Jednoduchá, mnohonásobná a parciální korelace

Pearsonův koeficient korelace

Nechť X, Y jsou náhodné veličiny se středními hodnotami E(X), E(Y) a rozptyly D(X), D(Y).
Číslo

$$R(X,Y)=\begin{cases} E\left(\dfrac{X-E(X)}{\sqrt{D(X)}}\cdot\dfrac{Y-E(Y)}{\sqrt{D(Y)}}\right)=\dfrac{C(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \text{ pro } \sqrt{D(X)}\sqrt{D(Y)}>0 \\ 0 \text{ jinak} \end{cases}$$

se nazývá Pearsonův koeficient korelace.
(Pro výpočet Pearsonova koeficentu korelace musíme znát simultánní distribuční funkci $\Phi(x,y)$ v obecném případě resp. simultánní hustotu pravděpodobnosti $\varphi(x,y)$ ve spojitém případě resp. simultánní pravděpodobnostní funkci $\pi(x,y)$ v diskrétním případě.)

Vlastnosti Pearsonova koeficientu korelace

a) $R(a_1, Y) = R(X, a_2) = R(a_1, a_2) = 0$

b) $R(a_1 + b_1X, a_2 + b_2Y) = \operatorname{sgn}(b_1b_2)\, R(X, Y) = \begin{cases} R(X,Y) \text{ pro } b_1b_2 > 0 \\ -R(X,Y) \text{ pro } b_1b_2 < 0 \end{cases}$

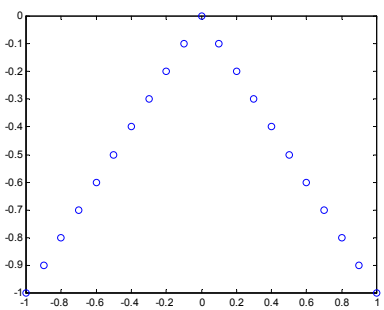
c) $R(X, X) = 1$ pro $D(X) \neq 0$, $R(X, X) = 0$ jinak

d) $R(X, Y) = R(Y, X)$

e) $|R(X,Y)| \leq 1$ a rovnost nastane tehdy a jen tehdy, když mezi veličinami X, Y existuje s pravděpodobností 1 úplná lineární závislost, tj. existují konstanty a, b tak, že pravděpodobnost $P(Y = a + bX) = 1$. Přitom $R(X, Y) = 1$, když $b > 0$ a $R(X, Y) = -1$, když $b < 0$. (Uvedená nerovnost se nazývá Cauchyova – Schwarzova – Buňakovského nerovnost.)

Z vlastností Pearsonova koeficientu korelace vyplývá, že se hodí pouze k měření těsnosti lineárního vztahu veličin X a Y. Při složitějších závislostech může dojít k paradoxní situaci, že Pearsonův koeficient korelace je nulový.

Ilustrace:



Definice nekorelovanosti

Je-li R(X, Y) = 0, pak řekneme, že náhodné veličiny jsou nekorelované. (Znamená to, že mezi X a Y neexistuje žádná lineární závislost. Jsou-li náhodné veličiny X,Y stochasticky nezávislé, pak jsou samozřejmě i nekorelované.)
Je-li R(X, Y) > 0, pak řekneme, že náhodné veličiny jsou kladně korelované. (Znamená to, že s růstem hodnot veličiny X rostou hodnoty veličiny Y a s poklesem hodnot veličiny X klesají hodnoty veličiny Y.)
Je-li R(X, Y) < 0, pak řekneme, že náhodné veličiny jsou záporně korelované. (Znamená to, že s růstem hodnot veličiny X klesají hodnoty veličiny Y a s poklesem hodnot veličiny X rostou hodnoty veličiny Y.)

Výběrový koeficient korelace

Nechť $(X_1, Y_1), ..., (X_n, Y_n)$ náhodný výběr rozsahu n z dvourozměrného rozložení daného distribuční funkcí $\Phi(x,y)$. Z tohoto dvourozměrného náhodného výběru můžeme stanovit:

výběrové průměry $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$, $M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i$,

výběrové rozptyly $S_1^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)^2$, $S_2^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - M_2)^2$,

výběrovou kovarianci $S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ a s jejich pomocí zavedeme

výběrový koeficient korelace $R_{12} = \begin{cases} \frac{1}{n-1}\sum_{i=1}^{n}\frac{X-M_1}{S_1}\cdot\frac{Y-M_2}{S_2} = \frac{S_{12}}{S_1 S_2} & \text{pro } S_1 S_2 > 0 \\ 0 & \text{jinak} \end{cases}$.

Vlastnosti Pearsonova koeficientu korelace se přenášejí i na výběrový koeficient korelace. (Výběrový koeficient korelace není nestranným odhadem skutečného koeficientu korelace, je odhadem vychýleným. Vychýlení je zanedbatekně malé pro rozsahy výběrů nad 30.)

Pearsonův koeficient korelace dvourozměrného normálního rozložení

Nechť náhodný vektor (X, Y) má dvourozměrné normální rozložení s hustotou

$$\varphi(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]},$$

přičemž $\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1^{\,2} = D(X)$, $\sigma_2^{\,2} = D(Y)$, $\rho = R(X,Y)$.

Marginální hustoty jsou:

$$\varphi_1(x) = \int_{-\infty}^{\infty}\varphi(x,y)dy = \ldots = \frac{1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{(x-\mu_1)^2}{2\sigma_1^{\,2}}},$$

$$\varphi_2(y) = \int_{-\infty}^{\infty}\varphi(x,y)dx = \ldots = \frac{1}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{(y-\mu_2)^2}{2\sigma_2^{\,2}}}.$$

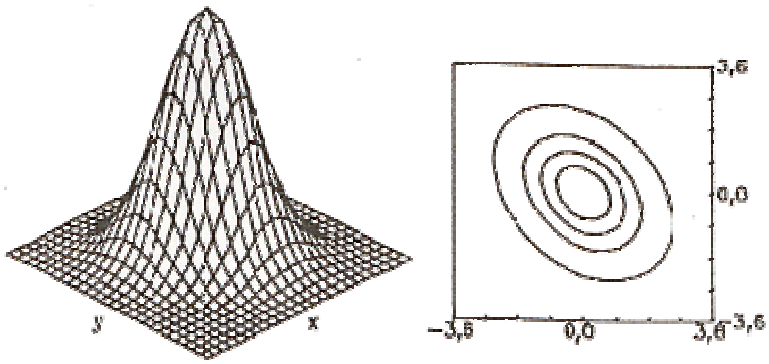
Je-li $\rho = 0$, pak pro $\forall (x,y) \in R^2 : \varphi(x,y) = \varphi_1(x)\varphi_2(y)$, tedy náhodné veličiny X, Y jsou stochasticky nezávislé. Jinými slovy: stochastická nezávislost složek X, Y normálně rozloženého náhodného vektoru je ekvivalentní jejich nekorelovanosti. Pro jiná dvourozměrná rozložení to neplatí!

Upozornění: nadále budeme předpokládat, že $(X_1, Y_1), ..., (X_n, Y_n)$ je náhodný výběr rozsahu n z dvourozměrného normálního rozložení

$$N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right).$$

Předpoklad dvourozměrné normality lze orientačně ověřit pomocí dvourozměrného tečkového diagramu: tečky by měly zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce. Vrstevnice hustoty dvourozměrného normálního rozložení jsou totiž elipsy:

Graf hustoty a vrstevnice dvourozměrného normálního rozložení s parametry $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = -0{,}75$:



Do dvourozměrného tečkového diagramu můžeme ještě zakreslit 100(1-α)% elipsu konstantní hustoty pravděpodobnosti. Bude-li více než 100α% teček ležet vně této elipsy, svědčí to o porušení dvourozměrné normality. Bude-li mít hlavní osa elipsy kladnou resp. zápornou směrnici, znamená to, že mezi veličinami X a Y existuje určitý stupeň přímé resp. nepřímé lineární závislosti.

Testování hypotézy o nezávislosti

Na hladině významnosti α testujeme H_0: X, Y jsou stochasticky nezávislé náhodné veličiny (tj. $\rho = 0$) proti
- oboustranné alternativě H_1: X, Y nejsou stochasticky nezávislé náhodné veličiny (tj. $\rho \neq 0$)
- levostranné alternativě H_1: X, Y jsou záporně korelované náhodné veličiny (tj. $\rho < 0$)
- pravostranné alternativě H_1: X, Y jsou kladně korelované náhodné veličiny (tj. $\rho > 0$).

Testová statistika má tvar: $T_0 = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$.

Platí-li nulová hypotéza, pak $T_0 \sim t(n-2)$.

Kritický obor pro test H_0 proti

- oboustranné alternativě: $W = \langle -\infty, -t_{1-\alpha/2}(n-2)\rangle \cup \langle t_{1-\alpha/2}(n-2), \infty)$,
- levostranné alternativě: $W = \langle -\infty, -t_{1-\alpha}(n-2)\rangle$,
- pravostranné alternativě: $W = \langle t_{1-\alpha}(n-2), \infty)$.

H_0 zamítáme na hladině významnosti α, když $t_0 \in W$.

Příklad: Testování hypotézy o nezávislosti proti oboustranné alternativě

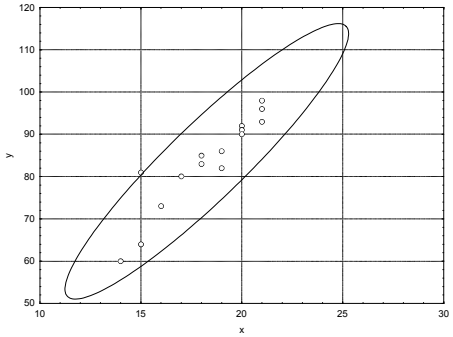
V dílně pracuje 15 dělníků. Byl u nich zjištěn počet směn odpracovaných za měsíc (náhodná veličina X) a počet zhotovených výrobků (náhodná veličina Y):

X 20 21 18 17 20 18 19 21 20 14 16 19 21 15 15

Y 92 93 83 80 91 85 82 98 90 60 73 86 96 64 81.

Orientačně ověřte dvourozměrnou normalitu dat, vypočtěte výběrový koeficient korelace mezi X a Y a na hladině 0,01 testujte hypotézu o nezávislosti X a Y proti oboustranné alternativě.

Řešení: Dvourozměrnou normalitu dat ověříme pomocí dvourozměrného tečkového diagramu.



Vidíme, že předpoklad dvourozměrné normality je oprávněný.

Vypočteme realizace

výběrových průměrů: $m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = 18{,}267$, $m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i = 83{,}6$,

výběrových rozptylů: $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_1)^2 = 5{,}6381$, $s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - m_2)^2 = 121{,}4$,

výběrové kovariance: $s_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_1)(y_i - m_2) = 24{,}2571$,

výběrového koeficientu korelace: $r_{12} = \frac{s_{12}}{s_1 s_2} = 0{,}927$.

Realizace testové statistiky: $t_0 = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} = 8{,}912$,

kritický obor $W = (-\infty, -t_{0,995}(13)\rangle \cup \langle t_{0,995}(13), \infty) = (-\infty, -3{,}012\rangle \cup \langle 3{,}012, \infty)$.

Protože $t_0 \in W$, hypotézu o nezávislosti veličin X a Y zamítáme na hladině významnosti 0,01. S rizikem omylu nejvýše 1% jsme tedy prokázali, že mezi počtem směn odpracovaných za měsíc a počtem zhotovených výrobků existuje závislost.

Výpočet pomocí systému STATISTICA

Vytvoříme datový soubor o dvou proměnných X, Y a 15 případech. Dvourozměrnou normalitu dat ověříme pomocí dvourozměrného tečkového diagramu – viz výše.

Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměn. – X, Y – OK – na záložce Možnosti vybereme Zobrazit detailní tabulku výsledků – Výpočet.

Prom. X & prom. Y	Korelace (smeny a vyrobky.sta) Označ. korelace jsou významné na hlad. p < ,05000 (Celé případy vynechány u ChD) Průměr	Sm.Odch.	r(X,Y)	r2	t	p	N	Konst. záv.: Y	Směr. záv: Y	Konst. záv.: X	Směrnic záv.: X
X	18,26667	2,37447									
X	18,26667	2,37447	1,000000	1,000000			15	0,000000	1,000000	0,000000	1,000000
X	18,26667	2,37447									
Y	83,60000	11,01817	0,927180	0,859663	8,923795	0,000001	15	5,010135	4,302365	1,562407	0,199812
Y	83,60000	11,01817									
X	18,26667	2,37447	0,927180	0,859663	8,923795	0,000001	15	1,562407	0,199812	5,010135	4,302365
Y	83,60000	11,01817									
Y	83,60000	11,01817	1,000000	1,000000			15	0,000000	1,000000	0,000000	1,000000

Výběrový koeficient korelace se realizoval hodnotou 0,92718, testová statistika nabyla hodnoty 8,924, odpovídající p-hodnota je 0,000001, tedy na hladině významnosti 0,01 zamítáme hypotézu o nezávislosti veličin X, Y.

Výpočet pomocí systému SPSS

Analyze – Correlate – Bivariate – Variables X, Y – OK

Correlations

		X	Y
X	Pearson Correlation	1,000	,927**
	Sig. (2-tailed)		,000
	N	15	15
Y	Pearson Correlation	,927**	1,000
	Sig. (2-tailed)	,000	
	N	15	15

**. Correlation is significant at the 0.01 level (2-tailed).

Vidíme, že na rozdíl od systému STATISTICA systém SPSS neposkytuje hodnotu testové statistiky, ve výstupní tabulce lze najít pouze realizaci výběrového koeficientu korelace, odpovídající p-hodnotu a rozsah náhodného výběru.

Interval spolehlivosti pro korelační koeficient

Náhodná veličina $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ má přibližně normální rozložení se střední hodnotou $E(Z) = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$ (2. sčítanec lze při větším n zanedbat) a rozptylem $D(Z) = \frac{1}{n-3}$. Standardizací veličiny Z dostaneme veličinu $U = \frac{Z - E(Z)}{\sqrt{D(Z)}}$, která má asymptoticky rozložení N(0,1). Tudíž 100(1-α)% asymptotický interval spolehlivosti pro $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ bude mít meze $Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}$. Interval spolehlivosti pro ρ pak dostaneme zpětnou transformací.

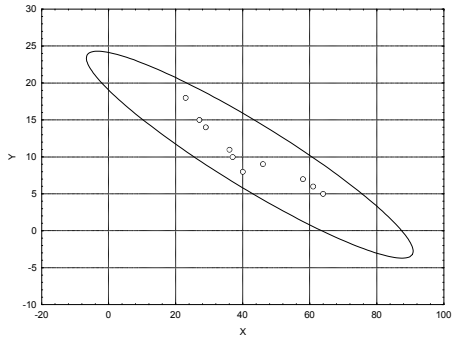
Poznámka: Jelikož Z = arctgh R_{12}, dostáváme R_{12} = tgh Z a meze intervalu spolehlivosti pro ρ můžeme psát ve tvaru $\text{tgh}\left(Z \pm \frac{u_{1-\alpha/2}}{\sqrt{n-3}}\right)$, přičemž $\text{tgh}\, x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

Příklad: Pracovník personálního oddělení určité firmy zkoumá, zda existuje vztah mezi počtem dní absence za rok (veličina Y) a věkem pracovníka (veličina X). Proto náhodně vybral údaje o 10 pracovnících.

Č.prac.	1	2	3	4	5	6	7	8	9	10
X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

Za předpokladu, že uvedené údaje tvoří číselné realizace náhodného výběru rozsahu 10 z dvourozměrného normálního rozložení, vypočtěte výběrový korelační koeficient a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny. Sestrojte 95% asymptotický interval spolehlivosti pro skutečný korelační koeficient ρ.

Řešení: Předpoklad o dvourozměrné normalitě dat ověříme orientačně pomocí dvourozměrného tečkového diagramu.



Vzhled diagramu svědčí o tom, že předpoklad je oprávněný.

Testujeme H_0: $\rho = 0$ proti H_1: $\rho \neq 0$. Vypočítáme $R_{12} = -0{,}9325$, tedy mezi věkem pracovníka a počtem dnů pracovní neschopnosti existuje silná nepřímá lineární závislost. Testová statistika: $T = -7{,}3053$, kvantil $t_{0,975}(8) = 2{,}306$, kritický obor $W = (-\infty, -2{,}306\rangle \cup \langle 2{,}306, \infty)$. Jelikož $T \in W$, zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y.

Vypočítáme $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}} = \frac{1}{2}\ln\frac{1-0{,}9325}{1+0{,}9325} = -1{,}6772$. Meze 95% asymptotického intervalu spolehlivosti pro ρ jsou $\mathrm{tgh}\left(-1{,}6772 \pm \frac{1{,}96}{\sqrt{7}}\right)$, tedy -0,9842 < ρ < -0,7336 s pravděpodobností přibližně 0,95.

Výpočet pomocí systému STATISTICA:

Ve STATISTICE vypočteme meze 100(1-α)% asymptotického intervalu spolehlivosti pro koeficient korelace ρ tak, že otevřeme nový datový soubor se dvěma proměnnými (pojmenujeme je DM a HM) a jedním případem.
Do Dlouhého jména proměnné DM zapíšeme příkaz
= TanH(0,5*log((1-0,9325)/(1+0,9325))-VNormal(0,975;0;1)/sqrt(7))
a do Dlouhého jména proměnné HM zapíšeme příkaz
= TanH(0,5*log((1-0,9325)/(1+0,9325))+VNormal(0,975;0;1)/sqrt(7))

	1 DM	2 HM
1	-0,98425	-0,73358

95% asymptotický interval spolehlivosti pro koeficient korelace ρ má tedy meze –0,98425 a -0,73358. (Protože nepokrývá hodnotu 0, zamítáme hypotézu o nezávislosti veličin X, Y na asymptotické hladině významnosti 0,05.)

Výpočet pomocí systému SPSS

Vzhledem k tomu, že v SPSS není dostupná funkce TanH, využijeme toho, že $\mathrm{tgh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Dále pro výpočet kvantilů rozložení N(0,1) použijeme funkci IDF.NORMAL(x, mean, std).
K datovému souboru přidáme 4 proměnné pojmenované pom1, pom2, dm, hm.
Transform – Compute Variable – Target Variable pom1 =
0.5*ln((1-0.9325)/(1+0.9325))-IDF.Normal(0.975,0,1)/sqrt(7)
OK – OK
Transform – Compute Variable – Target Variable pom2 =
0.5*ln((1-0.9325)/(1+0.9325))+IDF.Normal(0.975,0,1)/sqrt(7)
OK - OK
Transform – Compute Variable – Target Variable dm =
(exp(pom1)-exp(-pom1))/(exp(pom1)+exp(-pom1))
OK - OK

Transform – Compute Variable – Target Variable hm =
(exp(pom2)-exp(-pom2))/(exp(pom2)+exp(-pom2))
OK – OK
Dostaneme výsledek dm = –0,9842 a hm = -0,7336.

Varianční, kovarianční a korelační matice

Nechť $\mathbf{X} = (X_1, \ldots, X_p)'$ je náhodný vektor. Označme
$\mu_i = E(X_i)$ střední hodnotu náhodné veličiny X_i,
$\sigma_i^2 = D(X_i)$ rozptyl náhodné veličiny X_i,
$\sigma_{ij} = C(X_i, X_j)$ kovarianci náhodných veličin X_i, X_j
$\rho_{ij} = R(X_i, X_j)$ koeficient korelace náhodných veličin X_i, X_j

Vektor $E(\mathbf{X}) = (\mu_1, \ldots, \mu_p)'$ se nazývá vektor středních hodnot náhodného vektoru **X**.
Čtvercová matice řádu p $var(\mathbf{X}) = (\sigma_{ij})_{i,j=1,\ldots,p}$ se nazývá varianční matice náhodného vektoru **X**.
Čtvercová matice řádu p $cor(\mathbf{X}) = (\rho_{ij})_{i,j=1,\ldots,p}$ se nazývá korelační matice náhodného vektoru **X**.
Je zřejmé, že varianční matice a korelační matice jsou symetrické.

Nechť $\mathbf{X} = (X_1, \ldots, X_p)'$ a $\mathbf{Y} = (Y_1, \ldots, Y_q)'$ jsou náhodné vektory.
Matice typu pxq $cov(\mathbf{X},\mathbf{Y}) = (C(X_i, Y_j))$ se nazývá kovarianční matice vektorů **X**, **Y**.
Matice typu pxq $cor(\mathbf{X},\mathbf{Y}) = (\rho(X_i, Y_j))$ se nazývá korelační matice vektorů **X**,**Y**.

Odhady vektoru středních hodnot, varianční a korelační matice jednoho náhodného vektoru X

Nechť **X** je náhodný vektor, který má p-rozměrné rozložení s vektorem středních hodnot **μ**, varianční maticí $var(\mathbf{X})$ a korelační maticí $cor(\mathbf{X})$. Nechť je dán náhodný výběr $\mathbf{X}_1 = (X_{11}, \ldots, X_{1p})', \ldots, \mathbf{X}_n = (X_{n1}, \ldots, X_{np})'$ rozsahu n z tohoto rozložení.
Nestranný odhad vektoru **μ** je vektor výběrových průměrů $\mathbf{M} = (M_1, \ldots, M_p)'$,
kde $M_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ je výběrový průměr j-tého výběru, j = 1, …, p.
Nestranný odhad matice $var(\mathbf{X})$ je výběrová varianční matice $\mathbf{S} = (S_{ij}) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{M})(\mathbf{X}_i - \mathbf{M})'$ řádu p.
Vychýlený odhad matice $cor(\mathbf{X})$ je výběrová korelační matice $\mathbf{R} = (R_{ij})$, kde R_{ij} je výběrový korelační koeficient i-té a j-té složky vektoru **X**, tedy
$R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}$, i, j = 1, …, p. (Je zřejmé, že diagonální prvky matice **R** jsou jedničky a matice **R** je symetrická.)

Odhady kovarianční a korelační matice dvou náhodných vektorů X, Y

Nechť náhodný vektor $\mathbf{X}$ má p-rozměrné rozložení a nechť $\mathbf{X}_1, \ldots, \mathbf{X}_n$ je náhodný výběr z tohoto rozložení. Nechť náhodný vektor $\mathbf{Y}$ má q-rozměrné rozložení a nechť $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ je náhodný výběr z tohoto rozložení. Předpokládejme, že obě rozložení mají konečné druhé momenty. Nechť cov($\mathbf{X}$, $\mathbf{Y}$) je kovarianční matice těchto vektorů a cor($\mathbf{X}$, $\mathbf{Y}$) je korelační matice těchto vektorů. Označme

$$M_{Xj} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, j = 1,\ldots,p, M_{Yj} = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}, j = 1,\ldots,q,$$

$\mathbf{M_X} = (M_{X1}, \ldots, M_{Xp})'$, $\mathbf{M_Y} = (M_{Y1}, \ldots, M_{Yq})'$.

Nestranným odhadem kovarianční matice cov($\mathbf{X}$, $\mathbf{Y}$) vektorů $\mathbf{X}$, $\mathbf{Y}$ je výběrová kovarianční matice vektorů $\mathbf{X}$, $\mathbf{Y}$ definovaná vzorcem $\mathbf{S_{XY}} = (S_{ij}) =$

$$\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \mathbf{M_X})(\mathbf{Y}_i - \mathbf{M_Y})', i = 1, \ldots, p, j = 1, \ldots, q.$$

Vychýleným odhadem korelační matice cor($\mathbf{X}$, $\mathbf{Y}$) vektorů $\mathbf{X}$, $\mathbf{Y}$ je výběrová korelační matice vektorů $\mathbf{X}$, $\mathbf{Y}$ definovaná vzorcem $\mathbf{R_{XY}} = (R_{ij})$, kde R_{ij} je výběrový korelační koeficient i-té a j-té složky vektorů $\mathbf{X}$, $\mathbf{Y}$, i = 1, …, p, j = 1, …, q.

Koeficient mnohonásobné korelace a výběrový koeficient mnohonásobné korelace

Intenzitu lineární závislosti mezi náhodnou veličinou Y a náhodným vektorem $\mathbf{X} = (X_1, \ldots, X_p)'$ měříme pomocí koeficientu mnohonásobné korelace $\rho_{Y.\mathbf{X}}$. Jeho druhá mocnina je dána vzorcem

$$\rho_{Y.\mathbf{X}}^2 = \mathrm{cor}(Y, \mathbf{X})\, \mathrm{cor}(\mathbf{X})^{-1}\, \mathrm{cor}(\mathbf{X}, Y).$$

Má tyto vlastnosti:

a) $\rho_{Y.\mathbf{X}} \geq 0$

b) $\rho_{Y.\mathbf{X}} \geq |\rho(Y, X_i)|$ pro $\forall i = 1,\ldots,p$

c) $\rho_{Y.X_1\ldots X_p} \geq \ldots \geq \rho_{Y.X_1X_2} \geq \rho(Y, X_1)$

d) $\rho_{Y.\mathbf{X}} = 1 \Leftrightarrow$ existují konstanty $\beta_0, \beta_1, \ldots, \beta_p$ tak, že $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

Nechť náhodný vektor $(Y, X_1, \ldots, X_p)'$ má (p+1)-rozměrné rozložení s koeficientem mnohonásobné korelace $\rho_{Y.\mathbf{X}}$.

Nechť je dán náhodný výběr $(Y_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_n, X_{n1}, \ldots, X_{np})'$ rozsahu n z tohoto rozložení. Pak jako odhad $\rho_{Y.\mathbf{X}}$ slouží výběrový koeficient mnohonásobné korelace $r_{Y.\mathbf{X}}$, jehož druhá mocnina je dána vzorcem

$$r_{Y.\mathbf{X}}^2 = R_{Y\mathbf{X}}\, \mathbf{R}^{-1}\, R_{\mathbf{X}Y},$$

kde $R_{Y\mathbf{X}}$ je výběrová korelační matice veličiny Y a vektoru $\mathbf{X}$ (v tomto případě se redukuje na vektor $(r_{YX_1}, \ldots, r_{YX_p})$) a $\mathbf{R}$ je výběrová korelační matice vektoru $\mathbf{X}$.

Vlastnosti (a), (b), (c), (d) koeficientu mnohonásobné korelace se přenášejí i na výběrový koeficient mnohonásobné korelace.

Testování hypotézy o nezávislosti veličiny Y a vektoru X

Popis testu

Nechť náhodný výběr $(Y_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_n, X_{n1}, \ldots, X_{np})'$ pochází z (p+1)-rozměrného normálního rozložení, které má koeficient mnohonásobné korelace $\rho_{Y.\mathbf{X}}$. Musí platit $n > p+1$.
Testujeme hypotézu H_0: $\rho_{Y.\mathbf{X}} = 0$ proti H_1: $\rho_{Y.\mathbf{X}} \neq 0$. Vzhledem k tomu, že se jedná o výběr z (p+1)-rozměrného normálního rozložení, testujeme, zda existuje závislost mezi veličinou Y a vektorem **X**. (Je-li $\rho_{Y.\mathbf{X}} = 0$, pak z vlastnosti (b) plyne, že $\rho(Y,X_i) = 0$ pro všechna i = 1, …, p, tudíž náhodné veličiny Y a X_i jsou stochasticky nezávislé pro všechna i = 1, …, p.)

Testová statistika $F = \frac{n-p-1}{p} \cdot \frac{r_{Y.\mathbf{X}}^2}{1-r_{Y.\mathbf{X}}^2}$ se řídí rozložením F(p, n-p-1), pokud H_0 platí. Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$. Jestliže $F \in W$, H_0 zamítáme na hladině významnosti α.

Příklad

Při zkoumání závislosti hodinové výkonnosti dělníka (veličina Y – v kusech) na jeho věku (veličina X_1 – v letech) a době zapracovanosti (veličina X_2 – v letech) byly u 10 náhodně vybraných dělníků zjištěny tyto údaje:

Y	67	65	75	66	77	84	69	60	70	66
X_1	43	40	49	46	41	41	48	34	32	42
X_2	6	8	14	14	8	12	16	1	5	7

Za předpokladu, že uvedené hodnoty představují číselné realizace náhodného výběru rozsahu 10 ze třírozměrného normálního rozložení, testujte na hladině významnosti 0,05 hypotézu, že výkon dělníka nezávisí na jeho věku a době zapracovanosti.

Řešení:

$r_{YX_1} = 0{,}2287, r_{YX_2} = 0{,}4538, r_{X_1X_2} = 0{,}847$, $R_{Y\mathbf{X}} = (0{,}2287,\ 0{,}4538)'$,

$$\mathbf{R} = \begin{pmatrix} 1 & 0{,}847 \\ 0{,}847 & 1 \end{pmatrix},$$

$$r_{Y.\mathbf{X}}^2 = (0{,}2287,\ 0{,}4538) \begin{pmatrix} 1 & 0{,}847 \\ 0{,}847 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0{,}2287 \\ 0{,}4538 \end{pmatrix} = 0{,}2917$$

Interpretace: Variabilita výkonů dělníků je z 29% vysvětlena jejich věkem a dobou zapracovanosti.

$r_{Y.\mathbf{X}} = \sqrt{0{,}2917} = 0{,}5401$.

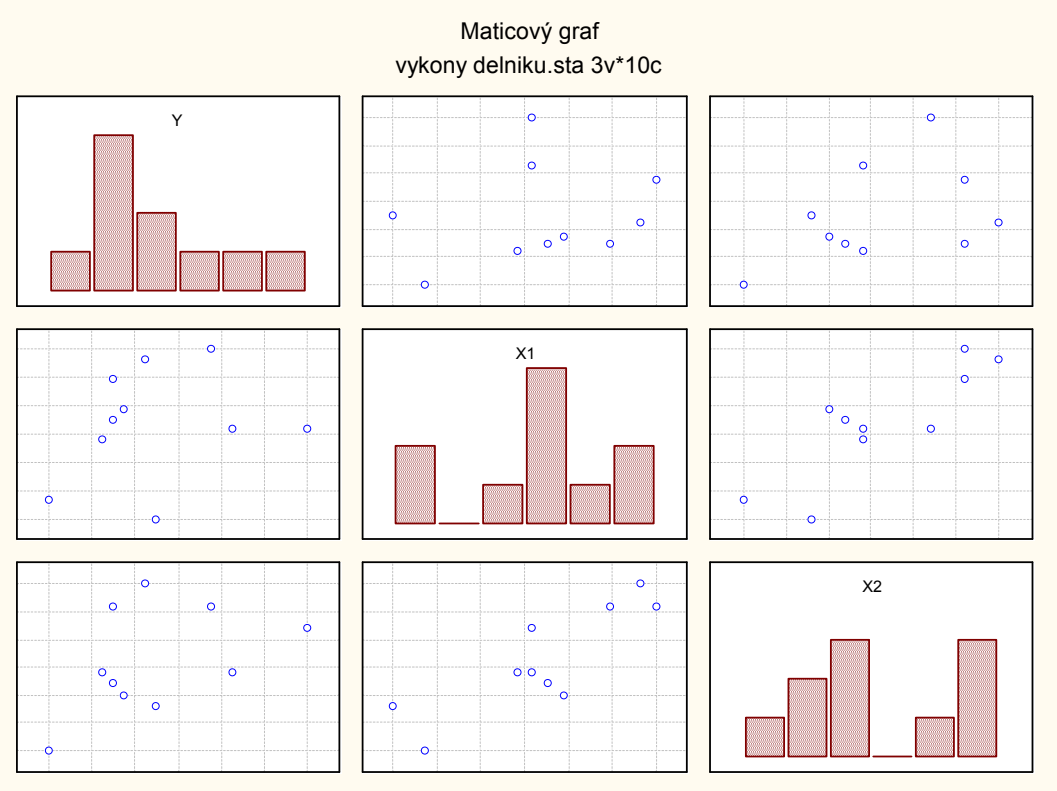
Testujeme hypotézu H_0: $\rho_{Y.\mathbf{X}} = 0$ proti H_1: $\rho_{Y.\mathbf{X}} \neq 0$.

Testová statistika: $F = \frac{10-2-1}{2} \cdot \frac{0{,}2917}{1-0{,}2917} = 1{,}441$, kvantil $F_{0,95}(2,7) = 4{,}737$, kritický obor $W = \langle 4{,}737, \infty)$, tedy $F \notin W$ a H_0 nezamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Závislost mezi dvojicemi proměnných posoudíme pomocí dvourozměrných tečkových diagramů:

Grafy – Maticové grafy – Proměnné – Vybrat vše – OK



Ve všech třech případech existuje mezi dvojicemi proměnných určitý stupeň přímé lineární závislosti.

Statistiky – Vícenásobná regrese – Proměnné – Závislá proměnná Y, seznam nezáv. proměnných X1, X2 – OK – OK.

Koeficient $r_{Y,(X_1,X_2)}$ najdeme v záhlaví výstupní tabulky pod označením Vícenás.

R = 0,54

Hodnota testové statistiky pro test nevýznamnosti koeficientu mnohonásobné korelace $\rho_{Y,(X_1,X_2)}$ je 1,4411, počet stupňů volnosti čitatele je 2, jmenovatele 7, odpovídající p-hodnota je 0,2991, tedy na hladině významnosti 0,05 nezamítáme hypotézu, že výkon dělníka není závislý na jeho věku a době zapracovanosti.

Výpočet pomocí systému SPSS

Analyze – Regression – Linear – Dependent Y, Independent X1, X2 – OK

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,540[a]	,292	,089	6,649

a. Predictors: (Constant), doba zapracovanosti, vek

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	127,425	2	63,712	1,441	,299[a]
	Residual	309,475	7	44,211		
	Total	436,900	9			

a. Predictors: (Constant), doba zapracovanosti, vek

b. Dependent Variable: vykon delnika

Koeficient parciální korelace

Nechť Y, Z jsou náhodné veličiny a $\mathbf{X} = (X_1, \ldots, X_p)'$ je náhodný vektor. Korelační koeficient $\rho(Y,Z)$ udává míru těsnosti lineárního vztahu mezi veličinami Y a Z. Ta však může být ovlivněna i tím, že mezi veličinami $X_1, \ldots, X_p$ existují veličiny, které silně korelují jak s Y, tak se Z. Zajímá nás proto, jaká je „čistá“ korelace mezi Y a Z, když se eliminuje vliv náhodného vektoru **X**.
Pokud se omezíme na lineární vztahy, můžeme vliv vektoru **X** na veličinu Y popsat lineární regresní funkcí

$\hat{Y} = \alpha + \boldsymbol{\beta}'\mathbf{X}$, kde $\boldsymbol{\beta} = \mathrm{var}(\mathbf{X})^{-1}\,\mathrm{cov}(\mathbf{X},Y)$, $\alpha = E(Y) - \boldsymbol{\beta}'E(\mathbf{X})$.

Tu část veličiny Y, kterou vektor **X** nevysvětlí, si můžeme představit jako reziduum $Y - \hat{Y}$. Analogicky pro veličinu Z dostáváme

$\hat{Z} = \gamma + \boldsymbol{\delta}'\mathbf{X}$, kde $\boldsymbol{\delta} = \mathrm{var}(\mathbf{X})^{-1}\,\mathrm{cov}(\mathbf{X},Z)$, $\gamma = E(Z) - \boldsymbol{\delta}'E(\mathbf{X})$,

tudíž reziduum $Z - \hat{Z}$ chápeme jako tu část veličiny Z, kterou vektor **X** nevysvětlí.
Korelační koeficient mezi rezidui $Y - \hat{Y}$ a $Z - \hat{Z}$ se nazývá parciální korelační koeficient mezi náhodnými veličinami Y a Z při pevně daném vektoru **X** a značí se $\rho_{Y,Z.\mathbf{X}}$. Tedy $\rho_{Y,Z.\mathbf{X}} = \rho(Y - \hat{Y}, Z - \hat{Z})$. Počítá se podle vzorce

$$\rho_{Y,Z.\mathbf{X}} = \frac{\rho(Y,Z) - \mathrm{cov}(Y,\mathbf{X})\mathrm{cor}(\mathbf{X})^{-1}\,\mathrm{cov}(\mathbf{X},Z)}{\sqrt{\left[1 - \mathrm{cov}(Y,\mathbf{X})\mathrm{cor}(\mathbf{X})^{-1}\,\mathrm{cov}(\mathbf{X},Y)\right]\left[1 - \mathrm{cov}(Z,\mathbf{X})\mathrm{cor}(\mathbf{X})^{-1}\,\mathrm{cov}(\mathbf{X},Z)\right]}}.$$

Nechť náhodný vektor $(Y, Z, X_1, \ldots, X_p)'$ pochází z (p+2)-rozměrného rozložení, které má parciální korelační koeficient $\rho_{Y,Z.\mathbf{X}}$. Nechť je dán náhodný výběr

$(Y_1, Z_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_n, Z_n, X_{n1}, \ldots, X_{np})'$ rozsahu n z tohoto rozložení. Musí platit n > p+2. Jako odhad $\rho_{Y,Z.\mathbf{X}}$ slouží výběrový parciální korelační koeficient $r_{Y,Z.\mathbf{X}}$:

$$r_{Y,Z.\mathbf{X}} = \frac{r_{YZ} - \mathbf{S}_{Y\mathbf{X}}\mathbf{R}_{\mathbf{XX}}^{-1}\mathbf{S}_{\mathbf{X}Z}}{\sqrt{\left[1 - \mathbf{S}_{Y\mathbf{X}}\mathbf{R}_{\mathbf{XX}}^{-1}\mathbf{S}_{\mathbf{X}Y}\right]\left[1 - \mathbf{S}_{Z\mathbf{X}}\mathbf{R}_{\mathbf{XX}}^{-1}\mathbf{S}_{\mathbf{X}Z}\right]}}$$

Testování hypotézy o nezávislosti veličin Y a Z při eliminaci vlivu vektoru X

Popis testu

Budeme předpokládat, že uvedený náhodný výběr pochází z (p+2)-rozměrného normálního rozložení. Testujeme hypotézu H_0: $\rho_{y,z.\mathbf{x}} = 0$ proti H_1: $\rho_{y,z.\mathbf{x}} \neq 0$. Vzhledem k tomu, že se jedná o výběr z normálního rozložení, testujeme, zda existuje závislost mezi Y a Z při eliminaci vlivu **X**.

Testová statistika $T_0 = \frac{r_{Y,Z.\mathbf{X}}\sqrt{n-p-2}}{\sqrt{1 - r_{Y,Z.\mathbf{X}}^2}}$ se řídí rozložením t(n-p-2), pokud H_0 platí.

Kritický obor: $W = \left(-\infty, t_{1-\alpha/2}(n-p-2)\right\rangle \cup \left\langle t_{1-\alpha/2}(n-p-2), \infty\right)$. Jestliže $T_0 \in W$, H_0 zamítáme na hladině významnosti α

Příklad

Pro data z příkladu o výkonnosti dělníků vypočtěte výběrové parciální korelační koeficienty $r_{Y,X_1.X_2}, r_{Y,X_2.X_1}$, interpretujte je, porovnejte je s obyčejnými výběrovými korelačními koeficienty r_{YX_1}, r_{YX_2} a pro α = 0,05 otestujte významnost uvedených parciálních korelačních koeficientů.

Řešení:

$r_{YX_1} = 0{,}2287$, $r_{YX_2} = 0{,}4538$, $r_{X_1X_2} = 0{,}847$

a) $r_{Y,X_1.X_2} = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{\sqrt{\left(1 - r_{YX_2}^2\right)\left(1 - r_{X_1X_2}^2\right)}} = -0{,}3286$

Interpretace: Korelační koeficient mezi výkonem a věkem vyšel 0,2287, tedy s rostoucím věkem roste výkon. Parciální korelační koeficient mezi výkonem a věkem při vyloučení vlivu doby zapracovanosti vyšel -0,3286, tedy u dělníků se stejnou dobou zapracovanosti klesá s rostoucím věkem výkon.

Testujeme H_0: $\rho_{Y,X_1.X_2} = 0$ proti H_1: $\rho_{Y,X_1.X_2} \neq 0$

Testová statistika: $T_0 = \frac{r_{Y,X_1.X_2}\sqrt{n-p-2}}{\sqrt{1 - r_{Y,X_1.X_2}^2}} = \frac{-0{,}3286\sqrt{7}}{\sqrt{1 - 0{,}3286^2}} = -0{,}9205$

Kvantil: $t_{0,975}(7) = 2{,}3646$
Kritický obor: $W = (-\infty, -2{,}3646\rangle \cup \langle 2{,}3646, \infty)$
$T_0 \notin W \Rightarrow H_0$ nezamítáme na hladině významnosti 0,05.

b) $r_{Y,X_2.X_1} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{(1 - r_{YX_1}{}^2)(1 - r_{X_1X_2}{}^2)}} = 0{,}5026$

Interpretace: Korelační koeficient mezi výkonem a dobou zapracovanosti vyšel 0,4538, tedy čím delší doba zapracovanosti, tím lepší výkon dělník podává. Parciální korelační koeficient mezi výkonem a dobou zapracovanosti při vyloučení vlivu věku vyšel 0,5026, tedy u stejně starých dělníků je poněkud silnější přímá lineární vazba mezi výkonem a dobou zapracovanosti.

Testujeme H_0: $\rho_{Y,X_2.X_1} = 0$ proti H_1: $\rho_{Y,X_2.X_1} \neq 0$

Testová statistika: $T_0 = \frac{r_{Y,X_2.X_1}\sqrt{n-p-2}}{\sqrt{1 - r_{Y,X_2.X_1}{}^2}} = \frac{0{,}5026\sqrt{7}}{\sqrt{1 - 0{,}5026^2}} = 1{,}538$

Kvantil: $t_{0,975}(7) = 2{,}3646$
Kritický obor: $W = (-\infty, -2{,}3646\rangle \cup \langle 2{,}3646, \infty)$
$T_0 \notin W \Rightarrow H_0$ nezamítáme na hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA

Statistiky – Základní statistiky/tabulky – Korelační matice – OK – na záložce Možnosti zaškrtneme Zobrazit r, úrovně p, počty N a zaškrtneme Zobrazit dlouhá jména proměnných, na záložce Detaily zvolíme Parciální korelace – 1. seznam proměnných Y, X1, druhý seznam proměnných X2 – OK

	Parciální korelace (vykony delniku) Označ. korelace jsou významné na hlad. p < ,05000 N=10 (Celé případy vynechány u ChD)	
Proměnná	Y	X1
Y: vykon delnika	1,0000	-,3286
	p= ---	p=,388
X1: vek	-,3286	1,0000
	p=,388	p= ---

Vidíme, že výběrový parciální korelační koeficient $r_{Y,X_1.X_2}$ je -0,3286, p-hodnota je 0,388, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nevýznamnosti $\rho_{Y,X_1.X_2}$.

Analogicky 1. seznam proměnných Y, X2, druhý seznam proměnných X1 – OK

Proměnná	Parciální korelace (vykony delniku) Označ. korelace jsou významné na hlad. p < ,05000 N=10 (Celé případy vynechány u ChD) Y	X2
Y: vykon delnika	1,0000	,5026
	p= ---	p=,168
X2: doba zapracovanosti	,5026	1,0000
	p=,168	p= ---

V tomto případě výběrový parciální korelační koeficient $r_{Y,X_2.X_1}$ je 0,5026, p-hodnota je 0,168, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nevýznamnosti $\rho_{Y,X_2.X_1}$.

Výpočet pomocí systému SPSS

Analyze – Correlate – Partial – Variables Y, X1 – Controlling for X2 – OK

Correlations

Control Variables			vykon delnika	vek
doba zapracovanosti	vykon delnika	Correlation	1,000	-,329
		Significance (2-tailed)	.	,388
		df	0	7
	vek	Correlation	-,329	1,000
		Significance (2-tailed)	,388	.
		df	7	0