M5120 – cvičení

Lineární regresní model

Ondřej Pokora (pokora@math.muni.cz)

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

(podzim 2015)

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

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

- náhodné chyby $\varepsilon_1, \ldots, \varepsilon_n$ 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
- $R^2 \in [0,1]$
- $R^2 =$ koeficient mnohonásobné korelace $=$ korelační poměr
- Korigovaný koeficient determinace *(adjusted R-squared, R-bar-squared)*

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

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

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`
$\widehat{R}^2$	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)
```

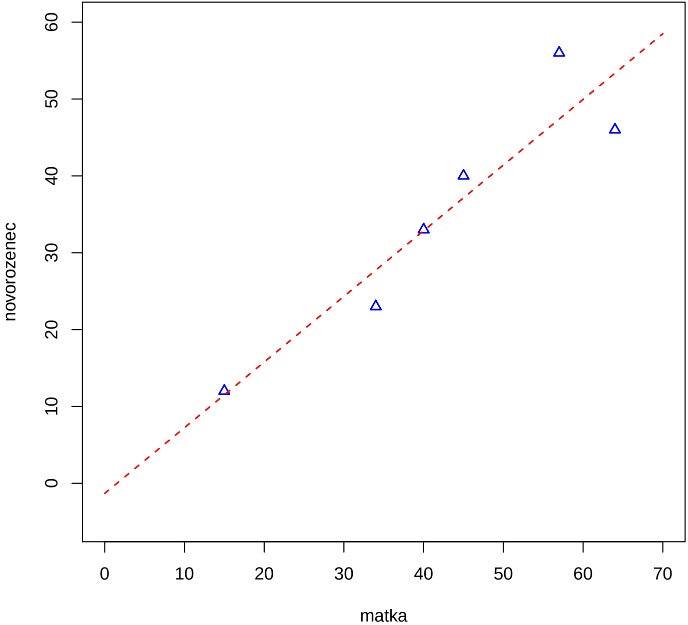
Dopočítáme k nim odpovídající y-ové souřadnice, tzn. $\widehat{\boldsymbol{Y}_*}$:

```
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 v adresáři `/Vyuka/R/M5120/data`, odkud si je zkopírujte. Pro jednotlivé soubory řešte lineární regresní modely, 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).

Příklad 1 (Datový soubor `C3H603.txt`)

Pomocí regresní přímky (a regresní paraboly) spočítejte MNČ-odhady vektoru parametrů $\widehat{\boldsymbol{\beta}}$*, aproximace* $\widehat{\mathbf{Y}}$*, reziduální součty čtverců* S_e *a* s^2 *v LRM* $(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$.

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

Tabulka uvádí výsledky 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)

Příklad 2 (Datový soubor `roztaznost.txt`)

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

Příklad 3 (Datový soubor `palivo.txt`)

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

Příklad 4 (`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.

Příklad 5 (`carbon_e.tab`)

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

Příklad 6 (`globtemp.txt`)

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

Příklad 7 (`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.

LRM – příklady (4)

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

Příklad 8 (`population.txt`)

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

Příklad 9 (`beh.txt`)

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

Příklad 10 (`deti.txt`)

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

LRM – příklady (5)

Příklad 11 (`domacnosti1957.txt`)

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

Příklad 12 (`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: $\boldsymbol{Y} \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2 I_n\right)$
- 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
- $\mathsf{T} = \dfrac{\boldsymbol{c}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}'\boldsymbol{\beta}^0}{s\ \sqrt{\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim \mathrm{t}(n-k)$ $\qquad (\boldsymbol{c} \in \mathbb{R}^k)$
- $\mathsf{F} = \dfrac{\left(\widehat{\boldsymbol{\beta}^*} - \boldsymbol{\beta}^{*0}\right)' \boldsymbol{W}^{-1} \left(\widehat{\boldsymbol{\beta}^*} - \boldsymbol{\beta}^{*0}\right)}{s^2\ m} \sim \mathrm{F}(m,\ n-k)$
  - $\boldsymbol{\beta}^*$ 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: $(\beta_1, \ldots, \beta_p) = \mathbf{0}$ proti H_1: $\exists i > 0$: $\beta_i \neq 0$:

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

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}:

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

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

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

Obecně platí implikace: X, Y stochasticky nezávislé $\Rightarrow \mathsf{R}_{XY} = 0$ (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):
    $$U = \Bigg(\underbrace{\frac{1}{2} \ln \frac{1 + r_{XY}}{1 - r_{XY}}}_{Z} - \frac{1}{2} \ln \frac{1 + \mathrm{R}_0}{1 - \mathrm{R}_0} - \frac{\mathrm{R}_0}{2(n-1)}\Bigg) \sqrt{n-3} \sim \mathrm{N}(0,1)$$

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 $\mathsf{R}_{Y\cdot\boldsymbol{X}}$
- $r_{Y\cdot\boldsymbol{X}} \approx \mathsf{R}_{Y\cdot\boldsymbol{X}} = \mathsf{R}(Y, \widehat{Y})$
- $\mathsf{R}_{Y\cdot\boldsymbol{X}} \in [0;1]$
- 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: $\mathsf{R}_{Y\cdot\boldsymbol{X}} = 0$ proti H_1: $\mathsf{R}_{Y\cdot\boldsymbol{X}} \neq 0$, 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 $r_{Y,Z\cdot \boldsymbol{X}}$:

- $r_{Y,Z\cdot \boldsymbol{X}} = \dfrac{r_{YZ} - \boldsymbol{R}_{YX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XZ}}{\sqrt{\left(1 - \boldsymbol{R}_{YX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XY}\right)\left(1 - \boldsymbol{R}_{ZX}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{XZ}\right)}}$
- 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 $\left((Y_1, \boldsymbol{X}_1, Z_1)', \ldots, (Y_n, \boldsymbol{X}_n, Z_n)'\right)$ z $(k+2)$-rozměrného rozdělení
- Test pro normálně rozdělené náhodné veličiny H_0: $\mathrm{R}_{Y,Z\cdot \boldsymbol{X}} = 0$ 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 $\boldsymbol{X}$:
  $T = \dfrac{r_{Y,Z\cdot \boldsymbol{X}}}{\sqrt{1 - r_{Y,Z\cdot \boldsymbol{X}}^2}} \sim \mathrm{t}(n-k-2)$

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)` $\Rightarrow r_{Y,Z\cdot X} = R(Y - \widehat{Y}, Z - \widehat{Z})$