Maximálně věrohodné odhady

Náhodný výběr $X_1, \dots, X_n$ rosahu n z rozdělení pravděpodobnosti P:

- $X_i \sim P$ $(i = 1, \dots, n)$
- $X_1, \dots, X_n$ jsou stochasticky nezávislé
- Co to znamená pro vztah mezi simultánní a marginální hustotou pravděpodobnosti $f(x)$ (pravděpodobnostní funkci $p(x)$) ?

Rozdělení pravděpodobnosti závislé na parametru (parametrech) θ:

- $f(x), p(x)$ jako funkce proměnné $\theta \Rightarrow L(\theta)$
- Věrohodnostní funkce $L(\theta)$ a logaritmická věrohodnostní funkce $l(\theta)$:

$$L(\theta) = L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

$$l(\theta) = l(\theta; x_1, \dots, x_n) = \ln L(\theta; x_1, \dots, x_n) = \sum_{i=1}^{n} \ln f(x_i; \theta) = \sum_{i=1}^{n} \ln p(x_i; \theta)$$

- Jak odhadnout θ ze znalosti $X_1, \dots, X_n$?

Maximálně věrohodné odhady

Myšlenka: parametr θ odhadneme hodnotou, která je při daném náhodném výběru ze známého rozdělení pravděpodobnosti nejvíce pravděpodobná.

Maximálně věrohodný odhad (MLE = maximum likelihood estimator) $\widehat{\theta}_{\mathrm{ML}}$ parametru θ se získá maximalizací věrohodnostní funkce $L(\theta)$:

$$\widehat{\theta}_{\mathrm{ML}} : L(\theta; x_1, \ldots, x_n) \to \max_\theta, \quad \text{resp.} \quad l(\theta; x_1, \ldots, x_n) \to \max_\theta$$

To znamená najít stacionární bod funkce $l(\theta)$ vzhledem k θ,

$$\frac{\partial}{\partial\theta} l(\theta) = 0 \qquad \text{(věrohodnostní rovnice)},$$

a ověřit 2. diferenciál, resp. derivaci,

$$\left.\frac{\partial^2}{\partial\theta^2} l(\theta)\right|_{\theta=\widehat{\theta}_{\mathrm{ML}}} < 0 \ .$$

Poznámka: v případě věktoru parametrů $\boldsymbol{\theta}$ řešíme soustavu věrohodnostních rovnic pro $\boldsymbol{\theta}$ z 1. derivací a ověřujeme negativní definitnost matice 2. derivací.

MLE – příklady (1)

1. Najděte ML-odhad parametru $\theta \in [0,1]$ pro náhodný výběr $X_1, \dots, X_n$ z binomického rozdělení $\mathrm{Bi}(m, \theta)$ ($m \in \mathbb{N}$) s pravděpodobností funkcí

$$p(x) = \binom{m}{x} \theta^x (1-\theta)^{m-x} \quad \text{pro} \quad x = 0, 1, \dots, m; \qquad p(x) = 0 \quad \text{jinak.}$$

2. Najděte ML-odhad parametru $\theta \in [0,1]$ pro náhodný výběr $X_1, \dots, X_n$ z geometrického rozdělení $\mathrm{Ge}(\theta)$ s pravděpodobností funkcí

$$p(x) = (1-\theta)^x \theta \quad \text{pro} \quad x \in \mathbb{N}_0; \qquad p(x) = 0 \quad \text{jinak.}$$

3. Najděte ML-odhady parametrů $\mu \in \mathbb{R}$ a $\sigma^2 > 0$ pro náhodný výběr $X_1, \dots, X_n$ z Gaussova rozdělení $\mathrm{N}(\mu, \sigma^2)$ s hustotou pravděpodobnosti

$$f(x) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\,\sigma^2}\right] .$$

4. Najděte ML-odhady parametrů $\mu \in \mathbb{R}$ a $\sigma^2 > 0$ pro náhodný výběr $X_1, \dots, X_n$ z logaritmického normálního rozdělení $\mathrm{LN}(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}\,x} \exp\left[-\frac{(\ln x-\mu)^2}{2\,\sigma^2}\right] \quad \text{pro} \quad x > 0; \qquad f(x) = 0 \quad \text{jinak.}$$

MLE – příklady (2)

5. Najděte ML-odhad parametru $\mu > 0$ pro náhodný výběr $X_1, \ldots, X_n$ z exponenciálního rozdělení $\mathrm{Ex}(\mu)$ s hustotou pravděpodobnosti

$$f(x) = \frac{1}{\mu} \exp\left[-\frac{x}{\mu}\right] \quad \text{pro} \quad x \geq 0; \qquad f(x) = 0 \quad \text{jinak.}$$

6. Najděte ML-odhad parametru $\lambda > 0$ pro náhodný výběr $X_1, \ldots, X_n$ z exponenciálního rozdělení $\mathrm{Ex}(\lambda)$ s hustotou pravděpodobnosti

$$f(x) = \lambda \exp[-\lambda x] \quad \text{pro} \quad x \geq 0; \qquad f(x) = 0 \quad \text{jinak.}$$

7. Najděte ML-odhady parametrů $\lambda > 0$ a $k > 0$ pro náhodný výběr $X_1, \ldots, X_n$ z Weibullova rozdělení $\mathrm{Wb}(\lambda, k)$ s hustotou pravděpodobnosti

$$f(x) = k\,\lambda\,x^{k-1} \exp\left[-\lambda\,x^k\right] \quad \text{pro} \quad x \geq 0; \qquad f(x) = 0 \quad \text{jinak.}$$

8. Najděte ML-odhad parametru $s > 0$ pro náhodný výběr $X_1, \ldots, X_n$ z Rayleighova rozdělení $\mathrm{Ra}(s)$ s hustotou pravděpodobnosti

$$f(x) = \frac{x}{s} \exp\left[-\frac{x^2}{2\,s}\right] \quad \text{pro} \quad x \geq 0; \qquad f(x) = 0 \quad \text{jinak.}$$

MLE – příklady (3)

9. Najděte ML-odhad parametru $\lambda > 0$ pro náhodný výběr $X_1, \ldots, X_n$ z Gamma rozdělení $\Gamma(\lambda, k)$ $(k > 0)$ s hustotou pravděpodobnosti

$$f(x) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} \exp[-\lambda x] \quad \text{pro} \quad x \geq 0; \qquad f(x) = 0 \quad \text{jinak.}$$

Najděte také věrohodnostní rovnici pro ML-odhad k.

Pomůcka: $\frac{\partial}{\partial k} \ln \Gamma(k) = \Psi(k) =$ digamma funkce.

⋆. Další příklady pro odvození ML-odhadů parametrů v rozděleních s podobnými tvary hustot naleznete na stránce dr. Forbelské.

MLE – řešení příkladů

1. $\widehat{\theta}_{\mathrm{ML}} = \dfrac{\overline{X}}{2\,\overline{X} - n}$
2. $\widehat{\theta}_{\mathrm{ML}} = \dfrac{1}{1 + \overline{X}}$
3. $\widehat{\mu}_{\mathrm{ML}} = \overline{X}$, $\qquad \widehat{\sigma^2}_{\mathrm{ML}} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_{\mathrm{ML}})^2$
4. $\widehat{\mu}_{\mathrm{ML}} = \dfrac{1}{n}\sum_{i=1}^{n} \ln X_i$, $\qquad \widehat{\sigma^2}_{\mathrm{ML}} = \dfrac{1}{n}\sum_{i=1}^{n}(\ln X_i - \widehat{\mu}_{\mathrm{ML}})^2$
5. $\widehat{\lambda}_{\mathrm{ML}} = \overline{X}$
6. $\widehat{\lambda}_{\mathrm{ML}} = \dfrac{1}{\overline{X}}$
7. $\widehat{\lambda}_{\mathrm{ML}} = \dfrac{n}{\sum_{i=1}^{n} X_i^k}$
8. $\widehat{s}_{\mathrm{ML}} = \dfrac{1}{2\,n}\sum_{i=1}^{n} X_i^2$
9. $\widehat{\lambda}_{\mathrm{ML}} = \dfrac{k}{\overline{X}}$, $\qquad$ ML-rovnice pro k: $\Psi(\widehat{k}_{\mathrm{ML}}) = \ln \lambda + \overline{X}$

Odhady parametrů momentovou metodou

Odhady parametrů tzv. metodou momentů spočívá ve vyjádření několika prvních (tolik, kolik potřebujeme) momentů M_p rozdělení pravděpodobnosti náhodné veličiny X,

$$M_p = \mathsf{E}(X^p) \qquad (p = 1, 2, \dots) \ .$$

Teoretické momenty M_p závisí na neznámých parametrech, které chceme odhadnout.

Momenty M_p v rovnici (rovnicích) nahradíme (aproximujeme) výběrovými momenty m_p,

$$m_p = \frac{1}{n} \sum_{i=1}^{n} X_i^p \qquad (p = 1, 2, \dots) \ ,$$

které závisí pouze na náhodném výběru. Algebraickým vyjádřením (příp. numerickým výpočtem) hledaných parametrů z rovnice (systému rovnic) pro m_p obdržíme odhady $\widehat{\theta}_{\mathrm{MM}}$ momentovou metodou.

Momentové odhady – příklady

Jako cvičení spočtěte odhady parametrů momentovou metodou v příkladech 1.–9. pro maximálně věrohodné odhady.

Nejdříve odvoďte (pro hustoty integrováním), nebo pomocí tabulek či počítače nalezněte, momenty daných rozdělení pravděpodobnosti,

$$M_p = \mathrm{E}(X^p) \qquad (p = 1, 2, \dots)$$

Pro kontrolu jsou první dva momenty M_1, M_2 uvedeny v tabulce vpravo.

Poté odvoďte momentové odhady parametrů a porovnejte je s maximálně věrohodnými odhady.

příklad	M_1	M_2
1.	$m\,\theta$	$m\,\theta\,(1-\theta)$
2.	$\frac{1-\theta}{\theta}$	$\frac{(1-\theta)^2+(1-\theta)}{\theta^2}$
3.	μ	$\mu^2+\sigma^2$
4.	$\exp\left[\mu+\frac{1}{2}\sigma^2\right]$	$\exp\left[2\mu+2\sigma^2\right]$
5.	μ	$2\mu^2$
6.	$\frac{1}{\lambda}$	$\frac{2}{\lambda^2}$
7.	$\frac{1}{k}\,\lambda^{-\frac{1}{k}}\,\Gamma\left(\frac{1}{k}\right)$	$\lambda^{-\frac{2}{k}}\,\Gamma\left(1+\frac{2}{k}\right)$
8.	$\sqrt{\frac{\pi s}{2}}$	$2\,s$
9.	$\frac{k}{\lambda}$	$\frac{k(k+1)}{\lambda^2}$

Momentové odhady – řešení příkladů

Označení: p-tý výběrový moment $= m_p = \frac{1}{n}\sum_{i=1}^{n} X_i^p \qquad (p = 1, 2, \dots)$

1. $\widehat{\theta}_{\mathrm{MM}} = \dfrac{m_1}{m}$
2. $\widehat{\theta}_{\mathrm{MM}} = \dfrac{1}{1 + m_1}$
3. $\widehat{\mu}_{\mathrm{MM}} = m_1$, $\qquad \widehat{\sigma^2}_{\mathrm{MM}} = m_2 - m_1^2$
4. $\widehat{\mu}_{\mathrm{MM}} = 2\ln m_1 - \frac{1}{2}\ln m_2$, $\qquad \widehat{\sigma^2}_{\mathrm{MM}} = \ln m_2 - 2\ln m_1$
5. $\widehat{\lambda}_{\mathrm{MM}} = m_1$
6. $\widehat{\lambda}_{\mathrm{MM}} = \dfrac{1}{m_1}$
8. $\widehat{s}_{\mathrm{MM}} = \dfrac{2\, m_1^2}{\pi}$, anebo $\widehat{s}_{\mathrm{MM}} = \dfrac{m_2}{2}$
9. $\widehat{\lambda}_{\mathrm{MM}} = \dfrac{m_1}{m_2 - m_1^2}$, $\qquad \widehat{k}_{\mathrm{MM}} = \dfrac{m_1^2}{m_2 - m_1^2}$

Náhodné vektory

Náhodný vektor $X = (X_1, \ldots, X_n)'$ (reálný) je měřitelná vektorová funkce $\boldsymbol{X} = (X_1, \ldots, X_n)' : \Omega \to \mathbb{R}^n$, jejíž složky jsou náhodné veličiny, $X_i : \Omega \to \mathbb{R}$, na stejném pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathrm{P})$.

Střední hodnota $\mathrm{E}(\boldsymbol{X})$ náhodného vektoru je definována po složkách:

$$\mathrm{E}(\boldsymbol{X}) = \big(\mathrm{E}(X_1), \ldots, \mathrm{E}(X_n)\big)' \ .$$

Kovarianční matice $\mathrm{cov}(\boldsymbol{X})$ *(variance-covariance matrix)* náhodného vektoru:

$$\mathrm{D}(\boldsymbol{X}) = \mathrm{cov}(\boldsymbol{X}) = \{c_{i,j}\}_{i,j=1}^{n} \quad , \quad \text{kde} \quad c_{i,j} = \mathrm{C}(X_i, X_j) \ .$$

- $\mathrm{cov}(\boldsymbol{X})$ je čtvercová řádu n, symetrická (proč?)
- hlavní diagonálu tvoří rozptyly $\mathrm{D}(X_i)$, ostatní složky jsou kovariance $\mathrm{C}(X_i, X_j)$
- pozitivně semidefinitní, tzn. $\forall \boldsymbol{u} \in \mathbb{R}^n : \boldsymbol{u}' \,\mathrm{cov}(\boldsymbol{X})\, \boldsymbol{u} \geq 0$

Poznámka: Anglická literatura často užívá společný pojem *variance* (var, Var). Potom rozumíme: $\mathrm{var}(X) = \mathrm{D}(X)$ pro náhodnou veličinu a $\mathrm{var}(\boldsymbol{X}) = \mathrm{cov}(\boldsymbol{X})$ pro náhodný vektor.

Jednoduché transformace náhodných vektorů

Lineární transformace $\mathbb{R}^n \to \mathbb{R}^m$:

$$\mathsf{E}(\boldsymbol{a} + B\boldsymbol{X}) = \boldsymbol{a} + B\,\mathsf{E}(\boldsymbol{X})$$
$$\operatorname{cov}(\boldsymbol{a} + B\boldsymbol{X}) = B\operatorname{cov}(\boldsymbol{X})\,B'$$

Lineární forma $\mathbb{R}^n \to \mathbb{R}$:

$$\mathsf{E}(a + \boldsymbol{b}'\boldsymbol{X}) = a + \boldsymbol{b}'\,\mathsf{E}(\boldsymbol{X})$$
$$\mathsf{D}(a + \boldsymbol{b}'\boldsymbol{X}) = \boldsymbol{b}'\operatorname{cov}(\boldsymbol{X})\,\boldsymbol{b}$$

Kvadratická forma $\mathbb{R}^n \to \mathbb{R}$:

$$\mathsf{E}(\boldsymbol{X}'A\boldsymbol{X}) = \mathsf{E}(\boldsymbol{X})'A\,\mathsf{E}(\boldsymbol{X}) + \operatorname{Tr}[A\operatorname{cov}(\boldsymbol{X})]$$

$\boldsymbol{a} \in \mathbb{R}^m$; $B \in \mathbb{R}^{n\times m}$; $a \in \mathbb{R}$; $\boldsymbol{b} \in \mathbb{R}^m$; $A \in \mathbb{R}^{n\times n}$ pozitivně definitní

Stopa *(Trace)* matice $= \operatorname{Tr}(C) = \sum_{i=1}^{n}\{c_{i,i}\}$
Platí: $\operatorname{Tr}(A\,B\,C) = \operatorname{Tr}(B\,C\,A) = \operatorname{Tr}(C\,A\,B)$

Náhodné vektory – příklady (1)

V následujících příkladech spočítejte $\mathrm{E}(\boldsymbol{X}), \mathrm{cov}(\boldsymbol{X})$ náhodného vektoru $\boldsymbol{X}$:

13. Znáte: $\mathrm{E}(X_i) = 10\,i,\ \mathrm{C}(X_i, X_j) = i\,j,\ (i, j = 1, 2)$.

14. Znáte: $\mathrm{E}(X_i) = 10\,i,\ \mathrm{D}(X_i) = i^2,\ (i, j = 1, 2, 3);\ \rho(X_i, X_j) = 0{,}5$ pro $i \neq j$.

15. $\boldsymbol{X}$ je náhodný výběr rozsahu 4 z rozdělení $\mathrm{N}(10, 4)$.

16. $\boldsymbol{X}$ je náhodný výběr rozsahu 5 z rozdělení $\mathrm{Ex}(\lambda)$.

17. Znáte: $\mathrm{E}\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 20 \\ 30 \\ 2 \end{pmatrix}, \mathrm{cov}\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 100 & 0 & 25 \\ ? & 9 & ? \\ ? & 9 & 16 \end{pmatrix}$. Spočítejte střední hodnoty, rozptyly, kovariance, korelační koeficienty náhodných veličin X_1, X_2, X_3. Které dvojice (trojice) veličin jsou stochasticky nezávislé?

18. V příkladu **17.**, s využitím vhodných transformací, spočítejte:

- $\mathrm{E}(10\,X_3)$
- $\mathrm{E}(2\,X_1 - 5\,X_3 - X_2)$
- $\mathrm{C}(10\,X_3, 2\,X_1 - 5\,X_3 - X_2)$
- $\mathrm{C}(X_1 + X_2, X_3 - X_2)$
- $\mathrm{D}(10\,X_3)$
- $\mathrm{D}(2\,X_1 - 5\,X_3 - X_2)$
- $\rho(10\,X_3, 2\,X_1 - 5\,X_3 - X_2)$
- $\rho(X_1 + X_2, X_3 - X_2)$

Náhodné vektory – příklady (2)

11. Dokažte vlastnosti kovarianční matice.

12. Dokažte vztahy pro střední hodnotu a kovarianční matici transformovaných náhodných vektorů.

19. Ověřte, pro že výběrový průměr $\overline{X}$ náhodného výběru rozsahu n z rozdělení se střední hodnotou μ a rozptylem σ^2 platí: $\mathrm{E}\left(\overline{X}\right) = \mu,\ \mathrm{D}\left(\overline{X}\right) = \sigma^2/n$.

20. Ověřte, že pro výběrový rozptyl S_X^2 náhodného výběru rozsahu n z rozdělení $\mathrm{N}(\mu, \sigma^2)$ platí: $\mathrm{E}\left(S_X^2\right) = \sigma^2,\ \mathrm{D}\left(S_X^2\right) = 2\sigma^4/(n-1)$.

21. Ověřte, pro že pro náhodný výběr rozsahu n z rozdělení se střední hodnotou μ a rozptylem σ^2 platí vztah $\mathrm{E}\left[3Q_1 - Q_2\right] = (n-3)\sigma^2$, kde $Q_1 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ a $Q_2 = (X_n - X_1)^2 + \sum_{i=2}^{n-1}(X_i - X_{i-1})^2$.

22. Spočítejte $\mathrm{E}(\boldsymbol{Y})$ a $\mathrm{cov}(\boldsymbol{Y})$ transformovaného náhodného vektoru

$$\boldsymbol{Y} = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ -1 & 2 \\ 0 & 10 \end{pmatrix} \boldsymbol{X}, \quad \text{když } \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 0 \\ 10 \end{pmatrix}, \mathrm{cov}(\boldsymbol{X}) = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \quad .$$

Náhodné vektory – příklady (3)

23. Spočtěte střední hodnotu $\mathrm{E}\left(Y_1^2 + Y_2^2 + Y_3^2\right)$ a kovarianční matici $\mathrm{cov}(\boldsymbol{Y})$, kde $Y_1 = X_1,\ Y_2 = X_1 + X_2,\ Y_3 = X_1 + X_2 + X_3$ jsou transformace vzájemně nezávislých náhodných veličin X_1, X_2, X_3,
$\mathrm{E}(X_1) = 10,\ \mathrm{E}(X_2) = 20,\ \mathrm{E}(X_3) = 30,\ \mathrm{D}(X_1) = 1,\ \mathrm{D}(X_2) = 4, \mathrm{D}(X_3) = 9$.

24. Spočtěte střední hodnotu $\mathrm{E}\left(Y_1 Y_2 + Y_2 Y_3 + Y_3 Y_1\right)$ a kovarianční matici $\mathrm{cov}(\boldsymbol{Y})$, kde $Y_1 = X_2 + X_3,\ Y_2 = X_1 + X_3,\ Y_3 = X_1 + X_2$ jsou transformace vzájemně nekorelovaných složek náhodného vektoru $\boldsymbol{X}$, $\mathrm{E}(\boldsymbol{X}) = (10, 10, 10)'$, $\mathrm{D}(X_i) = i^2$.

25. Spočtěte $\mathrm{E}(\boldsymbol{Y})$, $\mathrm{cov}(\boldsymbol{Y})$ a $\mathrm{E}\left(Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + 2Y_1 Y_4\right)$,
kde X_1, X_2, X_3, X_4 jsou náhodné veličiny, $\mathrm{E}(X_i) = 10$, $\mathrm{C}(X_i, X_j) = 1$.
Známe transformační vztahy
$X_1 = Y_1,\ X_2 = Y_2 - Y_1,\ X_3 = Y_3 - Y_2,\ X_4 = Y_4 - Y_3$.

26. Spočítejte střední hodnotu povrchu hranolu s podstavou tvaru čtverce. Délka hrany podstavy je náhodná veličina se střední hodnotou 10 a rozptylem 1, výška hranolu je náhodná veličina se střední hodnotou 20 a rozptylem 9 a její korelační koeficient s délkou hrany podstavy je 0,1.

Některé vztahy ve vektorové (maticové) formě

Když $\boldsymbol{X} = (X_1, \dots, X_n)'$ je náhodný vektor, jeho kovarianční matici lze vyjádřit následujícím způsobem:

$$\operatorname{cov}(\boldsymbol{X}) = \mathsf{E}\left\{[\boldsymbol{X} - \mathsf{E}(\boldsymbol{X})] \cdot [\boldsymbol{X} - \mathsf{E}(\boldsymbol{X})]'\right\} \quad .$$

Někdy je tímto vztahem kovarianční matice definována. Všimněte si podobnosti s definicí kovariance náhodných veličin: $\mathsf{C}(X_1, X_2) = \mathsf{E}\left\{[X_1 - \mathsf{E}(X_1)][X_2 - \mathsf{E}(X_2)]\right\}$. V našem případě však jde součin vektorů (sloupec · řádek → matice).

Některé známé vztahy přepsané vektorově (maticově):

- $X_i = \boldsymbol{e}_i' \boldsymbol{X}$, $\qquad \boldsymbol{e}_i = (0, \dots, 0, 1_i, 0, \dots, 0)$ je i-tý "jednotkový vektor"
- $n\,\overline{X} = \sum_{i=1}^n X_i = (1, \cdots, 1)\, \boldsymbol{X}$
- $\sum_{i=1}^n X_i^2 = \boldsymbol{X}'\, \boldsymbol{I}\, \boldsymbol{X}$, $\qquad \boldsymbol{I} = \operatorname{diag}(1, \dots, 1) =$ "jednotková matice"
- $n^2\, (\overline{X})^2 = \boldsymbol{X}'\, \boldsymbol{J}\, \boldsymbol{X}$, $\qquad \boldsymbol{J} = \{1\}_{i,j=1}^n =$ "jedničková matice"

Náhodné vektory – řešení příkladů

18.
- $\mathrm{E}(10\,X_3) = 20$
- $\mathrm{E}(2\,X_1 - 5\,X_3 - X_2) = 0$
- $\mathrm{C}(10\,X_3, 2\,X_1 - 5\,X_3 - X_2) = -390$
- $\mathrm{C}(X_1 + X_2, X_3 - X_2) = 25$
- $\mathrm{D}(10\,X_3) = 1600$
- $\mathrm{D}(2\,X_1 - 5\,X_3 - X_2) = 399$
- $\rho(10\,X_3, 2\,X_1 - 5\,X_3 - X_2) \approx -0{,}488$
- $\rho(X_1 + X_2, X_3 - X_2) \approx 0{,}905$

22. $\mathrm{E}(\boldsymbol{Y}) = \begin{pmatrix} 9 \\ 21 \\ 100 \end{pmatrix}, \quad \mathrm{cov}(\boldsymbol{Y}) = \begin{pmatrix} 7 & 8 & 50 \\ 8 & 13 & 70 \\ 50 & 70 & 400 \end{pmatrix}$

23. $\mathrm{E}(\cdots) = 4600 + 20 = 4620, \quad \mathrm{cov}(\boldsymbol{Y}) = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 5 & 5 \\ 1 & 5 & 14 \end{pmatrix}$

24. $\mathrm{E}(\cdots) = 1200 + 14 = 1214, \quad \mathrm{cov}(\boldsymbol{Y}) = \begin{pmatrix} 13 & 9 & 4 \\ 9 & 10 & 1 \\ 4 & 1 & 5 \end{pmatrix}$

25. $\mathrm{E}(\boldsymbol{Y}) = \begin{pmatrix} 10 \\ 20 \\ 30 \\ 40 \end{pmatrix}, \quad \mathrm{E}(\cdots) = 3800 + 38 = 3838, \quad \mathrm{cov}(\boldsymbol{Y}) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \\ 4 & 8 & 12 & 16 \end{pmatrix}$

26. Střední hodnota povrchu uvedeného hranolu je rovna $1000 + 3{,}2 = 1003{,}2$.

Regresní analýza

Regresní analýza:

- řeší problém odhadu parametrů v regresním modelu
- pevně dané (zvolené, nenáhodné) hodnoty vstupní veličiny, tzv. regresory
- odpověď je závislá na regresorech prostřednictvím známé funkce
- měřená odpověď je zatížena náhodnými chybami

Regresní model:

- regresory jsou nenáhodné veličiny, obvykle uspořádané do matice regresorů $\boldsymbol{X}$
- odpovědi $\boldsymbol{Y}$ tvoří náhodný vektor, obvykle dokonce náhodný výběr
- je dán funkční vztah $\boldsymbol{Y} = f(\boldsymbol{X}) + \boldsymbol{\epsilon}$
- $\boldsymbol{\epsilon}$ je vektor náhodných chyb
- tvar funkce f je dán, regresní analýza odhaduje parametry funkce

Lineární regresní model

Lineární regresní model (LRM) *(linear regression model)* $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$:

- vyjadřuje závislost tvaru

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\ , \text{ podrobněji } \underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ \vdots \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} x_{1,1} & \dots & x_{1,k} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,k} \end{pmatrix}}_{\boldsymbol{X}} \cdot \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \vdots \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\boldsymbol{\epsilon}}$$

- $\boldsymbol{Y}$ je náhodný vektor n pozorování
- regresory tvoří pevně danou (nenáhodnou) matici plánu *(design matrix)* $\boldsymbol{X}$
- $\boldsymbol{\beta}$ je vektor k neznámých parametrů
- závislost je lineární vzhledem k parametrům β_i
- náhodné chyby ϵ_i jsou vzájemně nezávislé a mají rozdělení $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$
- $n > k$, tj. počet pozorování je větší než počet parametrů
- $h(\boldsymbol{X}) = k$, tj. matice plánu je plné hodnosti

Metoda nejmenších čtverců (MNČ) (1)

Lineární regresní model $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, se obvykle řeší minimalizací jisté penalizační funkce. Při řešení metodou nejmenších čtverců *(least square method)* je kvadratický výraz

$$(\boldsymbol{Y} - \boldsymbol{X\beta})' (\boldsymbol{Y} - \boldsymbol{X\beta})$$

minimalizován vzhledem k neznámých hodnotám parametrů $\boldsymbol{\beta}$. Při splnění podmínek uvedeného lineárního regresního modelu existuje vždy právě jedno řešení této minimalizace, které označíme $\widehat{\boldsymbol{\beta}}$.

- Řešení lineárního regresního modelu metodou nejmenších čtverců, $\widehat{\boldsymbol{\beta}}$, lze vyjádřit jako řešení systému normálních rovnic:

$$\boldsymbol{X}'\boldsymbol{X\beta} = \boldsymbol{X}'\boldsymbol{Y} \quad .$$

MNČ-odhad parametrů je potom:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{Y}$$

Metoda nejmenších čtverců (MNČ) (2)

- Hodnoty sledované proměnné aproximované (predikované) zvoleným regresním modelem s parametry $\widehat{\boldsymbol{\beta}}$ určenými metodou nejmenších čtverců jsou

$$\widehat{Y} = X\widehat{\boldsymbol{\beta}} = X\left(X'X\right)^{-1}X'Y$$

- $\boldsymbol{r} = Y - \widehat{Y}$ jsou rezidua *(residuals)*, tj. odchylky skutečně pozorované hodnoty Y sledované veličiny od hodnoty $\widehat{Y}$ predikované regresním modelem
- Vhodnost modelu je kvantifikována reziduálním součtem čtverců

$$S_e = \boldsymbol{r}'\boldsymbol{r} = \left(Y - X\widehat{\boldsymbol{\beta}}\right)'\left(Y - X\widehat{\boldsymbol{\beta}}\right)$$

- Vhodnost modelu lze také posuzovat podle odhadu rozptylu náhodných chyb

$$s^2 = \frac{S_e}{n-k}$$

LRM – příklady (1)

31. V LRM $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$, $\boldsymbol{X} = \begin{pmatrix} 1 & -3 & 9 \\ 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix}$, $\boldsymbol{Y} = \begin{pmatrix} 7 \\ 4 \\ 2 \\ 2 \\ 5 \\ 8 \end{pmatrix}$ spočítejte MNČ-odhady vektoru parametrů $\widehat{\boldsymbol{\beta}}$, aproximace $\widehat{\boldsymbol{Y}}$, reziduální součty čtverců S_e a s^2. Jaké regresní funkci matice plánu odpovídá?

32. Ve dvou LRM $(\boldsymbol{Y}, \boldsymbol{X}_1, \boldsymbol{\beta})$ a $(\boldsymbol{Y}, \boldsymbol{X}_2, \boldsymbol{\beta})$,

$$\boldsymbol{X}_1 = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \quad \boldsymbol{X}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 4 \\ 9 \end{pmatrix}, \quad \boldsymbol{Y} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 2 \end{pmatrix},$$

na stejných datech spočítejte MNČ-odhady vektoru parametrů $\widehat{\boldsymbol{\beta}}$, aproximace $\widehat{\boldsymbol{Y}}$, reziduální součty čtverců S_e a s^2. Který model z obou modelů je vhodnější? Jakým regresním funkcím matice plánu odpovídají?

LRM – příklady (2)

V následujících příkladech spočítejte MNČ-odhady vektoru parametrů $\widehat{\boldsymbol{\beta}}$, aproximace $\widehat{Y}$, reziduální součty čtverců S_e a s^2 pro různé regresní funkce. Rozhodněte, která z funkcí je pro daná data nejvhodnější. Odhadnuté regresní funkce zobrazujte také graficky. Můžete též vyzoušet jeden bod z dat vynechat a sledovat, jak se parametry funkce změní.

- $y = \beta_0 + \beta_1 x$
- $y = \beta_0 + \beta_1 x + \beta_2 x^2$
- $y = \beta_0 + \beta_1 |x|$
- $y = \beta_1 x$
- $y = \beta_1 x + \beta_2 x^2$
- $y = \beta_0 + \beta_1 x + \beta_2 e^x$

33. Pro data

x	-2	-1	0	1	2
Y	-2	1	-2	1	-1

.

34. Pro data

x	-3	-1	0	1	3
Y	2	1	1	1	3

.

35. Pro data

x	-3	-2	-1	0	1	3
Y	-5	-3	-4	0	1	3

.

LRM – příklady (3)

36. Pro data

x	-2	-1	0	1	2
Y	-16	-6	0	4	-12

.

37. Pro data

x	-2	-1	0	1	2
Y	0	2	3	3	1

.

38. Pro data

x	-2	-1	0	1	2
Y	2	0	0	1	2

.

39. Pro data

x	-3	-1	0	1	2
Y	2	0	-2	-1	1

.

40. Pomocí regresní přímky procházející počátkem spočítejte MNČ-odhady vektoru parametrů $\widehat{\boldsymbol{\beta}}$, aproximace $\widehat{\boldsymbol{Y}}$, reziduální součty čtverců S_e a s^2 v LRM $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$ pro data

x	10	20	30	40	50	60
Y	0.18	0.35	0.48	0.65	0.84	0.97

Jedná se o měření koeficientu teplotní délkové roztažnosti měděné trubky. Teplotní rozdíl od 20 °C je x, prodloužení tyče je měřená veličina Y.

LRM – příklady (4)

41. Pomocí (obecné) regresní přímky spočítejte MNČ-odhady vektoru parametrů $\widehat{\boldsymbol{\beta}}$, aproximace $\widehat{\boldsymbol{Y}}$, reziduální součty čtverců S_e a s^2 v LRM $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$ pro data

x	40	64	34	15	57	45
Y	33	46	23	12	56	40

Jedná se o měření závislosti množství kyseliny mléčné ve 100 ml krve u matky-provorodičky, x, a jejího novorozence, Y. Obě veličiny jsou uváděny jako hmotnost v mg.

42. Pomocí (obecné) regresní paraboly spočítejte MNČ-odhady vektoru parametrů $\widehat{\boldsymbol{\beta}}$, aproximace $\widehat{\boldsymbol{Y}}$, reziduální součty čtverců S_e a s^2 v LRM $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$ pro data

x	40	50	60	70	80	90	100
Y	6.1	5.8	6.0	6.5	6.8	8.1	10.0

Jedná se o měření závislosti spotřeby paliva, Y v l/100 km motorového vozidla na rychlosti, x v km/h, při zařazeném stejném rychlostním stupni.

LRM v *R* prakticky – `lm`

- Uložíme data do proměnných:
  `x <- c (1, 2, 3, 4) ;   Y <- c (-1, 0, 2, 3)`
- Výpočet MNČ-odhadů (nejen) parametrů v *R*: funkce `lm` *(linear model)*
- *`formule`* je tvaru *`odpověď ~ matice plánu`* (regresní funkce), např.:
  - `Y ~ 1 + x`   obecná přímka (`1 +` není třeba psát)
  - `Y ~ 0 + x + I(x^2)`   parabola procházející počátkem (`0 +`)
  - `Y ~ I(abs(x))` $|x|$ (`1 +` si *R* doplní automaticky)
  - `Y ~ I(exp(x))` e^x (`1 +` si *R* doplní automaticky)
- `model <- lm (`*`formule`*`)`
- `summary(model)`   přehled výsledků modelu
- `coef(model)`, `model$coefficients` $\Rightarrow$ $\widehat{\beta}$
- `fitted(model)`, `model$fitted.values` $\Rightarrow$ $\widehat{Y}$
- `residuals(model)`, `model$residuals` $\Rightarrow$ $\boldsymbol{r}$
- `model$model` $\Rightarrow$ X (bez sloupce jedniček)
- Jak spočítáme S_e a s^2 `?`

LRM v *R* prakticky – obrázky

- Připravíme si prázdný graf, necháme více místa kolem zadaných dat:
  `plot (c(0,4), c(-2,4), type="n", xlab="osa x", ylab="osa y", main="nazev")`
- Vykreslíme data (x, Y) jako body (`pch`=číslo znaku, `col`=číslo barvy):
  `points (x, Y, pch=1, col=1)`
- Vykreslíme body predikované modelem $(x, \widehat{Y})$:
  `points (x, model$fitted.values, pch=18, col=2)`
- Predikce odpovědi modelu v libovolných bodech:
  `predict (model, data.frame (x=c(`*x-souřadnice bodů*`)))`
- Vykreslíme celou křivku odhadnuté regresní funkce (`lwd`=tloušťka čáry, `lty`=typ čáry, `from, to`=x-souřadnice intervalu, `add=TRUE` je důležité pro přidání křivky):
  `curve (predict (model, data.frame(x=x)), from=0, to=4, col=2, add=TRUE)`
- Opakováním předchozích kroků můžeme přikreslovat křivky dalších regresních funkcí, které máme spočítané.
- Kopie obrázku do PDF: `dev.copy2pdf (file="`*soubor*`.pdf")`

Rozdělení pravděpodobnosti v LRM

- máme bodové odhady $\widehat{\boldsymbol{\beta}}$, chceme intervalové odhady a testování hypotéz
- teorie klasického LRM předpokládá $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$
- Potom máme: $\boldsymbol{Y} \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2 \boldsymbol{I}\right)$ ($\boldsymbol{I}$ = jednotková matice řádu n)
- MNČ-odhad $\widehat{\boldsymbol{\beta}}$ vektoru parametrů je nestranný, $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}\left(\boldsymbol{\beta},\ \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$
- s^2 je nestranným odhadem rozptylu náhodných chyb, $\mathrm{E}(s^2) = \sigma^2$
- $\frac{S_e}{\sigma^2} \sim \chi^2(n-k)$
- náhodné veličiny $\widehat{\boldsymbol{\beta}}$ a s^2 jsou stochasticky nezávislé
- $$\mathrm{T} = \frac{\boldsymbol{c}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}'\boldsymbol{\beta}^0}{s\sqrt{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim \mathrm{t}(n-k) \qquad (\boldsymbol{c} \in \mathbb{R}^k)$$
- $$\mathrm{F} = \frac{\left(\widehat{\boldsymbol{\beta}^*} - \boldsymbol{\beta}^{*0}\right)' \boldsymbol{W}^{-1} \left(\widehat{\boldsymbol{\beta}^*} - \boldsymbol{\beta}^{*0}\right)}{s^2\ m} \sim \mathrm{F}(m,\ n-k)$$

• $\boldsymbol{\beta}^*$ je subvektor o m složkách

• $\boldsymbol{W}$ je tomuto subvektoru odpovídající blok $m \times m$ matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

• horní index 0 značí zvolený číselný vektor, např. při testování významnosti dosazujeme $\boldsymbol{\beta}^0 = (0, \cdots, 0_k)'$, resp. $\boldsymbol{\beta}^{*0} = (0, \cdots, 0_m)'$

Testy významnosti v LRM

Test významnosti koeficientu β_i, tj. test H_0: $\beta_i = 0$ proti H_1: $\beta_i \neq 0$:

- v T volíme $\boldsymbol{c} = \boldsymbol{e}_i$ (i–tý jednotkový vektor), $\boldsymbol{\beta}^0 = \boldsymbol{0}$
- $T_i = \dfrac{\widehat{\beta}_i - 0}{s\,\sqrt{v_{i,i}}} \sim \mathrm{t}(n-k)$, $\qquad v_{i,i} = \left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\}_{i,i}$

Test významnosti modelu, H_0: $(\beta_1, \ldots, \beta_{k-1}) = \boldsymbol{0}$ proti H_1: $\exists i > 0$: $\beta_i \neq 0$:

- v F volíme $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_0)$, $m = k-1$, $\boldsymbol{\beta}^{*0} = (0, \ldots, 0_{k-1})$
- $F = \dfrac{n-k}{k-1}\,\dfrac{S_T}{S_e} \sim \mathrm{F}(k-1, n-k)$
- rezidální součet čtverců: $S_e = \sum_{j=1}^{n}(Y_i - \widehat{Y}_i)^2$
- celkový součet čtverců: $S_T = \sum_{j=1}^{n}(Y_i - \overline{Y})^2$
- *R squared*: $R^2 = 1 - \frac{S_e}{S_T}$
- *R-bar squared, adjusted R squared*: $\overline{R}^2 = 1 - (1-R^2)\frac{n-1}{n-k}$
- formálně: R^2 = koeficient mnohonásobné korelace mezi $\boldsymbol{Y}$ a $\widehat{\boldsymbol{Y}}$

LRM v *R* prakticky – testy významnosti

- `summary (model)` ⇒ textový přehled
- `prehled <- summary (model)` ⇒ spočítá další parametry a testy v modelu
- `M <- coef(prehled)`, `M <- prehled$coefficients` ⇒ matice:
  - `M[,1]` ⇒ $\widehat{\beta}$, MNČ-odhady parametrů modelu
  - `M[,2]` ⇒ standardní chyby MNČ-odhadů parametrů modelu
  - `M[,3]` ⇒ T_i, t-statistiky významnosti koeficientů modelu
  - `M[,4]` ⇒ p-hodnoty testů $H_0\colon \beta_i = 0$ proti $H_1\colon \beta_i \neq 0$
- `prehled$r.squared` ⇒ R^2
- `prehled$adj.r.squared` ⇒ $\bar{R}^2$
- `prehled$fstatistic` ⇒ vektor $(F, k-1, n-k)$, F-statistika modelu
- `prehled$sigma` ⇒ s
- `prehled$df` ⇒ vektor $(k, n-k, k)$
- `model$df.residuals` ⇒ $n-k$

Obecné testy parametrů v LRM

Test lineární kombinace koeficientů H_0: $\boldsymbol{c}'\widehat{\boldsymbol{\beta}} = \boldsymbol{c}'\boldsymbol{\beta}^0$ proti H_1: $\boldsymbol{c}'\widehat{\boldsymbol{\beta}} \neq \boldsymbol{c}'\boldsymbol{\beta}^0$:

- $\boldsymbol{c} \in \mathbb{R}^k$ volíme podle požadované lineární kombinace
- $\boldsymbol{\beta}^0$ volíme tak, aby $\boldsymbol{c}'\boldsymbol{\beta}^0 \in \mathbb{R}$ byla testovaná hodnota
- $T = \dfrac{\boldsymbol{c}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}'\widehat{\boldsymbol{\beta}^0}}{s\sqrt{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim \mathrm{t}(n-k)$
- Příklad (parabola) • $\beta_0 + \beta_1 = 1$? $\Rightarrow$ volíme $\boldsymbol{c} = (1,1,0)'$, $\boldsymbol{c}'\boldsymbol{\beta}^0 = 1$
- Příklad (přímka) • $2\beta_0 - 3\beta_1 = 10$? $\Rightarrow$ volíme $\boldsymbol{c} = (2,3)'$, $\boldsymbol{c}'\boldsymbol{\beta}^0 = 10$

Vektorový test koeficientů H_0: $\widehat{\boldsymbol{\beta}^*} = \boldsymbol{\beta}^{*0}$ proti H_1: $\widehat{\boldsymbol{\beta}^*} \neq \boldsymbol{\beta}^{*0}$:

- testujeme subvektor $\boldsymbol{\beta}^*$ o m složkách ($m \leq k$)
- $F = \dfrac{\left(\widehat{\boldsymbol{\beta}^*} - \boldsymbol{\beta}^{*0}\right)' \boldsymbol{W}^{-1} \left(\widehat{\boldsymbol{\beta}^*} - \boldsymbol{\beta}^{*0}\right)}{s^2\ m} \sim \mathrm{F}(m,\ n-k)$
- $\boldsymbol{W}$ je testovanému subvektoru odpovídající blok $m \times m$ matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$
- Příklad • $(\beta_0, \beta_2)' = (0,0)'$?
- Příklad • $(\beta_0, \beta_1)' = (0,1)'$?

LRM – testy – příklady

V příkladech 33.–42. dále spočítejte:

- testujte významnost koeficientů β_i $(i = 0, \dots, k-1)$ pomocí statistik T_i
- testujte významnost modelu pomocí statistiky F
- porovnejte vhodnost regresních funkcí pomocí F, s^2, R^2 a $\overline{R}^2$

Poznámka: alternativně ke klasickému rozhodování o výsledku statistického testu na hladině významnosti α (porovnávání hodnoty testovací statistiky s příslušným kvantilem) lze o výsledku testu rozhodnout pomocí p-hodnoty *(p-value)* P:

- $P \leq \alpha \Rightarrow$ zamítáme H_0 ve prospěch H_1
- $P > \alpha \Rightarrow$ nezamítáme H_0

Korelační koeficient

Korelační koeficient (Pearsonův) ρ_{XY}:

- $\rho_{XY} = \dfrac{\mathrm{C}(X,Y)}{\sqrt{\mathrm{D}(X)}\sqrt{\mathrm{D}(Y)}}$
- $\rho_{XY} \in [-1;1]$
- je mírou lineární závislosti náhodných veličin X, Y (s kladnými rozptyly)
- Kovariance: $\mathrm{C}(X,Y) = \mathrm{E}[X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]$

Pro normálně rozdělené náhodné veličiny lze interpretovat pomocí LRM s regresní funkcí $\widehat{Y} = \beta_0 + \beta_1 X$:

- $\rho_{XY} = 1 \;\Rightarrow$ pozitivní lineární závislost, $\beta_1 > 0$ je významný
- $\rho_{XY} = -1 \;\Rightarrow$ negativní lineární závislost, $\beta_1 < 0$ je významný
- $\rho_{XY} = 0 \;\Rightarrow$ lineární nezávislost, β_1 není významný

Obecně platí implikace: X, Y stochasticky nezávislé $\Rightarrow \rho_{XY} = 0$ (nezávislost $\Rightarrow$ nekorelovanost)
Pro normálně rozdělení náhodně veličiny platí ekvivalence (nezávislost $=$ nekorelovanost)

Výběrový korelační koeficient

Výběrový korelační koeficient r_{XY}:

- je empirickou analogií korelačního koeficientu ρ_{XY}
- pro náhodný výběr $\big((X_1, Y_1)', \ldots, (X_n, Y_n)'\big)$
- $r_{XY} = \dfrac{S_{XY}}{S_X\, S_Y}$
- Výběrový rozptyl: $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\overline{X}^2 \right)$
- Výběrová kovariance:
  $S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n-1} \left(\sum_{i=1}^n X_i Y_i - n\overline{X}\,\overline{Y} \right)$
- Testy pro normálně rozdělené náhodné veličiny:
  - Test nezávislosti H_0: $\rho_{XY} = 0$ proti H_1: $\rho_{XY} \neq 0$:
    $$T = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sqrt{n-2} \sim \mathrm{t}(n-2)$$
  - Test H_0: $\rho_{XY} = \rho_0$ proti H_1: $\rho_{XY} \neq \rho_0$ (tzv. Z-transformace):
    $$U = \Bigg(\underbrace{\frac{1}{2} \ln \frac{1 + r_{XY}}{1 - r_{XY}}}_{Z} - \frac{1}{2} \ln \frac{1 + \rho_0}{1 - \rho_0} - \frac{\rho_0}{2(n-1)}\Bigg) \sqrt{n-3} \sim \mathrm{N}(0, 1)$$

Výběrový koeficient mnohonásobné korelace

Výběrový koeficient mnohonásobné korelace $r_{Y\cdot X}$:

- $r^2_{Y\cdot X} = \boldsymbol{R}_{YX}\boldsymbol{R}^{-1}_{XX}\boldsymbol{R}_{XY}$
- je empirickou analogií koeficientu mnohonásobné korelace $\rho_{Y\cdot X}$
- $r_{Y\cdot X} \approx \rho_{Y\cdot X} = \rho(Y, \widehat{Y})$
- $\rho_{Y\cdot X} \in [0; 1]$
- Koeficient determinace: $R^2 = r^2_{Y\cdot X}$
- pro náhodný výběr $\big((Y_1, \boldsymbol{X}_1)', \dots, (Y_n, \boldsymbol{X}_n)'\big)$ z $(k+1)$-rozměrného rozdělení
- Test pro normálně rozdělené náhodné veličiny H_0: $\rho_{Y\cdot X} = 0$ proti H_1: $\rho_{Y\cdot X} \neq 0$, tj. že Y nezávisí na komplexu náhodných veličin $\boldsymbol{X}$:
  $$F = \frac{n-k-1}{k}\,\frac{r^2_{Y\cdot X}}{1-r^2_{Y\cdot X}} \sim \mathrm{F}(k, n-k-1)$$

Výběrový koeficient parciální korelace

Výběrový koeficient parciální korelace $r_{Y,Z\cdot X}$:

- $$r_{Y,Z\cdot X} = \frac{r_{YZ} - \boldsymbol{R}_{YX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XZ}}{\sqrt{\left(1 - \boldsymbol{R}_{YX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY}\right)\left(1 - \boldsymbol{R}_{ZX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XZ}\right)}}$$
- je empirickou analogií parciálního korelačního koeficientu $\rho_{Y,Z\cdot X}$
- $r_{Y,Z\cdot X} \approx \rho_{Y,Z\cdot X} = \rho(Y - \widehat{Y}, Z - \widehat{Z})$
- $\rho_{Y,Z\cdot X} \in [-1; 1]$
- pro náhodný výběr $\left((Y_1, \boldsymbol{X}_1, Z_1)', \ldots, (Y_n, \boldsymbol{X}_n, Z_n)'\right)$ z $(k+2)$-rozměrného rozdělení
- Test pro normálně rozdělené náhodné veličiny H_0: $\rho_{Y,Z\cdot X} = 0$ proti H_1: $\rho_{Y,Z\cdot X} \neq 0$, tj. že Y, Z jsou nezávislé náhodné veličiny po odečtení (lineárního) vlivu komplexu náhodných veličin $\boldsymbol{X}$:
  $$T = \frac{r_{Y,Z\cdot X}}{\sqrt{1 - r^2_{Y,Z\cdot X}}} \sim \mathrm{t}(n-k-2)$$

Výběrové korelační koeficienty v *R*

- předpokládáme, že v proměnných `X, Y, Z` máme realizace náhodného výběru
- `cor (X, Y)` $\Rightarrow$ výběrový korelační koeficient r_{XY}
- Výběrový koeficient mnohonásobné korelace $r_{Z\cdot(X,Y)'}$
  - `mZ <- lm (1 + X + Y)` $\Rightarrow$ LRM na proměnných X, Y
  - `Zh <- mZ$fitted.values` $\Rightarrow \widehat{Z} = \widehat{\beta_0} + \widehat{\beta_1}X + \widehat{\beta_2}Y$
  - `cor (Z, Zh)` $\Rightarrow r_{Z\cdot(X,Y)'} = R(Z, \widehat{Z})$
- Výběrový koeficient parciální korelace $r_{Y,Z\cdot X}$
  - `mY <- lm (1 + X)` $\Rightarrow$ LRM na proměnné X
  - `mZ <- lm (1 + X)` $\Rightarrow$ LRM na proměnné X
  - `Yr <- mY$residuals` $\Rightarrow$ rezidua $Y - \widehat{Y} = Y - \widehat{\beta_0} - \widehat{\beta_1}X$
  - `Zr <- mZ$residuals` $\Rightarrow$ rezidua $Z - \widehat{Z} = Z - \widehat{\alpha}_0 - \widehat{\alpha}_1 X$
  - `cor (Yr, Zr)` $\Rightarrow r_{Y,Z\cdot X} = R(Y - \widehat{Y}, Z - \widehat{Z})$