Vícerozměrná rozdělení, odhady a testy měr a modelů „závislosti“

Dvourozměrné normální rozdělení

Hustota:

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right]$$

její marginály: $f(x_i)=\frac{1}{\sqrt{2\pi}\,\sigma_i}\exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right);i=1,2$ a odtud podmíněná rozdělení:

$$f(x_1/x_2)=\frac{f(x_1,x_2)}{f(x_2)}=\frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}\exp\left[-\frac{\left(x_1-\left(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)\right)\right)^2}{2\sigma_1^2(1-\rho^2)}\right]$$

$$f(x_2/x_1)=\frac{f(x_1,x_2)}{f(x_1)}=\frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}}\exp\left[-\frac{\left(x_2-\left(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\right)\right)^2}{2\sigma_2^2(1-\rho^2)}\right]$$

$$E(x_1/x_2)=\mu_1+\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2) \qquad \sigma^2(x_1/x_2)=\sigma_1^2(1-\rho^2)$$

Proto: $E(x_2/x_1)=\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)$ a $\sigma^2(x_2/x_1)=\sigma_2^2(1-\rho^2)$.

Normované a centrované dvourozměrné normální rozdělení: $\mu_1=\mu_2=0;\sigma_1=\sigma_2=1$

$$f(x_1,x_2)=\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(x_1^2+x_2^2-2\rho x_1x_2\right)\right]$$

Námět: Odvoďte marginály a podmíněná rozdělení pro normované a centrované normální rozdělení.

„Lineární“ míry „závislosti“ (korelace):

$E\{(x_1-\mu_1)^2\}=\sigma_1^2$, $E\{(x_2-\mu_2)^2\}=\sigma_2^2$ a $E\{(x_1-\mu_1)(x_2-\mu_2)\}=Cov(x_1,x_2)$

$\rho(x_1,x_2)=\frac{Cov(x_1,x_2)}{\sigma_1\sigma_2}\Rightarrow 0\le|\rho(x_1,x_2)|\le 1$.

Námět: dokažte tento vztah. Návod užijte Cauchy-Schwarzovy nerovnosti $\left|\sum_i x_iz_i\right|\le\sqrt{\sum_i x_i^2}\sqrt{\sum_i z_i^2}$ nebo její integrální tvar.

Pokud $\rho(x_1, x_2) = 0$ a x_1, x_2 **jsou normálně rozdělené (sdruženě),** jsou x_1, x_2 nezávislé.
Důkaz: Po dosazení $\rho = 0$ do

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} \right) \right]$$

dostaneme: $f(x_1, x_2) = f(x_1) f(x_2)$.

Poznámka: v tomto tvrzení je podstatný předpoklad **normálního rozdělení.** Příklad, kdy neplatí taková implikace lze nalézt v Alfréd Rényi: Teorie pravděpodobnosti, ACADEMIA, Praha 1972, str. 111.

Další vlastnosti korelačního koeficientu lze nalézt v: Alfréd Rényi: Teorie pravděpodobnosti, ACADEMIA, Praha 1972, str. 109-113.

Bodový odhad korelačního koeficientu

Mějme náhodný výběr z dvourozměrného (normálního = gaussovského) rozdělení

$$\begin{pmatrix} x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{1,n} \\ x_{2,1}, x_{2,2}, x_{2,3}, \ldots, x_{2,n} \end{pmatrix}, \text{ kde } \begin{aligned} \overline{x_1} &= \frac{1}{n}\sum_{i=1}^{n} x_{1,i} \\ \overline{x_2} &= \frac{1}{n}\sum_{i=1}^{n} x_{2,i} \end{aligned}$$

pak pod výběrovým koeficientem korelace bude rozuměno číslo:

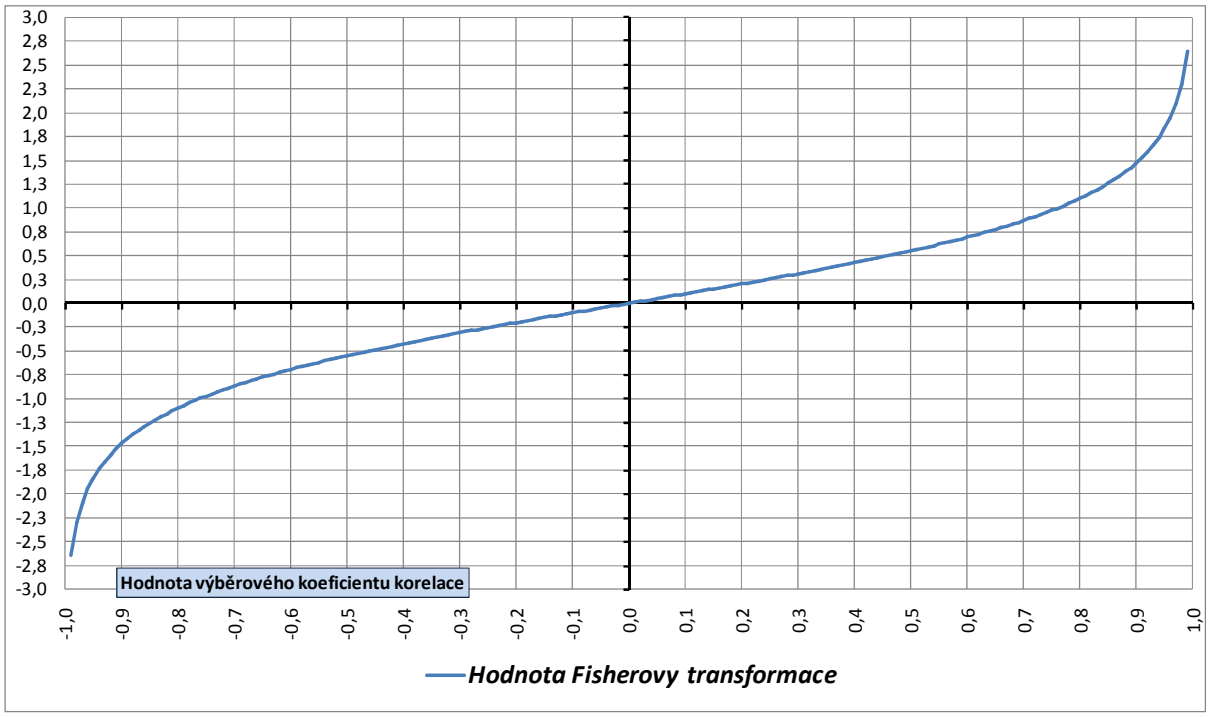
$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_{1,i} - \overline{x_1})(x_{2,i} - \overline{x_2})}{\sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} (x_{1,i} - \overline{x_1})^2 \right)} \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} (x_{2,i} - \overline{x_2})^2 \right)}} = \frac{\sum_{i=1}^{n} (x_{1,i} - \overline{x_1})(x_{2,i} - \overline{x_2})}{\sqrt{\left(\sum_{i=1}^{n} (x_{1,i} - \overline{x_1})^2 \right)\left(\sum_{i=1}^{n} (x_{2,i} - \overline{x_2})^2 \right)}}$$

Rozdělení výběrového koeficientu korelace je poměrně komplikované (viz Jaroslav Hátle, Jiří Likeš: Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974. Str. 183-184).

Proto navrhl R. A. Fisher transformaci: $z = \frac{1}{2}\lg\frac{1+r}{1-r}$, kdy pro transformovanou náhodnou proměnnou platí, pro dostatečně velké ***n*** *(n>10)* a $|\rho| < 1$ normalita (asymptotická) se střední hodnotou a rozptylem:

$$E\{z\} = \frac{1}{2}\lg\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}; \qquad \sigma^2\{z\} = \frac{1}{n-3}$$

Specielně v případě $\rho = 0 \Rightarrow E\{z\} = 0;\ \sigma^2\{z\} = \frac{1}{n-3}$.



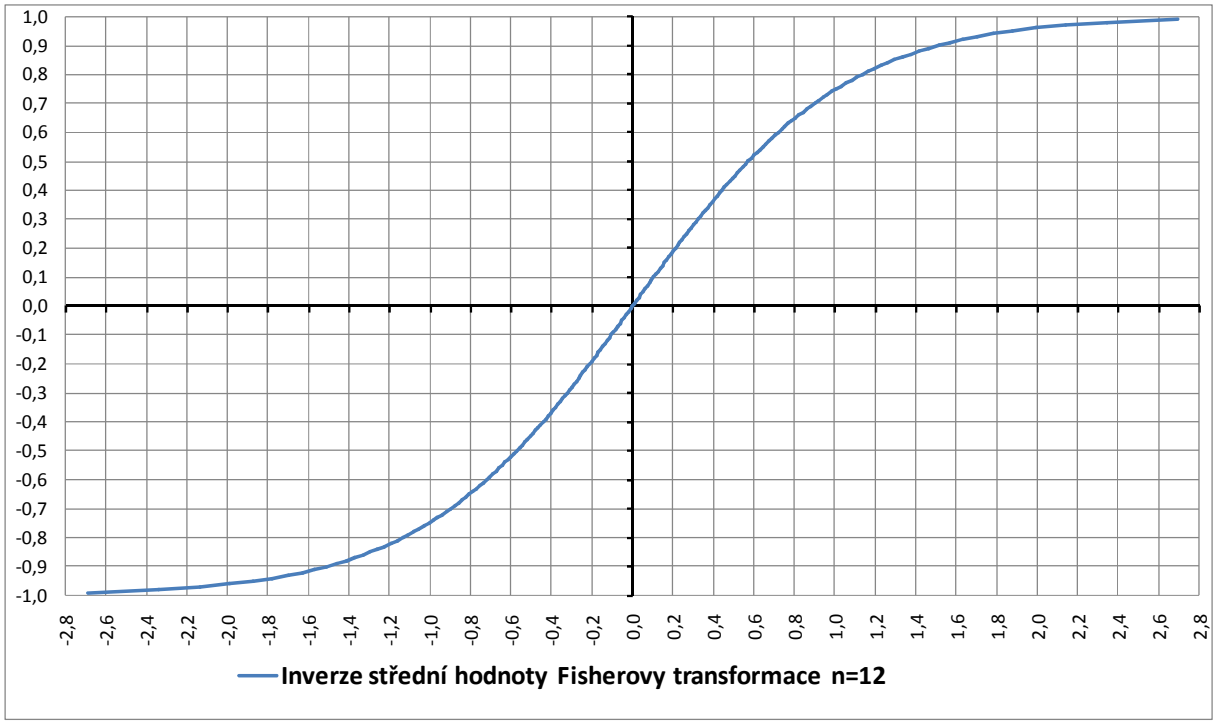
Odtud plyne pro:

Intervalový odhad koeficientu korelace

pro $\alpha = \alpha_1 + \alpha_2$

$$\frac{1}{2}\lg\frac{1+r}{1-r} - u_{1-\alpha_2}\frac{1}{\sqrt{n-3}} < \frac{1}{2}\lg\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)} < \frac{1}{2}\lg\frac{1+r}{1-r} + u_{1-\alpha_1}\frac{1}{\sqrt{n-3}}$$

Uvedené nerovnosti se vyřeší inverzí funkce $f_n(\rho) = \frac{1}{2}\lg\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$, jejíž průběh a průběh její inverze následuje:



Proto: $f_n^{-1}\left(\frac{1}{2}\lg\frac{1+r}{1-r}-u_{1-\alpha_2}\frac{1}{\sqrt{n-3}}\right)<\rho<f_n^{-1}\left(\frac{1}{2}\lg\frac{1+r}{1-r}+u_{1-\alpha_1}\frac{1}{\sqrt{n-3}}\right)$ je $100\alpha\%$ interval spolehlivosti pro koeficient korelace ρ. Pro dostatečně velká n lze člen $\frac{\rho}{2(n-1)}$ ve výrazu pro střední hodnotu zanedbat, pak není nutné dané nerovnosti řešit numericky a lze nalézt analytický výraz pro $f_n^{-1}(z)=f^{-1}(z)$. **Námět:** nalezněte takové vyjádření.

Test hypotézy o koeficientu korelace $H:\rho=0$ = test o nekorelovanosti

V Jaroslav Hátle, Jiří Likeš: Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974. Str. 183 je pro rozebíraný případ $\rho=0$ dokázáno, že statistika $t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ má Studentovo ***t-rozdělení*** s ***n-2*** stupni volnosti. To dává kritické obory:

Pravostranná alternativa	$\rho>0$	$W_\alpha=\left\{t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}};t\geq t_{1-\alpha}(n-2)\right\}$
Levostranná alternativa	$\rho<0$	$W_\alpha=\left\{t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}};t\leq -t_{1-\alpha}(n-2)\right\}$
Oboustranná alternativa	$\rho\neq 0$	$W_\alpha=\left\{t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}};\lvert t\rvert\leq t_{1-\alpha/2}(n-2)\right\}$

Námět: Lze efektivně odhadnout chybu druhého druhu a prověřovat výše uvedené testy na jejich sílu?

Námět: Odvoďte srovnatelné testy využívající Fisherovy transformace (fakticky se bude jednat o testy o střední hodnotě) a řešte problém uvedený v předchozím námětu.

Klasický „Pearsonův“ koeficient korelace „měří sílu“ případné „lineární závislosti“ viz Alfréd Rényi : Teorie pravděpodobnosti, ACADEMIA, Praha 1972, str. 109-113. Pokud potřebujeme „měřit sílu monotónní závislosti“ používá se více typů měr. Jednou z nich je „Spearmanův koeficient pořadové korelace“.

Spearmanův koeficient pořadové korelace

Jedná se o „měření závislosti“ založené na rankových (pořadových) statistikách.
Rankové statistiky:

Mějme náhodný výběr: $\begin{pmatrix} x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{1,n} \\ x_{2,1}, x_{2,2}, x_{2,3}, \ldots, x_{2,n} \end{pmatrix}$ a jejich pořádkové statistiky $\begin{pmatrix} x_{1,(1)}, x_{1,(2)}, x_{1,(3)}, \ldots, x_{1,(n)} \\ x_{2,(1)}, x_{2,(2)}, x_{2,(3)}, \ldots, x_{2,(n)} \end{pmatrix}$,
pak ke každému pozorování existuje přiřazené jeho pořadí (rank) v uspořádaném výběru:
$R(x_{1,i}) = R_i = j \Leftrightarrow x_{1,i} = x_{1,(j)}$ a $S(x_{1,i}) = S_i = j \Leftrightarrow x_{2,i} = x_{2,(j)}$. Nejprve se budeme zabývat jednou náhodnou proměnnou.
Předpokládáme, že náhodný výběr $x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{1,n}$ se řídí distribucí $F_1(x)$ a náhodný výběr $x_{2,1}, x_{2,2}, x_{2,3}, \ldots, x_{2,n}$ se řídí distribucí $F_2(x)$ a dvojice obou náhodných proměnných se řídí dvou rozměrnou distribucí $F(x_1, x_2)$. O všech třech $F_1(x), F_2(x), F(x_1, x_2)$ distribucích budeme předpokládat, že jsou spojité a rostoucí v celých jejich definičních oborech. Ze spojitosti distribučních funkcí plyne to, že žádná dvě pozorování nemohou být rovná.

Pro rankovou statistiku pak platí $P(R_i = j) = \frac{1}{n}$ a $P(R_i = j, R_k = l) = \frac{1}{n(n-1)}$. Odtud

$E\{R\} = \sum_{i=1}^{n} i \frac{1}{n} = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$, $E\{R^2\} = \sum_{i=1}^{n} i^2 \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} i^2 = \frac{1}{n} \frac{n(2n+1)(n+1)}{6} = \frac{(2n+1)(n+1)}{6}$ a

$\sigma^2(R) = E\{R^2\} - (E\{R\})^2 = \frac{(2n+1)(n+1)}{6} - \frac{(n+1)^2}{4} = (n+1)\frac{2(2n+1) - 3(n+1)}{12} = \frac{(n+1)(n-1)}{12} = \frac{(n^2-1)}{12}$.

Samozřejmě, totéž platí i pro druhou rankovou statistiku S. Tedy:

Ranková statistika zobrazuje pozorování libovolné náhodné proměnné na nová pozorování transformované náhodné proměnné, jejíž pravděpodobnostní popis nezávisí na pravděpodobnostním popisu původní náhodné proměnné. A rozdělení všech rankových statistik na náhodných výběrech stejného rozsahu je stejné.

To však už neplatí pro sdružená rozdělení. Proto sdružená rozdělení rankových statistik popisují „to, co obě náhodné proměnné spojuje“ nezávisle na tom, jak se individuálně chovají. A odtud lze odvodit „měření síly jejich propojení“, např.

$$r_S = corr(R,S) = \frac{Cov(R,S)}{\sqrt{\sigma^2(R)}\sqrt{\sigma^2(r)}} = \frac{12}{n^2-1} Cov(R,S) = \frac{12}{n^2-1} E\left\{\left(R - \frac{n+1}{2}\right)\left(S - \frac{n+1}{2}\right)\right\}.$$

Tj. klasický korelační koeficient pořadí (ranků) obou náhodných proměnných. Ten je nazýván podle svého autora Charlese Edwarda Spearmana (1863-1945), povoláním psychologa (General Intelligence, objectively determined and measured. American Journal of Psychology, 1904; Proof and measurement of association between two things. American Journal of Psychology, 1904).

Výše uvedenému vzorci pro Spearmanův koeficient korelace odpovídá vzorec pro jeho bodový odhad:

$$\overline{r_S} = \frac{12}{n^2-1}\frac{1}{n}\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right) = 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n(n^2-1)}$$

Taková úprava plyne z rovností: $\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} S_i = \frac{n(n+1)}{2}; \sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} S_i^2 = \frac{n(2n+1)(n+1)}{6}$.

Pokud platí hypotéza o nezávislosti = náhodné výběry $x_{1,1}, x_{1,2}, x_{1,3}, ..., x_{1,n}$ a $x_{2,1}, x_{2,2}, x_{2,3}, ..., x_{2,n}$ jsou pozorováním hodnot dvou nezávislých náhodných proměnných pak má statistika $t = \overline{r_S}\sqrt{\frac{n-2}{1-\overline{r_S}^2}}$ asymptoticky (n>20) ***t***-rozdělení s ***n-2*** stupni volnosti (srovnejte se stejnou statistikou pro testy o klasickém Pearsonově koeficientu korelace – Spearmanova korelace je klasickou korelací ale pořadí). Toho lze např. užít k testování hypotézy o nulovosti r_S =nezávislosti.

To dává kritické obory:

Pravostranná alternativa	$r_S > 0$	$W_\alpha = \left\{ t = \frac{\overline{r_S}\sqrt{n-2}}{\sqrt{1-\overline{r_S}^2}}; t \geq t_{1-\alpha}(n-2) \right\}$
Levostranná alternativa	$r_S < 0$	$W_\alpha = \left\{ t = \frac{\overline{r_S}\sqrt{n-2}}{\sqrt{1-\overline{r_S}^2}}; t \leq -t_{1-\alpha}(n-2) \right\}$
Oboustranná alternativa	$r_S \neq 0$	$W_\alpha = \left\{ t = \frac{\overline{r_S}\sqrt{n-2}}{\sqrt{1-\overline{r_S}^2}}; \lvert t \rvert \leq t_{1-\alpha/2}(n-2) \right\}$

(Výběrový) Spearmanův koeficient korelace má při hypotéze nezávislosti $E\{\overline{r_S}\} = 0$, $\sigma^2(\overline{r_S}) = \frac{1}{n-1}$. **Námět:** odvoďte. A je asymptoticky normální.

Existují i jiné statistiky a testy pro měření „monotónní závislosti“ a testování nezávislosti, jako jsou:
Kvadrantová statistika.
Kendallova statistika.
Wilcoxonova statistika.

O jejich vyjádření a užití se lze např. dočíst v:

Dagmar Blatná	Neparametrické metody. Testy založené na pořádkových a pořadových statistikách. Skripta VŠE, Praha 1996

Jejich teoretické vlastnosti jsou studovány např. v:

Jana Jurečková	Pořadové testy. Skripta MFFUK, SPN Praha 1981

Klasický lineární „regresní“ model:

Motivace:

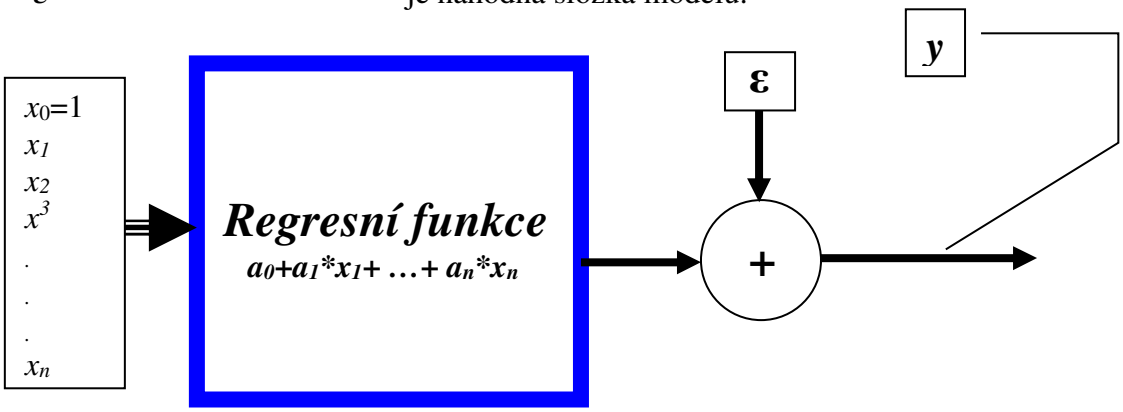
$$E(x_1 / x_2) = \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(x_2 - \mu_2) = \left(\rho \frac{\sigma_1}{\sigma_2}\right) x_2 + \left(\mu_1 - \rho \frac{\sigma_1}{\sigma_2} \mu_2\right)$$

$$E(x_2 / x_1) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1) = \left(\rho \frac{\sigma_2}{\sigma_1}\right) x_1 + \left(\mu_2 - \rho \frac{\sigma_2}{\sigma_1} \mu_1\right)$$

Předpokládáme vztah mezi vysvětlovanou a vysvětlujícími proměnnými ve tvaru:

$$y = \sum_{i=1}^{n} a_i x_i + a_0 + \varepsilon \quad \text{definujeme - li } x_0 = 1, \text{ pak } y = \sum_{i=0}^{n} a_i x_i + \varepsilon$$

y			je vysvětlovaná proměnná,
$x_i,$	$i = 0,1,...,n$	$x_0 = 1$	jsou vysvětlující proměnné a
ε			je náhodná složka modelu.



Dále předpokládáme, že máme k dispozici T pozorování:

$$y_1, x_{10}, x_{11}, ..., x_{1n}$$
$$y_2, x_{20}, x_{21}, ..., x_{2n}$$
$$.....$$
$$y_T, x_{T0}, x_{T1}, ..., x_{Tn} \quad x_{j0} = 1, \forall j = 1,...,T$$

O těchto pozorováních předpokládáme, že jsou pozorovány, měřeny bez chyby a jsou tedy nenáhodné. Jediná náhodnost je soustředěna do náhodné – nepozorovatelné – složky

$$\varepsilon_1, \varepsilon_2, ..., \varepsilon_T$$

Pozorování vysvětlované, vysvětlujících a náhodné složky jsou propojeny vztahy:

$$y_j = \sum_{i=0}^{n} a_i x_{ji} + \varepsilon_j \quad j = 1,...,T$$

Zapsáno pomocí aparátu vektorů a matic:

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \boldsymbol{\varepsilon} \tag{1}$$

Předpokládáme, další vlastnosti:

$E\{\boldsymbol{\varepsilon}\} = 0$ $var\{\boldsymbol{\varepsilon}\} = E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\} = \sigma^2\mathbf{I}$, kde σ^2 je další parametr modelu mimo $a_0, a_1, ..., a_n$ a $hodnost(\mathbf{X}) = n+1 \le T$. Pozor, předpoklad existence obou momentů $E\{\boldsymbol{\varepsilon}\}$ $var\{\boldsymbol{\varepsilon}\}$ je podstatný. Z uvedených předpokladů vyplývá, že veškerá náhodnost vysvětlující a vysvětlované proměnné je soustředěna do modelové proměnné ε - náhodné – nepozorované (nebo nepozorovatelné) složky.

Dále je nutné upozornit na fakt, že klasický model předpokládá ve vztahu $E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\} = \sigma^2\mathbf{I}$ jistou neproměnnost variability náhodné složky (stacionarita, …) a to, že náhodné složky různých pozorování jsou nekorelované.

Za uvedených předpokladů (pro její využití nejsou podstatné, podstatné jsou pro vlastnosti výsledků) lze využít metodu nejmenších čtverců“

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \boldsymbol{\varepsilon} \quad \Rightarrow \quad \mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{a} + \mathbf{X}^T\boldsymbol{\varepsilon} \quad \Rightarrow \quad (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{a} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon} \quad \Rightarrow$$

$$E\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\} = E\{\mathbf{a}\} + E\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\} \Rightarrow (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = E\{\mathbf{a}\} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E\{\boldsymbol{\varepsilon}\}$$

a protože $E\{\boldsymbol{\varepsilon}\} = 0$ dostáváme: $E\{\mathbf{a}\} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ a použijeme-li označení $E\{\mathbf{a}\} = \hat{\mathbf{a}}$ je optimálním odhadem ve smyslu nejmenších čtverců

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

Ze vztahu $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{a} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$ dostáváme: $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon} = \mathbf{a}$

odečteme-li od této rovnice rovnici (2), získáme: $-(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon} = \mathbf{a} - \hat{\mathbf{a}}$ a odtud:

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\,\mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1})^T = (\mathbf{a} - \hat{\mathbf{a}})(\mathbf{a} - \hat{\mathbf{a}})^T \Rightarrow$$

$$E\{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\,\mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1})^T\} = E\{(\mathbf{a} - \hat{\mathbf{a}})(\mathbf{a} - \hat{\mathbf{a}})^T\} \Rightarrow$$

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\}\mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1})^T = \boldsymbol{\Sigma}_{\mathbf{a}} \Rightarrow \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})((\mathbf{X}^T\mathbf{X})^{-1})^T = \boldsymbol{\Sigma}_{\mathbf{a}} \Rightarrow$$

$\sigma^2((\mathbf{X}^T\mathbf{X})^{-1})^T = \boldsymbol{\Sigma}_{\mathbf{a}}$. Při využití předpokladu symetrie matice $\boldsymbol{\Sigma}_{\mathbf{a}}$ dostaneme:

$$\boldsymbol{\Sigma}_{\mathbf{a}} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{3}$$

To je vztah pro kovarianční matici optimálního odhadu vektoru parametrů (2). Z kovarianční matice získáme standardním postupem korelace mezi jednotlivými komponentami odhadu parametrického vektoru $\mathbf{a}$. Ty jsou pak „mírou závislosti“ mezi jednotlivými komponentami tohoto vektoru. Na tomto místě je nezbytné upozornit, že pro znalost kovarianční matice potřebujeme znát neznámý parametr σ^2.

Námět pro přemýšlení: potřebujeme tento parametr znát pro stanovení korelací?

Použijeme označení: $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{a}}$, $\bar{y} = \frac{1}{T}\sum_{j=1}^{T} y_j$ a $\bar{\hat{y}} = \frac{1}{T}\sum_{j=1}^{T} \hat{y}_j$.

Vztahy pro vypočtená residua:

$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{a}}$. Označíme: $\mathbf{j}$ vektor ze samých jednotek, potom: $\sum_{j=1}^{T} \hat{\varepsilon}_j = \mathbf{j}^T \hat{\boldsymbol{\varepsilon}} = \mathbf{j}^T (\mathbf{y} - \mathbf{X}\hat{\mathbf{a}}) = 0$. Tj. součet vypočtených residuí je nulový. Plyne to z předpokladu: $E\{\boldsymbol{\varepsilon}\} = 0$. Dokažte! Je možno takový fakt dokázat bez předpokladu $x_{j0} = 1 \forall j = 1,...,T$ = regrese obsahuje absolutní člen? Přímým důsledkem je fakt $\mathbf{j}^{\mathbf{T}}\mathbf{y} = \mathbf{j}^{\mathbf{T}}\mathbf{X}\hat{\mathbf{a}} = \mathbf{j}^{\mathbf{T}}\hat{\mathbf{y}}$. Tj. součet pozorovaných hodnot vysvětlované proměnné je rovem součtu hodnot odhadů $\hat{y}_j$. Jinak - průměr pozorovaných hodnot vysvětlované proměnné je rovem průměru hodnot odhadů $\hat{y}_j$.

Definujeme:

$$TSS = \sum_{j=1}^{T} (y_j - \bar{y})^2 \quad \text{a} \quad SSE = \sum_{j=1}^{T} (y_j - \hat{y}_j)^2 \quad \text{a} \quad SSR = TSS - SSE$$

TSS	Total Sum of Squares
SSE	Sum of Squares Error
SSR	Sum of Squares Regression

Dále budeme používat výše uvedených výsledků: $\bar{y} = \bar{\hat{y}}$,

$$SSR = TSS - SSE = \sum_{j=1}^{T} \left[(y_j - \bar{y})^2 - \left(y_j - \hat{y}_j \right)^2 \right] = \sum_{j=1}^{T} \left[\left(y_j - \bar{\hat{y}} \right)^2 - \left(y_j - \hat{y}_j \right)^2 \right] = \ldots$$

Evidentně je $TSS \geq SSE$, tak byla $\hat{y}_j$ vypočítána. (neopomeňte předp. $x_{j0} = 1 \forall j = 1,...,T$ = regrese obsahuje absolutní člen). Odtud je $SSR \geq 0$. Pak nestranným odhadem parametru σ^2 je $s^2 = \frac{SSE}{T-n-1}$. Jedná se o běžný výběrový rozptyl odchylek pozorované hodnoty $\mathbf{y}$ od $\hat{\mathbf{y}}$.

Pak také můžeme stanovit odhad kovarianční matice odhadů $\hat{\mathbf{a}}$: $\mathbf{S}_{aa} = s^2 (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$. Obdobně lze určit i koeficient determinace $R^2 = \frac{SSR}{TSS} = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS}$. V naší úloze platí $0 \leq R^2 \leq 1$. Námět pro přemýšlení: Je i v tomto tvrzení předpoklad: $x_{j0} = 1 \forall j = 1,...,T$ = regrese obsahuje absolutní člen, podstatný? V tomto případě $\frac{T-1}{T-n-1} R^2$ je přibližně nestranným odhadem kvadrátu koeficientu mnohonásobné korelace.

Kvalita odhadnutých koeficientů („large sample“)

Ve skriptech [Cipra, str. 42-43] je dokázáno, že platí:

$\frac{\sqrt{T}(a_i - \hat{a}_i)}{\sqrt{q_{ii}}} \approx N(0,1)$ pro dostatečně velké T, ($T > \mathbf{30}$), kde q_{ii} je i-tý diag. prvek matice $s^2\left(\frac{\mathbf{X^T X}}{T}\right)^{-1}$. Tato asymptotika pak umožňuje konstruovat testy hypotéz a výroky o jednotlivých koeficientech a_i a jejich odhadech $\hat{a}_i$. Pomocí takových postupů lze pak, odkazem na [Cipra, str. 46-47] dokázat tvrzení odhad $\hat{\mathbf{a}}$ má rozdělení $N(\mathbf{a}, \sigma^2(\mathbf{X^T X})^{-1})$ - samozřejmě, zde, za předpokladu **normality rozdělení** vektoru náhodných odchylek $\boldsymbol{\varepsilon}$.
Z tohoto lze konstruovat další typy výroků a testů o kvalitě modelu. Odkaz na [Cipra, str. 46-55]. Samozřejmě lze odkázat i na další učebnice.

Vztah $\frac{\sqrt{T}(a_i - \hat{a}_i)}{\sqrt{q_{ii}}} \approx N(0,1)$ vlastně tvrdí:

$$\text{Prob}\left\{d_i \le a_i - \hat{a}_i \le h_i\right\} = \text{Prob}\left\{\frac{d_i\sqrt{T}}{\sqrt{q_{ii}}} \le \frac{\sqrt{T}(a_i - \hat{a}_i)}{\sqrt{q_{ii}}} \le \frac{h_i\sqrt{T}}{\sqrt{q_{ii}}}\right\} = \phi\left(\frac{h_i\sqrt{T}}{\sqrt{q_{ii}}}\right) - \phi\left(\frac{d_i\sqrt{T}}{\sqrt{q_{ii}}}\right) \text{ kde :}$$

$\phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{x^2}{2}}dx$ je normovaná distribuční funkce normálního (Gaussova) rozdělení pravděpodobnosti.

Některé heuristiky a pseudoheuristiky o vlivu jednotlivých vysvětlujících proměnných.

Označíme-li $\overline{x_i} = \frac{1}{T}\sum_{j=1}^{T} x_{ji}$ průměr z i-tého sloupce matice pozorování vysvětlujících proměnných $\mathbf{X}$, můžeme psát: $\overline{y} = \sum_{i=0}^{n} \hat{a}_i \overline{x_i}$. Pokud je $\overline{y} \neq 0$, dostáváme: $1 = \sum_{i=0}^{n} \frac{\hat{a}_i \overline{x_i}}{\overline{y}}$. Členy $\frac{\hat{a}_i \overline{x_i}}{\overline{y}}$ pak mohou být měřítkem vlivu i-té vysvětlující proměnné na vysvětlovanou proměnnou. Jejich znaménko pak signalizuje positivitu nebo negativitu tohoto vlivu.

Pro testy významnosti lze pak lze použít vztahu

$$\text{Prob}\left\{d_i \le a_i - \hat{a}_i \le h_i\right\} = \text{Prob}\left\{\frac{d_i\sqrt{T}}{\sqrt{q_{ii}}} \le \frac{\sqrt{T}(a_i - \hat{a}_i)}{\sqrt{q_{ii}}} \le \frac{h_i\sqrt{T}}{\sqrt{q_{ii}}}\right\} = \phi\left(\frac{h_i\sqrt{T}}{\sqrt{q_{ii}}}\right) - \phi\left(\frac{d_i\sqrt{T}}{\sqrt{q_{ii}}}\right)$$

pro $h_i = \varepsilon_i$ a $d_i = -\varepsilon_i$, kde ε_i je řešením rovnice

$\phi\left(\frac{\varepsilon_i\sqrt{T}}{\sqrt{q_{ii}}}\right)-\phi\left(\frac{-\varepsilon_i\sqrt{T}}{\sqrt{q_{ii}}}\right)=1-\alpha$, pro zvolenou hladinu významnosti α. Pokud dostaneme $\hat{a}_i-\varepsilon\le 0\le \hat{a}_i+\varepsilon$ můžeme vyslovit soud o nevýznamnosti koeficientu a_i a tím i o nevýznamnosti i-té vysvětlující proměnné pro vysvětlení proměnné vysvětlované. Jinak řečeno: pokud nula leží v symetrickém intervalu spolehlivosti pro příslušný koeficient, soudíme na jeho nulovost při hladině významnosti α. Přesněji: tuto hypotézu nevylučujeme.

Vztah $\phi\left(\frac{\varepsilon_i\sqrt{T}}{\sqrt{q_{ii}}}\right)-\phi\left(\frac{-\varepsilon_i\sqrt{T}}{\sqrt{q_{ii}}}\right)=1-\alpha$ má triviální řešení:

$\phi\left(\frac{\varepsilon_i\sqrt{T}}{\sqrt{q_{ii}}}\right)=1-\frac{\alpha}{2}\Leftrightarrow\frac{\varepsilon_i\sqrt{T}}{\sqrt{q_{ii}}}=\phi^{-1}\left(1-\frac{\alpha}{2}\right)\Leftrightarrow\varepsilon_i=\sqrt{\frac{q_{ii}}{T}}\phi^{-1}\left(1-\frac{\alpha}{2}\right)$, kde ϕ^{-1} je inverzní (kvantilový) operátor k operátoru ϕ. Poznámka: čím nižší je α tím širší je příslušný interval spolehlivosti a tím je „snazší dostat nulu do takového intervalu“ - tento fakt je nezbytné brát v úvahu při rozhodování o významnosti či nevýznamnosti.

Proto se používají také metody založené na následující úvaze: Pokud by i-tá vysvětlující proměnná neměla ovlivňovat vysvětlovanou proměnnou, znamenalo by to $a_i=0$ a pak by, podle vztahu $\frac{\sqrt{T}(a_i-\hat{a}_i)}{\sqrt{q_{ii}}}\approx N(0,1)$, měla náhodná veličina $\frac{\sqrt{T}\,\hat{a}_i}{\sqrt{q_{ii}}}$ rozdělení $N(0,1)$.

Některé heuristiky a pseudoheuristiky o kvalitě celého modelu.

Běžným kritériem celého modelu je koeficient determinace: $R^2=1-\frac{SSE}{TSS}$. Čím blíže je jedné tím lze považovat model za kvalitnější (opět neopomenout, mluví se o lineárním modelu s absolutním členem). Dále je dobré vzít do úvahy, že koeficient determinace je vlastně produktem rozkladu rozptylu vysvětlované proměnné na část regresní a zbytkovou. Tedy koeficient determinace jen měří „vysvětlení variability“ vysvětlované proměnné.

„Intervalový“ odhad celé regresní nadroviny:

$y=\sum_{i=0}^{n}a_ix_i+\varepsilon\Rightarrow\ \hat{y}=x^T\hat{a}$ prokonkrétní *x*, proto $\hat{y}=x^T\left(X^TX\right)^{-1}X^Ty=x^T\left(X^TX\right)^{-1}X^T(Xa+\varepsilon)$

A odtud: $E\{\hat{y}/x\}=x^T\left(X^TX\right)^{-1}X^TE\{y/x)=x^T\left(X^TX\right)^{-1}(X^TX)a=x^Ta$ Proto:

$\hat{y}-E\{\hat{y}/x\}=x^T(\hat{a}-a)\ \Rightarrow\ \sigma^2\{\hat{y}/x\}=E\left\{\left(x^T(\hat{a}-a)\right)^2\right\}=E\left\{x^T(\hat{a}-a)(\hat{a}-a)^Tx\right\}=x^T\Sigma_{\hat{a}}x$.

Protože: $odhad\{\sigma^2\{\hat{y}/x\}\}\equiv S^2\{y\}=s^2x^T(X^TX)^{-1}x$, lze konstatovat pro dostatečně velký počet pozorování $\hat{y}-E\{\hat{y}/x\}\approx N(0,s^2x^T(X^TX)^{-1}x)$ tj. odchylka odhadu regresní čáry od skutečné má asymptoticky normální rozdělení, s nulovou střední hodnotu a rozptylem $s^2x^T(X^TX)^{-1}x$. Tedy je heteroskedastická!

Doporučená a zdrojová literatura:

Jiří Reif	Metody matematické statistiky, ZČU v Plzni 2004
Jaroslav Hátle, Jiří Likeš	Základy počtu pravděpodobnosti a matematické statistiky. SNTL Praha 1974.
C. Radhakrishna Rao	Lineární metody statistické indukce a jejich aplikace, ACADEMIA, Praha 1978
Alfréd Rényi	Teorie pravděpodobnosti, ACADEMIA, Praha 1972
Cipra, T.	Ekonometrie, skripta MFF UK, Praha, 1984.
Mendenhall. W.	Introduction to Probability and Statistics. PWS-KENT, Publishing Company, Boston 1987.
Groebner, D.F., Shannon P. W.	Business Statistics. A Decision-Making Approach. Merrill Publishing Company, Columbus Ohio, 1989.
Jana Jurečková	Pořadové testy. Skripta MFFUK, SPN Praha 1981
Dagmar Blatná	Neparametrické metody. Testy založené na pořádkových a pořadových statistikách. Skripta VŠE, Praha 1996
Dagmar Blatná	Neparametrické metody II. Neparametrické odhady. Skripta VŠE, Praha 1999