní přímce

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$



Obr. 4.15 Odhady hodnot vysvětlující proměnné

Na obrázku 4.15 je znázorněn lineární regresní model, přímka byla odhadnuta pro hodnoty x z intervalu $\langle 1; 6\rangle$. Znázornili jsme dva odhady hodnot vysvětlované proměnné, interpolaci pro $x = 4$ (plná čára) a extrapolaci pro $x = 8$ (čárkovaně); hodnoty odhadů je možné nalézt na svislé ose. Odhadnutá regresní přímka má tvar $13{,}9 - 1{,}3x$, dostáváme tedy předpovědi

- $x = 4$: $\hat{y} = 13{,}9 - 1{,}3 \cdot 4 = 8{,}7$,
- $x = 8$: $\hat{y} = 13{,}9 - 1{,}3 \cdot 8 = 3{,}5$.

Všimněme si, že pro volbu vysvětlující proměnné vyšší než 10 by hodnota na regresní funkci $\hat{y}$ byla záporná i v případě, že z věcného hlediska by taková možnost vůbec nemohla nastat.

Tyto bodové odhady ještě doplníme směrodatnou chybou a sestrojíme intervaly spolehlivosti pro předpovědi. Podle Hebák a další (2013) je směrodatná chyba odhadu $\hat{y}$ střední hodnoty proměnné Y pro hodnotu vysvětlující proměnné x rovna

$$s_{\hat{y}} = s_R \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}}. \tag{4.66}$$

Potom lze $100(1-\alpha)\%$ interval spolehlivosti tuto střední hodnotu zapsat ve tvaru

$$\left(\hat{y} - t_{1-\alpha/2} s_{\hat{y}},\ \hat{y} + t_{1-\alpha/2} s_{\hat{y}}\right). \tag{4.67}$$

Meze intervalu (4.67) můžeme použít ke konstrukci pásu kolem regresní přímky, který bude intervalovým odhadem pro populační regresní přímku se spolehlivostí $1 - \alpha$.

Ještě nás bude zajímat interval spolehlivosti pro **individuální hodnotu proměnné** Y, tedy pro odhad individuální hodnoty vysvětlované proměnné pro hodnotu vysvětlující proměnné x. Interval spolehlivosti bude mít stejný střed $\hat{y}$ jako interval spolehlivosti pro střední hodnotu, musí být ale kvůli větší nejistotě, vyjádřené směrodatnou odchylkou, širší. Dostaneme

$$\left(\hat{y}-t_{1-\alpha/2}\sqrt{s_R^2+s_{\hat{y}}^2},\ \hat{y}+t_{1-\alpha/2}\sqrt{s_R^2+s_{\hat{y}}^2}\right). \tag{4.68}$$

Odpovídající pás kolem regresní přímky by měl v průměru obsahovat 95 % z n pozorování. Tento pás se dá použít také pro odhady individuálních hodnot, proto se mu někdy říká **predikční pás** kolem regresní přímky.

Ve vzorcích (4.67) a (4.68) si povšimněme, že poloviční šířka intervalu spolehlivosti $t_{1-\alpha/2}\, s_{\hat{y}}$ (resp. $t_{1-\alpha/2}\sqrt{s_R^2+s_{\hat{y}}^2}$) není pro všechny hodnoty vysvětlující proměnné x stejná, a tedy všechny odhady nejsou stejně přesné. Nejmenší šířky dosahují intervaly v případě, že $x = \bar{x}$. Bude-li se x vzdalovat od hodnoty $\bar{x}$ v libovolném směru, bude se šířka intervalu zvětšovat.

Vraťme se nyní k problému sdružených regresních přímek, kdy zaměníme vysvětlující a vysvětlovanou proměnnou. Budeme tedy uvažovat klasický regresní model takový, že vysvětlovaná proměnná je x a vysvětlující proměnná je y. Takový model můžeme zapsat ve tvaru

$$x_i = \alpha_0 + \alpha_1 y_i + \varepsilon'.$$

Náhodnou složku jsme označili ε' a budeme předpokládat, že splňuje předpoklady klasického lineárního modelu.

Probereme dvě mezní vzájemné polohy odhadnutých regresních přímek. Pokud je směrnice $\hat{\beta}_1$ nulová, je nulový i odhad směrnice $\hat{\alpha}_1$ a odhadnuté přímky mají tvar $\hat{y} = \bar{y}$, $\hat{x} = \bar{x}$. Sdružené přímky jsou v takovém případě navzájem kolmé a rovnoběžné s příslušnými osami souřadné soustavy. Tento případ odpovídá lineárně nezávislým veličinám x a y. V opačném případě, kdy všechna pozorování leží na jedné přímce, jsou všechna rezidua nulová a proměnné jsou deterministicky lineárně závislé. V takovém případě platí vztahy, které jsme odvodili v (4.30).

Indexy determinace v obou regresních modelech popsaných sdruženými regresními přímkami nabývají stejné hodnoty. Obdobně jsou také stejné hodnoty testové statistiky pro individuální test o směrnici a také o regresním modelu, obě statistiky mají stejné rozdělení, a tedy rozhodnutí jsou naprosto stejná. Všimněme si, že obě odhadnuté směrnice musejí mít také stejné znaménko a jejich součin je tedy nezáporný. Lze se přesvědčit, že tento součin je roven (společnému) indexu determinace, platí tedy

$$I^2 = \hat{\beta}_1\hat{\alpha}_1.$$

Příklad 4.10

Uvažujme závislost mezi cenou volně stojící myčky (šířka 60 cm) y v Kč a její hlučností x (dB). Máme k dispozici 32 dvojic pozorování. Lze očekávat nepřímou úměrnost, protože tišší přístroje jsou asi spíše dražší než přístroje hlučnější. V tomto případě má z věcného hlediska smysl uvažovat i sdruženou regresní přímku, která vysvětluje hlučnost myčky její cenou.

Z pozorovaných hodnot byla určena průměrná cena 13 096 Kč (výběrová směrodatná odchylka 3909 Kč) a průměrná hlučnost 44,8 dB (minimální hlučnost je 40 dB, maximální hlučnost 52 dB, výběrová směrodatná odchylka je rovna 2,6 dB). V tabulce 4.19 jsou uvedeny odhadnuté parametry regresní přímky pořízené pomocí statistického softwaru, jejich směrodatné chyby, hodnoty testového kritéria pro individuální testy o parametrech a p-hodnoty, příslušné těmto testům.

Odhadnutá regresní funkce má tvar

$$y = 51\,798 - 863{,}65x. \tag{4.69}$$

Pro test hypotézy, že směrnice β_1 je nulová, použijeme testové kritérium podle (4.62)

$$t=\frac{-863{,}65}{220{,}531}=-3{,}92,$$

a kritický obor je

$$W_{0,05}=\{t;|t|\geq t_{1-0{,}05/2}\}=(-\infty;\ -2{,}042\rangle\cup\langle 2{,}042;\infty).$$

Kritické hodnoty jsou podle (4.63) rovny $-t_{0,975}(30) = -2{,}042$ a $t_{0,975}(30) = 2{,}042$. Hodnota testového kritéria –3,92 je prvkem kritického oboru $W_{0,05}$, proto zamítáme (na hladině významnosti 0,05) hypotézu o nulové hodnotě směrnice populační přímky. Lineární závislost mezi proměnnými je tedy statisticky významná.

Tab. 4.19 Odhad parametrů a individuální testy o parametrech, závislost ceny na hlučnosti

Proměnná	Bodový odhad	Směrodatná chyba	t	p-hodnota
Posunutí	51 798	9 899,07	5,23	< 0,0001
Hlučnost	–863,65	220,53	–3,92	0,000 5

Tab. 4.20 Odhad parametrů a individuální testy o parametrech, závislost hlučnosti na ceně

Proměnná	Bodový odhad	Směrodatná chyba	t	p-hodnota
Posunutí	49,942 1	1,365 2	36,58	< 0,000 1
Cena	–0,000 4	0,000 1	–3,92	0,000 5

Hodnotu odhadnuté směrnice regresní přímky můžeme interpretovat tak, že pokud vzroste hlučnost o jeden decibel, klesne cena v průměru o 864 Kč. Pokud by nás zají-

mal interval spolehlivosti, podle vzorce (4.60) použijeme již nalezený kvantil $t_{0,975}(28) = 2,042$ a dostaneme

$$\hat{\beta}_1 \pm t_{1-\alpha/2} s_{\hat{\beta}_1} = -863,65 \pm t_{0,975} \cdot 220,53 = -863,65 \pm 450,32.$$

Dostáváme tedy interval (1314; −413) a můžeme říci, že tento interval pokrývá střední snížení ceny s pravděpodobností 0,95. Pokud tedy myčka je o 1 dB hlučnější, myčka bude v průměru o 413 až 1314 Kč levnější.

Tabulka 4.21, popisující rozklad celkového součtu čtverců odchylek S_y, obsahuje podrobnou informaci o regresním modelu. Například (použijeme-li tabulku 4.25) ze sloupce stupně volnosti můžeme vyčíst počet parametrů p modelu (2 = 1 + 1) a rozsah výběru (32 = 31 + 1 nebo 30 + 2). Testové kritérium pro test o regresním modelu z definice (4.64) dostaneme ve tvaru

$$F = \frac{\frac{S_T}{n-2}}{\frac{S_R}{32-2}} = \frac{160\,273\,754}{313\,507\,246} = 15,34.$$

Podle p-hodnoty je regresní model statisticky významný, neboli zamítáme nulovou hypotézu o nezávislosti ceny a hlučnosti. Stejný výsledek bychom dostali pomocí kritického oboru (volíme $\alpha = 0,05$ a použijeme kvantil $F_{0,95}(1,30) = 4,171$)

$W_\alpha = \{F; F \geq F_{1-0,05}\} = \langle 4,171; \infty)$.

Z tabulky 4.21 určíme také index determinace jako

$$I^2 = \frac{160\,273\,754}{473\,781\,000} = 0,338\,3 \ (33,83\,\%),$$

tedy model vysvětluje necelých 34 % variability ceny, zbylých 66 % vysvětleno nebylo a tato variabilita je zahrnuta v reziduálních čtvercích.

Ještě určíme výběrový korelační koeficient r_{xy} tak, že najdeme odmocninu z indexu determinace $\sqrt{I^2} = \sqrt{0,338\,3} = 0,581\,6$ a přidáme záporné znaménko směrnice odhadnuté přímky, tedy $r_{xy} = -0,581\,6$. Stejnou hodnotu dostaneme tak, že vynásobíme odhadnuté směrnice obou regresně sdružených přímek

$$\sqrt{-863,65 \cdot -0,000\,4} = 0,581\,6,$$

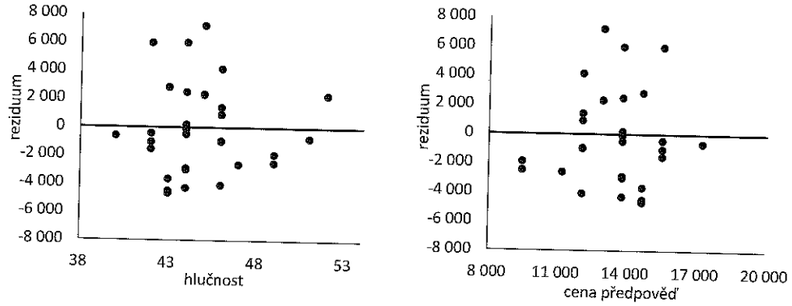
opět ale musíme přidat společné záporné znaménko směrnice. Třetí možností je vypočítat hodnotu koeficientu z definice (4.31).

Tab. 4.21 Analýza rozptylu pro regresní model z příkladu 4.10

Zdroj variability	Stupně volnosti	Součet čtverců	Průměrný čtverec	F	p-hodnota
Model	1	160 273 754	160 273 754	15,34	0,000 5
Reziduální	30	313 507 246	10 450 242		
Celkem	31	473 781 000			

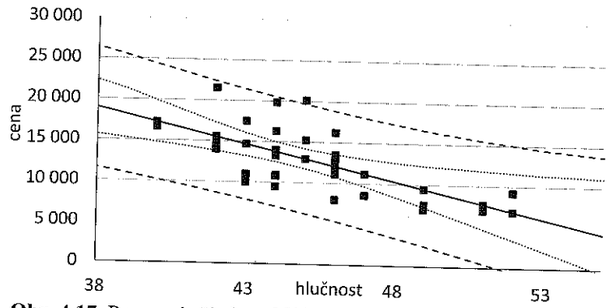
Z tabulky 4.21 také přímo získáme odhad rozptylu reziduí $s_R^2 = 10\,450\,242$, a tedy reziduální směrodatná odchylka je $s_R = \sqrt{10\,450\,242} = 3233$ Kč.

Na obrázcích 4.16 jsou dále ukázány základní možnosti, jak posoudit kvalitu modelu z hlediska splnění předpokladů o náhodné složce. Na prvním obrázku jsou rezidua znázorněna proti vysvětlující proměnné hlučnost, na druhém jsou pak na vodorovné ose hodnoty předpovědí hodnot vysvětlované proměnné. První obrázek je alternativou k obrázku 4.14. Na žádném obrázku rezidua modelu nevykazují jasné známky toho, že by byly porušeny předpoklady o náhodné složce.



Obr. 4.16 Grafy reziduí k příkladu 4.10

Na závěr příkladu jsou na obrázku 4.17 znázorněna pozorování, odhadnutá regresní přímka, předpovědi na regresní funkci a dále pás spolehlivosti pro regresní přímku (užší pás znázorněný tečkovaně) a pro předpovědi (vnější pás znázorněný čárkovaně). V tomto vnějším pásu by mělo ležet 95 % pozorování, z 32 leží na vnější hranici jedno pozorování (3 %), další dvě jsou na hranici uvnitř (6 %).



Obr. 4.17 Regresní přímka s 95% pásem spolehlivosti pro regresní přímku a s predikčním pásem

Ještě se pokusíme odhadnout střední cenu myčky a cenu jedné myčky, která má hlučnost 45 a 53 dB. Z údajů v zadání příkladu vyplývá, že v případě 45 dB jde o interpolaci a pro 53 dB o extrapolaci, neboť maximální hodnota hlučnosti v datech je 52 dB. Z odhadnuté regresní funkce (4.69) dostaneme bodové odhady

$\hat{y}$ = 51 798 – 863,65·45 = 12 934 Kč,

$\hat{y}$ = 51 798 – 863,65·53 = 6025 Kč.

Intervalové odhady se již budou lišit pro střední cenu myčky (budou užší) a pro individuální hodnotu (budou širší). Použijeme-li vztahy (4.66), (4.67) a (4.68), s_R = 3233 Kč, $\bar{x}$ = 44,8, $\sum(x_i-\bar{x})^2$ = 214,875 a $t_{0,975}(30)$ = 2,042, dostaneme pro obě volby vysvětlující proměnné směrodatné chyby předpovědí $s_{\hat{y}}$

$$\text{pro } x=42:\quad s_{\hat{y}} = s_R\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}} = 3\,233\cdot\sqrt{\frac{1}{32}+\frac{(42-44{,}8)^2}{214{,}875}} = 814 \text{ Kč},$$

$$\text{pro } x=53:\quad s_{\hat{y}} = s_R\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}} = 3\,233\cdot\sqrt{\frac{1}{32}+\frac{(53-44{,}8)^2}{214{,}875}} = 1897 \text{ Kč}.$$

Všimněme si, že směrodatná chyba pro 53 dB je větší než pro 42 dB. Pro hlučnost 42 dB sestrojíme také intervaly spolehlivosti pro individuální i průměrnou hodnotu ceny (v Kč). Pro průměrnou hodnotu je 95% interval spolehlivosti podle (4.67) roven

(12 934 – 2,042·814; 12 934 + 2,042·814) = (11 216; 14 652),

pro individuální cenu dostaneme širší interval, po dosazení do vztahu (4.68) je výsledkem interval

$$\left(12\,934-2{,}042\cdot\sqrt{814^2+3233^2};\ 12\,934+2{,}042\cdot\sqrt{814^2+3233^2}\right)=(6111;19\,757).$$

■

Pokud bychom uvažovali regresní modely z části (4.3.4), v případě funkcí lineárních v parametrech můžeme použít postupy uvedené výše pro přímku, jen se budou vztahovat k transformovaným proměnným. Můžeme například uvažovat klasický lineární regresní model s hyperbolickou regresní funkcí (4.41) ve tvaru

$$y=\beta_0+\beta_1\frac{1}{x}+\varepsilon,$$

nebo s logaritmickou funkcí (4.42)

$$y=\beta_0+\beta_1\ln x+\varepsilon.$$

Je ale třeba mít na paměti, že v prvním případě pracujeme se závislostí vysvětlované proměnné y na proměnné $1/x$ a ve druhém se závislostí na vysvětlující proměnné $\ln x$, nikoliv se závislostí na x.

Pokud je regresní funkce nelineární v parametrech (jako jsou například již probrané funkce mocnina, exponenciála a dále tyto funkce s posunutím), již se nejedná o lineární regresní model, ale o nelineární regresní model i v případě, že náhodné složky splňují předpoklady klasického lineárního regresního modelu. Práce s takovými modely a statistické úsudky o modelu a jeho parametrech přesahují možnosti tohoto textu, postupy lze najít ve specializovaných textech Hebák, Hustopecký, Malá (2005), Hebák a další (2013). Poznamenejme jen, že pokud použijeme linearizaci modelu (postup, který jsme představili pro mocninou funkci a exponenciální funkci), můžeme použít výše uvedené postupy lineární regrese pouze v linearizovaném modelu. Všimněme si, že například model

$$y=\beta_0\beta_1^x+\varepsilon$$

nelze přímo linearizovat, neboť je

$$\ln(y)=\ln(\beta_0\beta_1^x+\varepsilon)$$

a logaritmus součtu na pravé straně rovnice nelze snadno vypočítat. Linearizovaný lineární model

$$\ln(y)=\beta_0+x\ln(\beta_1)+\varepsilon',$$

se kterým umíme dobře pracovat, zřejmě není z hlediska předpokladů o rozdělení náhodných složek shodný. Proto lze dobře interpretovat vztah mezi logaritmem proměnné y a proměnnou x v případě exponenciální regresní funkce a mezi logaritmy obou proměnných v případě mocninné regresní funkce v základním souboru. Problematickou je nelineární zpětná transformace na vztah mezi vysvětlovanou proměnnou y a vysvětlující proměnnou x. Tato transformace způsobuje, že takto získané odhady parametrů v původním regresním modelu nemají vlastnosti uvedené v přehledu pro odhady v klasickém lineárním modelu, a to ani asymptoticky pro velké náhodné výběry.

4.3.6 Vícenásobná regrese a korelace

Doposud jsme se zabývali závislostí dvou proměnných a tuto závislost jsme charakterizovali regresní přímkou, která je popsána dvěma parametry. Podařilo se nám najít vhodný popis také pro závislosti, pro jejichž vyjádření se přímo nehodila přímka. Uvedli jsme možnost transformací vysvětlované nebo vysvětlující proměnné nebo zobecnění lineárního regresního modelu na nelineární regresní model, který je popsán regresní funkcí nelineární v neznámých parametrech. Další možností, jak postupovat při konstrukci regresního modelu, je použít více vysvětlujících proměnných, nebo místo přímky použít polynom vyššího řádu, tedy polynomickou regresi.

Vícenásobná regrese

Nyní budeme zkoumat závislost mezi jednou vysvětlovanou a více vysvětlujícími proměnnými, které také nazýváme **regresory**. Předpokládejme tedy, že máme k vysvětlujících kvantitativních proměnných, které označíme $x_1, x_2, \ldots, x_k$. Můžeme například sledovat závislost mzdy (v korunách) na třech vysvětlujících proměnných: věku, počtu let, které člověk strávil vzděláváním, a délky praxe v letech. Polynomická regrese je speciálním případem vícenásobné regrese, neboť například pro polynom řádu

dva máme dvě vysvětlující proměnné x a x^2, pro kubickou parabolu jsou vysvětlující proměnné tři – x, x^2 a x^3. Proto se všechny postupy, kterými se zde budeme zabývat, používají i při práci s klasickým lineárním modelem s polynomickou regresní funkcí.

Závislost mezi vysvětlovanou proměnnou y a k-rozměrným vektorem $\mathbf{x}$, do kterého umístíme k vysvětlujících proměnných $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$, popíšeme lineární regresní funkcí

$$f(\mathbf{x};\boldsymbol{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k. \qquad (4.70)$$

Zkoumání závislostí tohoto typu se nazývá **vícenásobnou (mnohonásobnou) regresí**. Předpokládejme, že data obsahují n pozorování vektoru $(y_i, \mathbf{x}_i) = (y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$, kde u vysvětlujících proměnných označuje první index pozorování ($i = 1, 2, \ldots, n$) a druhý index pořadí vysvětlující proměnné ($j = 1, 2, \ldots, k$). Regresní model má pak tvar

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i,$$

kde náhodné veličiny ε_i splňují předpoklady klasického lineárního modelu. V maticovém tvaru můžeme opět psát

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde regresní matice $\mathbf{X}$ má n řádků a $k + 1$ sloupců a platí

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}_{n\times(k+1)}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \ldots \\ \beta_k \end{pmatrix}_{(k+1)\times 1}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \ldots \\ y_n \end{pmatrix}_{n\times 1} \text{ a } \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \ldots \\ \varepsilon_n \end{pmatrix}_{n\times 1}.$$

Parametry $\beta_j, j = 1, 2, \ldots, k$ nazýváme také **dílčí regresní parametry**, neboť udávají změnu střední hodnoty vysvětlované proměnné y, jestliže hodnoty všech ostatních proměnných jsou stejné a vysvětlující proměnná x_j se změní o jednu jednotku. Parametr β_0 budeme nazývat **absolutní člen**. Celkem je třeba odhadnout $p = k + 1$ parametrů regresní funkce a rozptyl náhodné složky σ^2. Odhad $\hat{\boldsymbol{\beta}}$ vektoru parametrů metodou nejmenších čtverců získáme podle (4.40), pro hledání odhadů regresních parametrů používáme software. Odhadnutou regresní funkci pak lze zapsat jako

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \qquad (4.71)$$

Pro znázornění odhadnutého modelu ale již není možné použít graf, který znázorňuje x, y a odhadnutou regresní funkci, v případě pouze jedné vysvětlující proměnné v přímkové nebo polynomické regresi. Vhodným grafem je například obrázek, kdy na vodorovnou osu zobrazujeme hodnotu na regresní funkci $\hat{y}$, $i = 1, 2, \ldots, n$ a na osu svislou pozorovanou hodnotu y_i. V případě deterministické závislosti by všechny takové body ležely na ose prvního a třetího kvadrantu. Čím jsou tedy body blíže této přímce, tím model lépe vystihuje analyzovaná data.

Příklad 4.11

Uvažujme závislost mezi cenou volně stojící myčky (vysvětlovaná proměnná y v Kč) a roční spotřebou vody (x_1,v litrech), roční spotřebou elektřiny (x_2, v kW), hlučností (x_3, v dB) a počtem programů x_4, které nabízejí. Použijeme regresní model se čtyřmi vysvětlujícími proměnnými

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

a neznámé parametry odhadneme na základě vybraných $n = 32$ přístrojů. V příkladu je $k = 4$ a $p = 5$. Pomocí softwaru byly nalezeny odhady parametrů z tabulky 4.22.

Tab. 4.22 Odhadnuté parametry regresního modelu pro příklad 4.11

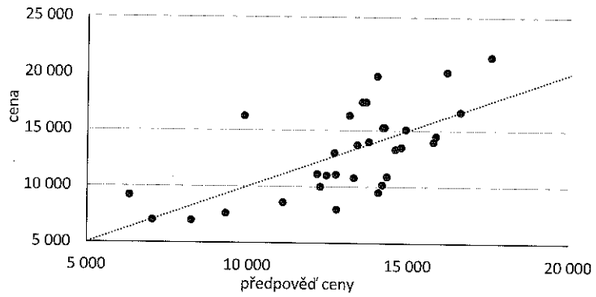
Proměnná	Odhad parametru	Směrodatná chyba	t	p-hodnota
konstanta	67 610	14 868	4,55	0,000 1
voda	−2,478	1,5326	−1,62	0,117 5
hlučnost	−925,493	345,899	−2,68	0,012 5
elektřina	−12,296	34,129	−0,36	0,721 4
programy	−422,531	324,706	−1,30	0,204 2

Odhadnutý regresní model má tvar

$$\hat{y} = 67\,610 - 2{,}48x_1 - 925{,}49x_2 - 12{,}30x_3 - 422{,}53x_4. \qquad (4.72)$$

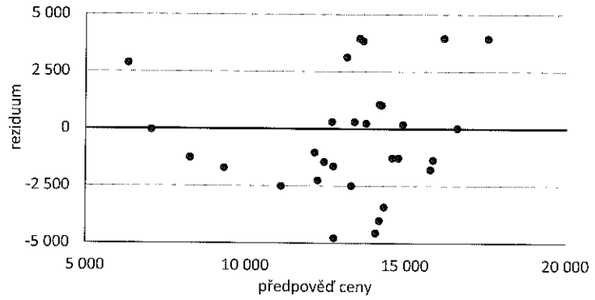
Odhadnuté regresní koeficienty jsou záporné, odhadnutá cena myčky tedy klesá s růstem všech vysvětlujících proměnných. Takovou vlastnost bychom mohli odvodit před analýzou dat pro obě spotřeby (vody a elektřiny) a pro hlučnost, nikoliv pro programy. Poslední výsledek odporuje intuitivnímu významu proměnných v modelu.

Na obrázku 4.18 jsou znázorněny hodnoty $\hat{y}_i$ na ose vodorovné a y_i na svislé ose. Osa prvního a třetího kvadrantu je znázorněna tečkovaně. Tento obrázek dobře popisuje kvalitu regresního modelu, pokud uvažujeme více vysvětlujících proměnných.



Obr. 4.18 Graf regresní funkce

Na obrázku 4.19 jsou znázorněna rezidua modelu, na vodorovné ose jsou předpovědi $\hat{y}_i$ hodnot vysvětlované proměnné.



Obr. 4.19 Graf reziduí modelu vícenásobné regrese

Testy hypotéz o parametrech v modelu vícenásobné regrese

Zabývejme se nyní testem hypotézy o parametrech regresních funkcí. Testujeme hypotézu

$$H_0: \beta_j = \beta_{0,j}$$

proti některé z alternativ:

$$H_1: \beta_j \neq \beta_{0,j}, \quad H_1: \beta_j > \beta_{0,j}, \quad H_1: \beta_j < \beta_{0,j}.$$

Za platnosti nulové hypotézy má statistika

$$T = \frac{\hat{\beta}_j - \beta_{0,j}}{s_{\hat{\beta}_j}} \tag{4.73}$$

Studentovo rozdělení o $n - p$ stupních volnosti. Kritické obory jsou v tabulce 4.23.

Tab. 4.23 Test hypotézy o regresních parametrech

H_0	H_1	Testové kritérium	Kritický obor
$\beta_j = \beta_{0,j}$	$\beta_j \neq \beta_{0,j}$ $\beta_j > \beta_{0,j}$ $\beta_j < \beta_{0,j}$	$T = \frac{\hat{\beta}_j - \beta_{0,j}}{s_{\hat{\beta}_j}}$ $T \sim t(n-p)$	$W_\alpha = \{t; \lvert t\rvert \geq t_{1-\alpha/2}\}$ $W_\alpha = \{t; t \geq t_{1-\alpha}\}$ $W_\alpha = \{t; t \leq -t_{1-\alpha}\}$

Zvláštním případem uvedeného testu je individuální test o parametru, který používáme pro testování významnosti jednotlivých vysvětlujících proměnných v regresním modelu, tedy nulová hypotéza má tvar ($j = 0, 1, \ldots, k$)

$$H_0: \beta_j = 0$$

a alternativa říká, že dílčí regresní parametr není roven nule. V takovém případě je použití příslušné nezávisle proměnné v modelu oprávněné. Testové kritérium z tabulky 4.23 má v takovém případě tvar

$$T = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}. \tag{4.74}$$

Nyní se zabývejme **testem o celém regresním modelu**. Nulová hypotéza říká, že neexistuje závislost mezi proměnnými popsaná modelem, alternativní hypotéza, že závislost existuje. Pokud bychom hypotézu nezávislosti chtěli formulovat pomocí regresních parametrů, znamená to, že všechny regresní parametry u vysvětlujících proměnných jsou nulové. Nenulový zůstane pouze regresní parametr β_0, který má v takovém případě význam průměru vysvětlované proměnné $\bar{y}$. Pokud nezamítneme nulovou hypotézu, není model dobře zvolený a je třeba hledat jiné vysvětlující proměnné nebo složitější závislost. Nulová hypotéza, že neexistuje závislost popsaná modelem, má tvar

$$H_0: \beta_0 = c, \beta_1 = \beta_2 = \ldots = \beta_k = 0.$$

Alternativní hypotéza říká, že aspoň jedno β je nenulové, tedy

$$H_1: \text{existuje aspoň jedno } j = 1, 2, \ldots, k \text{ takové, že } \beta_j \neq 0.$$

Test je založen na analýze rozptylu, je možné použít tabulku 4.24 nebo tabulku analýzy rozptylu v tabulce 4.25, která obsahuje přehledně všechny potřebné charakteristiky modelu a bývá výstupem ve statistickém softwaru.

Tab. 4.24 Test o regresním modelu

H_0	H_1	Testové kritérium	Kritický obor
$\beta_0 = c, \beta_1 = 0$ $j = 1, 2, \ldots, k$	neplatí H_0	$F = \frac{S_T/(p-1)}{S_R/(n-p)}$ $F \approx F(p-1, n-p)$	$W_\alpha = \{F; F \geq F_{1-\alpha}\}$

Tab. 4.25 Tabulka analýzy rozptylu pro test o regresním modelu

Zdroj variability	Stupně volnosti	Součet čtverců	Průměrný čtverec	F	p-hodnota
Model	$p-1$	S_T	$S_T/(p-1)$	$\frac{S_T/(p-1)}{S_R/(n-p)}$	
Reziduální	$n-p$	S_R	$S_R/(n-p)$		
Celkem	$n-1$	S_y			

Tabulka analýzy rozptylu 4.25 pro regresní model obsahuje také hodnotu reziduálního rozptylu s_R^2, který je odhadem rozptylu náhodné složky σ^2

$$s_R^2 = \frac{S_R}{n-p}. \tag{4.75}$$

Tabulku 4.25 můžeme použít k vyčíslení indexu determinace I^2 podle vzorce (4.54). Hodnota indexu je opět mezi 0 a 1, čím větší je hodnota indexu, tím větší část variability vysvětlované proměnné je vysvětlena regresním modelem, a tím lepší je regresní model. Všimněme si, že v případě vícenásobné regrese není možné hovořit o směru závislosti, směr můžeme uvažovat vždy jen vzhledem k vybrané vysvětlující proměnné. Směr závislosti je pak charakterizován znaménkem dílčího regresního koeficientu a vztahuje se vždy k ostatním vysvětlujícím proměnným konstantním.

Pokud bychom pomocí indexu determinace porovnávali kvalitu více modelů, například model přímky, paraboly a kubické paraboly nebo model vícenásobné regrese s vysvětlujícími proměnnými x_1, x_2, x_3, x_4 a například modely s vysvětlujícími proměnnými x_1, x_4 nebo x_1, x_2, x_3, vždy bychom volili model s větším počtem vysvětlujících proměnných, v daném příkladu kubickou parabolu nebo model se všemi čtyřmi vysvětlujícími proměnnými. Je to proto, že po přidání jakékoliv vysvětlující proměnné nikdy nedojde k poklesu indexu determinace. Proto používáme upravený index determinace I^2_{ADJ}, který zohledňuje počet parametrů modelu a také počet dat, která byla použita k odhadu parametrů modelu. **Upravený index determinace** I^2_{ADJ}, je definován jako

$$I^2_{ADJ} = 1 - (1 - I^2)\frac{n-1}{n-p}. \tag{4.76}$$

Tento index již nemůžeme interpretovat pomocí podílu vysvětlené variability, používáme ho jen při srovnání kvality různých regresních modelů, odhadnutých ze stejného souboru dat. Při úpravě indexu determinace dochází k snížení jeho hodnoty a pro velmi malé hodnoty I^2 může upravený index determinace nabývat i záporných hodnot. Často se tento index značí R^2_{ADJ}, stejně jako používáme R^2 místo I^2.

Příklad 4.11 (pokračování)

V příkladu vícenásobné regrese provedeme individuální testy o parametrech a test o vhodnosti celého modelu. Dále nalezneme index determinace a hodnotu reziduálního rozptylu.

Standardní výpis výsledků odhadu v lineárním modelu v tabulce 4.22 obsahuje bodové odhady parametrů, směrodatnou chybu odhadů (podle (4.56) a (4.55)), dále potom hodnoty testového kritéria individuálního testu o parametru (hodnota podle (4.74)) a p-hodnotu pro tento test. Pomocí posledního sloupce v tabulce lze tedy rozhodnout o individuální významnosti jednotlivých vysvětlujících proměnných v modelu bez dalších výpočtů.

Individuální testy o parametrech pro test hypotézy o nulové hodnotě jednotlivých parametrů (pro $j = 0, 1, \ldots, 5$) pracují s nulovou a alternativní hypotézou

$$H_0: \beta_j = 0, \quad H_1: \beta_j \neq 0.$$

Kritický obor je podle tabulky 4.23 roven ($t_{1-0,05/2}(27) = t_{0,975}(27) = 2,052$)

$$W_{0,05} = \{t; |t| \geq t_{1-0,05/2}\} = (-\infty; -2,052\rangle \cup \langle 2,052; \infty).$$

V kritickém oboru leží hodnoty kritéria T pro absolutní člen a proměnnou hlučnost myčky. Hodnoty testového kritéria pro ostatní vysvětlující proměnné jsou v oboru přijetí $V_{0,05} = (-2,052; 2,052)$. Stejné rozhodnutí umožňuje také poslední sloupec tabulky, který obsahuje p-hodnoty pro individuální testy, porovnáme-li $\alpha = 0,05$ a p-hodnotu.

Již jsme uvedli, že znaménko regresního parametru u proměnné počet programů neodpovídá intuitivní představě, že by myčky s velkým výběrem programů měly být dražší. Vzhledem k tomu, že koeficient u této vysvětlující proměnné není podle individuálního testu statisticky významně různý od nuly, je tedy docela dobře možné, že v případě jiného souboru dat by odhadnutá hodnota parametru byla kladná.

Používat individuální testy o parametrech pro výběr proměnných zařazených do regresního modelu se nedoporučuje, neboť by mohla být omylem vynechána proměnná, která do modelu z hlediska faktické významnosti patří nebo by naopak v modelu mohla být ponechána taková, která pro popis závislosti v celé populaci nutná není. Vhodnější krokové metody ukážeme v dalším textu.

Pomocí softwaru byla nalezena tabulka analýzy rozptylu regresního modelu 4.26.

Tab. 4.26 Tabulka analýzy rozptylu pro příklad 4.11

Zdroj variability	Stupně volnosti	Součet čtverců	Průměrný čtverec	F	p-hodnota
Model	4	210 039 319	52 509 830	5,38	0,002 6
Reziduální	27	263 741 681	9 768 210		
Celkem	31	473 781 000			

V našem modelu je

$$H_0: \beta_0 = c, \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.$$

Je tedy $n = 32$, $k = 4$, $p = 5$, $n - p = 27$ a pro $\alpha = 0,05$ je kritický obor

$$W_{0,05} = \{F; F \geq F_{0,95}\} = \langle 2,728; \infty).$$

Hodnota testového kritéria 5,38 je prvkem kritického oboru. Pokud bychom ještě chtěli určit p-hodnotu, skutečně je

$$P(X \geq 5,38) = 1 - F(5,38) = 0,002\,6,$$

kde X má rozdělení Fisherovo-Snedecorovo se stupni volnosti 4 a 27 a F je distribuční funkce tohoto rozdělení.

Z tabulky analýzy rozptylu snadno získáme hodnotu indexu determinace a opraveného indexu determinace

$$I^2 = \frac{210\ 039\ 319}{473\ 781\ 000} = 0{,}443\,3 \ \ (44{,}33\,\%), \tag{4.77}$$

$$I^2_{ADJ} = 1 - (1 - 0{,}443\,3)\frac{32-1}{32-5} = 0{,}360\,9\,.$$

Model tedy vysvětlil 44 % variability proměnné cena. Zbylých 56 % variability vysvětlující proměnné vysvětleno nebylo a říkáme, že jsou způsobeny vlivy, které v našem modelu zahrnuty nemáme.

Reziduální rozptyl je roven

$$s_R^2 = \frac{S_R^2}{n-p} = \frac{263\ 741\ 681}{32-5} = 9\ 768\ 210,$$

a tedy pro reziduální směrodatnou odchylku dostaneme

$$s_R = \sqrt{\frac{S_R^2}{n-p}} = \sqrt{\frac{263\ 741\ 681}{32-2}} = \sqrt{9\ 768\ 210} = 3125 \text{ Kč}.$$

■

Polynomická regrese

Nyní se vrátíme k další zmíněné možnosti, jak sestrojit vhodný regresní model. Místo přímky použijeme polynom vyššího řádu, tedy polynomickou regresi probíranou v části 4.3.3.

Vycházíme opět z toho, že hodnoty vysvětlující proměnné x_i pro $i = 1, 2, \ldots, n$, jsou pevné a závislost mezi vysvětlovanou proměnnou y a vysvětlující proměnnou x v celé populaci je popsána polynomickou regresní funkcí

$$f(x; \beta_0, \beta_1, \ldots, \beta_k) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k.$$

Klasický lineární regresní model má pak tvar

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k + \varepsilon_i,$$

kde náhodné veličiny $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ splňují předpoklady (4.49), tedy $\varepsilon_i \sim N(0, \sigma^2)$ a ε_i jsou nezávislé náhodné veličiny pro $i = 1, 2, \ldots, n$.

Libovolnou spojitou funkci můžeme libovolně přesně přiblížit polynomem dostatečně vysokého stupně. Pokud bychom jako regresní funkci zvolili polynom stupně n, procházel by všemi pozorovanými body. V praxi ale používáme pouze polynomy nižších řádů, neboť jinak je problematická interpretace parametrů a dále uvidíme, že v případě zahrnutí mnoha mocnin jedné vysvětlující proměnné obyčejně dochází k výskytu multikolinearity, kterou se budeme zabývat v dalším textu.

V případě polynomů vyšších řádů než jedna používáme pro odhad parametrů regresní funkce maticové vyjádření. Využijeme zápis modelu podle v (4.36)–(4.38) ve tvaru

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde jsme opět do vektoru $\mathbf{y}$ umístili pozorované hodnoty $y_1, y_2, \ldots, y_n$ a do vektoru $\boldsymbol{\varepsilon}$ náhodné složky $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Potom můžeme vektor odhadů $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)'$ parametrů regresního modelu nalezený metodou nejmenších čtverců zapsat jako

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

a odhadnutá regresní funkce má tvar

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_k x^k.$$

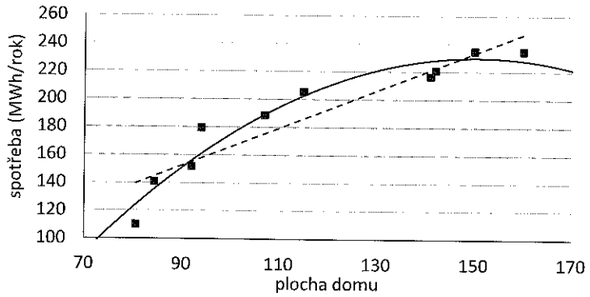
Model polynomické regrese je speciálním případem regresního modelu vícenásobné regrese. Pokud chceme proložit polynom stupně k, volíme v regresní funkci vícenásobné regrese (4.70) $x_1 = x$, $x_2 = x^2$, …, $x_k = x^k$. Proto můžeme použít všechny postupy, které jsme představili pro vícenásobnou regresi.

Příklad 4.12

Bylo pořízeno 10 pozorování plochy domů (v m^2) a roční spotřeby elektrické energie v MWh. Pozorované hodnoty jsou uvedeny v tabulce 4.12 a znázorněny na obrázku 4.20. Z rozložení bodů na obrázku je patrné, že regresní přímka nebude vhodným modelem pro závislost spotřeby elektrické energie a rozlohy domu (přímka je znázorněna čárkovaně) a je třeba zvolit nějakou polynomickou regresní funkci.

Tab. 4.27 Data pro příklad 4.12

i	spotřeba	plocha	plocha^2	$\hat{y}$	$\hat{\varepsilon}_i$
1	110,00	80,6	6 496,36	124,15	14,15
2	140,64	84,4	7 123,36	135,34	5,30
3	151,68	92,0	8 464,00	155,83	4,15
4	179,16	94,0	8 836,00	160,81	–18,35
5	188,52	107,0	11 449,00	188,94	0,42
6	205,32	115,0	13 225,00	202,62	–2,70
7	216,48	141,0	19 881,00	227,94	11,46
8	220,80	142,0	20 164,00	228,33	7,53
9	234,72	150,0	22 500,00	229,89	–4,83
10	234,48	160,0	25 600,00	227,95	–6,53



Obr. 4.20 Pozorované hodnoty a odhadnutá regresní parabola

Zvolili jsme regresní parabolu, odhadnuté parametry byly získány pomocí softwaru a jsou uvedeny v tabulce 4.28. Odhadnutá regresní funkce má tedy tvar

$$\hat{y} = -260,27 + 6,52x - 0,022x^2.$$

Tab. 4.28 Odhady parametrů modelu v příkladu 4.12

Proměnná	Odhad parametru	Směrodatná chyba	t	p-hodnota
posunutí	–260,27	91,76	–2,84	0,025 2
plocha	6,52	1,61	4,04	0,004 9
plocha^2	–0,022	0,007	–3,22	0,014 7

Z posledního sloupce tabulky obsahujícího p-hodnoty pro individuální testy o regresních parametrech plyne, že všechny parametry jsou statisticky významné. Stejný závěr učiníme pomocí kritického oboru. Podle tabulky 4.23 dostáváme

- pro parametr β_0: $H_0: \beta_0 = 0$, $H_1: \beta_0 \neq 0$: $t = \dfrac{-260,27}{91,76} = -2,84$,
- pro parametr β_1: $H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$: $t = \dfrac{6,52}{1,61} = 4,04$,
- pro parametr β_2: $H_0: \beta_2 = 0$, $H_1: \beta_2 \neq 0$: $t = \dfrac{-0,022}{0,007} = -3,22$.

Všechna testová kritéria mají za platnosti příslušných nulových hypotéz Studentovo rozdělení s $10 - 3 = 7$ stupni volnosti. Pro všechny tři testy má společný kritický obor tvar ($t_{0,975}(7) = 2,365$)

$W_{0,05} = \{t; |t| \geq t_{1-0,05/2}\} = (-\infty; -2,365\rangle \cup \langle 2,365; \infty)$

a všechny tři kritické hodnoty jsou prvkem tohoto intervalu. Proto zamítáme (na hladině významnosti 0,05) všechny tři nulové hypotézy.

V tabulce 4.29 je uveden postup testu o celém regresním modelu. Nulová hypotéza má tvar

$H_0: \beta_0 = c, \beta_1 = 0, \beta_2 = 0,$

alternativní hypotéza říká, že aspoň jeden parametrů β_1 a β_2 je nenulový. Model je statisticky významný, neboť p-hodnota je menší než desetina promile. Pokud bychom chtěli pro rozhodnutí použít kritický obor, podle tabulky 4.24 dostáváme (zvolíme $\alpha = 0,05$, $F_{0,95}(2,7) = 4,74$)

$W_{0,05} = \{F; F \geq F_{1-0,05}\} = \langle 4,74; \infty)$.

Testové kritérium 63,9 je prvkem kritického oboru, proto také zamítáme nulovou hypotézu, že závislost popsaná zvolenou regresní funkcí není statisticky významná.

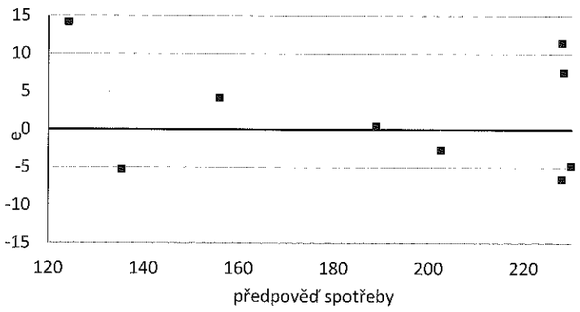
Tab. 4.29 Tabulka analýzy rozptylu pro regresní model z příkladu 4.12

Zdroj variability	Stupně volnosti	Součet čtverců	Průměrný čtverec	F	p-hodnota
Model	2	15 410	7 705,11	63,9	<0,0001
Reziduální	7	844,06	120,58		
Celkem	9	16 254			

Porovnáme-li indexy determinace pro model přímky (0,8714) a pro model paraboly (0,948 1), druhá hodnota je vždy větší, neboť model paraboly v sobě obsahuje také přímku jako speciální případ. Modely s různým počtem parametrů je třeba porovnávat upraveným indexem determinace, tedy po řadě podle (4.76) dostáváme ($n = 10$)

- přímka: $I^2_{ADJ} = 1 - (1 - 0,871 4)\dfrac{10-1}{10-2} = 0,855 3$,
- parabola: $I^2_{ADJ} = 1 - (1 - 0,948 1)\dfrac{10-1}{10-3} = 0,933 2$.

Ještě můžeme porovnat hodnoty reziduální směrodatné odchylky, který je 16,17 pro přímku a pro parabolu $\sqrt{120,58} = 10,98$. Podle všech těchto kritérií je parabola lepší než přímka. V tomto případě není třeba používat číselné charakteristiky, neboť z obrázku 4.20 je volba zřejmá. Ještě by bylo vhodné posoudit, zda náhodné chyby splňují předpoklady klasického lineárního modelu. Omezíme se na graf reziduí, na osu vodorovnou umístíme hodnotu předpovědi roční spotřeby $\hat{y}$ a na osu svislou hodnotu rezidua. Na obrázku nejsou zřejmé žádné výrazné odchylky od předpokladů klasického lineárního modelu na náhodnou složku.



Obr. 4.21 Rezidua regresního modelu z příkladu 4.12

Multikolinearita

Pokud zkoumáme závislost jedné vysvětlované proměnné na více vysvětlujících proměnných, nejkvalitnější odhady regresních parametrů bychom dostali tehdy, pokud by byla co nejsilnější závislost mezi jednotlivými vysvětlujícími proměnnými a proměnnou vysvětlovanou a co nejslabší závislost mezi vysvětlujícími proměnnými samými. Situace, kdy vysvětlující proměnné jsou mezi sebou příliš silně lineárně vázány, se nazývá **multikolinearita**. Tento jev způsobuje problémy při induktivních úsudcích v regresním modelu, například směrodatné chyby odhadů regresních parametrů jsou tak velké, že individuální testy o jednotlivých parametrech nejsou významné, a tedy se zdá, že žádný regresní koeficient není statisticky významně nenulový. Přitom ale celkový test o regresním modelu je statisticky vysoce významný. Pokud dojde k tomuto zdánlivě nelogickému výsledku, asi nejčastější příčinou je výskyt multikolinearity. V případě, že se v modelu multikolinearita vyskytuje, doporučuje se ji před použitím regresního modelu odstranit. V ekonomických aplikacích se často setkáváme se situací, že vysvětlující proměnné v regresních modelech jsou příliš silně vázány tak, jak ekonomické veličiny zpravidla bývají. Proto je třeba posouzení výskytu multikolinearity a jejímu následnému odstranění z modelu věnovat pozornost. Výskyt multikolinearity doprovází přirozeně také použití polynomické regrese, kdy vysvětlující proměnné jsou mocniny jediné vysvětlující proměnné x. To je jedním z důvodů, proč nepoužíváme polynomy s vysokou mocninou.

Při definici klasického lineárního modelu jsme předpokládali, že hodnoty vysvětlované proměnné jsou pevné. Pro posouzení multikolinearity jsme použili pojem lineární závislost, budeme s ní dále pracovat jen ve smyslu kapitoly 4.3.2 bez předpokladu o pravděpodobnostním rozdělení. Výjimkou budou statistické testy, kde předpoklad o rozdělení vysvětlujících proměnných klást musíme. Určitou informaci o závislostech nám poskytne **výběrová korelační matice R**, určená z hodnot všech k vysvětlujících proměnných $x_1, x_2, \ldots, x_k$ v případě vícenásobné regrese nebo všech použitých mocnin vysvětlující proměnné $x, x^2, \ldots, x^k$ v případě polynomické regrese. Výběrová korelační matice je čtvercová matice typu $k \times k$, která obsahuje výběrové korelační koeficienty (určené podle (4.31)) mezi vysvětlujícími proměnnými. Indexy v definici matice označují indexy u vysvětlujících proměnných v modelu a můžeme psát

$$\mathbf{R}=\begin{pmatrix} 1 & r_{12} & \ldots & r_{1k} \\ r_{21} & 1 & \ldots & r_{2k} \\ r_{31} & r_{32} & \ldots & r_{3k} \\ \ldots & \ldots & \ldots & \ldots \\ r_{k1} & r_{k2} & \ldots & 1 \end{pmatrix}. \tag{4.78}$$

Na místě v i-tém řádku a j-tém sloupci a také v j-tém řádku a i-tém sloupci je hodnota rovna výběrovému korelačnímu koeficientu mezi proměnnými x_i a x_j. Ukázali jsme, že hodnota korelačního koeficientu nezávisí na pořadí proměnných, proto je matice **R** symetrická. Vzhledem k tomu, že $r_{jj} = 1, j = 1, 2, \ldots, k$, má matice na hlavní diagonále vždy hodnoty jedna.

Důležitým indikátorem multikolinearity je hodnota determinantu korelační matice. Pokud jsou všechny dvojice vysvětlujících proměnných vzájemně nekorelované, všechny výběrové korelační koeficienty jsou přibližně rovny nule a determinant **R** je blízký hodnotě jedna (na hlavní diagonále jsou jedničky). Pokud jsou vysvětlující proměnné lineárně závislé, i korelační koeficienty jsou různé od nuly a lze ukázat, že determinant výběrové korelační matice je menší než jedna a s narůstající intenzitou závislostí se přibližuje nule. Rovná-li se nule, hovoříme někdy o **úplné multikolinearitě**. Znamená to, že aspoň jeden z korelačních koeficientů je roven v absolutní hodnotě jedné, a tedy některá vysvětlující proměnná je lineární funkcí jiné.

Při posouzení výskytu multikolinearity místo nějakého statistického testu poslouží i nalezení dvojic vysvětlujících proměnných, které jsou příliš silně lineárně závislé. Nejsnazší doporučení je uvažovat multikolinearitu za problém, pokud aspoň jeden párový korelační koeficient v matici **R** je v absolutní hodnotě větší než předem dané číslo, uvádí se 0,75 nebo případně 0,8, ještě benevolentnější hodnota 0,9. V takovém případě se pokusíme například vynechat jednu z příliš závislých vysvětlujících proměnných s předpokladem, že v lineární regresní funkci vliv vynechané proměnné zastoupí vysvětlovaná proměnná, kterou jsme v modelu ponechali.

V tabulce 4.30 je uvedena výběrová korelační matice pro vysvětlující proměnné, uvažované v příkladu 4.11. Všechny korelační koeficienty jsou v absolutní hodnotě menší než 0,7, korelace mezi vysvětlujícími proměnnými tedy nejsou příliš vysoké a z tohoto hlediska jsou tedy všechny vysvětlující proměnné do modelu zařazeny oprávněně.

Tab. 4.30 Výběrová korelační matice pro příklad 4.11

Proměnné	voda	hlučnost	elektřina	programy
voda	1	0,36737	0,39714	–0,23129
hlučnost	0,36737	1	0,60745	–0,68649
elektřina	0,39714	0,60745	1	–0,37983
programy	–0,23129	–0,68649	–0,37983	1

Volba regresní funkce

Hledání vhodné regresní funkce ve vícenásobné regresi je spojeno s volbou vysvětlujících proměnných. Pokud by měl model být skutečně validní a správně zvolený, musel by obsahovat všechny potřebné veličiny, které ovlivňují vysvětlovanou proměnnou, a neobsahovat žádné nadbytečné. V takovém případě by byl splněn základní předpoklad použití regresní analýzy. Takový cíl si ale v praxi můžeme těžko klást, obyčejně se spokojíme s modelem, který vyhovuje našim teoretickým i empirickým úvahám. Před formulováním regresního modelu je třeba nejprve posoudit všechny dostupné informace o charakteru závislosti mezi vysvětlovanou proměnnou a vysvětlujícími proměnnými v základním souboru. Takové úvahy obyčejně poskytují základní představu o vysvětlujících proměnných, které by do modelu měly být zařazeny. Další proměnné, které musíme do regresního modelu zařadit, jsou spojeny s konkrétními výzkumnými hypotézami a také s interpretací výsledného modelu, kdy budeme chtít odstranit vliv některých vlivů na vysvětlující proměnné, které jsou základním cílem modelování. Často se jedná o pohlaví, věk, vzdělání nebo další demografické charakteristiky osob, pokud zkoumáme závislost velikosti mzdy a délky praxe. V případě, že budeme sledovat ekonomické výsledky firem, budou takovými proměnnými například základní charakteristiky podnikání. Pokud jsou však vysvětlující proměnné kvalitativní, je nutné postupovat tak, jak je naznačeno v části 4.3.7 této knihy.

Pokud chceme využít empirické výsledky a shromážděná data, existují grafické postupy, které pomáhají s volbou vhodné funkce (Hebák a další, 2013). Pokud vybereme nějakou funkci, můžeme odhadnout parametry, a tím získat odhadnutou regresní funkci. V takovém případě můžeme porovnávat pozorované hodnoty y_i a odhadnuté hodnoty $\hat{y}_i$.

Před tím, než si vhodný model vybereme, je vhodné pomocí různých diagnostických postupů posoudit splnění předpokladů, kladených na náhodnou složku modelu. V případě neuspokojivých výsledků regresní diagnostiky (grafických metod či formálních statistických testů) je třeba hledat mezi dalšími typy regresních funkcí, zahrnout další vysvětlující proměnné nebo zeslabit předpoklady, kladené na náhodnou složku v regresním modelu (Hebák a další, 2013). V tomto textu probíraný klasický lineární regresní model je jen jednou z možností, jak regresní model formulovat.

Pokud chceme porovnat již vybrané regresní funkce, je možné použít upravený index determinace a vybrat model, pro který je hodnota indexu nejvyšší. Dalším kritériem může být malé kolísání kolem regresní funkce, proto si vybereme regresní model, který má nejmenší reziduální rozptyl. V příkladu 4.12 největší hodnota upraveného indexu determinace i nejmenší hodnota reziduálního rozptylu byly dosaženy pro jednu zkoumanou funkci, taková situace nemusí ale vždy nastat.

Předpokládejme nyní, že jsme již zvolili množinu vysvětlujících proměnných, o kterých se domníváme, že by měly být v našem modelu zahrnuty. Existují postupy, které nám pomohou vybrat proměnné, které by (na základě dostupných dat) měl regresní model skutečně obsahovat, protože obsahují informaci o vysvětlované proměnné. Problémem takových metod je, že skutečně pracují jen s analyzovanými daty, nikoliv přímo se závislostí v základním souboru. Vybrané vysvětlující proměnné tedy nemusejí poskytovat optimální množinu pro popis závislosti v celé populaci, ale pouze v analyzovaných datech.

Můžeme posoudit korelační koeficienty mezi vysvětlovanou proměnnou a vysvětlujícími proměnnými a určité základní vodítko nám poskytnou také individuální testy o jednotlivých parametrech. Další možností je použití **krokových metod**, které umožňují automatizované vybírání proměnných pro regresní model. Nejběžněji implementovanými metodami v softwaru jsou metoda **forward** (také **dopředný výběr**), metoda **backward** (také **zpětný výběr** proměnných) a jejich kombinace, nazývaná metoda **stepwise** (též **krokový výběr** proměnných). V metodě forward začínáme s modelem obsahujícím jen absolutní člen (konstantu) a přidáváme vysvětlující proměnné tak dlouho, až nezařazené proměnné již nepřinášejí žádnou další informaci pro vysvětlení proměnné y. Naopak při metodě backward hledání začíná s modelem obsahujícím všechny uvažované vysvětlující proměnné a postupně jsou vynechávány proměnné méně důležité, až všechny zařazené proměnné významně přispívají do regresního modelu, popisujícího závislost mezi vysvětlujícími proměnnými a vysvětlovanou proměnnou. Metoda stepwise začíná z prázdného modelu (obsahujícího jen konstantu) a v každém dalším kroku vždy hledá proměnnou, kterou by bylo vhodné přidat a současně pak zkoumá, zda nějaká proměnná již zařazená do modelu není již nadbytečná. Posouzení, kdy ještě model měnit a kdy ne, je založeno na různých kritériích, volbou parametrů procedury může postup řídit i uživatel. Je třeba podotknout, že použitím různých postupů můžeme dostat různý výsledek, jak ukazuje tabulka 4.31. Navíc se uvádí, že výsledný model nemusí obsahovat proměnné, které jsou pro závislost v celé populaci důležité a naopak mohou v modelu nechat proměnné, které do populačního modelu nepatří. Poznamenejme ještě, že krokové metody neodstraňují z modelu multikolinearitu. Vzhledem k dostupnosti výpočetní techniky je možné při výběru regresního modelu kombinovat více postupů a pracovat s větším počtem různých modelů.

Výsledky hledání regresního modelu krokovými metodami pro příklad 4.11 jsou uvedeny v tabulce 4.31. Všimněme si, že výše zmíněné metody poskytly různé výsledné modely. Regresní model získaný metodou backward má příliš malou hodnotu

indexu determinace; pokud bychom si měli vybrat mezi výsledky zbylých dvou metod, můžeme porovnat hodnoty upraveného indexu determinace a reziduálního rozptylu. V tabulce 4.32 jsou uvedeny odhadnuté parametry modelu s menším počtem proměnných. V modelu byla ponechána proměnná spotřeba vody, pro kterou dostáváme p-hodnotu větší než 0,05.

Tab. 4.31 Výsledky krokových metod výběru proměnných

Metoda	I^2	I^2_{ADJ}	s_R	Proměnné
Forward	44,06 %	0,381	3 076	voda, hlučnost, programy
Backward	21,08 %	0,156	2 590	elektřina, programy
Stepwise	40,45 %	0,363	3 119	voda, hlučnost

Tab. 4.32 Zmenšený regresní model

Proměnná	Odhad parametru	Směrodatná chyba	t	p-hodnota
Konstanta	52 399	9 557,413	5,48	< 0,000 1
Voda	–2,668	1,486	–1,80	0,083 1
Hlučnost	–712,764	228,788	–3,12	0,004 1

Pro tento model je $I^2 = 0{,}405$ a $I^2_{ADJ} = 0{,}363$. Index determinace je menší než index pro původní model (44 %); po odebrání dvou vysvětlujících proměnných jsou upravené indexy determinace srovnatelné.

4.3.7 Kvalitativní proměnné v lineární regresi

Do této chvíle jsme jako vysvětlující proměnné používali pouze kvantitativní proměnné. Pokud používáme regresní modely v praktických situacích, potřebujeme v regresních modelech použít také kategoriální proměnné jako je nejvyšší dosažené vzdělání nebo pohlaví, ale také třeba typ zájezdu (pobytový, poznávací) nebo obor činnosti firmy. V analýze časových řad je důležité do modelů zahrnout sezónní kolísání časové řady, sezónu (například jednotlivá čtvrtletí roku nebo dny v týdnu) budeme v páté kapitole této knihy považovat také za kvalitativní proměnnou. Pro zahrnutí kategoriálních proměnných do regresního modelu existují v podstatě dvě možnosti, mezi nimi si vybíráme podle toho, jak chceme výsledky interpretovat.

První možností je použít takzvané **dummy proměnné**. Jsou to pomocné vysvětlující proměnné v regresním modelu, které nabývají pouze hodnot 0 a 1 a slouží jako indikátory obměn kategoriální proměnné. Pokud má taková proměnná k různých kategorií, pro zahrnutí této proměnné do regresního modelu potřebujeme $k - 1$ dummy proměnných. Nejprve je třeba zvolit referenční kategorii, ke které budeme ostatní kategorie vztahovat. Dummy proměnné pak budou indikátory zbylých $k - 1$ kategorií,

referenční hodnota bude mít všechny hodnoty dummy proměnných rovné 0, a tím bude i ona jednoznačně určena.

Předpokládejme například, že mzda zaměstnance (vysvětlovaná proměnná y v Kč) bude lineárně záviset na době praxe (vysvětlující proměnná x_1 v letech); můžeme uvažovat lineární regresní model

$$y = \beta_0 + \beta_1 x_1 + \varepsilon. \tag{4.79}$$

Budeme-li chtít do regresního modelu zahrnout navíc kategoriální proměnnou nejvyšší dosažené vzdělání zaměstnance, která bude nabývat čtyř hodnot, a to Z (základní vzdělání), S (střední vzdělání), ÚS (střední vzdělání s maturitou) a VŠ (terciární vzdělání), je třeba použít tři dummy proměnné. Pokud zvolíme jako referenční kategorii úplné střední vzdělání s maturitou, kódování kategorií pomocí tří dummy proměnných základní (x_2), střední (x_3) a terciární (x_4), je ukázáno v tabulce 4.33.

Tab. 4.33 Kódování kvalitativních proměnných, dummy proměnné

Vzdělání	x_2	x_3	x_4
Z	1	0	0
S	0	1	0
ÚS	0	0	0
VŠ	0	0	1

Regresní model (4.79) potom přejde na model vícenásobné regrese

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon'. \tag{4.80}$$

V jedné rovnici jsou popsány regresní funkce pro všechny kategorie vzdělání:

- základní vzdělání $\beta_0 + \beta_2 + \beta_1 x_1$,
- střední vzdělání $\beta_0 + \beta_3 + \beta_1 x_1$,
- úplné střední vzdělání $\beta_0 + \beta_1 x_1$,
- vysokoškolské vzdělání $\beta_0 + \beta_4 + \beta_1 x_1$.

Všimněme si, že směrnice lineární závislosti na proměnné x_1 jsou stejné, liší se jen posunutí regresních přímek pro různé kategorie vzdělání. Při volbě tohoto modelu tedy předpokládáme, že všechny populační regresní přímky jsou rovnoběžné.

Tab. 4.34 Kódování kvalitativních proměnných, indikátorové proměnné

Vzdělání	x_2	x_3	x_4
Z	1	0	0
S	0	1	0
ÚS	–1	–1	–1
VŠ	0	0	1

Druhou možností je vztahovat úsudky nikoliv k vybrané referenční skupině, ale k průměru všech pozorování. V takovém případě použijeme opět kódování pomocí proměnných obsahujících 0 a 1, tyto nové proměnné nazveme (v souladu s Pecáková, 2011) indikátorové. Toto kódování je ukázáno v tabulce 4.34.

V tomto případě model vícenásobné regrese (4.80) obsahuje regresní přímky pro každou skupinu, definovanou nejvyšším dosaženým vzděláním:

- základní vzdělání $\beta_0 + \beta_2 + \beta_1 x_1$,
- střední vzdělání $\beta_0 + \beta_3 + \beta_1 x_1$,
- úplné střední vzdělání $\beta_0 - \beta_2 - \beta_3 - \beta_4 + \beta_1 x_1$,
- vysokoškolské vzdělání $\beta_0 + \beta_4 + \beta_1 x_1$.

Přímky jsou opět rovnoběžné, liší se pouze v odhadech posunutí.

Pokud bychom chtěli použít model, ve kterém by rychlost růstu mezd v závislosti na délce praxe byla odlišná pro různé stupně vzdělání, bylo by nutné, aby také směrnice regresních přímek byly různé. V takovém případě je nutné do regresního modelu vložit **interakci** mezi vzděláním a praxí. Interakce obecně umožňuje odlišit závislost pro různé kombinace proměnných. Všeobecně se ale jedná o velmi užitečný postup při konstrukci složitějších regresních modelů. V našem případě bychom například mohli očekávat rychlejší růst mezd pro zaměstnance s vysokoškolským vzděláním než se základním vzděláním. V takovém případě by model (4.80) měl tvar

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon'', \qquad (4.81)$$

kde proměnné x_1, x_2, x_3 a x_4 jsou proměnné z původního modelu (4.80), proměnná x_5 obsahuje interakci mezi základním vzděláním a délkou praxe (určíme ji jako součin proměnných x_1 a x_2, tj. x_1x_2), proměnná x_6 obsahuje interakci mezi středním vzděláním a délkou praxe (určíme ji jako součin proměnných x_1 a x_3, tj. x_1x_3) proměnná x_7 obsahuje interakci mezi vysokoškolským vzděláním a délkou praxe (určíme ji jako součin proměnných x_1 a x_4, tj. x_1x_4) a ε'' jsme označili náhodnou složku v regresním modelu. Model (4.81) pak je možné přepsat pro různé stupně vzdělání ve tvaru:

- základní vzdělání $\beta_0 + \beta_2 + (\beta_1 + \beta_5)x_1$,
- střední vzdělání $\beta_0 + \beta_3 + (\beta_1 + \beta_6)x_1$,
- úplné střední vzdělání $\beta_0 + \beta_1 x_1$,
- vysokoškolské vzdělání $\beta_0 + \beta_4 + (\beta_1 + \beta_7)x_1$.

Regresní přímky, popisující vztah mezi mzdou a dobou praxe, mají v tomto případě posunutí i směrnici závislou na skupině definované vzděláním.

4.3.8 Korelační analýza

Již během výkladu v předchozím textu jsme se opakovaně dotkli základů korelační analýzy. Na rozdíl od regresní analýzy se korelace zabývají spíše intenzitou než popisem závislosti a zkoumáním kauzality či vztahu příčiny a důsledku. Závislost uvažovaná v korelační analýze je vždy lineární. V případě párových korelačních koeficientů je oboustranná (vzájemná), proměnné mají ve vztahu stejné postavení a již nerozlišujeme vysvětlovanou a vysvětlující proměnnou.

Párový korelační koeficient

Nejpoužívanějším nástrojem pro popis lineární závislosti dvou kvantitativních proměnných (řekněme x a y) je korelační koeficient, jehož výběrová podoba r_{xy} byla zavedena v (4.31). Již víme, že (párový) korelační koeficient je charakteristikou intenzity i směru lineární závislosti mezi dvěma proměnnými. Proto je zřejmé, že můžeme hledat podobnost a společné vlastnosti a postupy s lineární (přímkovou) regresí.

Pokud je korelační koeficient kladný, je mezi proměnnými přímá úměrnost v tom smyslu, že rostou-li hodnoty jedné proměnné, rostou (v průměru) také hodnoty druhé proměnné. V případě záporného korelačního koeficientu je závislost nepřímá, tedy s růstem jedné veličiny hodnoty druhé (v průměru) klesají. Čím více se tedy blíží koeficient korelace v absolutní hodnotě jedné, tím považujeme danou lineární závislost za silnější, čím více se blíží nule, tím ji považujeme za volnější (slabší).

Zmiňme dále, že existuje souvislost mezi výběrovým korelačním koeficientem r_{xy} a odhadem směrnice regresní přímky $\hat{\beta}_1$ (podle (4.20)). Je zřejmé, že musejí mít stejné znaménko, existuje však postup, jak nalézt z regresního koeficientu korelační koeficient, případně obráceně. Pokud známe výběrové směrodatné odchylky s_x a s_y, lze ukázat, že platí

$$r_{xy} = \hat{\beta}_1 \frac{s_x}{s_y} \qquad (4.82)$$

a také

$$r_{xy} = \hat{\alpha}_1 \frac{s_y}{s_x}, \qquad (4.83)$$

kde $\hat{\beta}_1$ je odhad směrnice regresní přímky s vysvětlovanou proměnnou y a vysvětlující proměnnou x a $\hat{\alpha}_1$ je odhad směrnice sdružené přímky s vysvětlovanou proměnnou x a vysvětlující proměnnou y.

Podobně jako v případě regresní analýzy a práce s populační regresní funkcí předpokládejme, že existuje populační korelační koeficient ρ_{XY}, který popisuje lineární závislost mezi dvěma náhodnými proměnnými X a Y v celé zkoumané populaci a (X_i, Y_i), $i = 1, 2, \ldots, n$, je náhodný výběr z dvourozměrného pravděpodobnostního rozdělení obou náhodných veličin (Marek, 2012) s korelačním koeficientem ρ_{XY}. Pokud navíc rozdělení X a Y je dvourozměrné normální, výše uvedené vztahy (4.82) a (4.83) platí také mezi populačními charakteristikami ρ_{XY}, $\sqrt{D(X)}$ a $\sqrt{D(Y)}$.

Výběrový korelační koeficient je výběrovým protějškem populačního (teoretického koeficientu) ρ_{XY} a v případě náhodného výběru z dvourozměrného normálního

rozdělení je výběrový korelační koeficient r_{xy} bodovým odhadem populačního korelačního koeficientu ρ_{XY}. Tento odhad není nezkreslený, ale je konzistentní a asymptoticky nezkreslený (nezkreslený pro výběry o velkém rozsahu).

Na základě předchozích úvah můžeme sestrojit interval spolehlivosti pro populační korelační koeficient nebo testovat hypotézu o jeho hodnotě. Z možných hypotéz nás nejčastěji zajímá test hypotézy, že veličiny jsou lineárně nezávislé, tedy nulovou hypotézou je

$$H_0: \rho_{XY} = 0$$

proti alternativě, že veličiny jsou lineárně závislé, přímo úměrné nebo nepřímo úměrné, tedy

$$H_1: \rho_{XY} \neq 0, \quad H_1: \rho_{XY} > 0 \quad \text{nebo } H_1: \rho_{XY} < 0.$$

Průběh testu je shrnut v tabulce 4.35.

Tab. 4.35 Test o nulovosti párového korelačního koeficientu

H_0	H_1	Testové kritérium	Kritický obor
$\rho_{XY} = 0$	$\rho_{XY} \neq 0$	$T = \dfrac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ $T \sim t(n-2)$	$W_\alpha = \{t; \lvert t\rvert \geq t_{1-\alpha/2}\}$
	$\rho_{XY} > 0$		$W_\alpha = \{t; t \geq t_{1-\alpha}\}$
	$\rho_{XY} < 0$		$W_\alpha = \{t; t \leq -t_{1-\alpha}\}$

Příklad 4.13

V náhodném výběru o rozsahu 25 dvojic pozorování byl vypočten výběrový koeficient korelace $r = 0{,}23$. Statistickým testem posoudíme, zda z tohoto výsledku lze usuzovat na lineární nezávislost mezi proměnnými v základním souboru. Úvahu zopakujeme pro rozsah výběru 83 a stejnou hodnotu výběrového korelačního koeficientu.

Řešení

Nejprve uvažujme náhodný výběr o rozsahu 25 pozorování. V takovém případě určíme hodnotu testového kritéria podle tabulky 4.35 jako

$$t = \sqrt{25-2}\,\frac{0{,}23}{\sqrt{1-0{,}23^2}} = 0{,}480 \cdot 0{,}236 = 1{,}133.$$

Pokud bychom stejnou hodnotu korelačního koeficientu (0,23) dostali výpočtem z 83 pozorování, testové kritérium by mělo hodnotu

$$t = \sqrt{83-2}\,\frac{0{,}23}{\sqrt{1-0{,}23^2}} = 9 \cdot 0{,}236 = 2{,}127.$$

Zvolíme-li $\alpha = 0{,}05$, v prvním případě nulovou hypotézu na této hladině významnosti nezamítáme, ve druhém případě ji zamítneme, neboť pro 25 pozorování je (použijeme hodnotu kvantilu $t_{0{,}975}(23) = 2{,}069$)

$$W_{0{,}05} = \{t; |t| \geq t_{1-0{,}05/2}\} = (-\infty; -2{,}069\rangle \cup \langle 2{,}069; \infty)$$

a platí $1{,}133 \notin W_{0{,}05}$. Pro 81 stupňů volnosti již použijeme kvantil normovaného normálního rozdělení a dostaneme kritický obor

$$W_{0{,}05} = \{t; |t| \geq u_{1-0{,}05/2}\} = (-\infty; -1{,}96\rangle \cup \langle 1{,}96; \infty).$$

V tomto případě již $2{,}127 \in W_{0{,}05}$. Rozhodnutí, zda výběrový korelační koeficient je či není již statisticky významně odlišný od nuly, tedy závisí nejen na jeho velikosti, ale také na rozsahu náhodného výběru.

■

Nyní sestrojíme interval spolehlivosti pro populační korelační koeficient. Pokud je korelační koeficient v absolutní hodnotě nízký a rozsah výběru je velký, můžeme použít interval, založený na asymptotické normalitě výběrového korelačního koeficientu. Asymptotická normalita znamená, že výběrový korelační koeficient má pro velké výběry (asymptoticky) přibližně normální rozdělení. Přibližný $100(1-\alpha)\%$ interval spolehlivosti tedy můžeme zkonstruovat jako symetrický interval kolem bodového odhadu r_{xy}

$$r_{xy} \pm u_{1-\alpha/2}\frac{1-r_{xy}^2}{\sqrt{n}} = \left(r_{xy} - u_{1-\alpha/2}\frac{1-r_{xy}^2}{\sqrt{n}},\quad r_{xy} + u_{1-\alpha/2}\frac{1-r_{xy}^2}{\sqrt{n}}\right), \tag{4.84}$$

kde $u_{1-\alpha/2}$ je kvantil normovaného normálního rozdělení, odpovídající zvolené spolehlivosti $1-\alpha$.

Pokud je populační korelační koeficient odlišný od nuly, používá se pro konstrukci intervalu spolehlivosti Fisherova transformace

$$Z = \frac{1}{2}\ln\frac{1+r_{xy}}{1-r_{xy}}. \tag{4.85}$$

Statistika Z má pro velký rozsah výběru přibližně normální rozdělení se střední hodnotou $\frac{1}{2}\ln\frac{1+\rho_{XY}}{1-\rho_{XY}}$ a rozptylem $\frac{1}{n-3}$. Tuto statistiku můžeme použít pro test hypotézy o hodnotě korelačního koeficientu a pro konstrukci $100(1-\alpha)\%$ asymptotického intervalu spolehlivost pro tento koeficient. Ze vztahu (4.85) můžeme vyjádřit korelační koeficient jako funkci z ve tvaru

$$r_{xy} = \frac{e^{2z}-1}{e^{2z}+1}.$$

Pokud tedy najdeme krajní body z_D a z_H intervalu

$$(z_D, z_H) = \left(Z - u_{1-\alpha/2}\frac{1}{\sqrt{n-3}}, Z + u_{1-\alpha/2}\frac{1}{\sqrt{n-3}}\right), \tag{4.86}$$

dostaneme $100(1-\alpha)\%$ interval spolehlivosti pro korelační koeficient jako

$$\left(\frac{e^{2z_D}-1}{e^{2z_D}+1}, \frac{e^{2z_H}-1}{e^{2z_H}+1}\right). \tag{4.87}$$

Příklad 4.14

Předpokládejme, že z náhodného výběru o rozsahu 100 pozorování byl určen výběrový korelační koeficient $r_{xy} = 0{,}4$. Zkonstruujeme 95% interval spolehlivosti pro populační koeficient korelace ρ_{XY} pomocí Fisherovy transformace. Dosazením do vzorců (4.85), (4.86) a (4.87) postupně dostaneme, použijeme-li $n = 100$, $\alpha = 0{,}05$, $u_{0,975} = 1{,}96$, $r_{xy} = 0{,}4$

$$z = \frac{1}{2}\ln\frac{1+0{,}4}{1-0{,}4} = 0{,}424,$$

$$(z_D, z_H) = \left(0{,}424 - 1{,}96\frac{1}{\sqrt{100-3}};\ \ 0{,}424 + 1{,}96\frac{1}{\sqrt{100-3}}\right) = (0{,}225;\ \ 0{,}623),$$

$$\left(\frac{e^{2\cdot 0{,}225}-1}{e^{2\cdot 0{,}225}+1}, \frac{e^{2\cdot 0{,}623}-1}{e^{2\cdot 0{,}623}+1}\right) = (0{,}221;\ \ 0{,}553).$$

Všimněme si, že sestrojený interval spolehlivosti obsahuje hodnotu bodového odhadu $r_{xy} = 0{,}4$, bodový odhad zde ale není ve středu intervalu.

Párový korelační koeficient je vhodnou charakteristikou pro popis lineární závislosti proměnných, které jsou spojité a jejichž rozdělení není příliš vzdáleno od normálního rozdělení. Pokud to neplatí, je vhodné využít Spearmanův koeficient, definovaný v části 4.3.2. Statistické úsudky o tomto koeficientu je možné najít například v práci Hebák a další (2013).

■

Dílčí korelační koeficient

V ekonomických aplikacích se často setkáváme se situací, kdy veličiny vstupující do analýzy jsou silně lineárně závislé. Pokud tedy předpokládáme, že vysoká hodnota párového korelačního koeficientu mezi dvěma proměnnými x a y je způsobena nějakou další (skrytou) veličinou z, definovali jsme v (4.32) **dílčí korelační koeficient** $r_{xy.z}$, který umožňuje vyloučit vliv proměnné z (nebo obecně více proměnných $z_1, z_2, \ldots, z_k$) na lineární závislost mezi x a y. Pokud budeme předpokládat, že (X_i, Y_i, Z_i), $i = 1, 2, \ldots, n$, je náhodný výběr z trojrozměrného normálního rozdělení náhodných veličin (X, Y, Z), můžeme testovat hypotézu o nulové hodnotě populačního koeficientu dílčí korelace $\rho_{XY.Z}$. Nejčastěji nás zajímá test hypotézy, že veličiny X a Y jsou po eliminaci vlivu vysvětlující proměnné Z lineárně nezávislé, tedy

$$H_0: \rho_{XY.Z} = 0$$

proti alternativě, že veličiny jsou lineárně závislé, tedy

$$H_1: \rho_{XY.Z} \neq 0.$$

Postupujeme stejným způsobem jako u párového korelačního koeficientu (viz tabulka 4.35) s tím rozdílem, že se mění koeficient v čitateli testového kritéria T

$$T = \sqrt{n-3}\,\frac{r_{xy.z}}{\sqrt{1-r_{xy.z}^2}}.$$

Testové kritérium má Studentovo rozdělení s $n - 3$ stupni volnosti.

Koeficient dílčí korelace lze využít v regresní analýze, pokud se zajímáme o to, jaký přínos má zařazení nové proměnné do modelu. Máme-li tedy v modelu již proměnnou x_1 (uvažujeme tedy lineární závislost y na x_1), potom dílčí korelační koeficient $r_{yx_2.x_1}$ popisuje závislost, kterou nová proměnná x_2 přináší, pokud odstraníme vliv již zahrnuté proměnné x_1. Tento postup je možné zobecnit na situaci, kdy v regresním modelu je již k proměnných $x_1, x_2, \ldots, x_k$ a zvažujeme zahrnutí další proměnné x_{k+1}. Potom použijeme dílčí korelační koeficient $r_{yx_{k+1}.x_1,x_2,\ldots,x_k}$, který kvantifikuje sílu a směr závislosti mezi y a x_{k+1}, pokud vyloučíme vliv všech k proměnných za tečkou.

Koeficient vícenásobné korelace

Posledním korelačním koeficientem, kterým se zde budeme zabývat, je **vícenásobný korelační koeficient**. Uvažujme tedy situaci, kterou jsme zkoumali ve vícenásobné regresi, jednu vysvětlovanou proměnnou y a dvě vysvětlující proměnné x_1 a x_2. Budeme definovat korelační koeficient, který umožní posoudit intenzitu lineární závislosti mezi náhodnou veličinou Y a oběma náhodnými veličinami X_1 a X_2 dohromady. Tento koeficient nazýváme vícenásobný korelační koeficient. Jeho populační hodnotu, kterou označíme $\rho_{Y.X_1X_2}$, určíme z párových korelačních koeficientů ρ_{YX_1}, ρ_{YX_2} a $\rho_{X_1X_2}$ podle vzorce

$$\rho_{Y.X_1X_2} = \sqrt{\frac{\rho_{YX_1}^2 + \rho_{YX_2}^2 - 2\rho_{YX_1}\rho_{YX_2}\rho_{X_1X_2}^2}{1-\rho_{X_1X_2}^2}}.$$

Výběrovou hodnotu vícenásobného korelačního koeficientu $r_{y.x_1x_2}$ určíme obdobně nahrazením populačních korelačních koeficientů výběrovými korelačními koeficienty

$$r_{y.x_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}^2}{1-r_{x_1x_2}^2}}.$$

Vzhledem k tomu, že vícenásobný korelační koeficient je dán jako druhá odmocnina, může nabývat pouze nezáporných hodnot a lze ukázat, že maximální hodnotou je jed-

na. Nabývá tedy hodnot z intervalu $\langle 0;1\rangle$, proto popisuje jen intenzitu lineární závislosti, nikoli její směr. Koeficient posuzuje spojený vliv obou proměnných, proto je větší než oba párové korelační koeficienty r_{yx_1} a r_{yx_2}, maximálně jim je roven.

Chceme-li zkoumat lineární závislost jedné proměnné y na k vysvětlujících proměnných, použijeme koeficient vícenásobné korelace $\rho_{Y.X_1X_2\ldots X_k}$ popisující těsnost lineární závislosti vysvětlované proměnné Y a všech vysvětlujících proměnných $X_1, X_2, \ldots, X_k$ dohromady. Výběrovou hodnotu tohoto koeficientu určíme jako

$$r_{y.x_1x_2\ldots x_k} = \sqrt{1 - \frac{|R_{yx_1\ldots x_k}|}{|\mathbf{R}_{x_1\ldots x_k}|}}, \tag{4.88}$$

kde $\mathbf{R}_{yx_1\ldots x_k}$ je výběrová korelační matice typu $(k+1)\times(k+1)$ proměnných y a $x_1, x_2, \ldots, x_k$, $\mathbf{R}_{x_1\ldots x_k}$ je výběrová korelační matice typu $k\times k$ proměnných $x_1, x_2, \ldots, x_k$ a funkce $|.|$ značí determinant matice. Koeficient má obdobný význam i interpretaci jako párový korelační koeficient, jen z definice vyplývá, že nabývá pouze kladných hodnot, a tedy se vztahuje k intenzitě závislosti, nikoliv k jejímu směru. Ale již v oddílu věnovaném vícenásobné regresi jsme se zmínili, že je-li v regresním modelu více vysvětlujících proměnných, nelze sledovat celkový směr, jen směr závislosti v jednotlivých vysvětlujících proměnných.

Vícenásobný korelační koeficient popisuje těsnost lineárního vztahu mezi proměnnými a umožňuje posoudit kvalitu regresního modelu zkonstruovaného na základě vícenásobné regresní funkce. Lze ho také použít při hodnocení volby vysvětlujících proměnných. V případě, že jeho hodnota je malá, potom vybrané vysvětlující proměnné nepostačují k vysvětlení změn analyzované závisle proměnné. Z toho vyplývá souvislost s vícenásobnou regresí. Ve skutečnosti jde o korelační koeficient mezi vysvětlovanou proměnnou a nejlepší lineární kombinací vysvětlujících proměnných, která je výsledkem odhadu metodou nejmenších čtverců ve vícenásobné regresi. Potom platí

$$r_{y.x_1x_2\ldots x_k} = \sqrt{I^2},$$

kde I^2 je index determinace modelu vícenásobné regrese s vysvětlovanou proměnnou y a vysvětlujícími proměnnými $x_1, x_2, \ldots, x_k$.

Předpokládejme, že máme k dispozici náhodný výběr z $(k+1)$-rozměrného normálního rozdělení vektoru $(Y, X_1, X_2, \ldots, X_k)$. Test hypotézy o nulové hodnotě koeficientu vícenásobné korelace $\rho_{Y.X_1X_2\ldots X_k}$

$$H_0: \rho_{Y.X_1X_2\ldots X_k} = 0$$

proti alternativní hypotéze

$$H_1: \rho_{Y.X_1X_2\ldots X_k} \neq 0$$

je shrnut v tabulce 4.36.

Tab. 4.36 Test o nulové hodnotě koeficientu vícenásobné korelace

H_0	H_1	Testové kritérium	Kritický obor
$\rho_{Y.X_1X_2\ldots X_k} = 0$	$\rho_{Y.X_1X_2\ldots X_k} \neq 0$	$F = \frac{n-k-1}{k}\,\frac{r^2_{y.x_1x_2\ldots x_k}}{1-r^2_{y.x_1x_2\ldots x_k}}$ $F \approx F(k, n-k-1)$	$W_\alpha = \{F; F \geq F_{1-\alpha/2}\}$

Příklad 4.15

Uvažujme data použitá v příkladu 4.11. Najdeme koeficient vícenásobné korelace $r_{y.x_1x_2x_3x_4}$.

Řešení

V tabulce 4.37 je výběrová korelační matice všech proměnných, v prvním sloupci a řádku jsou uvedeny korelační koeficienty vysvětlujících proměnných x_1, x_2, x_3 a x_4 s vysvětlovanou proměnnou y.

Tab. 4.37 Výběrová korelační matice k příkladu 4.15

	cena	voda	hlučnost	elektřina	programy
cena	1	−0,45294	−0,58162	−0,44993	0,25525
voda	−0,45294	1	0,36737	0,39714	−0,23129
hlučnost	−0,58162	0,36737	1	0,60745	−0,68649
elektřina	−0,44993	0,39714	0,60745	1	−0,37983
programy	0,25525	−0,23129	−0,68649	−0,37983	1

Determinant matice 5×5 je roven 0,271 4, po vynechání prvního řádku a prvního sloupce dostaneme determinant 0,151 1. Po dosazení do (4.88) je

$$r_{y.x_1x_2x_3x_4} = \sqrt{1 - \frac{0,271\,4}{0,151\,1}} = 0,665\,8.$$

Všimněme si, že použijeme-li index determinace regresního modelu (4.72) uvedený v (4.77) $I^2 = 0,443\,3$, dostaneme stejnou hodnotu vícenásobného indexu determinace, neboť $\sqrt{I^2} = \sqrt{0,443\,3} = 0,665\,8$. Toto pozorování nabízí druhou možnost definice koeficientu, ve které nejdříve nalezneme index determinace pro vhodný model vícenásobné regrese a ten pak odmocníme.

■