Statistická analýza se zřídka zabývá pouze jednou izolovanou proměnnou. Častěji se zajímáme o srovnání několika rozdělení, o změny proměnné v čase nebo vztahy mezi proměnnými. V předchozí kapitole jsme se zabývali porovnáváním rozdělení jedné proměnné mezi dvěma skupinami nebo mezi dvěma různými časovými okamžiky. V této kapitole se koncentrujeme na základní metody pro studium závislosti mezi proměnnými. Takové metody budou užitečné, jestliže nás zajímá:

- predikce budoucích zisků z prodeje produktu v závislosti na jeho ceně;
- závislost redukce váhy na počtu týdnů, kdy se jedinec se podrobí dietnímu režimu s příjmem určitého energetické hodnoty;
- jak výška dítěte v šesti letech predikuje jeho výšku v šestnácti letech;
- jak ovlivňuje spotřeba alkoholu snížení tělesné teploty apod.

Vztahy, které jsme uvedli, nemají ryze funkčně deterministický charakter. Proto je nutné použít pro jejich analýzu statistické metody. Příslušná oblast statistiky se nazývá korelační a regresní analýza. Stejně jako v jednorozměrné analýze začneme s grafickou analýzou, pak přejdeme k shrnujícím numerickým charakteristikám dat. Při výkladu budeme klást důraz na ty vztahy mezi proměnnými, jež lze popsat přímkou (lineární závislostí). Výklad později rozšíříme na jednoduché nelineární vztahy. Postupy této kapitoly tvoří základ pro metody vícerozměrné statistiky (kap. 10, resp. kap. 13). Hodnocením vztahů mezi kategoriálními proměnnými se budeme zabývat v následující kapitole 8.

Korelační analýza zkoumá vztahy proměnných graficky a pomocí různých měr závislosti, které nazýváme korelační koeficienty. **Regresní analýza** dává odpovědi na otázky typu: jaký vztah existuje mezi proměnnými X a Y (lineární, kvadratický atd.), lze proměnnou Y odhadnout pomocí proměnné X a s jakou chybou? Statistická analýza v těchto souvislostech má následující cíle:

a) poskytnout číselné míry vztahu dvou proměnných podobným způsobem, jako průměr a směrodatná odchylka popisují chování jedné proměnné;
b) najít vzorce pro optimální predikci proměnné, kterou považujeme za závisle proměnnou;

c) ohodnotit chybu predikce;
d) ověřovat různé hypotézy o zkoumaném vztahu.

Korelační a regresní analýza má intelektuální kořeny v práci Francise Galtona (1894). Inspirován částečně dílem svého bratrance Charlese Darwina *O původu druhů* snažil se Galton odhalit dědičné vlastnosti talentu, pohybových schopností a intelektu. Jeho výzkum začal studiem velikosti a váhy bílého hrachu ve dvou generacích. Ten vedl k odhalení fenoménu regrese, kterou Galton nazýval „reverse" (návrat). Popíšeme ho podrobněji v kapitole 7.4. Galton zjistil, že potomci extrémně velikých hrachů nebyli v průměru tak extrémně velicí, jako jejich „rodiče". Později zaznamenával fyzické charakteristiky tisíců dobrovolníků a nalezl podobnou „regresi" k průměru při mezigeneračním srovnání. Jeho koncepty pak rozvinul jeho žák K. Pearson a další statistici. Tradiční název „regrese" se stále používá v důsledku tradice, ačkoli se tato statistická technika spíše používá pro predikci. Metodu nejmenších čtverců pro odhad regresních funkcí používali již A. M. Legendre a C. F. Gauss.

7.1 Zobrazení dvojrozměrných dat

Základní postup dvojrozměrné analýzy dat je podobný jako v jednorozměrném případě:

1. Nejdříve se pokusíme zobrazit data graficky.
2. Hledáme základní konfigurace a tendence v datech.
3. Přidáváme numerické charakteristiky různých aspektů dat.
4. Často se nám podaří vystihnout stručným způsobem základní konfiguraci dat pomocí pravděpodobnostního modelu.

Data máme v podobě určitého počtu číselných dvojic údajů (x_i, y_i), které jsme získali měřením proměnných X a Y. Vždy je na místě provést před dalším zpracováním jejich grafickou interpretaci. Vynesením dat do souřadnicového systému (např. na milimetrový papír nebo zobrazením na displeji počítače pomocí vhodného programu) získáme základní představu o společném rozdělení obou proměnných. Každý bod odpovídá jednomu páru měření. Takovému grafickému znázornění říkáme dvojrozměrný bodový graf. Jeho prohlídkou odhadneme, zda je mezi proměnnými přesná funkcionální závislost, případně volnější vztah, jejž nazýváme statistická závislost, anebo jestli jsou na sobě evidentně nezávislé. Na obrázku 7.1 jsou znázorněny hodnoty měření výšky a váhy deseti studentů. Graf zobrazuje údaje z tabulky 7.1. Data se také zobrazují pomocí tzv. korelační tabulky (tab. 7.2). Dochází přitom k určité ztrátě informací rozdělením dat do intervalů.

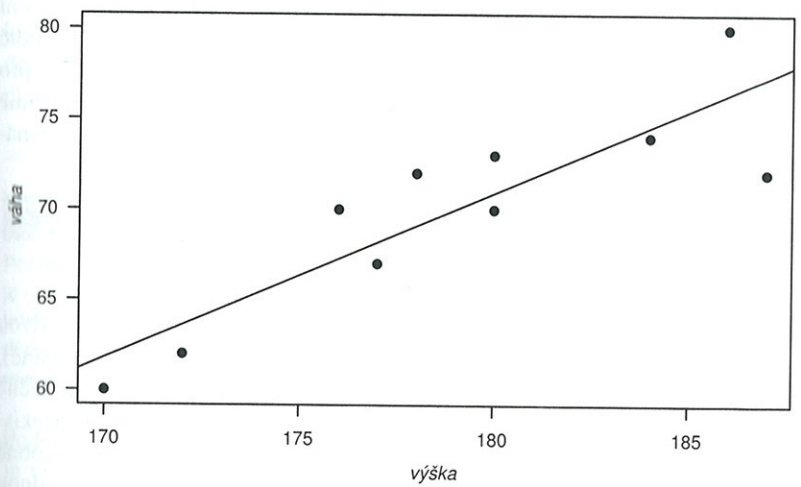
Tab. 7.1 Příklad dat, jejichž závislost chceme posoudit

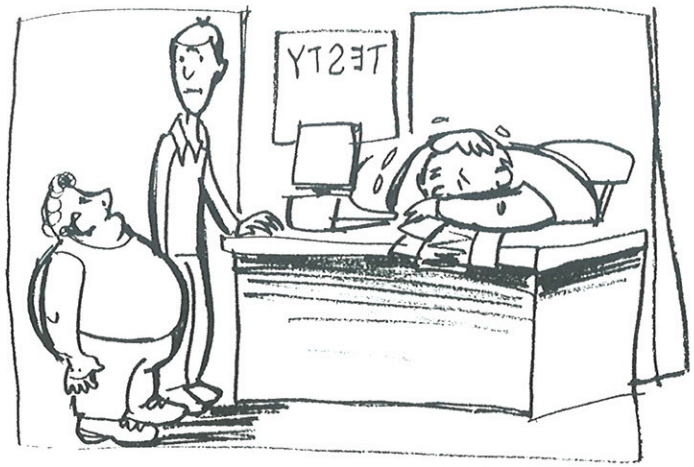
Výška [cm]	187	170	180	184	178	180	172	176	186	177
Váha [kg]	72	60	73	74	72	70	62	70	80	67

Tab. 7.2 Příklad korelační tabulky – korelace zjištěných hodnot výšky a váhy deseti studentů

Váha [kg]	**Výška [cm]**			**Celkem**
	do 170	**170–180**	**180–190**	
do 60	1	0	0	1
60–70	0	4	0	4
70–80	0	2	3	5
Celkem	1	6	3	10

Obr. 7.1 Bodový graf pro posouzení závislosti mezi váhou a výškou studentů





„Prý jsme mu zkazili jeho pozitivní korelaci mezi výškou a váhou."

Cílem regresní a korelační analýzy je popis statistických vlastností vztahu dvou nebo více proměnných. Dvojrozměrný bodový graf nebo korelační tabulka dávají první představu o rozdělení sledovaných proměnných. Graf často indikuje překvapivé vlastnosti dat jako nelinearitu vztahu, nehomogenitu nebo přítomnost odlehlých hodnot. Na obrázku 7.1 je rovněž vynesena přímka, která byla proložena body metodou nejmenších čtverců. Vliv třetí proměnné na rozložení bodů můžeme zachytit různým tvarem nebo barvou bodů v závislosti na hodnotě této proměnné (např. u dat o výšce a váze bychom mohli použít různé značky pro body odpovídající chlapcům a dívkám, pokud bychom tuto informaci o proměnné *pohlaví* měli k dispozici). Některé možné konfigurace dat v grafu popíšeme v následujícím odstavci.

7.2 Korelační analýza

V nejobecnějším smyslu, slovo „korelace" označuje míru stupně asociace dvou proměnných. Říká se, že dvě proměnné jsou korelované (resp. asociované), jestliže určité hodnoty jedné proměnné mají tendenci se vyskytovat společně s určitými hodnotami druhé proměnné. Míra této tendence může sahat od neexistence korelace (všechny hodnoty proměnné Y se vyskytují stejně pravděpodobně s každou hodnotou proměnné X) až po absolutní korelaci (s danou hodnotou

proměnné X, se vyskytuje právě jedna hodnota proměnné Y). Pro měření korelace byla navržena řada koeficientů. Liší se podle typů proměnných, pro které se využívají. Statistické usuzování o korelačních koeficientech se opírá o teorii pravděpodobnosti pro společné rozdělení dvou nebo více náhodných proměnných.

Při zkoumání korelačních vztahů má rozhodující význam kvalitativní rozbor příslušného materiálu. Nemá smysl měřit závislost tam, kde na základě logické úvahy nemůže existovat. Často je zbytečné měřit závislosti i z jiných důvodů. Je to zejména tehdy, když je korelace způsobena: a) formálními vztahy mezi proměnnými; b) nehomogenitou studovaného základního materiálu; c) působením společné příčiny.

Formální korelace vzniká např. tehdy, když se zjišťuje korelace procentuálních charakteristik, jež se navzájem doplňují do 100 % (např. korelace procentního zastoupení bílkovin a tuku v potravinách).

Jestliže populace, kterou studujeme, obsahuje subpopulace, pro něž se průměrné hodnoty proměnných X a Y liší, vypočtené korelační vztahy jsou touto **nehomogenitou** silně ovlivněny a jejich hodnoty nepopisují skutečný vztah mezi uvažovanými proměnnými. Nehomogenita materiálu se projeví na bodovém grafu tak, že shluky bodů pro subpopulace se budou nacházet v různých oblastech souřadnicového systému. Na obrázku 7.2 je modelově ukázáno působení nehomogenity. Ta má za důsledek, že korelačním koeficientem hodnotíme bez diferenciace najednou dva shluky bodů, které přísluší k různým populacím. Na obrázku a) to vede k nenulovém korelačními koeficientu i přesto, že v obou shlucích jsou proměnné nekorelované, naopak proměnné na obrázku b) jsou v obou shlucích proměnné korelované, ale celková korelace je nulová.

Příkladem **korelací způsobených společnou příčinou** jsou vztahy mezi některými mírami těla, např. mezi délkou pravé a levé ruky. Jiným známým příkladem jsou zdánlivé korelace způsobené časovým faktorem nebo faktorem modernizace u dvou řad údajů.

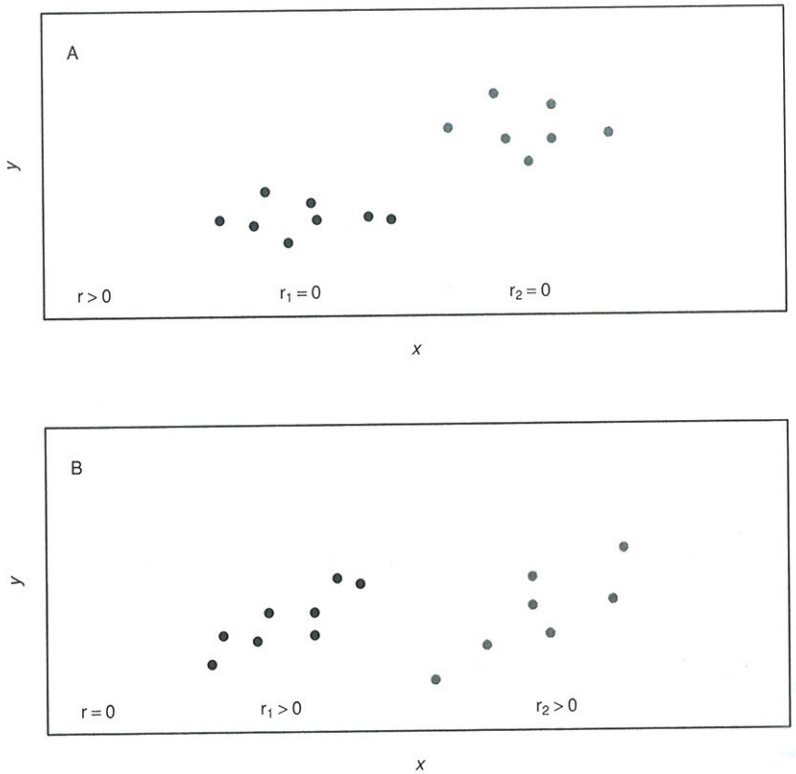
PŘÍKLAD 7.1

Zdánlivé korelace

Počet televizních přístrojů na osobu koreluje s očekávanou délkou života. Ve státech, kde je mnoho televizních přístrojů, dosahují obyvatelé vysokého věku. Je možné změnou počtu televizních přístrojů dosáhnout prodloužení věku v oblastech světa, kde je nižší očekávaná délka života?

Podobným korelacím se někdy říká „nesmyslné" korelace. Hodnota korelace je vysoká. Nesmyslný by byl závěr o příčinném působení. Korelační závislost

Obr. 7.2 Příklad kladné (A) a nulové (B) korelace, které jsou způsobené nehomogenitou dat

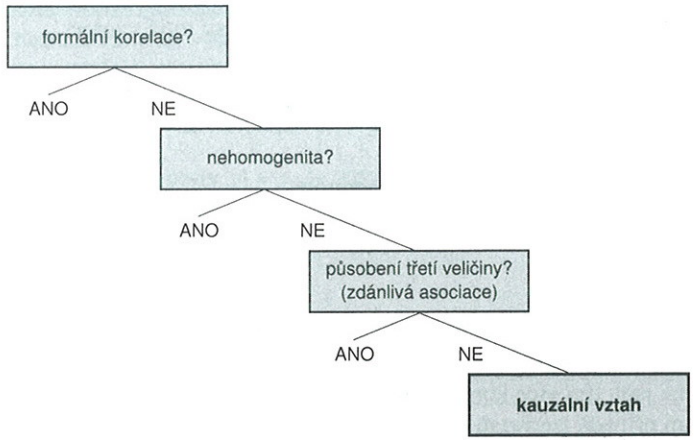


Korelační koeficient r je vypočtený pro všechny body, koeficienty r_1 a r_2 odděleně pro každý shluk zvlášť

je zdůvodněna proměnnou „národní důchod", jež je společnou příčinou obou proměnných.

Kromě tohoto působení proměnné jako „společné příčiny" mohou působit matoucí (rušivé) proměnné, které korelují jak s cílovou proměnnou, tak s proměnnou ovlivňující. Proměnná v tomto případě znesnadňuje interpretaci, protože nelze rozlišit vliv matoucí a sledované ovlivňující proměnné na cílovou proměnnou. Uvádíme pořadí, v němž máme vylučovat nezajímavé korelace, než se dostaneme do fáze, kdy by velká korelace mohla indikovat kauzální vztah (obr. 7.3).

Obr. 7.3 Postup pro ověření kauzálního vztahu



7.2.1 Pearsonův korelační koeficient

Přes některé své nedostatky zůstává Pearsonův korelační koeficient r nejdůležitější mírou síly vztahu dvou náhodných spojitých proměnných X a Y. Počítáme jej z n párových hodnot $\{(x_i, y_i)\}$ změřených na n jednotkách náhodně vybraných z populace. Korelační koeficient r nabývá hodnot z intervalu $[-1; 1]$. Jestliže má hodnotu 1 nebo -1, pak y-souřadnici bodu lze přesně spočítat pomocí lineárního vztahu z jeho x-souřadnice. Korelační koeficient r počítáme pomocí tzv. kovariance s_{xy} a směrodatných odchylek s_x a s_y obou proměnných:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Vzorec s kovariancí pomáhá porozumět tomu, že r má kladnou hodnotu, pokud asociace proměnných je pozitivní. Dejme tomu, že studujeme korelaci výšky a váhy studentů. Jedinci, kteří mají hodnotu výšky nad průměrem, mívají nadprůměrnou i hodnotu váhy. Oba rozdíly od průměru, jež spolu násobíme při výpočtu kovariance, budou mít u vyšších a těžších jedinců kladnou hodnotu. Jedinci, kteří mají menší výšku, mají obvykle i menší váhu. U nich jsou oba

rozdíly od průměrů záporné, a proto je součin rozdílů od průměru rovněž kladný. Protože je většina sčítanců kladných, musí být kladná i výsledná hodnota kovariance a tedy i korelačního koeficientu. Tuto interpretaci lze ještě lépe pochopit při výpočtu r pomocí standardizovaných hodnot. Platí totiž vzorec

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\sum_{i=1}^{n} x_i' y_i'}{n-1},$$

kde x' a y' označují standardizované hodnoty.

Důležité vlastnosti Pearsonova korelačního koeficientu r shrneme pomocí několika tvrzení:

1. Platí $-1 \leq r \leq 1$.
2. Jestliže $|r| = 1$, leží všechny body na nějaké přímce.
3. Jestliže $r = 0$, nazýváme X a Y nekorelované proměnné. Dvě náhodné proměnné jsou tím více korelovány, čím blíže je hodnota r k číslům 1 nebo -1. V tom případě lze vztah obou proměnných dobře vyjádřit přímkou.
4. Jestliže $r < 0$, resp. $r > 0$, tak se Y v průměru zmenšuje, resp. zvětšuje při zvětšování proměnné X. Říkáme, že je asociace je záporná, resp. kladná.
5. Pearsonův korelační koeficient vyjadřuje pouze sílu *lineárního* vztahu. Špatně měří jiné vztahy, ať jsou jakkoli silné.
6. Korelační koeficient se nezmění, když změníme jednotky měření proměnných X a Y.
7. Podobně jako průměr nebo směrodatná odchylka, je korelační koeficient r velmi ovlivněn odlehlými hodnotami.
8. Korelační koeficient r nerozlišuje mezi závisle a nezávisle proměnnou.
9. Korelační koeficient r není úplným popisem dat i při velmi silném lineárním vztahu. Pro úplnější popis potřebujeme znát rovnici přímky, která vyjadřuje tvar vztahu.
10. Pokud jedna z proměnných nemá náhodný charakter (její hodnoty jsou pevně určeny), není vhodné korelační koeficient použít.
11. Korelace, ať je jakkoli silná, *neznamená* sama o sobě průkaz příčinného vztahu, tedy toho, že změny proměnné X skutečně působí změny proměnné Y.

Mezi proměnnými mohou existovat nejrůznější vztahy a máme i různé způsoby, jak je měřit. Některé z nich popíšeme v dalších odstavcích. Ačkoli korelační koeficient se používá velmi často, je nutné mít na paměti jeho omezení.

PŘÍKLAD 7.2

Výpočet korelačního koeficientu

Budeme hodnotit závislost výšky a váhy, jejichž hodnoty jsme naměřili u 10 studentů. Vypočítáme korelační koeficient pro párové hodnoty, které jsou uvedeny spolu s potřebnými dopočítanými hodnotami v tabulce 7.3. Hodnoty jsou zobrazeny na obr. 7.1 (s. 239).

Součet v posledním sloupci je základem pro výpočet kovariance

$$\mathrm{cov}(x, y) = s_{xy} = 259/(10 - 1) = 28{,}8.$$

Dále jsme zjistili: $\bar{x} = 1790/10 = 179$; $\bar{y} = 700/10 = 10$; $s_x = 5{,}61$; $s_y = 5{,}83$. Korelační koeficient má tedy hodnotu $r = 28{,}8/(5{,}61 \times 5{,}83) = 0{,}88$.

Tab. 7.3 Příklad postupu výpočtu korelačního koeficientu

	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
	187	72	8	2	16
	170	60	−9	−10	90
	180	73	1	3	3
	184	74	5	4	20
	178	72	−1	2	−2
	180	70	1	0	0
	172	62	−7	−8	56
	176	70	−3	0	0
	186	80	7	10	70
	177	67	−2	−3	6
Součet	1790	700	0	0	259

Někdy se zařazují hodnoty korelace do pásem podle síly asociace. V tabulkce 7.4 uvádíme jeden z návrhů. Interpretace hodnot korelačního koeficientu není tak přímočará, jako je tomu u většiny jednorozměrných charakteristik. Proto se doporučuje dopočítat další charakteristiky, jako jsou parametry proložené přímky nebo směrodatná chyba odhadu při regresi (viz další kapitola).

Tab. 7.4 Pásma síly asociace podle velikosti korelačního koeficientu r

Síla asociace	$\lvert r \rvert$
malá	0,1–0,3
střední	0,3–0,7
velká	0,7–1,0

Hodnota korelačního koeficientu je bohužel silně ovlivňována odlehlými hodnotami ve výběru. Zkreslení také nastane, když se při výběru objektů omezíme pouze na ty, jejichž hodnota proměnné X nebo Y musí ležet v určitém intervalu. Korelační koeficient r má pak tendenci být menší než korelace r' vypočítaná bez omezení kladeného na data. Pro úpravu zkresleného korelačního koeficientu vlivem omezení rozsahu měření proměnné X použijeme vzorec

$$r' = \frac{Ur}{\sqrt{(U^2 - 1)r^2 + 1}},$$

kde $U = s/s'$ je poměr směrodatné odchylky s měření X ve studii a směrodatné odchylky s' v populaci bez restrikce.

Korelační koeficient je také ovlivněn nepřesností metod, kterými měříme obě proměnné. Jestliže známe r_{yy} a r_{xx} koeficienty spolehlivosti měření obou proměnných (jedná se korelace opakovaných měření), lze se přiblížit hodnotě korelačního koeficientu bezchybně změřených proměnných $r_{x'y'}$ pomocí úpravy

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}.$$

PŘÍKLAD 7.3

Význam exploračního zobrazení dvojrozměrných dat

Jednoduchým příkladem toho, jakou důležitou roli hraje explorační zobrazení dat, je zkoumání čtyř sérií modelových dat podle Anscomba (1973), které uvádí tabulka 7.5. Základní statistické charakteristiky proměnných X a Y a jejich korelační koeficient mají pro první sérii dat hodnoty $\bar{x} = 9{,}0$; $s_x = 3{,}31$; $\bar{y} = 7{,}5$; $s_y = 2{,}03$ a $r = 0{,}816$. Pokud spočteme tyto charakteristiky pro ostatní série, zjistíme, že jsou stejné. Pokud však všechny čtyři série zobrazíme graficky (viz obr. 7.5a-d, s. 260), výsledek je dost překvapivý.

Tab. 7.5 Série modelových dat se stejnými základními statistickými charakteristikami a korelačními koeficienty

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,1	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,1	4	5,39	19	12,5
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

7.2.2 Pravděpodobnostní rozdělení dvou náhodných proměnných

Teorie pravděpodobnosti popisuje nejen rozdělení jedné náhodné proměnné, ale i společná pravděpodobnostní rozdělení dvou nebo více náhodných proměnných. Této teorie je zapotřebí tehdy, když chceme navrhovat pravděpodobnostní modely vztahu proměnných a zdůvodnit procedury pro statistické usuzování v korelační a regresní analýze. V našem jednoduše pojatém výkladu budeme postupovat tak, abychom mohli získané výsledky využít i v kapitole o analýze závislosti kategoriálních proměnných.

Zatím jsme se seznámili s jednou dvojrozměrnou charakteristikou, s Pearsonovým korelačním koeficientem r. Teoretickou hodnotu Pearsonova korelačního koeficientu v populaci označujeme ρ. Získali bychom ji výpočtem z údajů o všech prvcích populace. Výběrový koeficient r je bodovým odhadem této hodnoty. S rostoucím rozsahem výběru n se hodnota výběrového korelačního koeficientu r_n blíží ke své teoretické hodnotě ρ.

Teoretickou hodnotu ρ lze přímo odvodit podobně jako teoretickou střední hodnotu μ, když známe společné pravděpodobnostní rozdělení náhodných proměnných, pro které korelační koeficient počítáme. Koncept dvojrozměrného pravděpodobnostního rozdělení a techniku výpočtu teoretických hodnot ozřejmíme pomocí jednoduchého příkladu. Postupujeme podobně jako v jednorozměrném případě (viz kap. 4.2).

Představme si, že v daném pokusu můžeme získat pro hodnoty proměnných X a Y pouze tři různé hodnoty: $x \in (7; 15; 2)$, $y \in (3; 6; 9)$. Společné pravděpodobnostní p_{xy} rozdělení proměnných X a Z je popsáno tabulkou pravděpodobností všech možných kombinací uvedených hodnot (tab. 7.6). Poslední sloupec resp. poslední řádek tabulky 7.6 a) obsahuje jednorozměrná rozdělení p_x a p_y náhodných proměnných X a Y. Tyto hodnoty jsme dostali jako součet pravděpodobností v daném řádku, resp. sloupci. Nazýváme je **marginální rozdělení**. Z tabulky je vidět, že dvojici hodnot $(x, y) = (6; 7)$ lze dostat v náhodném pokusu s pravděpodobností 0,1, avšak pravděpodobnost výskytu jevu $y = 7$ je 0,2. Pomocí marginálních rozdělení spočítáme očekávané hodnoty pro proměnné X a Y a také dopočítáme teoretické hodnoty rozptylů obou proměnných pro proměnnou X a Y – (tab. 7.6 b, c).

Tab. 7.6 Příklad dvojrozměrného prevděpodobnostního rozdělení a výpočet jeho charakteristik

a) Dvojrozměrné rozdělení

	y			
x	7	15	2	p_x
3	0,1	0,2	0,0	0,3
6	0,1	0,0	0,3	0,4
9	0,0	0,1	0,2	0,3
p_y	0,2	0,3	0,5	1,0

b) Výpočet průměrných hodnot pro proměnnou X

x	p_x	xp_x	x^2p_x
3	0,3	0,9	2,7
6	0,4	2,9	14,4
9	0,3	2,7	24,3
Součet	1,0	6,0	41,4
		$= E(x)$	$= E\left(x^2\right)$

c) Výpočet průměrných hodnot pro proměnnou Y

y	p_y	yp_y	y^2p_y
7	0,2	1,4	9,8
15	0,3	4,5	67,5
2	0,5	1,0	2,0
Součet	1,0	6,9	79,3
		$= E(y)$	$= E\left(y^2\right)$

Teoretické hodnoty průměru a rozptylu náhodných proměnných X a Y jsou tedy:

$$\mu_x = E(X) = \sum_i x_i p_i(x) = 6{,}0 \qquad \sigma_x^2 = E(x^2) - \mu_x^2 = 41{,}4 - 6^2 = 5{,}4$$

$$\mu_y = E(Y) = \sum_j y_j p_j(y) = 6{,}9 \qquad \sigma_y^2 = E(y^2) - \mu_y^2 = 79{,}3 - 6{,}9^2 = 31{,}7$$

To znamená, že $\sigma_x = \sqrt{5{,}4} = 2{,}32$ a $\sigma_y = \sqrt{31{,}7} = 5{,}63$.

Teoretickou kovarianci σ_{xy} vypočteme modifikací vzorce pro výběrovou kovarianci:

$$\sigma_{xy} = E\left[(X - \mu_x)(Y - \mu_y)\right] = E(XY) - \mu_x\mu_y = \sum_i \sum_j x_i y_j \, p_{ij}(x, y) - \mu_x\mu_y$$

Nejdříve spočítáme hodnoty $E(XY)$:

$$\begin{aligned} E(XY) &= \sum_i \sum_j x_i y_j \cdot p_{ij}(x, y) \\ &= 0{,}1 \times 3 \times 7 + 0{,}2 \times 3 \times 15 + 0{,}0 \times 3 \times 2 \\ &+ 0{,}1 \times 6 \times 7 + 0{,}0 \times 6 \times 15 + 0{,}3 \times 6 \times 2 \\ &+ 0{,}0 \times 9 \times 7 + 0{,}1 \times 9 \times 15 + 0{,}2 \times 9 \times 2 \\ &= 36{,}0 \end{aligned}$$

Takže $\sigma_{xy} = E(XY) - \mu_x\mu_y = 36{,}0 - 6{,}0 \times 6{,}9 = -5{,}4$. Teoretickou hodnotu koeficientu korelace pak dostaneme dosazením teoretických hodnot do vzorce pro výběrový koeficient korelace

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{-5{,}4}{2{,}32 \times 5{,}63} = -0{,}41.$$

Teoretická korelace −0,41 indikuje sílu závislosti mezi oběma proměnnými.

V případě spojitých náhodných proměnných jsou tyto výpočty sice komplikovanější, ale stejně jako u jednorozměrných charakteristik se koncepčně moc neliší.

Dále připomeneme pojem nezávislosti náhodných proměnných a ukážeme, že pokud náhodné proměnné jsou nezávislé, pak se jejich korelační koeficient rovná nule. V dřívějším výkladu jsme **nezávislost náhodných proměnných** vymezili požadavkem, že realizace jedné náhodné proměnné neovlivňuje chování druhé náhodné proměnné (např. hodnota jedné proměnné u určité osoby neovlivňuje hodnotu měření jiné proměnné u téže ani jiné osoby).

Definice nezávislosti dvou náhodných proměnných vychází z počítání pravděpodobností podmnožin dvojrozměrného prostoru $R \times R$. Nechť množiny A_x, resp. A_y mají pravděpodobnosti $P_x(A_x)$, resp. $P_y(A_y)$ vzhledem k rozdělení proměnné X, resp. Y. Pak X a Y jsou **stochasticky nezávislé**, nebo prostě nezávislé, pokud pravděpodobnost množiny $A_x \times A_y$ vzhledem k uvažovanému dvojrozměrnému rozdělení lze vypočítat vynásobením pravděpodobností obou množin $P(A_x \times A_y) = P_x(A_x)P_y(A_y)$. Tato podmínka musí platit pro všechny podmnožiny A_x a A_y. Je patrné, že se jedná o převedení pojmu nezávislosti náhodných jevů na chování náhodných proměnných.

Pro náš příklad popíšeme dvojrozměrné rozdělení tabulkou 7.6a. Pravděpodobnosti p_{xi}, resp. p_{yj} vznikly součtem pravděpodobností v řádku, resp. v sloupci tabulky. Nazýváme je **marginální pravděpodobnosti**. Definují marginální rozdělení, které popisuje náhodné chování izolovaných proměnných X a Y. Jestliže proměnné X a Y jsou nezávislé, pak z definice plyne, že pravděpodobnosti v tabulce jsou součiny marginálních pravděpodobností (viz tab. 7.7b).

Pojem stochastické nezávislosti dále ilustruje výpočet podmíněné pravděpodobnosti $p(x = i \mid y = j)$, tedy pravděpodobnosti, že náhodná proměnná X bude mít hodnotu i, za předpokladu, že náhodná proměnná Y má hodnotu j. Protože platí $p(x = i \mid y = j) = p_{xiyj}/p_{yj}$, má tabulka hledaných podmíněných pravděpodobností tvar jako tabulka 7.7c. Hodnoty v ní vyjadřují, že fixujeme-li

Tab. 7.7 Obecné dvojrozměrné rozdělení a nezávislost proměnných

a) Dvojrozměrné rozdělení proměnných X a Y

	y			
x	7	15	2	p_x
3	p_{11}	p_{12}	p_{13}	p_{x1}
6	p_{21}	p_{22}	p_{23}	p_{x2}
9	p_{31}	p_{32}	p_{33}	p_{x2}
p_y	p_{y1}	p_{y2}	p_{y3}	1

b) Podmínka pro nezávislost X a Y

	y			
x	7	15	2	p_x
3	$p_{x1}p_{y1}$	$p_{x1}p_{y2}$	$p_{x1}p_{y3}$	p_{x1}
6	$p_{x2}p_{y1}$	$p_{x2}p_{y2}$	$p_{x2}p_{y3}$	p_{x2}
9	$p_{x3}p_{y1}$	$p_{x3}p_{y2}$	$p_{x3}p_{y2}$	p_{x2}
p_y	p_{y1}	p_{y2}	p_{y3}	1

c) Podmíněná rozdělení proměnné X za podmínky $y = y_i$

	y			
x	7	15	2	p_x
3	p_{x1}	p_{x1}	p_{x1}	p_{x1}
6	p_{x2}	p_{x2}	p_{x2}	p_{x2}
9	p_{x3}	p_{x3}	p_{x3}	p_{x2}
p_y	p_{y1}	p_{y2}	p_{y3}	1

Tab. 7.8 Rozdělení pravděpodobnosti pro dvě nezávislé proměnné

	Proměnná y			
Proměnná x	7	15	2	p_x
3	0,06	0,09	0,15	0,3
6	0,08	0,12	0,2	0,4
9	0,06	0,09	0,15	0,3
p_y	0,2	0,3	0,5	1

proměnnou Y, je podmíněné rozdělení náhodné proměnné X stejné pro všechny hodnoty proměnné Y a toto rozdělení se shoduje s příslušným marginálním rozdělením proměnné X. Pojmenujme očekávanou hodnotu náhodné proměnné X při fixované hodnotě náhodné proměnné Y „podmíněná očekávaná hodnota". Z tabulky 7.7c je patrné, že podmíněné očekávané hodnoty náhodné proměnné X jsou stejné pro všechny hodnoty náhodné proměnné Y.

Pro ilustraci, jak se nezávislost projevuje na hodnotě korelačního koeficientu, vytvoříme z původního dvojrozměrného rozdělení proměnných X a Y v naší tabulce 7.7a nové rozdělení, aby proměnné byly nezávislé. Postupujeme tak, že zachováme podobu jednorozměrných marginálních pravděpodobnostních rozdělení proměnných X a Y a dopočítáme ostatní pravděpodobnosti podle předpisu pro nezávislost. Nové dvojrozměrné pravděpodobnostní rozdělení popisuje tabulka 7.8. Ukážeme, že se v tomto případě – to je u nezávislých náhodných proměnných – očekávaná hodnota jejich součinu rovná součinu jejich očekávaných hodnot, tedy $E(XY) = E(X)E(X)$. Protože $p_{ij} = p_i p_j$, platí:

$$E(XY) = \sum_i \sum_j x_i y_j p_{ij} = \sum_i \sum_j x_i y_j p_i p_j = \sum_i x_i p_i \sum_j y_j p_j = E(X)E(Y)$$

Důležitá je okolnost, že uvedený vztah lze zobecnit a dokázat i pro spojité náhodné proměnné.

Označili jsme $E(X) = \mu_x$ a $E(Y) = \mu_y$. Protože pro nezávislé proměnné platí:

$$E[(X - \mu_x)(Y - \mu_y)] = E(XY) - \mu_x\mu_y = E(X)E(Y) - \mu_x\mu_y = \mu_x\mu_y - \mu_x\mu_y = 0,$$

plyne z toho, že kovariance σ_{xy} a tedy i (teoretický) korelační koeficient ρ dvou stochasticky nezávislých náhodných proměnných jsou vždy rovné nule (čtenář se může přesvědčit přímým výpočtem pro hodnoty v poslední tabulce). Neplatí to však obráceně. Nulová hodnota korelačního koeficientu neznamená vždy, že proměnné jsou stochasticky nezávislé. Pro jednu významnou třídu rozdělení však i toto obrácené tvrzení platí. Jedná se o tzv. dvojrozměrné normální rozdělení

náhodných proměnných (X, Y). Jde o rozšíření pojmu normálního rozdělení, které jsme poznali v kap. 4.5.3, na systém dvou proměnných. Dvojrozměrné normální rozdělení je jednoznačně určeno průměry a rozptyly obou proměnných a jejich korelačním koeficientem ρ_{xy}. Zobecnění pro vícerozměrné normální rozdělení se provádí analogicky.

7.2.3 Odhad a testování korelačního koeficientu

Popíšeme testy a intervaly spolehlivosti pro Pearsonův korelační koeficient. Tyto metody lze použít za předpokladu, že společné rozdělení obou proměnných lze modelovat dvojrozměrným normálním rozdělením nebo – jinak vyjádřeno – rozdělení obou proměnných je normální a jejich vztah je přibližně lineární.

Při posuzování, zda se vypočítaná hodnota korelačního koeficientu významně liší od nuly, použijeme tabulku IX z přílohy B, kde jsou hodnoty kritických mezí pro výběrový korelační koeficient v závislosti na rozsahu výběru. Jestliže bylo k dispozici n párových hodnot, má vypočtený korelační koeficient $n - 2$ stupňů volnosti. Přesahuje-li v absolutní hodnotě hodnotu v tabulce pro požadovanou hladinu významnosti, můžeme vztah považovat za prokázaný na dané hladině významnosti. Snadno nahlédneme, že s rostoucím počtem pozorování prokážeme statistickou významnost i velmi malého korelačního koeficientu.

Jestliže chceme testovat obecnější hypotézu $H_0: \rho_{xy} = \rho_0$ proti alternativě $H_1: \rho_{xy} \neq \rho_0$, kde $\rho_0 \neq 0$, musíme použít tzv. Fisherovu z-transformaci (arctanh – „arkustangens hyperbolický"):

$$z = \dot{z}(r) = \operatorname{arctanh}(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

kde $\dot{z}$ označujeme Fisherovu transformaci. Touto transformací jsme rozšířili interval hodnot $-1 \leq r \leq +1$ na interval $-\infty \leq z \leq +\infty$. Nová proměnná má přibližně průměr

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$$

a směrodatnou odchylku

$$s_z = \sqrt{\frac{1}{n-3}},$$

takže pro test nulové hypotézy lze použít interval spolehlivosti ve tvaru

$$z - ts_z \leq \mu_z \leq z + ts_z,$$

kde t je kritická hodnota pro dvoustranný test zjištěná pomocí t-rozdělení o $n - 2$ stupních volnosti na odpovídající hladině významnosti.

Zpět do měřítka korelačního koeficientu převedeme oba krajní body intervalu pomocí inverzní transformace $\dot{z}^{-1}$:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Získáme tak interval spolehlivosti pro ρ_{xy}.

PŘÍKLAD 7.4

Testování hodnoty korelačního koeficientu

Test hypotézy $H_0: \rho_{xy} = 0{,}5$ proti $H_1: \rho_{xy} \neq 0{,}5$ pro náš případ, kdy $n = 10$, $r = 0{,}88$, provedeme pomocí intervalu spolehlivosti s hladinou 0,95. Vypočteme nejdříve Fisherovu z-transformaci (protože ρ_0 se nerovná nule)

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0{,}88}{1-0{,}88}\right) = \frac{1}{2}\ln\left(\frac{1{,}88}{0{,}12}\right) = 1{,}375$$

a směrodatnou odchylku

$$s_z = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{10-3}} = \sqrt{\frac{1}{7}} = 0{,}37796 \doteq 0{,}378.$$

Kritická hodnota t-rozdělení s 8 stupni volnosti má pro zvolenou hladinu spolehlivosti hodnotu 2,306. Interval spolehlivosti má tedy tvar

$$1{,}375 - 2{,}306 \times 0{,}378 \leq \mu_z \leq 1{,}375 + 2{,}306 \times 0{,}378 \Rightarrow (0{,}504; 2{,}247).$$

Pomocí zpětné transformace $\dot{z}^{-1}$ převedeme tento interval do měřítka pro r a dostáváme (0,465; 0,977). Protože hodnota 0,5 leží v tomto intervalu, nemůžeme nulovou hypotézu zamítnout.

Pokud chceme testovat významnost rozdílu dvou korelačních koeficientů r_1 a r_2, získaných změřením dvojic proměnných ve dvou rozdílných skupinách r_1 a r_2, transformujeme oba korelační koeficienty Fisherovou transformací na hodnoty $\dot{z}_1$ a $\dot{z}_2$. Přibližně platný 95% interval spolehlivosti pro rozdíl Δ_z má pak tvar

$$\dot{z}_1 - \dot{z}_2 - 1{,}96\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}} \leq \Delta_z \leq \dot{z}_1 - \dot{z}_2 + 1{,}96\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}.$$

Na meze tohoto intervalu následně uplatníme zpětnou transformaci $\dot{z}^{-1}(\Delta_z)$, abychom získali interval spolehlivosti pro hodnotu $\Delta = \rho_1 - \rho_2$.

7.2.4 Problém třetí proměnné v korelační analýze

Korelace mezi dvěma proměnnými je někdy zavádějící a obtížně se interpretuje. Musíme zohlednit, že korelace dvou proměnných může být ovlivněna několika dalšími proměnnými. Mnoho atributů – jako např. výška, váha, síla, mentální schopnost, slovní zásoba, dovednost číst atd. – roste v rozmezí 6 až 18 let s věkem. Korelace těchto proměnných budou určitě pozitivní. Když z nich však vyloučíme působení věku, pravděpodobně klesnou k nule. Vliv rušivého faktoru „věk" kontrolujeme dvěma způsoby. Buď měříme vztah proměnných pouze pro vybranou věkovou kategorii, nebo použijeme tzv. parciální korelační koeficient. Podívejme se na tuto druhou možnost podrobněji. Budeme uvažovat rušivý vliv pouze jedné proměnné, ačkoli postup výpočtu parciálního korelačního koeficientu lze zobecnit pro libovolné množství rušivých parametrů. Pro jeho užití platí stejné předpoklady a omezení jako v případě normálního korelačního koeficientu.

Předpokládáme lineární asociace mezi proměnnými X, Y a Z zachycené korelačními koeficienty ρ_{xy}, ρ_{xz}, ρ_{yz}. **Parciální korelační koeficient** $\rho_{xy.z}$ měřící sílu vztahu proměnných X, Y po vyloučení vlivu parametru Z vypočítáme podle vzorce:

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

Jeho odhad pomocí naměřených hodnot (x_i, y_i, z_i) získáme dosazením výběrových hodnot korelačních koeficientů za teoretické. Výpočty a odhady parciálních korelačních koeficientů $\rho_{yz.x}$ a $\rho_{xz.y}$ dostaneme příslušnou cyklickou záměnou korelačních koeficientů.

Při testování nulové hodnoty parciálního korelačního koeficientu postupujeme stejně jako v případě jednoduchého korelačního koeficientu. Abychom však nalezli správnou kritickou mez, použijeme počet stupňů volnosti $n - 3$, kde n je počet trojic dat ve výběru.

PŘÍKLAD 7.5

Problém třetí proměnné v korelační analýze

V rámci screeningové akce bylo vyšetřeno 142 starších žen, u kterých byly také zaznamenávány parametry věk (v), krevní tlak (t) a koncentrace cholesterolu (c) v krvi. Pro ně se vypočítaly korelační koeficienty $r_{vt} = 0{,}33$; $r_{vc} = 0{,}5$; $r_{tc} = 0{,}25$. Protože zvýšené hodnoty krevního tlaku by mohly souviset se zvýšeným množstvím cholesterolu na stěnách cév, byla tato otázka důkladněji statisticky zkoumána. Parametry t a c s věkem rostou, tážeme se proto, zda jejich poměrně slabší korelace není způsobena efektem parametru věk. Vliv věku jako rušivého parametru se eliminuje zjištěním parciálního korelačního koeficientu $r_{tc.v}$:

$$r_{tc.v} = \frac{0{,}25 - 0{,}33 \times 0{,}50}{\sqrt{(1-0{,}33^2)(1-0{,}50^2)}} = 0{,}1$$

Pro 139 = (142 – 3) stupňů volnosti se nedá na hladině významnosti 5 % prokázat významnost tohoto korelačního koeficientu. Tímto statistickým zkoumáním jsme neukázali, že pro každou věkovou kategorii je krevní tlak pozitivně korelován s hladinou cholesterolu v krvi.

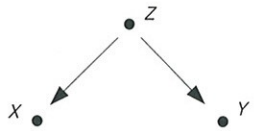
Výpočet parciálního korelačního koeficientu provádíme ve studiích, v nichž nás zajímá hlubší analýza vztahu mezi proměnnými a ověřování hypotéz o příčinných vztazích. Uvedeme v přehledu různé konfigurace korelačních vztahů proměnných X, Y, Z, přičemž budeme uvažovat i směr možné kauzality (obr. 7.4). Uvedená kauzální schémata implikují hodnoty korelačních koeficientů (v praktické analýze ovšem předpokládáme rovnost nule pouze přibližnou). Naopak to jednoznačně

Obr. 7.4 Různé konfigurace korelačních vztahů

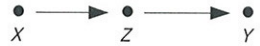
a) X, Y, Z jsou nekorelovány
$r_{xy} = 0$
$r_{yz} = 0$
$r_{xz} = 0$

b) X a Y jsou dvě nekorelované příčiny pro proměnnou Z
$r_{xy} = 0$
$r_{yz} \neq 0$
$r_{xz} \neq 0$

c) Z je společná příčina X a Y
$r_{xy} \neq 0$
$r_{yz} \neq 0$
$r_{xz} \neq 0$
ale $r_{xy.z} = 0$



d) vztah X a Y je zprostředkován Z
$r_{xz} \neq 0$
$r_{yz} \neq 0$
$r_{xy} = r_{xz}r_{yz}$
ale $r_{xy.z} = 0$



neplatí. Například $X \to Z \to Y$ má stejné koeficienty jako $Y \to Z \to X$. Stejně tak situace c) a d) jsou empiricky neodlišitelné. V těchto případech interpretujeme vztahy na základě dosavadních teoretických poznatků a pomocí základních kritérií pro ověřování kauzálního vztahu: a) silná asociace mezi proměnnými; b) prokázání této asociace v různých podmínkách (konzistence asociace); c) prokázaní změny hodnoty jedné proměnné při změně hodnoty druhé proměnné; d) působení proměnné klasifikované jako příčina předchází efektu v čase; e) existence věrohodného teoretického modelu působení.

7.2.5 Vliv dvou nezávisle proměnných na závisle proměnnou

Mnohonásobný koeficient korelace se používá v situacích, kdy chceme zjistit celkovou sílu vztahu mezi zvolenou proměnnou na jedné straně a několika dalšími (predikujícími) proměnnými $X_2, X_3, \ldots, X_k$ na straně druhé. Hodnotí se jím význam kumulativního vlivu více proměnných na zvolenou cílovou proměnnou. Mnohonásobný korelační koeficient, který pro tři proměnné značíme $\rho_{x.yz}$, je číselnou mírou možnosti predikce cílové proměnné X pomocí proměnných Y a Z:

$$\rho_{x.yz} = \sqrt{\frac{\rho_{xy}^2 + \rho_{xz}^2 - 2\rho_{xz}\rho_{xy}\rho_{yz}}{1 - \rho_{yz}^2}}$$

Jeho odhad získáme dosazením příslušných výběrových korelačních koeficientů do tohoto vzorce. Nulovou hypotézu, že $\rho_{x.yz} = 0$, testujeme pomocí F-testu provedeným transformovanou hodnotou $r_{x.yz}$:

$$F = \frac{r_{x.yz}^2(n-3)}{2(1 - r_{x.yz}^2)}$$

V tomto statistickém testu zjišťujeme, zda je hodnota F větší než kritická mez F-rozdělení se stupni volnosti 2 a $n - 3$. (V kapitole o mnohonásobné lineární regresní analýze se budeme tímto problémem zabývat podrobněji.)

PŘÍKLAD 7.6

Výpočet mnohonásobného korelačního koefiecientu

Výzkum vycházel ze zkušenosti sportovní praxe, že osvojení motorické dovedností závisí komplexně na různých znacích jedince. Na závěr základního lyžařského kurzu pro šestnáctileté účastníky se změřil čas ve slalomu u 36 dívek. Také se u nich zjišťovaly další charakteristiky. V tabulce 7.9 uvádíme korelace dosaženého času ve slalomu a dvou vybraných parametrů z této studie, abychom mohli spočítat, jak silně dosažený čas na těchto parametrech závisí.

Mnohonásobný korelační koeficient mezi dosaženým časem ve slalomu jako cílovou proměnnou a prediktory Y a Z má hodnotu:

$$r_{X.ZY} = \sqrt{\frac{0{,}34^2 + 0{,}46^2 - 2(-0{,}34)(0{,}46)(0{,}45)}{1 - 0{,}45^2}} = 0{,}77$$

Tab. 7.9 Korelační matice pro tři proměnné charakterizující skupinu účastnic lyžařského kurzu

	X	Y	Z
Čas ve slalomu (X)	1,00	−0,34	0,46
Test rovnováhy (Y)	−0,34	1,00	0,45
Test sociální úzkosti (Z)	0,46	0,45	1,00

7.2.6 Spearmanův korelační koeficient pořadí

Anglický psycholog Charles Edward Spearman (1863–1945) navrhl svůj koeficient korelace tak, že koreloval postupem podle Pearsona *pořadí* jednotlivých měření obou proměnných. Význam tohoto kroku spočívá v tom, že jeho koeficient zachycuje monotónní vztahy (ne pouze lineární, ale obecně rostoucí nebo klesající); je rezistentní vůči odlehlým hodnotám.

Spearmanovým korelačním koeficientem, jehož teoretickou hodnotu značíme ρ_s, měříme sílu vztahu X a Y, když nemůžeme předpokládat linearitu očekávaného vztahu nebo normální rozdělení proměnných X a Y. Závislost proměnných může mít obecně vzestupný nebo sestupný charakter. Jestliže $r_s = 1$, resp. $r_s = -1$, párové hodnoty (x_i, y_i) leží na nějaké vzestupné, resp. klesající funkci. Hodnoty r_s nemění jakákoli vzestupná transformace původních dat. Pro malé rozsahy je jeho výpočet méně pracný než výpočet Pearsonova korelačního koeficientu.

Odhadem ρ_s, je výběrový koeficient korelace r_s $(-1 \le r_s \le 1)$, který pro daný výběr (x_i, y_i) spočteme podle vzorce

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2 - 1)},$$

kde D_i jsou rozdíly pořadí R_x a R_y hodnot x_i a y_i vzhledem k ostatním hodnotám seřazeného výběru podle velikosti. Před výpočtem je nutno oběma řadám čísel

x_i a y_i tato pořadí přiřadit. Jestliže dvě čísla v řadě hodnot x_i, resp. y_i jsou stejná, přiřadíme jim průměrnou hodnotu příslušných pořadí. Obdobně provedeme tuto úpravu pro více stejných hodnot. V každé řadě nesmí být více než 1/5 pozorování stejných. Pokud se tak stane, musíme celý výpočet upravit.

PŘÍKLAD 7.7

Výpočet Spearmanova korelačního koeficientu

Výpočet r_s si ukážeme pro hodnoty z tabulky 7.10:

$$r_s = 1 - \frac{6 \times 26}{10(100 - 1)} = 0{,}84$$

Tab. 7.10 Příklad postupu při výpočtu Spearmanova korelačního koeficientu pořadí

x	y	R_x	R_y	$D = R_x - R_y$	$D \times D$
187	72	10,00	6,50	3,50	12,25
170	60	1,00	1,00	0,00	0,00
180	73	6,50	8,00	−1,50	2,25
184	74	8,00	9,00	−1,00	1,00
178	72	5,00	6,50	−1,50	2,25
180	70	6,50	4,50	2,00	4,00
172	62	2,00	2,00	0,00	0,00
176	70	3,00	4,50	−1,50	2,25
186	80	9,00	10,00	−1,00	1,00
177	67	4,00	3,00	1,00	1,00
Součet					26,00

Pro posouzení statistické významnosti koeficientu r_s slouží tabulka X z přílohy B. Přesahuje-li hodnota $|r_s|$ tabulkovou hodnotu pro daný počet párů měření n a hladinu významnosti, můžeme vztah považovat za prokázaný. Pro náš příklad, testujeme-li dvoustrannou hypotézu $\rho_s = 0$ na hladině 1 %, je tabulková hodnota 0,746 (tabulka obsahuje kritické hodnoty pro dvoustranné testy). Vztah mezi oběma proměnnými z příkladu je tedy prokázán. U větších výběrů ($n \geq 30$) lze na hladině α použít přibližný z-test hypotézy $\rho_s = 0$:

$$z = |r_s| \sqrt{n - 1}$$

Spearmanův koeficient r_s někdy používáme pro odhad Pearsonova korelačního koeficientu, resp. r, jelikož pro dvojrozměrně normálně rozdělené proměnné X a Y platí přibližný vztah $\rho = 2\sin(0{,}523\rho_s)$. Tento vzorec je upřesněním přibližně platného vztahu $\rho = \rho_s$. Podle Spearmana lze jeho koeficient korelace s výhodou uplatnit v situacích, kdy:

- potřebujeme rychlý a rezistentní odhad korelačního koeficientu r;
- testujeme schopnost zkoumané osoby správně řadit objekty nebo vlastnosti podle určitých hledisek tak, že ji necháme seřadit tyto objekty nebo vlastnosti a toto seřazení pak srovnáme se standardem;
- testujeme možnost přítomnosti monotónního trendu v časové řadě měření.

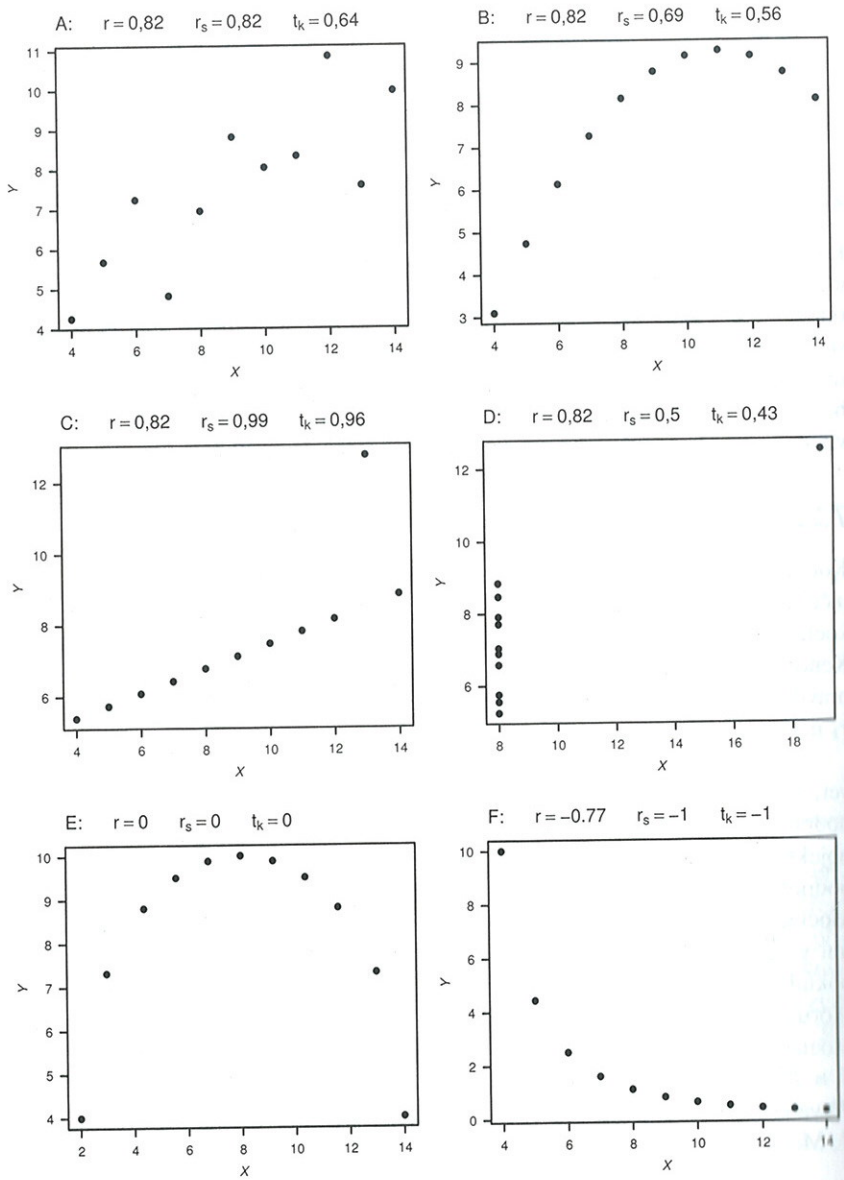
Pro usnadnění interpretace jsou na obrázku 7.5 znázorněna data z příkladu 7.3 (s. 246, množina 1 = A, 2 = B, 3 = C, 4 = D) a uvedeny k nim vypočtené korelační koeficienty podle Pearsona, Spearmana a Kendalla, aby bylo umožněno srovnání chování těchto koeficientů (viz odstavec o Pearsonově koeficientu). Obrázek ukazuje, jak zachytí Spearmanův koeficient vztah reprezentovaný různými bodovými konfiguracemi. Graf F dokumentuje jeho schopnost měřit monotónní vztahy, graf C ukazuje jeho rezistenci vůči odlehlým hodnotám.

7.2.7 Kendallův koeficient pořadové korelace

Korelační koeficient má měřit „sílu vztahu" dvou proměnných. Ale různé korelační koeficienty ho měří různým způsobem. Pearsonův i Spearmanův korelační koeficient mohou mít hodnotu 0,3, ale pokaždé to znamená něco trochu jiného. Kendallův korelační koeficient má na rozdíl od předchozích dvou jednoduchou pravděpodobnostní interpretaci. Jeho teoretickou hodnotu v populaci označujeme τ_k nebo Kendallovo *tau*.

Zatímco Spearman koreloval pořadí, Kendall založil svoji statistiku na inverzích v pořadí. Vycházíme z dat, která se týkají metrického nebo ordinálního hodnocení n objektů ($i = 1, 2, \ldots, n$) podle dvou kritérií X a Y. Ke každému objektu i získáme ohodnocení (x_i, y_i). Nejdříve seřadíme dvojice (x_i, y_i) tak, že hodnoty x_i budou tvořit rostoucí posloupnost. Jestliže mezi kritérii X a Y je kladná asociace, pak také y_i budou mít vzestupnou tendenci. Při záporné asociaci budou mít y_i sestupnou tendenci. Kendall proto rozlišuje vztah $y_j > y_i$, resp. $y_j < y_i$, pokud $j > i$ ($i = 1, 2, \ldots, n-1$). V prvním případě nastává tzv. **konkordance**, jež skóruje pro kladnou asociaci, ve druhém **diskordance**, která skóruje pro negativní asociaci. Počet všech konkordancí, resp. diskordancí označíme P, resp. Q. Rozdíl $S = P - Q$ někdy nazýváme Kendallovo S a je jednoduchou mírou závislosti. Převaha konkordancí, resp. diskordancí vede ke kladné, resp. záporné hodnotě S. Možná škála hodnot S závisí na rozsahu výběru n. Jednoduchá úprava však

Obr. 7.5 Zobrazení různých bodových konfigurací a k nim dopočítaného Pearsonova (r), Spearmanova (r_s) a Kendallova (t_k) korelačního koeficientu



tento problém vyřeší. S se totiž může pohybovat mezi hodnotami $-0{,}5n(n-1)$ a $0{,}5n(n-1)$. Proto se Kendallův koeficient *tau* t_k počítá podle formule

$$t_k = \frac{S}{D} = \frac{P-Q}{D},$$

kde jmenovatel D je maximální možný počet konkordancí, resp. diskordancí a má hodnotu $n(n-1)/2$.

PŘÍKLAD 7.8

Výpočet konkordancí a Kendallova koeficientu pořadové korelace

Vypočítáme počet diskordancí a konkordancí pro data v tabulce 7.11. Protože počty P a Q jsou přibližně stejné, mezi proměnnou X a Y není pravděpodobně žádná asociace. S má hodnotu −2.

Kendallův koeficient $t_k = -2/36 = -0{,}05$.

Tab. 7.11 Příklad výpočtu Kendallova koeficientu pořadové korelace

Věk (X)	Cholesterol (Y)	Konkordance	Diskordance
41	274	1	7
45	209	4	3
50	194	5	1
51	270	1	4
54	165	4	0
59	234	2	1
62	281	0	2
68	238	0	1
71	208	0	0
Součet		$P = 17$	$Q = 19$

Platí $-1 \leq t_k \leq 1$ a hodnot právě ± 1 nabývá t_k ve stejných situacích jako Spearmanův koeficient. Kritické hodnoty pro rozhodování, kdy je možné zamítnout hypotézu nezávislosti X a Y (H_0: $\tau_k = 0$), nalezneme pomocí speciálních tabulek. Některé programy dokáží spočítat přesnou p-hodnotu pro test nulové hodnoty τ_k. Pro velká n má t_k přibližně normální rozdělení se střední hodnotou 0 a směrodatnou odchylkou s_τ

$$s_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}},$$

pokud proměnné X a Y jsou nezávislé. Rozhodování o nulové hodnotě τ_k vychází z testovací z-statistiky $z = t_k/s_\tau$, kterou porovnáváme s kritickými hodnotami standardizovaného normálního rozdělení.

Interpretace τ_k je přímočařejší než u Spearmanova koeficientu ρ_s. Jestliže $\tau_k = p$, můžeme u dvou náhodně vybraných jedinců očekávat s pravděpodobností p, že jejich seřazení podle kritéria X bude stejné jako seřazení podle kritéria Y. Většinou oba koeficienty mají přibližně stejnou velikost.

V kapitole 8.4 poznáme využití Kendallova korelačního koeficientu při hodnocení závislosti v kontingenčních tabulkách, jež vznikly klasifikací objektů podle dvou ordinálních znaků.

Jestliže v údajích existují shody ($x_j = x_i$, resp. $y_j = y_i$), musíme výpočet modifikovat, protože v tomto případě nemůže koeficient dosáhnout hodnoty -1, resp. 1. Modifikaci uplatňujeme při větším počtu shod a týká se jmenovatele D ve vzorci pro výpočet Kendallova *tau*. Označme symboly u, resp. v počty shodných pořadí mezi x_i, resp. y_i postupně v jednotlivých skupinách shodných pořadí a symboly U a V součty, které mají tvar:

$$U = 0{,}5 \sum u(u-1),$$
$$V = 0{,}5 \sum v(v-1)$$

Modifikace výpočtu spočívá v nahrazení D číslem $D' = \sqrt{(D-U)(D-V)}$. Takto modifikovaný výpočet Kendallova *tau* nazýváme **Kendallovo *tau-b***, značíme t_b. Kendallovo t_b lze interpretovat jako korelaci mezi hodnotami dx a dy, kde dx se rovná 1, resp. -1, pokud pro $j > i$ je $x_j > x_i$, resp. $x_j < x_i$, a nule v ostatních případech. Hodnoty dy počítáme obdobně. Jak hodnoty dx, tak hodnoty dy spočítáme pro všechny možná srovnání, kterých je $n(n-1)/2$. (Zvára, 2000)

7.2.8 Bodově biseriální korelační koeficient a koeficient ϕ

Vztah mezi spojitou metrickou proměnnou a binární proměnnou se měří biseriálním korelačním koeficientem r_{pb} tak, že n dvojic měření se rozdělí na dvě skupiny podle hodnoty alternativního parametru a spočte se hodnota r_{pb} podle vzorce

$$r_{pb} = \frac{(\bar{x}_1 - \bar{x}_2)}{s}\sqrt{\frac{n_1 n_2}{n(n-1)}},$$

kde n_i, resp. $\bar{x}_i$ jsou počty, resp. průměrná hodnota spojitého parametru v obou skupinách a s je společná směrodatná odchylka. Tento koeficient r_{pb} testujeme podobně jako normální korelační koeficient. Jestliže $r_{pb} > 1$, resp. $r_{bk} < -1$,

dosadíme za něj hodnotu 1, resp. -1. Uvedený vzorec se v praxi nepoužívá, protože stejnou hodnotu dostaneme použitím algoritmu pro Pearsonův koeficient korelace pro dvojice hodnot obou proměnných, přičemž binární proměnnou zastupují nuly a jedničky. Jestliže binární proměnná vznikla dichotomizací spojité normálně rozdělené proměnné, můžeme spočítat odhad Pearsonova korelačního koeficientu obou spojitých proměnných pomocí tzv. biseriálního korelačního koeficientu (viz Howell 1992, s. 270).

Koeficient ϕ je Pearsonův korelační koeficient vypočítaný pro dvě alternativní proměnné, které kódujeme pomocí hodnot 0 a 1. (Existuje i jednodušší výpočet, ale ten nemá v době počítačů opodstatnění.) Platí, že $\phi^2 = \chi^2/n$, kde χ^2 je testovací statistika nezávislosti v čtyřpolní tabulce a n je počet dvojic, z nichž se počítá korelační koeficient. Test nulové hodnoty koeficientu ϕ se provádí stejně jako test nezávislosti pro čtyřpolní tabulku, která je tvořena četnostmi kombinací hodnot obou proměnných (viz kap. 8.3.1).

7.2.9 Korelační koeficient v klasickém modelu teorie měření

V úvodní kapitole jsme popsali základní způsoby hodnocení kvality měřicích instrumentů. K hlavním konceptům hodnocení kvality měření patří *reliabilita* (spolehlivost), *validita* a *objektivita*. Při praktickém uplatňování těchto konceptů hraje velkou roli Pearsonův koeficient korelace. Upozorníme na nejdůležitější situace jeho použití. Nepůjde však o ucelený výklad, jak postupovat při řešení úkolu evaluace instrumentu měření (viz kap. 13.8.3). Nejdříve uvedeme základní prvky klasické teorie měření (teorie testů), která má využití především v behaviorálních vědách.

Klasická teorie měření pojednává vztah mezi naměřenou hodnotou a hodnotou konstruktu, již považujeme za „skrytou", latentní proměnnou, pomocí matematicko-statistického modelu. Hodnota této latentní proměnné se považuje za „správnou hodnotu" T (*true score*), která je jenom nedostatečně zachycena instrumentem X (např. hodnota jednoho indikátoru nebo součet hodnot jednotlivých indikátorových položek z dotazníku). Přijímá se několik hypotéz. První je, že se napozorovaná hodnota X skládá aditivně ze správné hodnoty a z chyby měření. Model pro měření instrumentem X proto zapisujeme ve tvaru

$$X = T + E,$$

který vyjadřuje, že napozorovaná hodnota (skór) je součtem hypotetické „správné hodnoty" a chybové komponenty E. Chyba E může vzniknout mnoha způsoby, resp. také u ní lze identifikovat různé komponenty: např. technickou chybu vlastní

procedury a chybu závislou na externích podmínkách. Do chybové komponenty započítáváme někdy i průměrnou přirozenou intraindividuální variabilitu hodnoty T u měřených jedinců.

Klasický model měření vychází dále z těchto předpokladů (Blahuš, 1975; Řehák, 1998):

1. Chyba E nekoreluje s hodnotou T: $\rho_{ET} = 0$.
2. Chyba E neobsahuje systematické vychýlení: $\mu_E = 0$.
3. Při různých měřeních jsou chyby nekorelované: $\rho_{E_1E_2} = 0$.
4. Také nekoreluje správná hodnota jednoho měření (jedné položky testu) stejného konstruktu s chybou druhého měření (druhé položky testu) stejného konstruktu: $\rho_{T_1E_2} = 0$.

Symbol ρ označuje Pearsonův koeficient korelace.

Tímto způsobem je definován základní model teorie měření nebo testování. Uvedená formalizace umožňuje definovat kritéria kvality měřicího instrumentu a odvodit vztahy mezi reliabilitou a validitou. Předpokládá se, že uvedené postuláty jsou v praxi splněny.

Reliabilitu neboli konzistenci a opakovatelnost měření zachycujeme obecně dvěma způsoby: relativně a absolutně. **Relativní reliabilita** se odhaduje relativními bezrozměrnými hodnotami. **Absolutní reliabilita** se udává přímo v jednotkách, v nichž se daná proměnná měří.

Všimneme si nejdříve relativního konceptu reliability. Koeficient reliability $Rel(X)$ pro měřicí metodu je definován poměrem $Var(T)/Var(X)$, kde $Var()$ označuje jako obvykle teoretický rozptyl náhodné proměnné. Uvedená definice říká, že:

$$reliabilita = \frac{rozptyl\ pravdivého\ skóru}{rozptyl\ pravdivého\ skóru + chybový\ rozptyl}$$

Tento vztah lze také přepsat do tvaru

$$Rel(X) = (Var(X) - Var(E)) \ / \ Var(X) = \rho^2_{TX}\,.$$

Lze ukázat, že když provedeme u zvolených jedinců X_1 a X_2 nezávislá měření latentní proměnné T se shodnou hodnotou $Var(E)$, jejich korelace se rovná $Rel(X)$. Tento důležitý vztah se využívá ve třech přístupech k odhadu reliability:

Test-retest reliabilita. V tomto přístupu odhadujeme $Rel(X)$ Pearsonovým korelačním koeficientem dvou měření n objektů danou metodou ve dvou časových okamžicích.

Reliabilita paralelních měření. Provedeme měření n objektů dvěma nezávislými metodami. O metodách předpokládáme, že mají stejnou reliabilitu, a $Rel(X)$ odhadujeme korelačním koeficientem získaných dvou řad měření.

Reliabilita zjištěná půlením testu. Tuto metodu použijeme, jestliže měření X získáváme jako součet parciálních hodnot, které např. obdržíme jako odpovědi na různé položky dotazníku, jež měří stejnou charakteristiku. Počítáme korelační koeficient $r_{1/2}$ mezi dvěma polovinami položek dotazníku zadaného n osobám. $Rel(X)$ se pak spočte podle vzorce $Rel(X) = 2r_{1/2}/(1 + r_{1/2})$.

Korelace položky s celkovým skórem. Při korelování každé položky testu, resp. navrhované škály s celkovým skórem (hodnotou testu) dostáváme další míru internální konzistence ukazující, jak každá položka souhlasí se sumou odpovědí na ostatní položky. Ze sumy vyloučíme hodnotu pro hodnocenou položku.

O koeficientech pro hodnocení vnitrotřídní korelace viz kapitola 9.3.2.

Popsali jsme různé přístupy, které hodnotí **relativní reliabilitu** neboli relativní konzistenci měření. **Absolutní reliabilita** je koncept, jímž se posuzují změny hodnot na jejich škále. Jinak řečeno, tento typ konzistence vyjadřuje velikost variability, která se očekává u naměřené hodnoty. Její určení vychází ze směrodatné chyby měření s, již lze odhadnout číslem $s = s_x\sqrt{[1 - Rel(X)]}$, kde s_x označuje rozptýlenost dat ve skupině. Hodnota $3s$ udává tzv. **kritickou diferenci**. Její význam je následující: jestliže máme dvě měření x_1 a x_2 u stejné osoby, pak pouze v 5 % případů bude jejich rozdíl $(x_1 - x_2)$ v absolutní hodnotě větší než $3s$, pokud mezi měřeními a při měření nedošlo k žádné změně. Kritickou mez diference lze aplikovat i na měření od dvou osob. Pokud jejich rozdíl je větší než kritická diference, můžeme tvrdit, že správné hodnoty obou osob se skutečně liší. Absolutní reliabilitou (konzistencí) se zabýváme také v kapitole 7.3.7.

■ ■ ■

Také **validitu** měřicí metody hodnotíme pomocí korelačního koeficientu. Z různých typů validity měřicího procesu máme na mysli *kriteriální* validitu, která zahrnuje *prediktivní* a *souběžnou* validitu. Hlavním rozdílem mezi oběma validitami je čas provedení měření. Při hodnocení kriteriální validity korelujeme hodnoty posuzovaného měření s hodnotami měření standardem. Při hodnocení prediktivní validity korelujeme hodnoty testových výsledků s kriteriálními hodnotami získanými po uplynutí určité doby a odhadujeme tak schopnost predikovat tyto hodnoty hodnoceným testem.

Objektivitu měřicího prostředku někdy hodnotíme tak, že korelujeme výsledky vyhodnocení dvěma hodnotiteli. Tím dostáváme relativní míru objektivity. Dnes je tendence používat spíše absolutní míry shody, jako je *kappa koeficient* (viz kap. 8.3.4). V tomto směru se uplatňuje také přístup podle Blanda a Altmana, který popisujeme v kapitole 7.3.7.

V teorii testů se **obtížnost položky testu** obvykle definuje jako podíl jedinců, kteří položku správně zodpoví. **Diskriminační síla položky testu** se odhaduje jako korelace mezi celkovým skóre a hodnotou dané položky. Čím větší je tato

Tab. 7.12 Využití Pearsonova koeficientu korelace při hodnocení metod měření

Korelační koeficient r_{xy}		
X	Y	Aplikace/interpretace
měření v čase I	měření v čase II	odhad reliability
první polovina testu	druhá polovina testu	odhad reliability
paralelní forma testu I	paralelní forma testu II	odhad reliability
hodnocený test	cílové kritérium	souběžná validita
hodnocený test	měření kritéria v budoucnu	prediktivní validita
hodnotitel I	hodnotitel II	odhad objektivity

korelace, tím je zřejmě hodnocená položka více konzistentní s celým testem. Pokud má položka jenom dvě hodnoty, počítáme pomocí Pearsonovy formule bodově biseriální korelační koeficient.

■ ■ ■

Podaný stručný přehled využití korelačního koeficientu ukazuje, jak významnou roli hraje tato charakteristika v oblasti, která byla ovlivněna klasickou teorií testů. V biomedicínských vědách má Pearsonův korelační koeficient mnoho kritiků. V kapitole 7.3.7 o alternativních způsobech prokládání přímky zmíníme některé nedostatky Pearsonova korelačního koeficientu, hrající roli v kontextu hodnocení měřicích metod. Tabulka 7.12 zachycuje přehled uplatnění Pearsonova korelačního koeficientu při posuzování kvality metod měření v behaviorálních vědách. Pro oblast sportu popisuje Blahuš (1975) podrobně teorii aplikace korelace při návrhů testů pohybové výkonnosti.

7.3 Regresní analýza

V této kapitole se převážně zabýváme metodami, jež nám umožňují zkoumat vztahy mezi dvěma proměnnými. Dosud jsme mluvili hlavně o tom, jak vyšetřovat existenci vztahu a popsat sílu asociace pomocí měr závislosti – korelačních koeficientů. V regresní analýze nám půjde o to přesněji popsat tvar vztahu mezi proměnnými X a Y a charakterizovat jeho vhodnost pro predikci hodnot závisle proměnné pomocí hodnot nezávisle proměnné. Může jít např. o následující situace:

- Korelační koeficient i graf prokazují lineární vztah mezi spotřebou zemního plynu v bytě v závislosti na venkovní teplotě. Otázka zní, jak přesně můžeme predikovat spotřebu pomocí teploty.

- V medicíně měříme fyziologické parametry různými metodami. Některé jsou dokonalejší a dražší, jiné jsou méně přesné, ale levnější. Často se řeší problém, zda je možné nahradit dražší metodu levnější měřením, jak přesně to dokážeme a kolik informace ztratíme.
- Ve sportovním výzkumu máme např. data o rychlosti skokanů na hraně můstku a dosažené délce skoku. Zajímá nás, jaký je mezi nimi vztah: lze pomocí rychlosti predikovat délku skoku, s jakou přesností, je vztah lineární?

Podobně lze analyzovat vztahy mezi dvojicemi proměnných: testové skóre z matematiky v přípravě, testové skóre při závěrečné zkoušce; množství hnojiva, velikost úrody; velikost domu, cena domu apod.

PŘÍKLAD 7.9

Úloha regresní analýzy

Dobří běžci zrychlují frekvenci kroků podle toho, jakou běží rychlostí. V tabulce 7.13 uvádíme průměrné počty kroků za vteřinu pro skupinu vrcholových sprinterů při různých rychlostech. Jestliže u vrcholového běžce chceme predikovat jeho rychlost pomocí krokové frekvence, zkontrolujeme linearitu vztahu a proložíme body přímku. Pomocí její rovnice pak jednoduše vypočítáme pro každou hodnotu krokové frekvence příslušnou rychlost.

Tab. 7.13 Příklad regrese – průměrná frekvence kroků skupiny sprinterů

Rychlost [m/s]	5,24	5,67	5,82	6,03	6,84	7,09	7,55
Počet kroků za vteřinu	3,05	3,12	3,17	3,25	3,36	3,46	3,55

V regresní analýze obecně analyzujeme vztah mezi jednou proměnnou zvanou cílová nebo závislá proměnná, a několika dalšími, které nazýváme nezávislé nebo ovlivňující proměnné. **Cílovou proměnnou** nazýváme někdy **regresand** (označujeme ji symbolem Y), nezávislou **regresor** (označujeme ji X). Vztah reprezentujeme matematickým modelem, což je rovnice, jež svazuje regresand s regresorem a pravděpodobnostní předpoklady, které by měl vztah splňovat. Závisle proměnná je spojena s nezávisle proměnnými funkcí nazývanou **regresní funkce**, jež obsahuje několik neznámých parametrů. Říkáme, že provádíme regresi závislé proměnné na nezávislých proměnných. Jestliže tato funkce je lineární v těchto parametrech (nemusí být lineární v proměnných), mluvíme o **lineárním regresním modelu**. Jinak nazýváme model nelineární. Statistické problémy, které nás zajímají v regresní analýze, jsou:

a) získání statistických odhadů neznámých parametrů regresní funkce;
b) testování hypotéz o těchto parametrech;
c) ověřování předpokladů regresního modelu.

Budeme se zabývat jednoduchou lineární regresní analýzou, kdy dvě proměnné jsou spolu spojeny lineární funkcí.

7.3.1 Prokládání dat přímkou a metoda nejmenších čtverců

Vycházíme z datového materiálu v podobě uspořádaných dvojic číselných údajů $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$ pro proměnné X a Y. Jestliže graf ukáže lineární vztah mezi proměnnými, usilujeme o zachycení vztahu tím, že body proložíme přímku. Různí lidé však mohou „podle oka" proložit stejnými body dosti rozdílné přímky. Stane se to zvláště tehdy, jestliže je korelace mezi proměnnými menší. Žádná přímka pak neprotne všechny body. Hledáme tedy přímku, jež je experimentálním bodům co možná nejblíže. Snažíme se určit takovou přímku, která bude co nejlépe predikovat y-hodnoty pomocí x-hodnot. Z tohoto požadavku plyne, že přímka by měla být nejblíže k bodům ve vertikálním směru, protože chyby predikce se týkají y-hodnot.

PŘÍKLAD 7.10

Proložení dat regresní přímkou

Při predikci hmotnosti pomocí výšky můžeme zjistit rozdíl mezi predikcí a naměřenou hodnotou, pokud jsme znali u jedince jak jeho výšku, tak hmotnost. Přímka z obrázku 7.1 (s. 239) je dána rovnicí

$$\textit{hmotnost}\,[\text{kg}] = 0{,}912 \times \textit{výška}\,[\text{cm}] - 93{,}24.$$

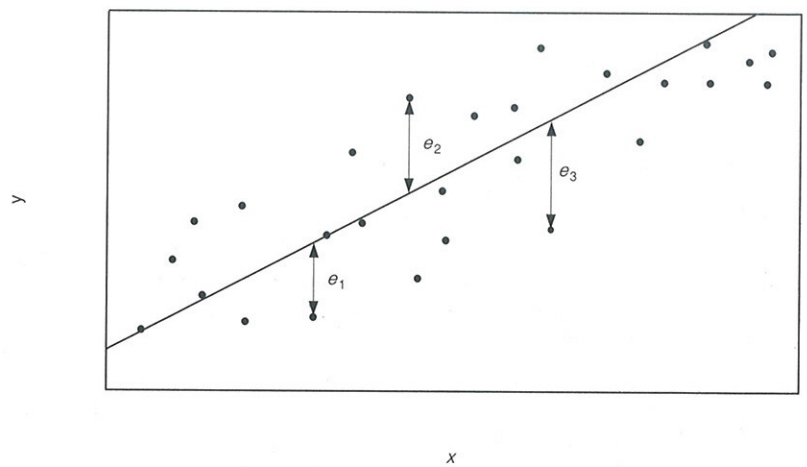
Pomocí této rovnice vypočítáme predikci hmotnosti jedince, který je např. 170 cm vysoký:

$$\textit{hmotnost}\,[\text{kg}] = 0{,}912 \times 170 - 93{,}24 = 61{,}79.$$

U jednoho studenta ze souboru byly naměřeny údaje (*váha* = 60; *výška* = 170), což znamená, že při predikci jeho váhy uvedenou funkcí jsme se dopustili chyby −1,79 = 60 − 61,79.

Rozdílu mezi naměřenou a predikovanou hodnotu říkáme **reziduální hodnota predikce** nebo chyba predikce a značíme ji symbolem e. Dobře proložená přímka $y = a + bx$ minimalizuje velikosti reziduálních hodnot pro hodnoty $\{(x_i; y_i)\}$, kterými přímku prokládáme. Je mnoho způsobů, jak to provést. Nejčastěji se používá **metoda nejmenších čtverců**.

Obr. 7.6 Vzdálenosti bodů od regresní přímky



Hodnoty parametrů a, b přímky $y = a + bx$ získáme metodou nejmenších čtverců tak, aby byl minimální součet druhých mocnin reziduálních hodnot

$$s_r^2 = \sum e_i^2 = \sum (y_i - a - bx_i)^2$$

vzhledem k parametrům a, b. Grafickou interpretaci odchylek od regresní přímky pro metodu nejmenších čtverců ukazuje obrázek 7.6. Minimalizujeme sečtené čtverce úseček, které vyznačují vzdálenost bodu od proložené přímky ve směru osy Y. Výpočet tohoto minima vede k optimálním hodnotám

$$b = r\frac{s_y}{s_x}, \quad a = \bar{y} - b\bar{x},$$

kde r je korelace obou proměnných a s_x, s_y jsou směrodatné odchylky naměřených hodnot proměnných X a Y. Nalezenou přímku nazýváme regresní přímka. Hodnota $\hat{y}_i$ je odhad cílové proměnné pomocí regresního vztahu ($\hat{y}_i = a + bx_i$):

reziduální hodnota = naměřená hodnota y − predikovaná hodnota $\hat{y}$

Rozptýlenost bodů kolem přímky je charakterizována **zbytkovým (reziduálním) rozptylem** $s_{y.x}^2$, případně směrodatnou chybou odhadu při regresi $s_{y.x}$:

$$s_{y.x}^2 = \frac{s_r^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

Poznamenejme, že ve jmenovateli je číslo $n - 2$, což odpovídá stupňům volnosti reziduálních hodnot. Jestliže známe $n - 2$ reziduálních hodnot, pak při zadaných hodnotách regresní přímky jsou zbývající dvě reziduální hodnoty plně určeny.

Někdy zjišťujeme regresi také v obráceném směru. Vztahy pro regresi X na Y získáme vhodnou záměnou ve vzorcích (např. $b_x = r\, s_x / s_y$, kde r je korelační koeficient). Mezi směrnicemi obou regresních přímek b_x a b_y existuje vztah $r = \sqrt{b_x b_y}$. Můžeme tedy nalézt dvě regresní přímky $y = a_y + b_y x$ a $x = a_x + b_x y$. Obě se budou protínat v bodě $(\bar{x}, \bar{y})$ a tvoří jakési nůžky. Čím větší je korelace, tím více jsou nůžky stisknuty. Proměnné na levé straně obou rovnic považujeme za závisle proměnné (predikované). Korelační koeficient vystupuje i v následujících dvou přibližných vztazích (jež se mění v rovnost s rostoucím n):

$$r^2 \doteq \frac{s_y^2 - s_{y.x}^2}{s_y^2}$$

$$r^2 \doteq \frac{s_x^2 - s_{x.y}^2}{s_x^2}$$

Matematicky vyjadřují tvrzení, že číslo $100r^2$ udává v procentech tu část celkové variability proměnné Y, resp. X, která je vysvětlena znalostí hodnoty nezávislé proměnné X, resp. Y. Tuto hodnotu nazýváme **koeficient determinace**. Koeficient determinace je poměr vysvětlené variability k celkové variabilitě proměnné Y:

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

Jestliže jsme proměnné standardizovali pomocí jejich průměrů a směrodatných odchylek na hodnoty x' a y', výpočet regresní přímky mezi standardizovanými hodnotami se překvapivě zjednoduší. Regresní přímka pak totiž prochází počátkem a regresní koeficient se rovná korelačnímu koeficientu: $y' = rx'$ s chybou odhadu $s_{x.y} = \sqrt{1 - r^2}$. Koeficient $s_{y.x}$ je mírou chyby, které se dopouštíme při odhadu nebo předpovědi závisle proměnné Y pomocí X.

PŘÍKLAD 7.11

Příklad proložení regresní přímky

Pro tabulku hodnot (x, y) z příkladu 7.2 o korelačním koeficientu mezi výškou a hmotností studentů získáme výpočtem regresní přímku ve tvaru $y = a + bx$ a reziduální směrodatnou odchylku $s_{y.x}$:

$$b = 0{,}878 \times 5{,}83/5{,}83 = 0{,}912$$

$$a = 10 - 0{,}912 \times 179 = -93{,}24$$

$$s_{y.x} = 5{,}83 \times \sqrt{1 - 0{,}87^2} = 2{,}87$$

Tato přímka je znázorněna na obrázku 7.1 (s. 239).

Koeficient determinace má hodnotu $100 \times 0{,}87^2 = 75{,}7\%$. Znamená to, že nezávisle proměnná je schopna vysvětlit skoro 76 % variability závisle proměnné. Predikční rovnice pro váhu pomocí výšky má tvar přímky: *váha* = 0,912 × *výška* − 93,2.

7.3.2 Grafická analýza reziduálních hodnot

Regresní přímka jednoduchým způsobem vystihuje vztah mezi závisle a nezávisle proměnnou. Odchylky od tohoto vztahu jsou také důležité. Například se zajímáme o hodnoty, které jsou v dané konfiguraci dat neobvyklé, nebo se chceme přesvědčit, zda je vztah opravdu lineární. K tomu nám slouží především grafická analýza dvojrozměrného bodového grafu údajů $\{(x_i; y_i)\}$. Také analýza reziduálních hodnot nám pomáhá ověřit kvalitu proložení dat přímkou a odhalit neobvyklé hodnoty. Sestrojujeme histogram reziduálních hodnot e_i nebo dvojrozměrný bodový graf, jenž zachycuje jejich vztah k hodnotám nezávisle proměnné X. Také si všímáme ovlivnění velikosti reziduálních hodnot jinými proměnnými, např. časovým okamžikem, v němž byly hodnoty získány (viz obr. 7.7).

7.3.3 Statistické usuzování v lineárním regresním modelu

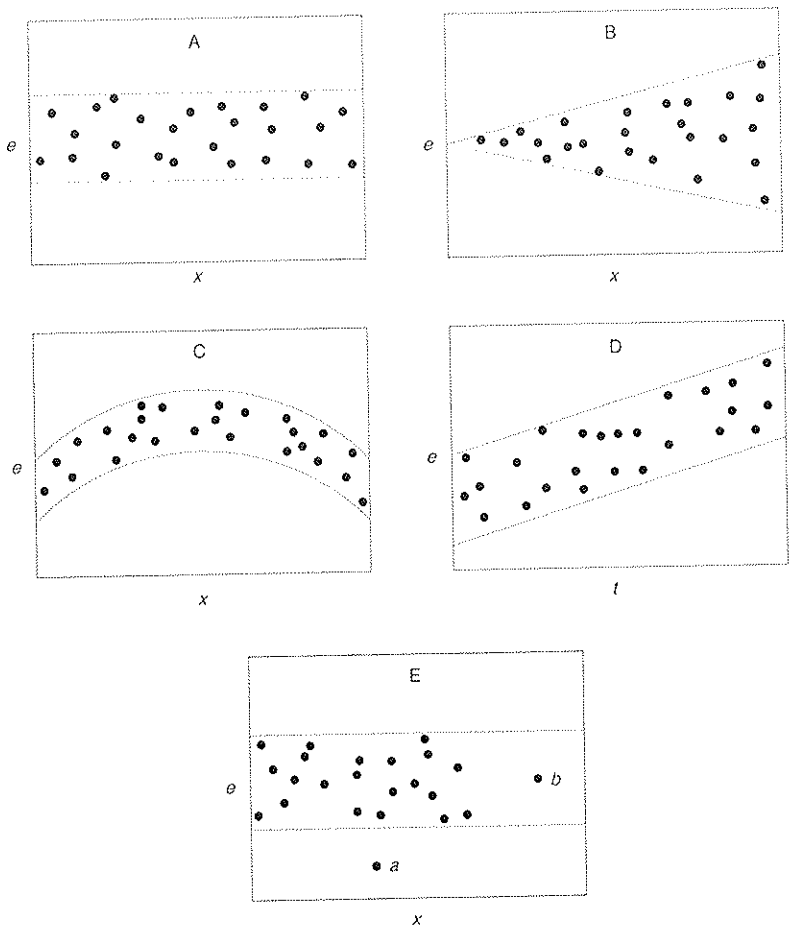
Předpokládáme, že máme k dispozici výběr párových hodnot $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ změřených na statistických jednotkách z populace W. Jestliže jsme ověřili graficky lineární vztah a proložili body regresní přímku, chceme data dále zkoumat v rámci statistického modelu, který zachycuje úsporným způsobem jejich proměnlivost. Při statistickém modelování závislosti závisle proměnné Y na nezávisle proměnné X vycházíme v jednoduché lineární regresní analýze z předpokladu, že pro naměřené údaje platí rovnice

$$y = \alpha + \beta x + e,$$

kde chybová hodnota e reprezentuje náhodnou proměnnou s nulovou střední hodnotou a směrodatnou odchylkou $\sigma_{y.x}$ stejnou pro všechny hodnoty proměnné X. O hodnotách chyby e, které představují chybovou složku predikce proměnné Y pomocí regresní rovnice, navíc předpokládáme, že mají normální rozdělení a jsou na sobě nezávislé. Vztah mezi proměnnými musí splňovat podmínku

$$E(Y \mid X = x) = \alpha + \beta x,$$

Obr. 7.7 Schematické příklady grafů reziduálních hodnot regrese



Na obrázku jsou různé konfigurace reziduálních hodnot e_i v závislosti na hodnotách x_i. Jestliže konfigurace bodů má tvar vyznačený na obrázku A, můžeme tvrdit, že předpoklad lineárního vztahu je dobře splněn. Na obrázku B je bodová konfigurace, jež indikuje, že rozptyl bodů kolem regresní přímky se zvyšuje s rostoucím X. Konfigurace na obrázku C naznačuje nutnost použití některé nelineární regresní křivky. Při zobrazení párových hodnot (t_i, e_i), kde t_i je časový okamžik i-tého měření, můžeme dostat konfiguraci, která je na obrázku D a která nás upozorňuje, že časový faktor by měl být součástí regresního modelu. Konfigurace E obsahuje dva neobvyklé body, jež mohou být klasifikovány jako odlehlá hodnota. **Odlehlý bod** je takový, který leží mimo základní konfiguraci bodů v grafu. Údaj může být odlehlý ve směru Y, ve směru X nebo v obou směrech. Odlehlý údaj ve směru nezávisle proměnné se nazývá vybočující. Bod nazýváme **vlivný**, pokud se po jeho odstranění podstatně změní poloha regresní přímky. Body, jež jsou odlehlé ve směru X, jsou často vlivné. Na obrázku je takovým bodem bod b.

což znamená, že podmíněná střední hodnota proměnné Y je lineární funkcí hodnot x. Parametry regresního modelu jsou regresní koeficient β, absolutní člen α a směrodatná chyba odhadu $\sigma_{y.x}$.

Regresní analýza vychází z jednosměrného konceptu: $x \to y$, kde x, resp. y je hodnota nezávislé, resp. závislé proměnné. Tento koncept vede k použití metody nejmenších čtverců při vyrovnávání dat. Hodnoty regresní funkce odpovídají průměrným hodnotám proměnné Y při zadaných hodnotách nezávislé proměnné. Hodnota teoretické regresní funkce je v určitém smyslu „dobrou" predikcí závisle proměnné.

Nyní si musíme uvědomit, že párové hodnoty, s nimiž pracujeme, jsme mohli získat v zásadě dvěma rozdílnými způsoby:

1. Hodnoty nezávisle proměnné x_i jsme sami zvolili. Na příslušných jednotkách se zvolenou hodnotou proměnné X jsme měřili hodnoty y_i proměnné Y. V této situaci je pouze Y náhodnou proměnnou.
2. Párové hodnoty (x_i, y_i) jsme zjišťovali na n jednotkách náhodně vybraných z populace W. Obě proměnné X a Y považujeme za náhodné proměnné.

Vzorce, které si dále uvedeme, jsou odvozeny za předpokladu první výběrové situace, ale platí také pro druhou výběrovou situaci.

Prvním krokem při statistické inferenci je nalezení odhadů parametrů regresní přímky $\alpha + \beta x$. Platí, že odhady a a b metodou nejmenších čtverců jsou dobrými odhady obou parametrů α a β.

Třetí parametr regresního modelu, směrodatná chyba odhadu při regresi $\sigma_{y.x}$, popisuje variabilitu proměnné Y okolo regresní přímky a velikost reziduálních chyb. Výběrový hodnota $s_{y.x}$ je dobrým odhadem hodnoty $\sigma_{y.x}$. Na statistice $s_{y.x}$ je založeno v regresní analýze několik metod statistického usuzování. Je např. složkou směrodatných chyb odhadů SE_a a SE_b parametrů α a β:

$$SE_a = \sqrt{\frac{s_{y.x}^2 \sum x_i^2}{s_x^2(n-1)n}}, \qquad SE_b = \sqrt{\frac{s_{y.x}^2}{s_x^2(n-1)}}.$$

Testy významnosti o parametrech regresní přímky α a β můžeme provést pomocí intervalů spolehlivosti $a \pm t_{n-2}SE_a$, $b \pm t_{n-2}SE_b$, kde t_{n-2} označuje kritickou hodnotu t-rozdělení s $n-2$ stupni volnosti pro zvolenou hladinu spolehlivosti. Testujeme např. hypotézu, že $\beta = 1$. Jestliže interval spolehlivosti pro hodnotu β neobsahuje hodnotu 1, můžeme tuto hypotézu zamítnout.

V obou vztazích je výraz s odmocninou hodnotou směrodatné chyby odhadu SE příslušného koeficientu (SE_a a SE_b). Tedy test významnosti nulové hypotézy $H_0: \beta = 0$ můžeme také provést pomocí statistiky $z = b/SE_b$, která má za platnosti nulové hypotézy asymptoticky standardizované normální rozdělení.

Poznamenejme, že hypotéza $\beta = 0$ je ekvivalentní hypotéze $\rho = 0$. Jestliže Pearsonův korelační koeficient mezi proměnnými X a Y má hodnotu blízkou nule, nemá cenu počítat příslušnou regresní přímku.

V souvislosti s predikcí pomocí regresní přímky rozlišujeme dva různé intervalové odhady.

Interval spolehlivosti pro hodnotu $y = \alpha + \beta x$ regresní přímky při zadané hodnotě x má tvar

$$\left(\hat{y} - t s_{y.x}\sqrt{Q_1}, \hat{y} + t s_{y.x}\sqrt{Q_1}\right),$$

kde

$$Q_1 = \frac{1}{n} + \frac{(x - \bar{x})}{s_x^2(n-1)}.$$

Číslo t je kritická hodnota Studentova t-rozdělení s $n - 2$ stupni volnosti určující $p\%$ interval spolehlivosti. Pro větší n se hodnota Q_1 blíží k nule.

Predikční interval pro budoucí pozorování y při zadané hodnotě x je

$$\left(\hat{y} - t s_{y.x}\sqrt{Q_2}, \quad \hat{y} + t s_{y.x}\sqrt{Q_2}\right),$$

kde $Q_2 = 1 + Q_1$ a t je kritická hodnota Studentova t-rozdělení s $n - 2$ stupni volnosti určující $p\%$ interval spolehlivosti. Zřejmě platí, že pro větší n jsou Q_2, resp. t přibližně rovny 1, resp. 2 (pokud $p = 95\%$). Pro lepší pochopení obou intervalů provedeme jejich srovnání.

Srovnání predikčního intervalu a intervalů spolehlivosti, jestliže nezávisle proměnná má hodnotu $x = x^*$:

- Oba intervaly jsou nejužší v místě $x^* = \bar{x}$.
- Interval spolehlivosti pro dané $x = x^*$ je vždy užší než odpovídající predikční interval.
- Predikční interval je určen pro individuální pozorování, kdežto interval spolehlivosti je určen pro hodnoty regresní přímky.
- S rostoucím n se zmenšuje šířka intervalu spolehlivosti i predikčního intervalu, ale u intervalu spolehlivosti se šířka blíží k nule, kdežto u predikčního intervalu k $2s_{y.x}$ (pro $p = 95\,\%$).
- Poloha x^* vůči $\bar{x}$ ovlivňuje šířku obou intervalů.

Přibližně platí (při větším rozsahu výběru), že dvě rovnoběžky k vypočítané regresní přímce vedené ve vzdálenosti $2s_{y.x}$ na každé straně (ve směru osy Y) tvoří pás, ve kterém je 95 % všech pozorování (platí i pro budoucí pozorování).

PŘÍKLAD 7.12

Statistické usuzování v lineární regresi

Na obrázku 7.8 uvádíme vybrané části výstupu počítačového programu typické pro jednoduchou regresní analýzu. V příkladu se snažíme najít predikční přímku pro výkon ve skoku dalekém pomocí výkonu v běhu na 75 m. Tabulku dat s podrobnějším popisem uvádíme v tabulce 2.9, s. 77.

Výtah z výpočtů obsahuje korelační koeficient (−0,75) a údaje, jež jsou nutné pro sestavení regresní přímky. Ty najdeme v prvním sloupci čísel. Regresní přímka má tvar *skok daleký* = 6,29 − 0,21 × (*běh* 75 m).

Testovací t-statistika ve třetím sloupci se vypočítá jako podíl hodnot v prvním a druhém sloupci. Čtvrtý numerický sloupec obsahuje pravděpodobnosti, které indikují, že je nepravděpodobná platnost hypotézy, že směrnice i průsečík s osou y mají hodnotu nula. To je potvrzeno v předposledním sloupci. Směrodatná chyba odhadu $s_{y\,x}$ vyjadřuje přesnost predikce regresní rovnicí. Jestliže chceme získat přibližnou mez, jež ohraničuje 95 % chyb, vynásobíme hodnotu 0,19 číslem 1,96. Výkon ve skoku dalekém tedy odhadujeme výkonem v běhu na 75 m pomocí regresní přímky s přibližnou chybou 38 cm.

Graf 7.9 ukazuje jednotlivé body a pásy spolehlivosti pro regresní přímku (vnitřní pás) a predikční pásy pro jednotlivé body (vnější pás). Predikční pás se vypočítá pomocí směrodatné chyby odhadu $s_{y\,x}$. Oba pásy jsou nejužší v oblasti průměru dat.

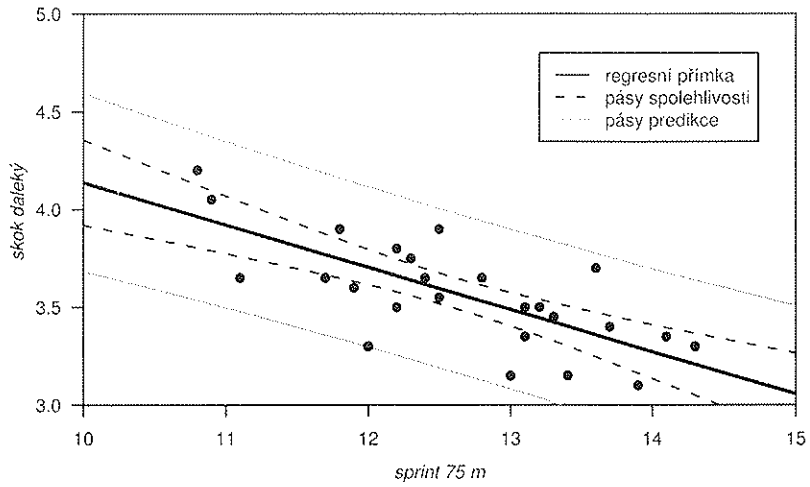
Obr. 7.8 Příklad výstupu pro úlohu regresní analýzy

```
Zpráva o mnohonásobné regresi
Závisle prom. SkokDaleký

Korelační matice
            Běh75m      SkokDaleký
Běh75m      1,000000    -0,753232
SkokDaleký  -0,753232   1,000000

Sekce pro regresní rovnici
Nezávislá   Regresní    Směr.          T-hodn.     Prav.      Rozhodnutí
proměnná    koeficient  chyba          (Ho: B=0)   hladina    (5,0%)
Průsečík    6,29631     0,4851871      12,9771     0,000000   Zamítni Ho
Běh75m      -0,2159598  3,849484E-02   -5,6101     0,000009   Zamítni Ho
R**2 0,567359
Směrodatná chyba při regresi 0,1933914
```

Obr. 7.9 Regresní závislost a pásy predikce a spolehlivosti



7.3.4 Ověřování předpokladů regresní analýzy

Oprávněnost použití modelu lineární regrese, prováděných statistických testů a predikce pomocí predikčního intervalu je podmíněna tím, že zkoumaný regresní vztah přibližně splňuje následující předpoklady:

1. Regresní vztah mezi proměnnými Y a X má lineární charakter.
2. Pro celý rozsah uvažovaných x je hodnota reziduální směrodatné odchylky $\sigma_{y.x}$ konstantní. Této vlastnosti říkáme **homoskedascita** standardních chyb odhadu při regresi. Znamená, že rozptýlenost bodů kolem regresní přímky je stejná pro všechny uvažované hodnoty proměnné X.
3. Hodnoty y_i mají normální rozdělení pro dané hodnoty přesně určených x_i a jsou na sobě nezávislé (stochasticky). Normální rozdělení je předpoklad, který zaručuje oprávněnost použití tabulek t-rozdělení nebo F-rozdělení. Předpoklad nezávislosti požaduje, aby jednotlivé body $(x_i; y_i)$ se zachycovaly na sobě nezávisle. Souvisí s uspořádáním sběru dat.

Malé odchylky od předpokladu homoskedascity a normality je možno tolerovat. Podmínky 1 a 2 implikují, že reziduální hodnoty $(y_i - \hat{y}_i)$ mají normální rozdělení s nulovou střední hodnotou a jsou na sobě stochasticky nezávislé. Ověřování těchto předpokladů můžeme provést mnoha způsoby:

- Grafickou analýzou rozdělení hodnot e_i lze odhalit vlastnosti studované závislosti, jež použitý model nebere v úvahu, a podle toho modifikovat další postup naší statistické analýzy. Její zásady jsme uvedli v kapitole 7.3.2.
- Předpoklad normality rozdělení hodnot e_i můžeme testovat testy dobré shody nebo pouhou optickou analýzou jejich histogramu. Celkově by hodnoty e_i měly mít nulovou střední hodnotu a rozptyl, který je odhadnutý hodnotou $s^2_{y.x}$. Jestliže absolutní hodnota e_i je větší než $3s_{y.x}$, snažíme se zjistit příčinu tak velké chyby predikce, případně dotyčné měření $(x_i; y_i)$ posuzujeme jako odlehlou hodnotu a regresní rovnici znovu přepočítáme bez této hodnoty.
- Jestliže zobrazení párových hodnot $(z_i; e_i)$ odhalí závislost regrese na třetí proměnné Z, můžeme si toto tvrzení ověřit vypočtením příslušného mnohonásobného korelačního koeficientu a testem významnosti zlepšení predikce F-testem. Testovací F-statistika má tvar

$$F = \frac{(r^2_{y.xz} - r^2_{yx})(n-3)}{(1 - r^2_{y.xz})},$$

přičemž F-testovací hodnotu srovnáme s kritickou mezi F-rozdělení se stupni volnosti 1 a $n - 3$. Podobně postupujeme, jestliže ověřujeme testem linearitu vztahu. Za proměnnou Z pak dosadíme hodnoty x^2. Zjišťujeme tak, zda místo úsporného modelování pomocí přímky není vhodnější popsat analyzovaný vztah pomocí kvadratické funkce.

PŘÍKLAD 7.13

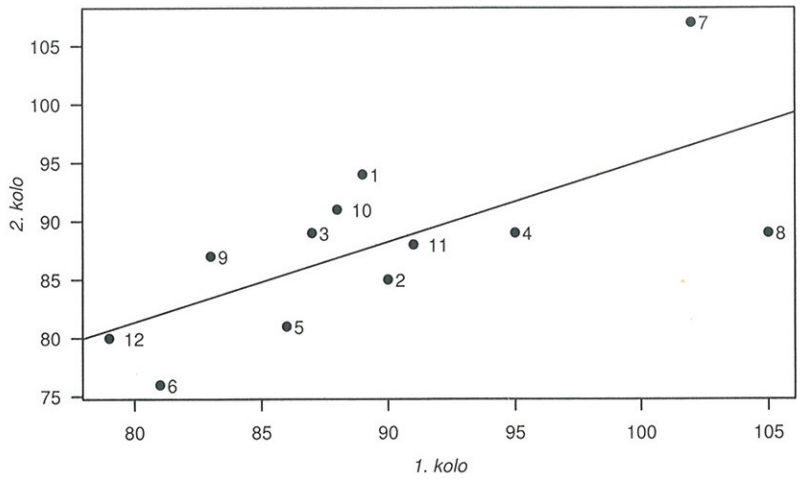
Ověřování předpokladů lineární regrese

Jak dobře predikují výsledky v prvním turnajovém kole golfu výsledky ve druhém kole? V tabulce 7.14 uvádíme data, která popisují výsledky ženského vysokoškolského týmu v turnaji počtem úderů potřebných na dokončení jednoho kola. Bodový dvojrozměrný graf ukazuje lineární vztah mezi výsledky prvního a druhého kola. Jsou na něm patrné dva nezvyklé body pro hráčky 7 a 8 (viz obrázek 7.10).

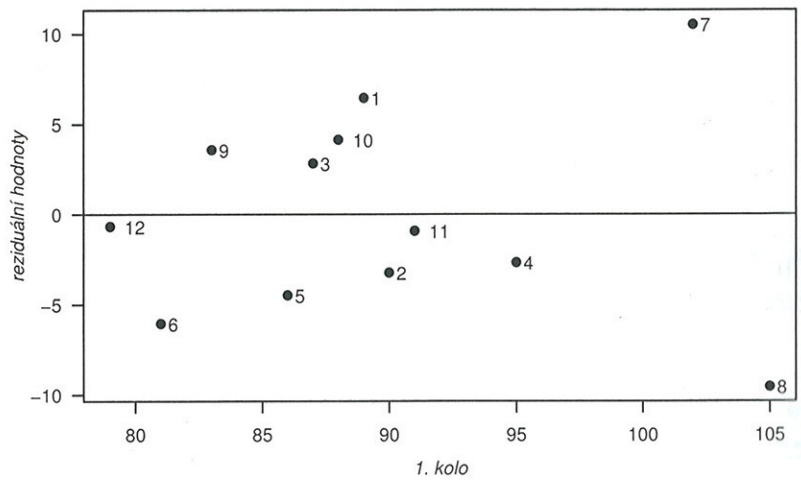
Tab. 7.14 Příklad regrese – závislost počtu úderů ve druhém kole golfového turnaje na úspěšnosti v prvním kole

Hráčka	1	2	3	4	5	6	7	8	9	10	11	12
První kolo (X)	89	90	87	95	86	81	102	105	83	88	91	79
Druhé kolo (Y)	94	85	89	89	81	76	107	89	87	91	88	80
Reziduální hodnota	6.45	−3.23	2.83	−2.63	−4.47	−6.03	10.51	−9.54	3.58	4.14	−0.91	−0.66

Obr. 7.10 Závislost výsledků ve druhém kole na výsledcích v prvním kole ze soutěže v golfu



Obr. 7.11 Graf reziduálních hodnot pro regresi výsledků ze soutěže v golfu



Pro všechny body jsme nalezli regresní přímku $y = 26{,}33 + 0{,}69x$. Hodnota t-statistiky pro test hypotézy $H_0: \beta = 0$ je 2,99 a vede k p-hodnotě pro dvoustranný test 0,0136. Reziduální hodnoty regrese uvádíme v posledním řádku tabulky 7.14 a v závislosti na hodnotách výsledků prvního kola je ukazuje obrázek 7.11. Prohlídkou grafu zjistíme, že konfigurace reziduí nevykazuje jasný trend, takže lineární model je pro popis dat vyhovující. Dva nezvyklé body 7 a 8 mají veliké reziduální hodnoty. Prozkoumejme pomocí grafu stonku a listu, zda má empirické rozdělení reziduálních hodnot symetrický tvar (normální tvar). Předtím je zaokrouhlíme na nejbližší celé číslo.

```
−1 | 0
−0 | 643311
 0 | 3446
 1 | 1
```

Rozdělení vypadá symetricky. To znamená, že předpoklad o normálním rozdělení povede při statistickém usuzování k aproximativně správným výsledkům. Regrese je vždy ovlivněna extrémními body. Přesvědčíme se o tom vynecháním bodu 7 a pak bodu 8 v případě směrnice:

$b = 0{,}69$ pro všechny body;
$b = 0{,}41$, jestliže vynecháme bod 7;
$b = 1{,}1$ jestliže vynecháme bod 8.

Přezkoumáním se zjistilo, že všechny hodnoty byly správně zaregistrovány. Výsledky jsou silně ovlivněny hodnotami 7 a 8. Abychom mohli dojít ke spolehlivějším závěrům, museli bychom pracovat s větším počtem výsledků.

7.3.5 Test náhodnosti

Před pokusem proložit křivku časovou posloupností numerických údajů nás může zajímat, zda lze vůbec zamítnout nulovou hypotézu, že data se mění zcela náhodně. Podobnou otázku si klademe také při přezkoumávání, zda má výběr skutečně náhodný charakter, v případě, že jsme byli nějak omezeni při jeho realizaci a data vznikla v časové sekvenci. Existuje několik metod, jak testovat hypotézu náhodnosti výběru na základě pořadí, jak data vznikají. Tyto metody nám umožňují rozhodnout, zda konfigurace v sérii dat odpovídají náhodě, nebo ne. Popíšeme asymptoticky platný z-test, založený na počítání tzv. iterací. Pojem **iterace** vysvětlíme na příkladu. Mějme zaznamenáno pohlaví (M nebo Ž) u zájemců o lístek na koncert. K pokladně se zájemci dostavili v pořadí:

M Ž M Ž M M M Ž M Ž M M M Ž Ž M M M M Ž Ž M Ž M M M
Ž M M M Ž Ž Ž M Ž M M M Ž M Ž M M M M Ž Ž M

Zajímá nás, zda zájemci M (muži) jsou zcela náhodně promíchání se zájemci Ž (ženy). Podtržení označuje jednu iteraci, tedy skupinu hodnot stejného druhu. Zjišťujeme počet m iterací v sérii. Začněme od začátku naší řady. První čtyři znaky chápeme jako čtyři jednočlenné skupiny – určují 4 iterace. Další skupina tří znaků M tvoří jednu iteraci. V uvedené sérii je dohromady 27 iterací. Pokud by mechanismus tvoření série měření byl náhodný – při daném počtu hodnot M(n_a) a hodnot Ž(n_b), pak průměrný počet iterací a jejich směrodatná odchylka mají hodnotu:

$$\mu_r = \frac{2n_a n_b}{n_a + n_b} + 1\,, \quad \sigma_r = \sqrt{\frac{2n_a n_b(2n_a n_b - n_a - n_b)}{(n_a + n_b)^2(n_a + n_b - 1)}}$$

Test náhodnosti se provádí pomocí z-statistiky

$$z = \frac{m - \mu_m}{\sigma_m},$$

která má za platnosti nulové hypotézy asymptoticky standardizované normální rozdělení. V našem případě je $m = 27$, $n_a = 30$, $n_b = 18$. Po krátkém počítání dostaneme testovací statistiku ve tvaru

$$z = \frac{27 - 23{,}5}{3{,}21} = 1{,}09.$$

Protože testovací statistika v absolutní hodnotě nepřesahuje hodnotu 1,96, nemůžeme hypotézu náhodnosti zamítnou na hladině významnosti 0,05. Poznamenejme, že takto konstruovaný test nepředpokládá rovnost n_a a n_b.

Test iterací není omezen na testování náhodnosti časové řady hodnot alternativní proměnné. Jakákoli posloupnost čísel může být zkoumána podobně, jestliže čísla rozdělíme podle toho, zda leží pod, nebo nad jejich mediánem. Tento přístup lze použít také pro test trendů v reziduálních hodnotách, jestliže u časové řady, jíž jsme proložili nějakou křivkou, posoudíme příslušné reziduální hodnoty a určíme, které mají kladné a které záporné znaménko. Pokud použitá regresní závislost je vyhovující, pak počet iterací kladných a záporných znamének by neměl být ani moc velký, ani moc malý. Test náhodnosti lze v tomto případě doplnit znaménkovým testem rovnosti počtu kladných a záporných znamének. Oba testy však zkoumají jinou hypotézu.

7.3.6 Nelineární regresní analýza

V některých případech vyplyne z úvahy nebo z analýzy grafického znázornění bodů $(x_i; y_i)$, že regresní vztah proměnných X a Y nelze popsat přímkou. Potom

hledáme jiné jednoduché regresní křivky. Často je možné transformací proměnných X nebo Y získat lineární vztah mezi transformovanými proměnnými. Nové proměnné pak můžeme analyzovat známými metodami lineární regresní analýzy. Cílem další interpretace analýzy je vyložit získané výsledky pro původní netransformované hodnoty.

Přibližme si ideu linearizační transformace na případu deterministické závislosti, kdy se závisle proměnná mění exponenciálně v čase t: $y = a\,e^{bt}$, $(a; b > 0)$. Jestliže zlogaritmujeme tuto rovnici, dostaneme

$$\ln(y) = \ln(a) + bt.$$

Vidíme, že logaritmus závisle proměnné je lineární funkcí času. Jestliže považujeme za novou závisle proměnnou $z = \ln(y)$, celý vztah se zjednodušil. Tento důsledek se využívá při hledání parametrů regresního vztahu, když chceme použít běžné vzorce pro odhady parametrů metodou nejmenších čtverců. (Místo přirozeného logaritmu můžeme samozřejmě použít např. dekadický logaritmus – nezmění se tím princip, jen číselné hodnoty regresních koeficientů.)

V dalším výkladu se omezíme pouze na jednoduché soustavy takto odvozených regresních křivek, které lze obecně vyjádřit rovnicí:

$$y = f(x, a, b)$$

kde a a b jsou vhodné parametry a f je předpis pro regresní vztah. Při analýze postupujeme takto:

1. Nejdříve zvolíme tvar regresního vztahu, což je funkce f, která pravděpodobně dobře vystihne průběh našich empirických dat nebo jejíž tvar je odůvodněn teoretickými úvahami.
2. Hodnoty y a x transformujeme linearizačními transformacemi, jejichž předpisy jsou dané vybranou regresní funkcí f. Dostaneme nové hodnoty y' a x'
3. Pro hodnoty y' a x' vypočteme regresní přímku $z' = a' + b'x'$.
4. Zpětnou transformací hodnot a' a b' získáme hodnoty a, b jako odhady parametrů původního regresního vztahu $y = f(x, a, b)$. Tato zpětná transformace je opět závislá na zvolené funkci f.
5. Testy můžeme provádět pro parametry a' a b'. Smysl hypotéz o těchto parametrech musíme interpretovat ve vztahu k parametrům a, b.

V tabulce 7.15 uvádíme některé typy funkcí f, příslušné linearizační transformace a vztahy mezi koeficienty regrese a', b' a parametry a, b.

Tab. 7.15 Vybrané linearizační transformace

Regresní vztah je dán funkcí	Linearizační transformace pro x a y		Vztahy pro získání koeficientů a, b	
	$y' =$	$x' =$	$a =$	$b =$
$y = a + b/x$	y	$1/x$	a'	b'
$y = ab^x$	$\ln y$	x	$\exp(a')$	$\exp(b')$
$y = ax^b$	$\ln y$	$\ln x$	$\exp(a')$	b'
$y = a\exp(b/x)$	$\ln y$	$1/x$	$\exp(a')$	b'
$y = a + bx^n$	y	x^n	a'	b'

PŘÍKLAD 7.14

Nelineární regrese

Hodnoty světových rekordů v plavání volným způsobem (podle tabulek z roku 1991) jsou uvedeny v tabulce 7.16.

Různá zkoumání ukázala, že dosažené časy nejsou lineárně závislé na délce trati. Aby se vztah linearizoval, doporučuje se provést logaritmickou transformaci dosaženého času i délky tratě. Po provedení transformací získáme regresní rovnice (s uvedením směrodatné chyby pro regresní koeficient) – tab. 7.17.

Test rozdílnosti regresních koeficientů neprokázal statisticky významný rozdíl, proto můžeme obě regrese přepočítat s tím, že uvažujeme společný regresní koeficient. Nakonec transformujeme výsledky do původního měřítka (tab. 7.18).

Tab. 7.16 Příklad nelineární regrese – světové rekordy v plavání volným způsobem

Vzdálenost [m]	50	100	200	400	800	1500
Čas mužů	21,81	48,42	1:46,69	3:46,69	7:47,85	14:50,36
Čas žen	24,98	54,73	1:57,55	4:03,85	8:16,22	15:52,10

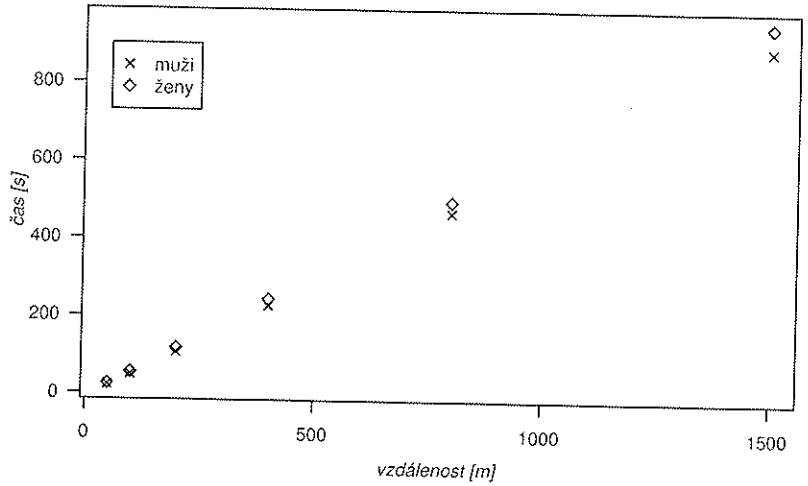
Tab. 7.17 Příklad regresních rovnic pro nelineární data

Muži	$\ln(\text{čas}) = -1{,}146 + 1{,}091 \times \ln(\text{vzdálenost})$	$r = 0{,}999$	0,0146
Ženy	$\ln(\text{čas}) = -0{,}922 + 1{,}067 \times \ln(\text{vzdálenost})$	$r = 0{,}999$	0,0146

Tab. 7.18 Příklad nelineární regrese – rovnice transformované do původního měřítka

Muži	Ženy
$\text{Čas} = 0{,}339 \times \text{Vzdálenost}^{1{,}079}$	$\text{Čas} = 0{,}372 \times \text{Vzdálenost}^{1{,}079}$

Obr. 7.12 Údaje o světových rekordech v plavání pro muže a ženy



Vidíme, že poměrový index *čas (ženy)/čas (muži)* činí 1,097. To znamená, že ženy potřebují o 9,7 % více času než muži. Graficky je situace znázorněna na obrázku 7.12.

7.3.7 Porovnání metod měření a Blandův-Altmanův graf

V tomto odstavci upozorníme na situace, v nichž prokládáme data přímkou jiným způsobem, než je standardní metoda nejmenších čtverců. Půjde o určité zobecnění metody nejmenších čtverců a jednu rezistentní metodu. Tyto alternativní přístupy objasníme v souvislostech hodnocení kvality metod měření. Přitom připomeneme některé statistické charakteristiky a jejich aplikaci při srovnávání

metod měření. Uvažujeme měření prováděná v biomedicíně, kde se klade důraz na posuzování velikosti systematické chyby a samozřejmě i náhodné chyby měření, avšak popisované postupy mají použití v mnoha dalších problémových situacích a nalézají v poslední době uplatnění i v psychometrii. Na rozdíl od běžné regresní analýzy se nepoužívají pro řešení problému predikce.

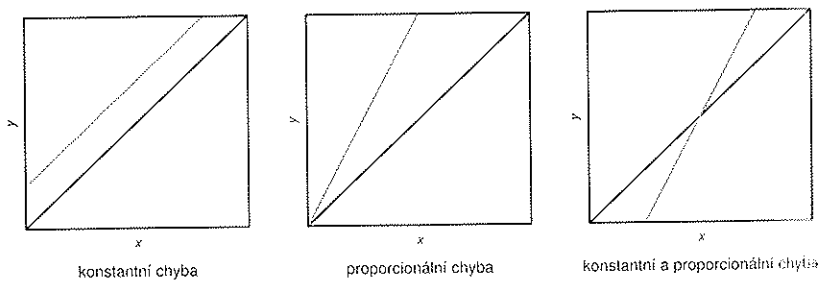
Zabýváme se případem, kdy porovnáváme metody měření X a Y, které mají měřit stejnou veličinu. V praxi jde o srovnání nově navržené metody (označíme ji Y) s referenční metodou (označíme ji X) nebo se *zlatým standardem*, což je metoda, která představuje aktuálně nejlepší metodu vzhledem k velikosti systematické chyby a úrovni reliability (opakovatelnost výsledků). Při srovnávání vycházíme z n dvojic měření (x_i, y_i), které jsme získali změřením objektů i metodou X a Y v rámci tzv. srovnávacího experimentu.

Při posuzování kvality měřicí metody vycházíme z modelu

$$Z = T + S + \varepsilon,$$

kde Z, resp. T je naměřená, resp. správná hodnota, S systematická chyba, ε náhodná chyba s nulovou střední hodnotou a rozptylem $Var_Y(\varepsilon)$. U systematické chyby rozlišujeme konstantní složku, jež je stejná v celém rozsahu měření, a proporcionální složku, která je úměrná hladině měření (např. koncentraci měřené látky). Obě tyto složky mohou být způsobeny jinými mechanismy v měřicím procesu. Jednotlivé typy systematických chyb a jejich vztahy jsou zobrazeny na obrázku 7.13. Také rozptyl $Var_Y(\varepsilon)$ často závisí na hodnotě měření (hladině koncentrace). Je důležité si uvědomit, že tento model lze uplatnit jak pro srovnávanou metodu Y, tak pro referenční metodu X. Pokud $S = 0$, říkáme někdy, že metoda je správná.

Obr. 7.13 Typy systematických chyb měřicích metod



Stručně zhodnotíme, jak se standardně používané statistiky uplatňují při hodnocení vztahu porovnávané a referenční měřicí metody. Mezi nejpoužívanější patří Pearsonův korelační koeficient r, párový t-test pro přezkoušení přítomnosti systematické odchylky mezi metodami a výpočet parametrů regresní přímky. Korelační koeficient je citlivý k náhodné chybě. Proto se používá k jejímu odhalení. Je však citlivý také k rozmezí měření. Často zvětšením rozsahu měření dosáhneme značného přiblížení korelačního koeficientu k jedničce. Snad největší chyba spočívá v tom, že přisuzujeme důležitost tomu, že korelační koeficient je významně různý od nuly. Ve srovnávacích experimentech není tento typ uvažování na místě, přesto se údaje o této významnosti pravidelně objevují v hodnotících zprávách. Závažná je skutečnost, že korelačním koeficientem neodhalujeme ani přítomnost proporcionální chyby, ani chyby konstantní. Odpůrci korelačního koeficientu tvrdí, že tato statistika by se při hodnocení dat při srovnávání metod měření neměla nikdy používat. Někdy se doporučuje ho nahradit koeficientem konkordance r_{LIN} podle Lina (Lin 1989)

$$r_{\mathrm{LIN}} = \frac{2\,\mathrm{cov}(x, y)}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2},$$

který má jako popisná statistika srovnatelnosti metod (nebo opakovatelnosti měření) několik výhodných vlastností:

a) $-1 \le -|r| \le |r_{\mathrm{LIN}}| \le |r| \le 1$;
b) $r_{\mathrm{LIN}} = r$ tehdy a jen tehdy, jestliže $s_x^2 = s_y^2$ a také $\bar{x} = \bar{y}$;
c) $r_{\mathrm{LIN}} = 0$ tehdy a jen tehdy, jestliže $r = 0$;
d) $r_{\mathrm{LIN}} = \pm 1$, pokud $r = 1$ a také $s_x^2 = s_y^2$ a $\bar{x} = \bar{y}$.

Pomocí tohoto koeficientu posuzujeme, jak jsou dvojice měření těsně rozloženy kolem přímky dané rovnicí $y = x$.

Testovací statistiku párového t-testu počítáme pomocí směrodatné odchylky s_d diferencí párů měření a průměrné diference m_d podle vzorce

$$t = \frac{m_d}{(s_d/\sqrt{n})},$$

kde n je počet měřených objektů.

Tyto statistiky se uvádějí ve zprávách o srovnávacích experimentech měřicích metod spolu s dopočítanou p-hodnotou. Průměrná diference m_d poskytuje hodnověrný odhad systematické chyby pouze v případě, kdy proporcionální chyba není přítomna. Také testování pomocí t-statistiky má význam pouze v této souvislosti. Charakteristika s_d kvantifikuje náhodnou chybu způsobenou náhodnými chybami srovnávané i referenční metody. Neodráží specificky náhodnou chybu

srovnávané metody, protože ji zvětšuje náhodná chyba referenční metody. Jestliže se velikosti náhodných chyb mění podle úrovně měření, ovlivňuje to silně hodnotu s_d. Tato charakteristika pak vyjadřuje průměrnou variabilitu, která se obtížně interpretuje. Bohužel je nutné přihlédnout ke skutečnosti, že s_d také odráží proporcionální systematickou chybu, jíž je srovnávaná metoda zatížená. Lepší je nahradit t-statistiku intervalem spolehlivosti pro odhad systematické chyby.

Běžně se určuje vztah mezi měřením X a měřením Y nalezením takové regresní přímky $y = a + bx$, která minimalizuje vzhledem k parametrům regresní přímky součet čtverců odchylek bodů $(x_i; y_i)$ od hledané přímky ve směru kolmém na osu X. Tímto způsobem získáme odhad absolutního členu a přímky, odhad b pro směrnici regresní přímky a odhad chyby při regresi $s_{y.x}$. Každá z těchto statistik je citlivá k jinému typy chyb. Chyba $s_{y.x}$ odhaduje náhodnou chybu mezi metodami a má stejné omezení jako s_d až na to, že $s_{y.x}$ není rušivě ovlivněna přítomností proporcionální systematické chyby. Konstantní chyba se může odhadovat pomocí průsečíku s osou Y (hodnota a) a proporcionální chyba pomocí hodnoty $1 - b$. Celkovou chybu lze odhadnout 95% mezí, která má hodnotu $a + (1 - b)x + 2s_{y.x}$. Tato mez určuje maximální velikost rozdílů naměřených hodnot od skutečné hodnoty s 95% spolehlivostí. Je patrné, že závisí na úrovni proměnné X.

Lineární regresní analýza se může použít jenom v pásmu lineárního vztahu mezi oběma metodami měření. Nelinearita v datech znehodnocuje odhady absolutního členu i směrnice, což vede ke špatnému odhadu systematické chyby na jednotlivých rozhodovacích hladinách, kde se měření používá. Odhady koeficientů regresní přímky jsou velmi citlivé k odlehlým hodnotám. Proto bychom měli data vždy překontrolovat graficky. Jednoduchá regrese má další dvě omezení: a) neuvažuje náhodnou chybu u referenční metody; b) jedním z jejích předpokladů je konstantnost směrodatné chyby odhadu v celém rozmezí. S porušením předpokladu o stálosti rozptylu kolem regresní přímky se vyrovnáme použitím vážené regrese, pro kterou potřebujeme znát profil změn rozptýlenosti.

Popsané metody prokládání přímkou vycházejí z toho, že referenční metoda má mnohem menší velikost náhodné chyby než hodnocená metoda. Protože náhodné chyby ovlivňují jak srovnávanou, tak referenční metodu, není model jednoduché regrese zcela adekvátní. Deming (1943) navrhl hledat vztah mezi hodnocenou a referenční metodou pomocí vážené regrese, kdy předpokládáme, že průměrné hodnoty jsou vázané lineárním vztahem a výsledky měření oběma metodami jsou zatíženy náhodnými chybami:

$$x_i = \hat{x}_i + \varepsilon_i$$

$$y_i = \hat{y}_i + \delta_i$$

Optimální řešení dostaneme, když minimalizujeme vážený součet čtverců chyb vzhledem k parametrům regresní přímky (a, b)

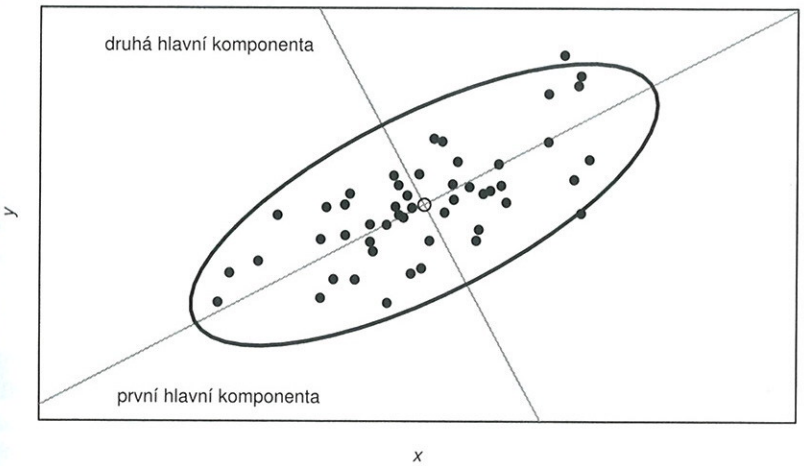
$$\min S = \sum w_i(y_i - \hat{y}_i)^2 + v_i(x - \hat{x}_i)^2,$$

kde $\hat{y}_i = a + b\hat{x}_i$ a w_i a v_i jsou hodnoty vah kvadratických odchylek. Velikost vah w_i a v_i je v nepřímé v závislosti na velikosti chyby metody, jež jsou charakterizovány chybovými rozptyly $Var_x(\varepsilon)$ a $Var_y(\delta)$. Z analýzy celé procedury plyne, že Demingova metoda hledá přímku, jež minimalizuje součet čtverců vzdáleností bodů od přímky, přičemž vzdálenosti se měří pod úhlem od regresní přímky, který je závislý na poměru rozptylů charakterizujících náhodnou chybu obou metod $(Var_x(\varepsilon)/Var_y(\delta) = \lambda)$. Odhad parametru b má tvar

$$b = G + \sqrt{G^2 + 1/\lambda}, \quad \text{kde} \quad G = \frac{s_y^2 - (1/\lambda)s_x^2}{2rs_xs_y}.$$

Jestliže $Var_x(\varepsilon)$ má nulovou hodnotu (nebo relativně malou vůči $Var_y(\delta)$), získáme přímku odpovídající jednoduché regresi. Jestliže koeficient λ má hodnotu rovnou jedné, pak jsme získali přímku odpovídající hlavní komponentě, jež spojuje vrcholy konfidenční elipsy, popisující rozptýlenost bodů $(x_i; y_i)$. Ve sku-

Obr. 7.14 Shluk bodů proložený první a druhou hlavní komponentou



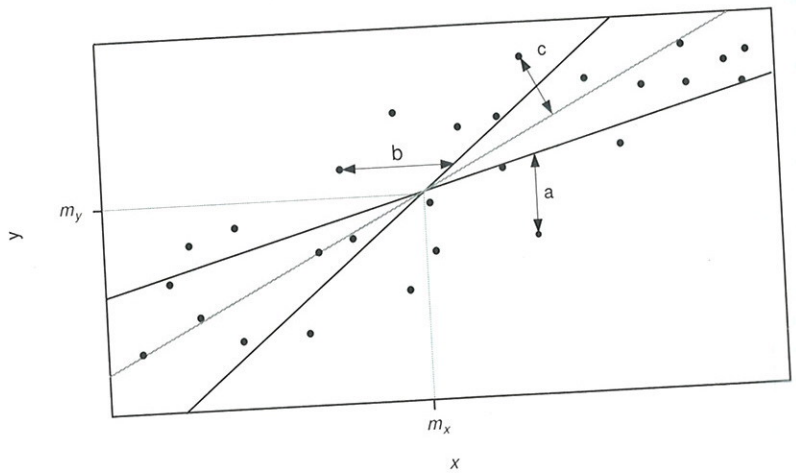
tečnosti jde o první hlavní komponentu, známou z vícerozměrné statistické analýzy (viz kap. 13.7). Obrázek 7.14 ukazuje první hlavní komponentu sestrojenou pro shluk bodů odpovídající naměřeným váhám a výškám u skupiny 50 chlapců. Konfidenční elipsa pokrývá v tomto případě přibližně 95% bodů. V tomto případě měříme vzdálenost kolmo na hledanou přímku. Z tohoto důvodu se celé proceduře někdy říká ortogonální regrese.

Na obrázku 7.15 jsou modelově znázorněny tři přímky proložené body třemi metodami:

a) jednoduchá regrese, kdy y je závisle proměnná;
b) jednoduchá regrese, kdy x je závisle proměnná;
c) ortogonální regrese.

Zvolený směr predikce v metodě b) způsobuje, že dostaneme jinou prokládací přímku než v případě a), protože počítáme (podobně jako v případě ortogonální regrese) s jinými vzdálenostmi bodů od proložené přímky, jejichž čtverce sčítáme, když se opíráme o princip nejmenších čtverců.

Obr. 7.15 Proložení shluku bodů třemi metodami nejmenších čtverců s označením směru určení vzdálenosti bodů od hledané přímky, z kterých se počítají druhé mocniny (čtverce)



a) vzdálenost pro regresi, kde y je závisle proměnná
b) vzdálenost pro regresi, kde x je závisle proměnná
c) vzdálenost pro ortogonální regresi

Jestliže lze přijmout, že $Var_x(\varepsilon)/Var_y(\delta) = s_x^2/s_y^2$, získáme přímku odpovídající standardizované hlavní komponentě, která spojuje vrcholy konfidenční elipsy pro shluk bodů se standardizovanými souřadnicemi $(x_i'; y_i')$. Explicitní tvar odhadu směrnice přímky má v tomto případě jednoduchou podobu $b = s_y/s_x$. Protože navíc uvedený předpoklad odpovídá často se vyskytujícím poměrům v datech, doporučuje se tento způsob proložení jako dobrá alternativa pro jednoduchou regresi. Všechny přímky získané Demingovou metodou procházejí těžištěm $(m_x; m_y)$. Z této vlastnosti odvodíme tvar pro odhad absolutního členu $a = \bar{y} - b\bar{x}$ (podrobněji Zvára, 1989, s. 187; Carrol, 1996). Poznamenejme, že výpočty Demingovou metodou předpokládají přesný lineární vztah mezi „průměrnými" hodnotami obou srovnávaných metod.

Uvedené způsoby regrese používají v explicitních výrazech pro odhad regresního koeficientu korelační koeficient a směrodatné odchylky. Naneštěstí jsou tyto statistiky velmi citlivé vůči odlehlým hodnotám. To vede k úvaze, zda by bylo možno pro parametry regresní přímky navrhnout neparametrickou proceduru. Tímto směrem postupovali Passing a Bablok (1983). Jejich procedura odhaduje regresní koeficient jako medián částečných odhadů b_{ij} tohoto koeficientu:

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j} \quad \text{pro} \quad i < j$$

Absolutní člen regresní přímky se odhaduje jako medián hodnot $a_i = y_i - bx_i$. Passing a Bablok navrhli svoji neparametrickou metodu tak, aby jejich odhad byl kvalitní pro případ, že obě metody jsou zatíženy náhodnou chybou.

V případě potřeby testujeme pomocí vhodných statistik specifické hypotézy o parametrech regresní přímky. Nejčastěji se zaměřujeme na hypotézu $(a, b) = (0,1)$, která odpovídá identitě obou metod. Příslušné statistiky najdeme v citované literatuře. Pro nejvíce potřebné výpočty odhadů regresní přímky, intervalů spolehlivost i sestrojení grafu pro všechny zde uvedené eventuality lze využít volně dostupný program P. Marquise (2002).

Výhodou Demingovy i Passingovy-Bablokovy metody je okolnost, že obě dávají – na rozdíl od jednoduché regrese metodou nejmenších čtverců – jedinou prokládající přímku a zohledňují přítomnost náhodné chyby u hodnocené i referenční metody.

PŘÍKLAD 7.15

Porovnání různých přístupů k proložení přímky daty

V tabulce 7.20 uvádíme výsledky proložení přímkou získané jednotlivými přístupy pro modelová data v tabulce 7.19, přičemž metoda 1 představuje proměnnou X. Data byla generována tak, aby přímka vyhovovala vztahu $y = x$. Je patrné, že v tomto případě přístup

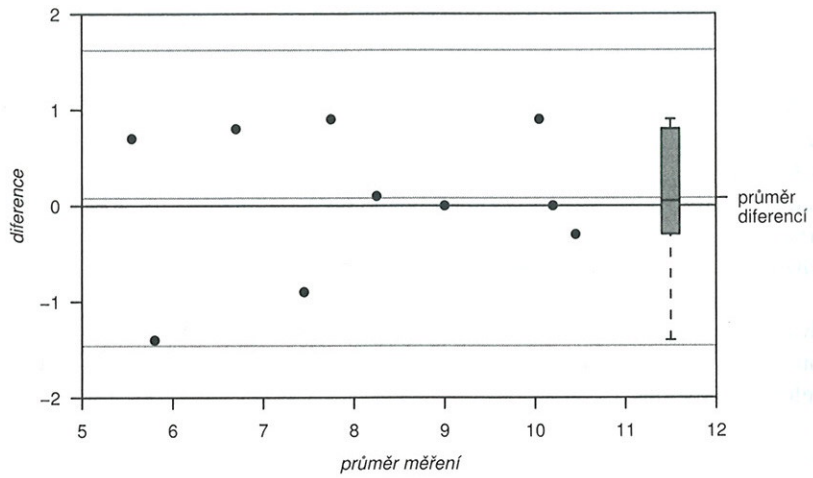
Tab. 7.19 Příklad porovnávání dvou metod měření – modelová data

Vzorek	1	2	3	4	5	6	7	8	9	10
Metoda 1	7,0	8,3	10,5	9,0	5,1	8,2	10,2	10,3	7,1	5,9
Metoda 2	7,9	8,2	9,6	9,0	6,5	7,3	10,2	10,6	6,3	5,2

Tab. 7.20 Příklad porovnání dvou metod měření – různé přístupy k proložení přímky

	b	*a*
„Klasická" lineární regrese	0,862	1,049
Metoda standardní hlavní komponenty	0,947	0,352
Metoda hlavní komponenty	1,001	0,088
Passingova-Bablokova regrese	1,000	−0,050

Obr. 7.16 Blandův-Altmanův graf spolu s krabicovým grafem rozdílů mezi dvěma metodami



pomocí hlavní komponenty a Passingova-Bablokova regrese dávají nejlepší proložení. Obrázek 7.16 ukazuje Blandův-Altmanův graf spolu s krabicovým grafem rozdílů mezi oběma metodami.

Numerickou analýzu vztahu obou metod měření doporučují Bland a Altman (1986) nahradit grafickou analýzou pomocí modifikace grafu reziduálních hodnot pro regresi. Standardně nanášíme na osu Y reziduální hodnoty a na osu X hodnoty prediktoru. Tito autoři tvrdí, že co nás zajímá při srovnání metod, není tvar přímky, který váže průměrné hodnoty obou metod, ale pouze chování rozdílů hodnot při měření jednotlivých objektů oběma metodami. Modifikace podle Blanda a Altmana spočívá v tom, že na osu Y nanášíme rozdíl $x - y$ hodnot získaných referenční a srovnávanou metodou a na osu X jejich průměr $(x + y)/2$, abychom vyrušili jev regrese k průměru (umělou korelaci mezi hodnotami $x - y$ a x). Bland-Altmanův graf, nazývaný též rozdílový graf, adekvátněji hodnotí nepodobnost měření oběma metodami. Ve srovnávacích experimentech nás zajímají především rozdíly $(x - y)$, a ne rozdíly hodnot srovnávané metody od regresní přímky. Graf je doplněn o tři kontrolní čáry, jež reprezentují průměr rozdílů, od něhož ještě zakreslíme přímky ve vzdálenosti 1,96 s_d na obě strany. Bland a Altman doporučují doplnit tento graf výpočtem intervalu spolehlivosti pro průměrný rozdíl, průměrem rozdílů m_d a jejich směrodatnou odchylkou s_d. Jestliže prohlídka odhalí trend v rozdílech $(x - y)$, počítáme ještě korelaci mezi hodnotami $(x - y)$ a $(x + y)/2$ a její statistickou významnost. Tato analýza posuzuje přítomnost proporcionální chyby. Proměnlivost velikosti náhodné složky chyb hodnotíme obdobně grafem závislosti *absolutních hodnot* rozdílů $(x - y)$ na hodnotách $(x + y)/2$. Článek autorů patří k nejcitovanějším metodologickým pracím v lékařské literatuře a jeho recepce je značná i v metodologické literatuře behaviorálních věd.

7.4 Regrese k průměru

Regrese k průměru patří k jevům, které mohou negativně ovlivnit interní validitu výzkumné studie. Tímto fenoménem se poprvé zabýval v roce 1886 F. Galton. Ozřejmíme ho na příkladu, jenž zajímal také Galtona. Předpokládáme, že korelace mezi výškou otce a syna je 0,8, a dále, že průměrná výška otců a synů je stejná. Výška otce je 190 cm, jaký bude odhad výšky syna pomocí lineární regrese? Očekávali bychom, že odhad bude 190 cm, ale není tomu tak. Abychom rozdíl jednoduše vysvětlili, budeme předpokládat, že jsme standardizovali naměřené hodnoty výšek a že hodnota 190 leží jednu směrodatnou odchylku nad průměrem výšek otců. Jestliže uvažujeme standardizované hodnoty, regresní koeficient se rovná korelačnímu koeficientu. Tedy $b = 0{,}8$ a predikce výšky má hodnotu 0,8 vynásobenou výškou otce ve standardizovaných hodnotách. Z toho je patrné, že predikce výšky syna je menší než výška otce. Tuto skutečnost nazýváme „regrese k průměru". Musíme ji zohlednit v některých typech statistické analýzy.

Regrese k průměru, resp. regrese ke středu, se projevuje, když analyzujeme výsledky získané na nenáhodně vybraném souboru jednotek, tedy třeba v kvaziexperimentálních výzkumech. S tímto fenoménem se setkáváme především v situacích, kdy výběr jedinců do experimentální skupiny se děje podle hodnoty proměnné, která má být zvolenou intervencí ovlivněna.

PŘÍKLAD 7.16

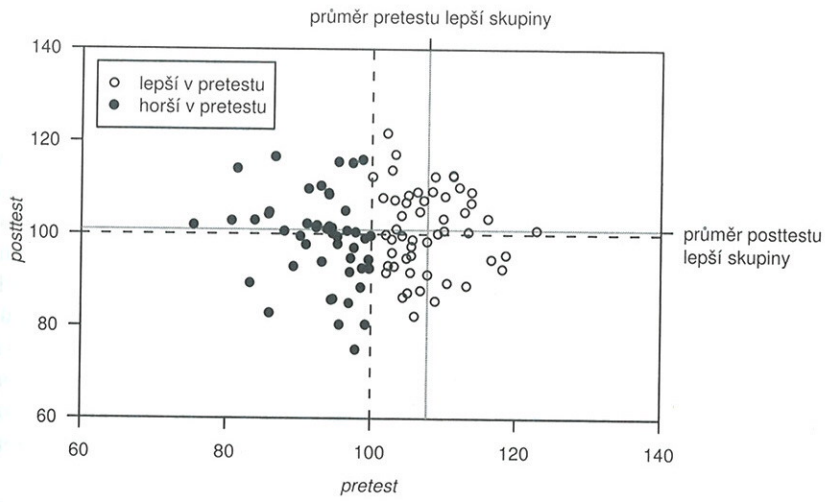
Situace, kdy nastává regrese k průměru

- V rámci medicínského výzkumu předpokládáme, že intervence má snížit v průměru hodnotu sledované proměnné. Do skupiny vybereme jedince, u nichž jsme pretestem naměřili zvýšené hodnoty. Lze ukázat, že průměr z výsledků druhého měření (z posttestu) bude nižší než průměr z prvních měření pretestem, ať jsme intervenci provedli, nebo ne.
- Trenér zařadí do reprezentačního družstva pro olympiádu jedince s nejlepšími výsledky z posledního závodu. Po olympiádě zjistí, že průměrný výkon byl horší, a přisuzuje to podmínkám při závodě.
- Studenti s horším výsledkem testu z matematiky jsou zařazeni do doučovacího kurzu. Po jeho absolvování jsou u těchto studentů zjištěny v průměru lepší hodnoty u testu sestaveného podobným způsobem. Tato změna se zdůvodňuje působením kurzu.

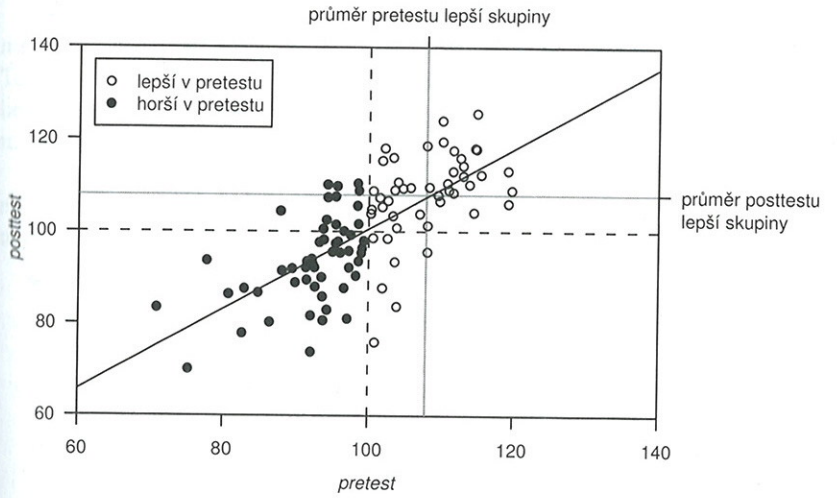
I v těchto příkladech může dojít k popsané změně v důsledku efektu regrese k průměru. Uváděné důvody pro změnu sportovního výkonu, resp. zlepšení hodnot testu z matematiky tudíž nemusí být opodstatněné.

Jak regrese k průměru v této souvislosti vzniká, ukážeme graficky na jednoduchém příkladu. Obrázek 7.17 znázorňuje pomocí bodového dvojrozměrného grafu výsledky pretestu a posttestu zcela nekorelovaných testů. Je patrné, že průměr hodnot pretestu je totožný s průměrem hodnot posttestu. Mezi provedením obou testů nedošlo k žádné změně průměrné hodnoty u celé skupiny. Představme si, že vybereme skupinu jedinců, u kterých hodnoty pretestu jsou větší než průměr. Snadno vidíme, že průměr hodnot posttestu má menší hodnotu, než byl průměr hodnot pretestu pro vybranou skupinu. Průměr hodnot posttestu je totiž totožný s původní hodnotou průměru. Jestliže mezi pretestem a posttestem existuje dokonalá shoda, tedy $r = 1$, pak průměry hodnot u obou skupin měření získaných od osob, jejichž hodnota pretestu ležela nad průměrem, budou zcela totožné. Jestliže korelace měření má hodnotu mezi nulou a jedničkou, pak vypočítané průměry jak hodnot pretestu, tak hodnot posttestu budou ležet mezi dvěma popsanými extrémy hodnot průměrů. To znamená, že průměr hodnot měření posttestu bude vždy menší než průměr hodnot pretestu u té skupiny jedinců, jejichž hodnota pretestu byla větší než průměr. Tento případ je ukázán na obrázku 7.18.

Obr. 7.17 Regrese k průměru pro nekorelované hodnoty pre- a posttestu



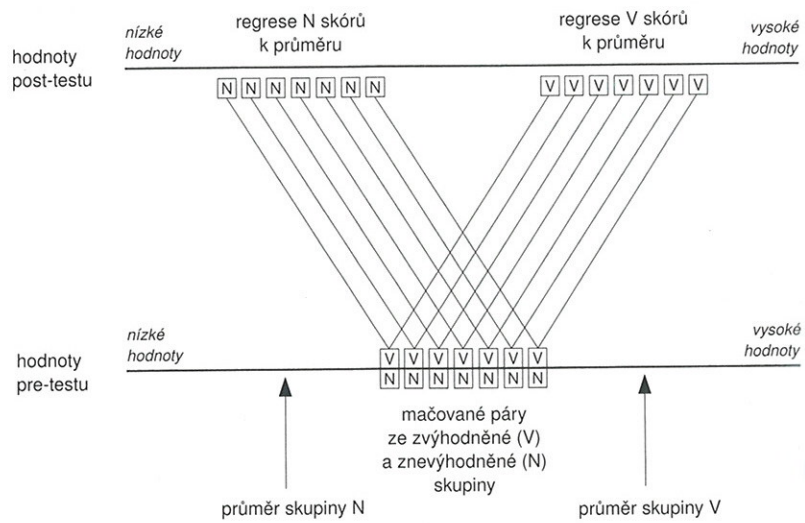
Obr. 7.18 Regrese k průměru pro korelované hodnoty pre- a posttestu



Regresi k průměru zdůvodněme ještě jiným způsobem. Jestliže uvažujeme jedince s hodnotami pretestu nad průměrem, pokud nenastala žádná jiná změna, některé hodnoty posttestu u této skupiny budou pod průměrem původního rozdělení hodnot posttestu. Na druhou stranu nelze očekávat, že hodnoty posttestu budou rozloženy více směrem k větším hodnotám. Proto průměr hodnot posttestu bude menší než průměr hodnot pretestu.

Regrese k průměru také způsobuje, že proces vyrovnávání (mačování) nevede často k validním výsledkům. Předpokládejme, že porovnáváme studijní výkony socioekonomicky „zvýhodněné" a „znevýhodněné" skupiny studentů. Pretesty u znevýhodněné populace budou mít nižší hodnoty než u zvýhodněné populace studentů. Aby byly vytvořené „srovnatelné" skupiny, přistoupilo se k vytvoření dvojic studentů, kteří měli stejné výsledky v pretestu. Chceme porovnat jejich studijní výkony po absolvování semestru pomocí posttestu. Vlivem regrese k průměru budou výsledky v obou skupinách značně odlišné i bez působení nějaké intervence. Tato okolnost může vést k závěru, že znevýhodnění studenti nejsou schopni dosáhnout takových pokroků jako zvýhodnění studenti. Samozřejmě že efekt regrese k průměru může být zastřen, zesílen nebo zmírněn působeními dalších faktorů. Celou situaci zobrazuje obrázek 7.19.

Obr. 7.19 Efekt regrese k průměru u extrémních skupin



Souhrn

Jednoduchá korelační analýza a lineární regrese patří mezi základní metody statistické analýzy vztahu proměnných.

Pearsonův korelační koeficient r měří sílu a směr asociace dvou spojitých proměnných. Tento korelační koeficient můžeme vypočítat pro každý shluk bodů, ale adekvátně změří asociaci pouze pro lineární vztahy, jestliže proměnné měříme bez omezení na určitý interval hodnot.

Speciální korelační koeficienty (parciální a mnohonásobný) jsou vhodné pro posouzení působení třetí proměnné na sledovaný vztah nebo posouzení vhodnosti predikce závisle proměnné pomocí dvou nezávislých proměnných. Korelačními koeficienty Spearmanovým a Kendallovým nahrazujeme Pearsonův korelační koeficient v případech přítomnosti odlehlých hodnot nebo nelineárního vztahu.

Regresní přímka získaná standardní metodou nejmenších čtverců minimalizuje součet čtverců vertikálních vzdáleností pozorovaných y-hodnot od regresní přímky. Interpretaci dat provádíme také pomocí grafického znázornění, kontrolou linearity a působení extrémních bodů.

Vysvětlili jsme základní metody statistického usuzování, které jsou vhodné pro jednoduchý lineární regresní model. Jsou použitelné pouze tehdy, jestliže studovaný vztah je skutečně lineární, reziduální odchylky mají normální rozdělení a jejich rozptyl je stejný v celém rozsahu proměnné X.

Korelace a regrese jsou úzce spojené koncepty. Pearsonův korelační koeficient r je směrnicí b regresní přímky, jestliže měříme hodnoty x i y ve standardizovaných jednotkách.

Jednoduchá regresní analýza je nejjednodušším případem obecného lineárního modelu vícenásobné regrese, jímž se budeme zabývat ve zvláštní kapitole. Techniky, které jsme poznali v této kapitole, jako analýza reziduálních hodnot nebo transformace proměnných s cílem linearizovat vztah, lze použít i v obecném modelu lineární regrese.

Další informace o statistické analýze pomocí jednoduchého lineárního regresního modelu uvádějí např. Havránek (1993), Meloun (1994), Procházka (1999) nebo Zvára (1997).