M5120 Lineární statistické modely I
Poznámky do cvičení

Ondřej Pokora (pokora@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

podzim 2013 (aktualizace 16.12.2013)

Maximálně věrohodné odhady

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

- $X \sim P$ ($i = 1, \ldots, n$)
- $X_1, \ldots, X_n$ 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 $l(\theta)$:

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

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

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

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 $L(\theta)$:

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

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 $\mathrm{Bi}(N, \theta)$ ($N \in \mathbb{N}$ známé) s pravděpodobnostní funkcí

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

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

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

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 $\mathrm{LN}(\mu, \sigma^2)$:

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

MLE – příklady (2)

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

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

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

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

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

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

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

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

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

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

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. $\widehat{\theta}_{\mathrm{ML}} = \dfrac{\overline{X}}{N}$

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. $\widehat{\lambda}_{\mathrm{ML}} = \dfrac{k}{\overline{X}}\ ,$ 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 $\widehat{\theta}_{\mathrm{M}}$ 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,

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

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.	$N\theta$	$N\theta(1-\theta)$
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.	$\sqrt{\frac{\pi s}{2}}$	$2s$
9.	$\frac{k}{\lambda}$	$\frac{k(k+1)}{\lambda^2}$

Momentové odhady – řešení příkladů

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

1. $\widehat{\theta}_{\mathrm{M}} = \dfrac{m_1}{N}$

2. $\widehat{\theta}_{\mathrm{M}} = \dfrac{1}{1 + m_1}$

3. $\widehat{\mu}_{\mathrm{M}} = m_1 \;, \qquad \widehat{\sigma^2}_{\mathrm{M}} = m_2 - m_1^2$

4. $\widehat{\mu}_{\mathrm{M}} = 2 \ln m_1 - \frac{1}{2} \ln m_2 \;, \qquad \widehat{\sigma^2}_{\mathrm{M}} = \ln m_2 - 2 \ln m_1$

5. $\widehat{\lambda}_{\mathrm{M}} = m_1$

6. $\widehat{\lambda}_{\mathrm{M}} = \dfrac{1}{m_1}$

8. $\widehat{s}_{\mathrm{M}} = \dfrac{2\, m_1^2}{\pi}$, anebo $\widehat{s}_{\mathrm{M}} = \dfrac{m_2}{2}$

9. $\widehat{\lambda}_{\mathrm{M}} = \dfrac{m_1}{m_2 - m_1^2} \;, \qquad \widehat{k}_{\mathrm{M}} = \dfrac{m_1^2}{m_2 - m_1^2}$

Náhodné vektory

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

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

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

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

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

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

Poznámka: Anglická literatura často užívá společný pojem *variance* (var, Var). Potom rozumíme: $\operatorname{var}(X) = \mathsf{D}(X)$ pro náhodnou veličinu a $\operatorname{var}(\boldsymbol{X}) = \operatorname{cov}(\boldsymbol{X})$ 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 $\mathbb{R}^n \to \mathbb{R}$:

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

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

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

Známé vzorce ve vektorové (maticové) formě

- i-tá složka náhodného vektoru

$$X_i = \boldsymbol{e}_i' \boldsymbol{X}, \qquad \text{kde } \boldsymbol{e}_i = (0_1, \dots, 0_{i-1}, 1_i, 0_{i+1}, \dots, 0_n)' \text{ je jednotkový vektor}$$

- výběrový průměr

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} (1, \cdots, 1) \boldsymbol{X} = \left(\tfrac{1}{n}, \cdots, \tfrac{1}{n} \right) \boldsymbol{X}$$

- druhý výběrový moment

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 = \frac{1}{n} \boldsymbol{X}' I_n \boldsymbol{X}, \qquad \text{kde } I_n \text{ je jednotková matice řádu } n$$

- čtverec výběrového průměru

$$(\overline{X})^2 = \overline{X}' \overline{X} = \frac{1}{n^2} \boldsymbol{X}' J_n \boldsymbol{X}, \qquad \text{kde } J_n = \{1\}_{i,j=1}^{n} \text{ je matice jedniček}$$

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: $\mathrm{E}(X_i) = 10\,i,\ \mathrm{D}(X_i) = i^2,\ (i, j = 1, 2, 3);\ \mathrm{R}(X_i, X_j) = 0{,}5$ 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:

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

Náhodné vektory – příklady (3)

23. Spočtěte střední hodnotu $m = \mathsf{E}\left(Y_1^2 + Y_2^2 + Y_3^2\right)$ 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, $\mathsf{E}(X_1) = 10,\ \mathsf{E}(X_2) = 20,\ \mathsf{E}(X_3) = 30,\ \mathsf{D}(X_1) = 1,\ \mathsf{D}(X_2) = 4, \mathsf{D}(X_3) = 9$.

24. Spočtěte střední hodnotu $m = \mathsf{E}\left(Y_1Y_2 + Y_2Y_3 + Y_3Y_1\right)$ 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 $\boldsymbol{X}$, $\mathsf{E}(\boldsymbol{X}) = (10, 10, 10)'$, $\mathsf{D}(X_i) = i^2$.

25. Spočtěte $\mathsf{E}(\boldsymbol{Y})$, $\mathrm{cov}(\boldsymbol{Y})$ a $m = \mathsf{E}\left(Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + 2Y_1Y_4\right)$, kde X_1, X_2, X_3, X_4 jsou náhodné veličiny, $\mathsf{E}(X_i) = 10$, $\mathsf{C}(X_i, X_j) = 1$. 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.

Náhodné vektory – příklady (3)

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^2_X náhodného výběru rozsahu n z rozdělení $\mathrm{N}(\mu, \sigma^2)$ platí: $\mathrm{E}\left(S^2_X\right) = \sigma^2$.
Bonusová těžší úloha: ověřte, že $\mathrm{D}\left(S^2_X\right) = 2\sigma^4/(n-1)$.

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

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

Náhodné vektory – řešení příkladů

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

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

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

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

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

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

Lineární regresní model (LRM)

- popisuje lineární závislost

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

- Y je závislá proměnná, náhodná veličina
- $x_1, \dots, x_p$ jsou nenáhodné vysvětlující proměnné, tzv. regresory
- ε je náhodná chyba, $\mathsf{E}(\varepsilon) = 0$, s neznámým konstatním rozptylem $\mathsf{D}(\varepsilon) = \sigma^2$
- $\beta_0, \beta_1, \dots, \beta_p$ jsou neznámé parametry
- úkol regresní analýzy: na základě opakovaných měření závislé proměnné za různých hodnot regresorů optimálně určit parametry $\beta_0, \beta_1, \dots, \beta_p$ modelu
- předpokládáme, že měření je opakováno n krát, tzn. pro $i = 1, \dots, n$ máme:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i$$

- náhodné chyby $\varepsilon_1, \dots, \varepsilon_n$ jsou nezávislé stejně rozdělené náhodné veličiny (i.i.d.)

LRM $(Y, \mathbf{X}\beta, \sigma^2 I)$ plné hodnosti

- $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i$ v maticovém tvaru:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ \vdots \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix}}_{\mathbf{X}} \cdot \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

- $\boldsymbol{Y}$ je náhodný vektor n pozorování
- regresory tvoří nenáhodnou $n \times k$ matici plánu *(design matrix)* $\mathbf{X}$
- $\boldsymbol{\beta}$ je vektor $k = p + 1$ neznámých parametrů
- závislost je lineární vzhledem k parametrům β_j
- vektor náhodných chyb má kovariační matici $\mathrm{D}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$
- $n > k = p + 1$, tj. počet pozorování je větší než počet parametrů
- matice plánu je plné hodnosti: $h(\mathbf{X}) = k = p + 1$

Metoda nejmenších čtverců (MNČ)

- V LRM $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ chceme najít vektor parametrů $\boldsymbol{\beta}$ tak, aby naměřené hodnoty $\boldsymbol{Y}$ byly optimálně aproximovány vektorem $\mathbf{X}\boldsymbol{\beta}$.
- Metoda nejmenších čtverců *(ordinary least-square method)* stanovuje odhad $\widehat{\boldsymbol{\beta}}$ jako bod minima penalizační funkce, která je součtem čtverců odchylek:

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} x_{i,j}\,\beta_j \right)^2 = (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) \longrightarrow \min_{\boldsymbol{\beta}} \Longrightarrow \widehat{\boldsymbol{\beta}}$$

- Při splnění podmínek pro LRM plné hodnosti existuje vždy právě jedno řešení této minimalizační úlohy. To lze nalézt vyřešením soustavy normálních rovnic

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\boldsymbol{Y}$$

- Odhad vektoru parametrů v LRM metodou nejmenších čtvreců je tedy tvaru

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

Poznámka: při numerických výpočtech se inverzní matice nepočítá přímo, ale využívá se např. Q-R rozkladu.

Veličiny v LRM (1)

- Aproximované hodnoty závislé proměnné *(fitted values, Y-hat)*

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

- Rezidua *(residuals)*: $\boldsymbol{r} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$
- Reziduální součet čtverců

$$S_e = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \boldsymbol{r}'\boldsymbol{r} = (\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

  kvantifikuje velikost variability, kterou se nepodařilo LRM vysvětlit.
- Odhad rozptylu σ^2 náhodných chyb

$$\widehat{\sigma^2} = s^2 = \frac{S_e}{n-k} = \frac{S_e}{n-p-1}$$

- Standardní reziduální chyba *(residual standard error)*: $\widehat{\sigma} = s = \sqrt{s^2}$

Veličiny v LRM (2)

- Regresní součet součet čtverců

$$S_r = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$$

  kvantifikuje velikost variability, kterou se LRM podařilo zachytit. Je dán součtem kvadrátů odchylek aproximovaných hodnot od výběrového průměru.
- Celkový součet součet čtverců je násobkem výběrového rozptylu:

$$S_t = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = (n-1)S_Y^2$$

- Platí: $S_t = S_r + S_e$

Koeficient determinace v LRM

- Koeficient determinace *(coefficient of determination, R-squared)*

$$R^2 = 1 - \frac{S_e}{S_t} = \frac{S_r}{S_t} = \frac{S_r/n}{S_t/n} = \frac{\text{ML-odhad vysvětleného rozptylu}}{\text{ML-odhad celkového rozptylu}}$$

  se používá ke kvantifikaci poměrné části variability, kterou se LRM podařilo vysvětlit
- $R^2 \in [0, 1]$
- V LRM platí: R^2 = koeficient mnohonásobné korelace = korelační poměr
- Korigovaný koeficient determinace *(adjusted R-squared, R-bar-squared)*

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} =$$

$$= 1 - \frac{S_e/(n-k)}{S_t/(n-1)} = 1 - \frac{s^2}{S_Y^2} = 1 - \frac{\text{odhad rozptylu chyb}}{\text{výběrový celkový rozptyl}}$$

  lze použít pro porovnání různých podmodelů LRM

Řešení LRM pomocí *R*

V *R* řeší LRM příkaz `lm` (*linear model*):

`model <- lm (`*formule*`, data=`*tabulka*`)`

Máme-li zvlášť vektor regresorů ($\boldsymbol{x}$) a pozorování ($\boldsymbol{Y}$), datovou tabulku (*data frame*) vytvoříme příkazem: `tabulka <- data.frame (x, Y)`

Zápis tzv. *formule* pro některé regresní funkce:

$Y = \beta_0 + \beta_1 x$	`Y ~ x`, nebo `Y ~ 1 + x`, abolutní člen je vkládán implicitně
$Y = \beta_1 x$	`Y ~ 0 + x`
$Y = \beta_0 + \beta_1 x + \beta_2 x^2$	`Y ~ x + I(x^2)`
$Y = \beta_1 \lvert x \rvert$	`Y ~ 0 + I(abs(x))`
$Y = \beta_0 + \beta_1 \mathrm{e}^x$	`Y ~ I(exp(x))`
$Y = \beta_0 + \beta_1 \ln x$	`Y ~ I(log(x))`

Detailní výsledky a další číselné charakteristiky získáme příkazem `vysledky <- summary (model)`,
příp. s parametrem `correlation=TRUE` pro výpočet výběrové korelační matice parametrů.

Veličiny v LRM pomocí *R*

$\widehat{\boldsymbol{\beta}}$	MNČ-odhady parametrů	`model$coefficients` `coef(model)`
$\left(\widehat{\beta}_j, \mathrm{SD}(\widehat{\beta}_j), T_j, p_j\right)$	odhady, směrodatné odchylky, testy významnosti, p-hodnoty	`vysledky$coefficients` `coef(vysledky)`
$\widehat{\boldsymbol{Y}}$	aproximované hodnoty	`model$fitted.values` `fitted.values(model)`
$\boldsymbol{r}$	rezidua	`model$residuals` `residuals(model)`
$n-k$	stupně volnosti	`model$df.residual`
$\mathbf{X}$	matice plánu	`model.matrix(model)`
s	odhad směrod. odchylky chyb	`vysledky$sigma`
R^2	koeficient determinace	`vysledky$r.squared`
$\overline{R}^2$	korigovaný koef. determinace	`vysledky$adj.r.squared`
$(F, k-1, n-k)$	celkový F-test	`vysledky$fstatistic`
$(k, n-k, k)$	stupně volnosti	`vysledky$df`
$\mathrm{R}(\widehat{\boldsymbol{\beta}})$	korelační matice pro $\widehat{\boldsymbol{\beta}}$	`vysledky$correlation`

Grafy regresních funkcí v *R*

Připravíme souřadnicový systém vhodných rozměrů (první dva vektory min–max), do nějž zatím nic kreslit nebudeme (`type="n"`), přidáme popisky:

```
plot (c(0,70), c(-5,60), type="n", xlab="osa x", ylab="osa y")
```

Pomocí bodů vykreslíme data z tabulky, tzn. $(\mathbf{x}, \boldsymbol{Y})$. První dva parametry jsou vektory x-ových a y-ových souřadnic, následují grafické parametry:

```
points (tabulka$x, tabulka$Y, col=4, pch=24, lwd=1.5, cex=1.0)
```

Zvolíme si dostatečně hustou síť x-ových souřadnic $(\mathbf{x}_*)$:

```
xx <- seq (0, 70, by=0.1)
```

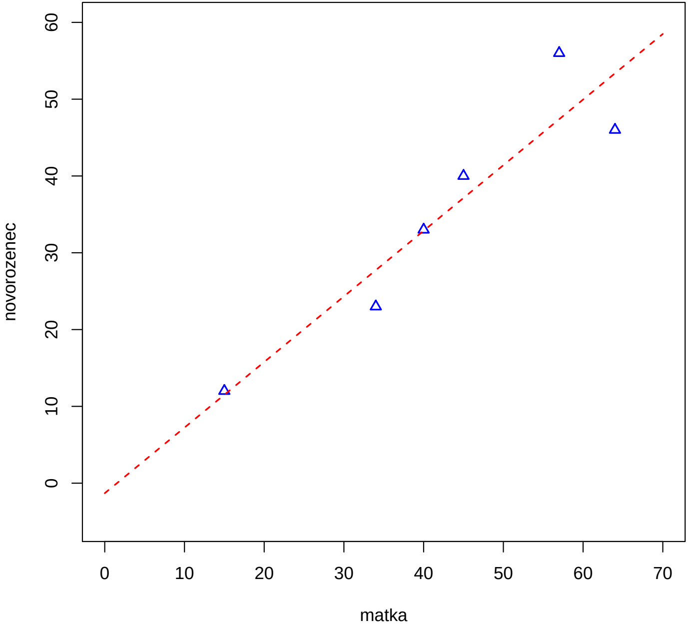
Dopočítáme k nim odpovídající y-ové souřadnice, tzn. $\widehat{\boldsymbol{Y}_*}$:

```
YY <- predict (model, data.frame (x=xx))
```

Vykreslíme graf funkce jako křivku $(\mathbf{x}_*, \widehat{\boldsymbol{Y}_*})$, podobně jako body:

```
lines (xx, YY, col=2, lwd=1.5, lty=2)
```

Obrázek můžeme uložit mj. příkazem `dev.copy2pdf(file="`*`obrazek`*`.pdf")`


novorozenec
matka
0
10
20
30
40
50
60
70

LRM – příklady (1)

Datové soubory k následujícím příkladům jsou dostupné na serveru `bart` v adresáři `/erko/M5120/data`, odkud si je můžete zkopírovat. Pro jednotlivé soubory postupně navrhněte několik různých regresních modelů, spočítejte odhady parametrů a další číselné charakteristiky, a modely porovnejte. Vykreslete do jednoho obrázku data i grafy spočítaných regresních funkcí. Později přidejte testování významnosti modelu (F-test) a testy významnosti parametrů (t-testy).

31. Datový soubor `C3H603.txt`. 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 $(\boldsymbol{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$ 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.

LRM – příklady (2)

32. Datový soubor `roztaznost.txt.` 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 $(\boldsymbol{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$ 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.

33. Datový soubor `palivo.txt`. 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 $(\boldsymbol{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$ 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 – příklady (3)

34. `carb_dio.tab`: Zkoumejte závislost koncentrace CO_2 v atmosféře v letech 1764–1995 pomocí regresní funkce tvaru vhodného polynomu a exponeneciální funkce. Predikujte hodnoty pro 21. století.

35. `carbon_e.tab`: Zkoumejte závislost uhlíkových emisí v letech 1950–1995 pomocí regresní funkce tvaru vhodného polynomu. Predikujte hodnoty pro 21. století.

36. `globtemp.txt`: Zkoumejte závislost průměrné teploty v letech 1866–1996 pomocí regresní funkce tvaru vhodného polynomu. Predikujte hodnoty pro 21. století.

37. `oil_prod.txt`: Zkoumejte závislost objemu vytěžené ropy v letech 1880–1988 pomocí regresní funkce tvaru vhodného polynomu a exponeneciální funkce. Predikujte hodnoty pro 21. století.

38. `population.txt`: Zkoumejte závislost velikosti populace na Zemi na čase pomocí regresní funkce tvaru vhodného polynomu a exponeneciální funkce. Predikujte hodnoty pro 21. století.

LRM – příklady (4)

U vícerozměrných dat dále spočítejte výběrové korelační koeficienty a výběrové parciální korelační koeficienty. Obdržené hodnoty intepretujte.

39. `beh.txt`: Příklad na vícerozměrnou regresi (maratonské běžkyně, 1977). Pomocí regresních přímek zkoumejte závislosti mezi třemi veličinami: rychlost běhu, koroková frekvence, délka kroku.

40. `deti.txt`: Příklad na vícerozměrnou regresi. Pomocí regresních přímek zkoumejte závislosti mezi třemi veličinami: hmotností dítěte, věkem a počtem bodů z diktátu.

41. `domacnosti1957.txt`: Příklad na vícerozměrnou regresi (ČSR, 1957). Pomocí regresních přímek zkoumejte závislosti mezi třemi veličinami: počtem členů domácnosti, příjmy a výdaji.

42. `enrollment.txt`: Příklad na vícerozměrnou regresi (VŠ v USA). Pomocí regresních přímek zkoumejte závislosti mezi veličinami: počet přihlášek na vysokou školu (ROLL), míra nezaměstnanosti (UNEM), počet absolventů střední školy (HGRAD) a průměrný příjem (INC).

Rozdělení pravděpodobnosti v LRM

- máme bodové odhady $\widehat{\boldsymbol{\beta}}$, chceme intervalové odhady a testování hypotéz
- teorie klasického LRM předpokládá $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 I_n)$
- Potom máme: $\boxed{\boldsymbol{Y} \sim \mathrm{N}\left(\boldsymbol{X\beta},\ \sigma^2 I_n\right)}$
- 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
- $\boxed{\mathrm{T}} = \dfrac{\boldsymbol{c}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}'\boldsymbol{\beta}^0}{s\sqrt{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} \boxed{\sim \mathrm{t}(n-k)}$ $\qquad (\boldsymbol{c} \in \mathbb{R}^k)$
- $\boxed{\mathrm{F}} = \dfrac{\left(\widehat{\boldsymbol{\beta}^*} - \boldsymbol{\beta}^{*0}\right)' \boldsymbol{W}^{-1} \left(\widehat{\boldsymbol{\beta}^*} - \boldsymbol{\beta}^{*0}\right)}{s^2\, m} \boxed{\sim \mathrm{F}(m,\ n-k)}$
  - $\boldsymbol{\beta}^*$ 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), $\boldsymbol{\beta}^0 = \mathbf{0}$
- $T_i = \dfrac{\widehat{\beta}_i - 0}{s\,\sqrt{v_{i,i}}} \sim \mathrm{t}(n-k), \qquad v_{i,i} = \left\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\right\}_{i,i}$

Test významnosti modelu, H_0: $(\beta_1, \dots, \beta_p) = \mathbf{0}$ proti H_1: $\exists i > 0$: $\beta_i \neq 0$:

- v F volíme $\boldsymbol{\beta}^* = (\beta_1, \dots, \beta_p)$, $m = k - 1$, $\boldsymbol{\beta}^{*0} = (0, \dots, 0_p)$
- $F = \dfrac{n-k}{k-1}\,\dfrac{S_r}{S_e} = \dfrac{n-k}{k-1}\left(\dfrac{S_t}{S_e} - 1\right) \sim \mathrm{F}(k-1, n-k)$

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 • $(\beta_0, \beta_1)' = (0,1)'$?

Korelační koeficient

Korelační koeficient (Pearsonův) R_{XY}:

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

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

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

Obecně platí implikace: X, Y stochasticky nezávislé $\Rightarrow R_{XY} = 0$ (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 R_{XY}
- pro náhodný výběr $\big((X_1, Y_1)', \dots, (X_n, Y_n)'\big)$
- $r_{XY} = \dfrac{S_{XY}}{S_X\, S_Y}$
- 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: $\mathrm{R}_{XY} = 0$ proti H_1: $\mathrm{R}_{XY} \neq 0$:
    $$T = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sqrt{n-2} \sim \mathrm{t}(n-2)$$
  - Test H_0: $\mathrm{R}_{XY} = \mathrm{R}_0$ proti H_1: $\mathrm{R}_{XY} \neq \mathrm{R}_0$ (tzv. Z-transformace):
    $$U = \left(\underbrace{\frac{1}{2} \ln \frac{1 + r_{XY}}{1 - r_{XY}}}_{Z} - \frac{1}{2} \ln \frac{1 + \mathrm{R}_0}{1 - \mathrm{R}_0} - \frac{\mathrm{R}_0}{2(n-1)} \right) \sqrt{n-3} \sim \mathrm{N}(0, 1)$$

Výběrový koeficient mnohonásobné korelace

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

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

Výběrový koeficient parciální korelace

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

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

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` $\Rightarrow \widehat{Z} = \widehat{\beta_0} + \widehat{\beta_X} X + \widehat{\beta_Y} Y$
  - `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 $Y - \widehat{Y} = Y - \widehat{\beta_0} - \widehat{\beta_X} X$
  - `Zr <- mZ$residuals` $\Rightarrow$ rezidua $Z - \widehat{Z} = Z - \widehat{\alpha_0} - \widehat{\alpha_Z} X$
  - `cor (Yr, Zr)` $\Rightarrow r_{Y,Z\cdot X} = R(Y - \widehat{Y}, Z - \widehat{Z})$