Matematická statistika 2

Mirko Navara
Centrum strojového vnímání
katedra kybernetiky FEL ČVUT
Karlovo náměstí, budova G, místnost 104a
http://cmp.felk.cvut.cz/~navara

6. listopadu 2012

Obsah

1 Podmíněná rozdělení pravděpodobnosti

1.1 Motivační úloha

Příklad (počet es)**:**

1. Hráč má 10 z balíčku 32 karet, v němž jsou 4 esa. Náhodná veličina Y je počet es v jeho ruce. Pravděpodobnost, že má y es, je

$$p_Y(y) = \frac{\binom{4}{y}\binom{32-4}{10-y}}{\binom{32}{10}}, \qquad y \in \{0,1,2,3,4\},$$

(hypergeometrické rozdělení),

y	0	1	2	3	4
$p_Y(y)$	0.203	0.428	0.289	0.073	0.006

$$\mathrm{E}Y = \frac{10}{32}\,4 = \frac{5}{4}.$$

2. Podívám se na svých 10 karet (ze stejného balíčku) a vidím, že nastal jev B: mám 2 esa. Tím se pravděpodobnost počtu es prvního hráče mění na

$$p_{Y|B}(y) = \frac{\binom{2}{y}\binom{22-2}{10-y}}{\binom{22}{10}}, \qquad y \in \{0,1,2\},$$

(vybíráme z 22 karet, mezi nimiž jsou 2 esa),

y	0	1	2
$p_Y(y\|B)$	$\frac{2}{7} \doteq 0.286$	$\frac{40}{77} \doteq 0.519$	$\frac{15}{77} \doteq 0.195$

$$\mathrm{E}\,(Y|B) = \frac{10}{22}\,2 = \frac{10}{11}.$$

3. Obecněji: Náhodná veličina X je počet es v mé ruce a x její realizace („skutečná hodnota“), tj. nastal jev $X = x$.
Pravděpodobnost počtu es prvního hráče se mění na

$$p_{Y|X=x}(y) = p_{Y|X}(y|x) = \frac{\binom{4-x}{y}\binom{22-4+x}{10-y}}{\binom{22}{10}}, \quad x \in \{0,\dots,4\},\ y \in \{0,\dots,4-x\},$$

(vybíráme z $22-4+x$ karet, mezi nimiž je $4-x$ es),

$$\mathrm{E}\,(Y|X=x)=\frac{10}{22}\,(4-x)=\frac{5}{11}\,(4-x)\,,\qquad x\in\{0,\dots,4\}\,.$$

Tento výsledek popisuje rozdělení náhodné veličiny

$$\mathrm{E}\,(Y|X)=\frac{5}{11}\,(4-X)\,,$$

která nabývá hodnot

$$\mathrm{E}\,(Y|X=x)=\frac{5}{11}\,(4-x)\in\left\{\frac{20}{11},\frac{15}{11},\frac{10}{11},\frac{5}{11},0\right\}$$

s pravděpodobnostmi

$$p_X(x)=p_Y(x)=\frac{\binom{4}{x}\binom{32-4}{10-x}}{\binom{32}{10}}\,,\qquad x\in\{0,1,2,3,4\}\,.$$

x	0	1	2	3	4
$\mathrm{E}\,(Y\|X=x)$	$\frac{20}{11}$	$\frac{15}{11}$	$\frac{10}{11}$	$\frac{5}{11}$	0

Náhodná veličina $\mathrm{E}\,(Y|X)$ má střední hodnotu

$$\begin{aligned}\mathrm{E}\,(\mathrm{E}\,(Y|X))&=\frac{20}{11}\,p_X(0)+\frac{15}{11}\,p_X(1)+\frac{10}{11}\,p_X(2)+\frac{5}{11}\,p_X(3)\\&=\mathrm{E}\left(\frac{5}{11}\,(4-X)\right)=\left(\frac{5}{11}\,(4-\mathrm{E}X)\right)=\left(\frac{5}{11}\,\left(4-\frac{5}{4}\right)\right)=\frac{5}{4}=\mathrm{E}Y\,.\end{aligned}$$

1.2 Opakování podmíněné pravděpodobnosti

Pravděpodobnost (=pravděpodobnostní míra) P
Značení: $P(\ .\ )$ pravděpodobnostní míra,
$P[\ .\ ]$ pravděpodobnost jevu popsaného v závorce
(je-li argument jev, obě značení se shodují)
Jev B, $P(B)\neq 0\neq P(\bar{B})$
$\forall$ jev A : $P(A)=c\;P(A|B)+c'\;P(A|\bar{B})$,
kde $c=P(B)$, $c'=P(\bar{B})$, tj. $c+c'=1$.
P je lineární (dokonce **konvexní**) kombinace podmíněných pravděpodobností $P(\ .\ |B)$, resp. $P(\ .\ |$
což jsou obyčejné pravděpodobnosti popisující jiné modely (kde B je jev jistý, resp. nemožný).
Nadále předpoklad $P(B)\neq 0$ budeme potřebovat,
předpoklad $P(\bar{B})\neq 0$ nikoli.

1.3 Rozdělení podmíněné jevem

Podmíněná pravděpodobnostní funkce $p_{Y|B}$ *diskrétní* náhodné veličiny Y za podmínky B:

$$p_{Y|B}(y)=P[Y=y|B]=\frac{P[Y=y,B]}{P[B]}\,,$$

kde $P[Y=y,B]=P\left(\{\omega\in\Omega\mid Y(\omega)=y,\omega\in B\}\right)=P\left(\{\omega\in B\mid Y(\omega)=y\}\right)$.

Podmíněná distribuční funkce $F_{Y|B}$ náhodné veličiny Y za podmínky B:

$$F_{Y|B}(y) = P[Y \le y|B] = \frac{P[Y \le y, B]}{P[B]}\ ,$$

kde $P[Y \le y, B] = P\left(\{\omega \in B \mid Y(\omega) \le y\}\right)$.
Pro daný jev B má všechny vlastnosti distribuční funkce.
Definována, i když Y není diskrétní
⇒ **podmíněná kvantilová funkce** $q_{Y|B}$
⇒ **podmíněná hustota** $f_{Y|B}$ *spojité* náhodné veličiny

$$F_{Y|B}(y) = \int_0^y f_{Y|B}(t)\,\mathrm{d}t$$

⇒ **podmíněná střední hodnota** $\mathrm{E}(Y|B)$
⇒ **podmíněný rozptyl** $\mathrm{D}(Y|B)$

1.4 Rozdělení podmíněné hodnotou diskrétní náhodné veličiny

Diskrétní náhodnou veličinu Y lze podmiňovat jevem $X = x$, pokud $p_X(x) = P[X = x] \neq 0$; pak

$$p_{Y,X}(y,x) = P[Y = y, X = x] = P[Y = y \mid X = x] \cdot P[X = x] = p_{Y|X}(y|x) \cdot p_X(x)\,,$$

$$\begin{aligned} F_{Y,X}(y,x) &= \sum_{t:\, t\le x} \sum_{u:\, u\le y} p_{Y,X}(u,t) = \sum_{t:\, t\le x} \sum_{u:\, u\le y} p_{Y|X}(u|t) \cdot p_X(t) \\ &= \sum_{t:\, t\le x} F_{Y|X}(y|t) \cdot p_X(t)\,. \end{aligned}$$

Obecněji (pro *libovolnou* náhodnou veličinu Y)

$$F_{Y,X}(y,x) = \sum_{t:\, t\le x} F_{Y|X}(y|t) \cdot p_X(t)\,.$$

1.5 Rozdělení podmíněné hodnotou spojité náhodné veličiny

Nelze podmiňovat jevem $X = x$, neboť $P[X = x] = 0$; použijeme místo toho *hustotu* $f_X(x)$ (je-li nenulová) a definujeme analogicky podmíněnou hustotu $f_{Y|X} \colon \mathbb{R}^2 \to \langle 0, \infty)$ jako libovolnou funkci splňující

$$f_{Y,X}(y,x) = f_{Y|X}(y|x) \cdot f_X(x)\,,$$

přesněji a obecněji

$$F_{Y,X}(y,x) = \int_{-\infty}^{x} F_{Y|X}(y|t) \cdot f_X(t)\,\mathrm{d}t = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{Y|X}(u|t) \cdot f_X(t)\,\mathrm{d}u\,\mathrm{d}t\,.$$

1.6 Vlastnosti podmíněných rozdělení

Může existovat
podmíněná střední hodnota $\mathrm{E}(Y|X=x)$
resp. **podmíněný rozptyl** $\mathrm{D}(Y|X=x)$
jako *funkce* $\mathrm{E}(Y|X)$, resp. $\mathrm{D}(Y|X)$ proměnné x.
Pokud argument x neznáme, ale nahradíme náhodnou veličinou X, pak podmíněná střední hodnota, resp. podmíněný rozptyl je náhodná veličina a může mít obvyklé charakteristiky, např. $\mathrm{E}(\mathrm{E}(Y|X)), \mathrm{D}(\mathrm{E}(Y|X)), \mathrm{E}(\mathrm{D}(Y|X))$... Ty, pokud existují, splňují rovnosti

$$\begin{aligned}\mathrm{E}(\mathrm{E}(Y|X)) &= \mathrm{E}Y\,,\\ \mathrm{D}Y &= \mathrm{D}(\mathrm{E}(Y|X)) + \mathrm{E}(\mathrm{D}(Y|X))\,.\end{aligned}$$

1.7 Příklady podmíněných rozdělení

Příklad: První hráč hodí čtyřmi stejnými mincemi; ty, na kterých padne líc, si ponechá jako výhru (náhodná veličina X). Poté druhý hráč hodí zbylými mincemi; ty, na kterých padne líc, tvoří jeho výhru (náhodná veličina Y). Popište sdružené rozdělení X, Y, marginální rozdělení X a Y, podmíněná rozdělení $Y|X$ a $X|Y$. U všech rozdělení stanovte (podmíněnou) střední hodnotu a (podmíněný) rozptyl.
Řešení: (Čísla za tabulkami jsou pravděpodobnostní funkce marginálních rozdělení, což v případě sdruženého rozdělení jsou řádkové, resp. sloupcové součty.)

$p_{X,Y}$ x \ y	0	1	2	3	4	$p_X(x)$
0	$\frac{1}{256}$	$\frac{1}{64}$	$\frac{3}{128}$	$\frac{1}{64}$	$\frac{1}{256}$	$\frac{1}{16}$
1	$\frac{1}{32}$	$\frac{3}{32}$	$\frac{3}{32}$	$\frac{1}{32}$	0	$\frac{1}{4}$
2	$\frac{3}{32}$	$\frac{3}{16}$	$\frac{3}{32}$	0	0	$\frac{3}{8}$
3	$\frac{1}{8}$	$\frac{1}{8}$	0	0	0	$\frac{1}{4}$
4	$\frac{1}{16}$	0	0	0	0	$\frac{1}{16}$
$p_Y(y)$	$\frac{81}{256}$	$\frac{27}{64}$	$\frac{27}{128}$	$\frac{3}{64}$	$\frac{1}{256}$	

$\mathrm{E}X = 2\,,\qquad \mathrm{E}Y = \frac{27}{64}\,1 + \frac{27}{128}\,2 + \frac{3}{64}\,3 + \frac{1}{256}\,4 = 1$

$p_{Y\|X}$ x \ y	0	1	2	3	4	$p_X(x)$	$\mathrm{E}(Y\|X=x)$
0	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{16}$	2
1	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	0	$\frac{1}{4}$	$\frac{3}{2}$
2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0	0	$\frac{3}{8}$	1
3	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	$\frac{1}{4}$	$\frac{1}{2}$
4	1	0	0	0	0	$\frac{1}{16}$	0
$p_Y(y)$	$\frac{81}{256}$	$\frac{27}{64}$	$\frac{27}{128}$	$\frac{3}{64}$	$\frac{1}{256}$		

$\mathrm{E}(\mathrm{E}(Y|X)) = \frac{1}{16}\,2 + \frac{1}{4}\,\frac{3}{2} + \frac{3}{8}\,1 + \frac{1}{4}\,\frac{1}{2} = 1 = \mathrm{E}Y$

$p_{X\|Y}$ y x	0	1	2	3	4	$p_X(x)$
0	$\frac{1}{81}$	$\frac{1}{27}$	$\frac{1}{9}$	$\frac{1}{3}$	1	$\frac{1}{16}$
1	$\frac{8}{81}$	$\frac{2}{9}$	$\frac{4}{9}$	$\frac{2}{3}$	0	$\frac{1}{4}$
2	$\frac{8}{27}$	$\frac{4}{9}$	$\frac{4}{9}$	0	0	$\frac{3}{8}$
3	$\frac{32}{81}$	$\frac{8}{27}$	0	0	0	$\frac{1}{4}$
4	$\frac{16}{81}$	0	0	0	0	$\frac{1}{16}$
$p_Y(y)$	$\frac{81}{256}$	$\frac{27}{64}$	$\frac{27}{128}$	$\frac{3}{64}$	$\frac{1}{256}$	
$\mathrm{E}(X\|Y=y)$	$\frac{8}{3}$	2	$\frac{4}{3}$	$\frac{2}{3}$	0	

$\mathrm{E}(\mathrm{E}(X|Y)) = \frac{81}{256}\,\frac{8}{3} + \frac{27}{64}\,2 + \frac{27}{128}\,\frac{4}{3} + \frac{3}{64}\,\frac{2}{3} = 2 = \mathrm{E}X$

Příklad [Wassermann 3.8.22, rozšířeno]**:** Nechť $0 < a < b < 1$. Náhodná veličina X má spojité rovnoměrné rozdělení na $\langle 0, 1\rangle$, pomocí ní jsou definovány náhodné veličiny Y, Z:

$$Y = \begin{cases} 1 & \text{pro } X < b, \\ 0 & \text{jinak,} \end{cases}$$

$$Z = \begin{cases} 1 & \text{pro } X > a, \\ 0 & \text{jinak.} \end{cases}$$

Jsou veličiny Y, Z nezávislé? (Zdůvodněte.)
Najděte $\mathrm{E}(Y|Z), \mathrm{D}(Y|Z), \mathrm{E}(\mathrm{E}(Y|Z)), \mathrm{E}(\mathrm{D}(Y|Z)), \mathrm{D}(\mathrm{E}(Y|Z))$ a ověřte rovnost $\mathrm{D}Y = \mathrm{D}(\mathrm{E}(Y|Z)) + \mathrm{E}(\mathrm{D}(Y|Z))$.
Řešení:

$p_{Y,Z}$ y z	0	1	$p_Z(z)$
0	0	a	a
1	$1-b$	$b-a$	$1-a$
$p_Y(y)$	$1-b$	b	

$p_{Y\|Z}$ y z	0	1	$p_Z(z)$	$\mathrm{E}(Y\|Z=z)$	$\mathrm{D}(Y\|Z=z)$
0	0	1	a	1	0
1	$\frac{1-b}{1-a}$	$\frac{b-a}{1-a}$	$1-a$	$\frac{b-a}{1-a}$	$\frac{(1-b)\,(b-a)}{(1-a)^2}$

$\mathrm{E}(\mathrm{E}(Y|Z)) = 1 \cdot a + \frac{b-a}{1-a} \cdot (1-a) = b = \mathrm{E}Y$

$\mathrm{E}(\mathrm{D}(Y|Z)) = 0 \cdot a + \frac{(1-b)\,(b-a)}{(1-a)^2} \cdot (1-a) = \frac{(1-b)\,(b-a)}{1-a}$

$\mathrm{D}(\mathrm{E}(Y|Z)) = \mathrm{E}((\mathrm{E}(Y|Z))^2) - \underbrace{(\mathrm{E}(\mathrm{E}(Y|Z)))^2}_{(\mathrm{E}Y)^2} = 1^2 \cdot a + \left(\frac{b-a}{1-a}\right)^2 \cdot (1-a) - b^2 = \frac{a\,(1-b)^2}{1-a}$

$\mathrm{D}(\mathrm{E}(Y|Z)) + \mathrm{E}(\mathrm{D}(Y|Z)) = \frac{(1-b)\,(b-a)}{1-a} + \frac{a\,(1-b)^2}{1-a} = b\,(1-b) = \mathrm{D}Y$

2 Regrese

neboli náhrada pozorované závislosti vhodnou funkcí.

2.1 Odhad konstanty

Úloha 1: Odhadujeme spojitou náhodnou veličinu Y, předpokládáme

$$Y = \vartheta + \mathcal{E}\,,$$

kde $\vartheta \in \mathbb{R}$ je neznámý parametr,
$\mathcal{E}$ je náhodná veličina (chyba, šum) s rozdělením $\mathrm{N}(0, \sigma^2)$,
σ^2 je (konstantní, známý nebo neznámý) rozptyl,

$$f_{\mathcal{E}}(t) = \frac{1}{\sigma\sqrt{2\,\pi}}\,\exp\left(-\frac{t^2}{2\,\sigma^2}\right)\,,$$
$$f_Y(t) = \frac{1}{\sigma\sqrt{2\,\pi}}\,\exp\left(-\frac{(t-\vartheta)^2}{2\,\sigma^2}\right)\,.$$

Vstupy: posloupnost realizací vzniklých nezávislými pokusy (realizace náhodného výběru z rozdělení n. v. Y)

$$\boldsymbol{y} = (y_1, \ldots, y_n)\,.$$

Předpokládáme

$$y_j = \vartheta + e_j\,,$$

kde $(e_1, \ldots, e_n)$ je realizace náhodného výběru z rozdělení $\mathrm{N}(0, \sigma^2)$.

2.2 Odhad konstanty metodou maximální věrohodnosti

$$\Lambda(\vartheta) = \prod_j f_Y(y_j) = \prod_j \frac{1}{\sigma\sqrt{2\,\pi}}\,\exp\left(-\frac{(y_j-\vartheta)^2}{2\,\sigma^2}\right)\,,$$
$$\lambda(\vartheta) = \ln\Lambda(\vartheta) = \sum_j \ln f_Y(y_j) = \sum_j \left(-\ln\sigma\sqrt{2\,\pi}\,-\,\frac{(y_j-\vartheta)^2}{2\,\sigma^2}\right) =$$
$$= \underbrace{-n\,\ln\sigma\sqrt{2\,\pi}}_{\text{konst.}} - \underbrace{\frac{1}{2\,\sigma^2}}_{\text{konst.}}\,\underbrace{\sum_j (y_j-\vartheta)^2}_{\kappa(\vartheta)}\,,$$

$$\hat{\vartheta} = \operatorname*{argmax}_{\vartheta}\Lambda(\vartheta) = \operatorname*{argmax}_{\vartheta}\lambda(\vartheta) = \operatorname*{argmin}_{\vartheta}\kappa(\vartheta) =$$
$$= \operatorname*{argmin}_{\vartheta}\sum_j (y_j-\vartheta)^2 = \operatorname*{argmin}_{\vartheta}\sum_j e_j^2$$

$\Rightarrow$ **metoda nejmenších čtverců**

$$0 = \frac{\partial\kappa(\hat{\vartheta})}{\partial\hat{\vartheta}} = \frac{\partial}{\partial\hat{\vartheta}}\left(\sum_j y_j^2 - 2\,\hat{\vartheta}\sum_j y_j + n\,\hat{\vartheta}^2\right) = -2\sum_j y_j + 2\,n\,\hat{\vartheta}\,,$$
$$\hat{\vartheta} = \frac{1}{n}\sum_j y_j = \bar{y}\,.$$

Realizace výběrového průměru $\bar{y}$ = odhad metodou nejmenších čtverců = max. věrohodný odhad (za předpokladu normálního rozdělení chyb)

2.3 Řešení bez předpokladu normality

Obecněji (pokud $\mathcal{E}$ má střední hodnotu)

$$\mathrm{E}Y = \vartheta + \mathrm{E}\mathcal{E}\,.$$

Realizace výběrového průměru $\bar{y}$ = odhad metodou nejmenších čtverců $\neq$ max. věrohodný odhad (ten může být vychýlený)

2.4 Odhad funkce: obecný nelineární model

Úloha 2: Odhadujeme spojitou náhodnou veličinu Y, předpokládáme

$$Y = g(X) + \mathcal{E}\,,$$

kde X je nezávislá **vysvětlující** náhodná veličina,jejíž hodnoty můžeme měřit (spojitá nebo diskrétní),
$g\colon \mathbb{R} \to \mathbb{R}$ je neznámá funkce (závislá na neznámých parametrech),
$\mathcal{E}$ je náhodná veličina s rozdělením $\mathrm{N}(0, \sigma^2)$,
σ^2 je (konstantní, známý nebo neznámý) rozptyl,

$$f_{Y|X}(y|x) = \frac{1}{\sigma\sqrt{2\,\pi}} \exp\left(-\frac{(y - g(x))^2}{2\,\sigma^2}\right).$$

Vstupy: realizace náhodného výběru ze sdruženého rozdělení n. vektoru (Y, X)

$$(\boldsymbol{y}, \boldsymbol{x}) = ((y_1, x_1), \ldots, (y_n, x_n))\,.$$

Předpokládáme $y_j = g(x_j) + e_j$,
kde $(e_1, \ldots, e_n)$ je realizace náhodného výběru z rozdělení $\mathrm{N}(0, \sigma^2)$.
Řešeni: Pokud víme, že $X = x$, přejdeme k podmíněným pravděpodobnostem a jejich středním hodnotám:

$$\mathrm{E}(Y|X = x) = \mathrm{E}(g(X) + \mathcal{E}|X = x) = \mathrm{E}(g(X)|X = x) = g(x)\,,$$

Odhad (hodnot) funkce $g(x)$ = odhad podmíněných středních hodnot $\mathrm{E}(Y|X = x)$.
Problém: Jak je odhadnout z realizace $((y_1, x_1), \ldots, (y_n, x_n))$?
Řešeni: Odhad $\hat{g}(x)$ = aritmetický průměr těch y_j, pro která $x_j = x$, tj.

$$\hat{g}(x) = \frac{1}{|M(x)|} \sum_{j \in M(x)} y_j\,, \quad \text{kde } M(x) = \{j \in \{1, \ldots, n\} : x_j = x\}\,.$$

Problém 1: Lze použít pouze pro velký rozsah výběru ve srovnání s počtem možných hodnot n. v. X, rozhodně pouze pro X diskrétní!
Problém 2: I pro rozsáhlá data není postup vhodný, nabývá-li X mnoha hodnot – přetrénování: do modelu zahrnujeme i statistické chyby jednotlivých výsledků.
Řešeni: Omezíme se na modely (funkce g), které mají mnohem méně parametrů, než je hodnot n. v. X.
Tím je umožněna též **statistická indukce**, tedy zobecnění výsledků i na ty hodnoty n. v. X, které dosud nebyly pozorovány.

2.5 Odhad náhodné veličiny: lineární model dimenze 1

Úloha 3: Odhadujeme spojitou náhodnou veličinu Y, předpokládáme

$$Y = \vartheta_0 + \vartheta_1 X + \mathcal{E}\,,$$

kde $\vartheta_0, \vartheta_1 \in \mathbb{R}$ jsou neznámé parametry (**regresní koeficienty**),
$\mathcal{E}$ je náhodná veličina s rozdělením $\mathrm{N}(0, \sigma^2)$,
σ^2 je (konstantní, známý nebo neznámý) rozptyl,
X je nezávislá **vysvětlující** náhodná veličina,

$$f_{Y|X}(y|x) = \frac{1}{\sigma\sqrt{2\,\pi}} \exp\left(-\frac{(y-\vartheta_0-\vartheta_1\,x)^2}{2\,\sigma^2}\right).$$

Vstupy: realizace náhodného výběru ze sdruženého rozdělení n. vektoru (Y, X)

$$(\boldsymbol{y}, \boldsymbol{x}) = ((y_1, x_1), \ldots, (y_n, x_n))\,.$$

Předpokládáme $y_j = \vartheta_0 + \vartheta_1\,x_j + e_j$, kde $(e_1, \ldots, e_n)$ je realizace náhodného výběru z rozdělení $\mathrm{N}(0, \sigma^2)$.

2.6 Řešení metodou maximální věrohodnosti: lineární model dimenze 1

Při známém rozptylu σ^2:

$$\Lambda(\vartheta_0, \vartheta_1) = \prod_j f_{Y|X}(y_j|x_j) = \prod_j \frac{1}{\sigma\sqrt{2\,\pi}} \exp\left(-\frac{(y_j-\vartheta_0-\vartheta_1\,x_j)^2}{2\,\sigma^2}\right),$$

$$\lambda(\vartheta_0, \vartheta_1) = \ln\Lambda(\vartheta_0, \vartheta_1) = \underbrace{-n\,\ln\sigma\sqrt{2\,\pi}}_{\text{konst.}} - \underbrace{\frac{1}{2\,\sigma^2}}_{\text{konst.}} \underbrace{\sum_j (y_j-\vartheta_0-\vartheta_1\,x_j)^2}_{\kappa(\vartheta_0,\vartheta_1)},$$

$$\begin{aligned}\hat{\vartheta} &= \operatorname*{argmax}_{\vartheta_0,\vartheta_1} \Lambda(\vartheta_0, \vartheta_1) = \operatorname*{argmax}_{\vartheta_0,\vartheta_1} \lambda(\vartheta_0, \vartheta_1) = \operatorname*{argmin}_{\vartheta_0,\vartheta_1} \kappa(\vartheta_0, \vartheta_1) \\ &= \operatorname*{argmin}_{\vartheta_0,\vartheta_1} \sum_j (y_j-\vartheta_0-\vartheta_1\,x_j)^2 = \operatorname*{argmin}_{\vartheta_0,\vartheta_1} \sum_j e_j^2\end{aligned}$$

$\Rightarrow$ **metoda nejmenších čtverců**
odhad metodou nejmenších čtverců = max. věrohodný odhad (za předpokladu normálního rozdělení chyb)

$$0 = \frac{\partial\kappa(\hat{\vartheta}_0, \hat{\vartheta}_1)}{\partial\hat{\vartheta}_0} = 2\,n\,\hat{\vartheta}_0 - 2\sum_j y_j + 2\,\hat{\vartheta}_1 \sum_j x_j\,,$$

$$\hat{\vartheta}_0 = \frac{1}{n}\sum_j y_j - \hat{\vartheta}_1\,\frac{1}{n}\sum_j x_j\,,$$

$$\hat{\vartheta}_0 = \bar{y} - \hat{\vartheta}_1\,\bar{x}\,.$$

Dosadíme do

$$0 = \frac{\partial \kappa(\hat{\vartheta}_0, \hat{\vartheta}_1)}{\partial \hat{\vartheta}_1} = 2\,\hat{\vartheta}_1 \sum_j x_j^2 - 2 \sum_j x_j\,y_j + 2\,\hat{\vartheta}_0 \sum_j x_j =$$
$$= 2\,\hat{\vartheta}_1 \sum_j x_j^2 - 2 \sum_j x_j\,y_j + 2\,n\,\bar{x}\,\underbrace{\left(\bar{y} - \hat{\vartheta}_1\,\bar{x}\right)}_{\hat{\vartheta}_0},$$

$$\hat{\vartheta}_1 = \frac{\sum_j x_j\,y_j - n\,\bar{x}\,\bar{y}}{\sum_j x_j^2 - n\,\bar{x}^2} = \frac{\sum_j (x_j - \bar{x})\,(y_j - \bar{y})}{\sum_j (x_j - \bar{x})^2}$$

a do odhadů chyb

2.7 Regresní přímka 1

Je tvořena body (x, y), které splňují rovnici

$$y = \hat{\vartheta}_0 + \hat{\vartheta}_1\,x\,.$$

Věta: Bod $(\bar{x}, \bar{y})$ leží na regresní přímce,

$$\bar{y} = \hat{\vartheta}_0 + \hat{\vartheta}_1\,\bar{x}\,.$$

Důkaz: Už jsme odvodili $\hat{\vartheta}_0 = \bar{y} - \hat{\vartheta}_1\,\bar{x}\,.$

Odečtením dostáváme rovnici regresní přímky ve tvaru

$$y - \bar{y} = \hat{\vartheta}_1\,(x - \bar{x})\,.$$

Sklon (směrnice) regresní přímky $\hat{\vartheta}_1$ se nezmění, pokud přičteme konstantu ke všem hodnotám názávisle, resp. závisle proměnné. Mohli jsme od nich např. odečíst realizace výběrových průměrů $\bar{x}$, resp. $\bar{y}$ a zjednodušit si výrazy.

2.8 Regresní přímka 2

Poznámka: Lze také odhadnout lineární závislost X na Y dle modelu

$$X = \vartheta_0^* + \vartheta_1^*\,Y + \mathcal{E}^*\,,$$

směrnice regresní přímky vyjde

$$\hat{\vartheta}_1^* = \frac{\sum_j (x_j - \bar{x})\,(y_j - \bar{y})}{\sum_j (y_j - \bar{y})^2}\,.$$

Výsledkem je obecně **jiná** regresní přímka, jejíž rovnici lze psát ve tvarech

$$x = \hat{\vartheta}_0^* + \hat{\vartheta}_1^*\,y\,,$$
$$x - \bar{x} = \hat{\vartheta}_1^*\,(y - \bar{y})\,,$$
$$y - \bar{y} = \frac{1}{\hat{\vartheta}_1^*}\,(x - \bar{x})\,.$$

2.9 Interpretace regresních koeficientů

Označme odhady rozptylů a kovariance:

$$\hat\sigma^2_{\boldsymbol{x}} := \frac{1}{n}\sum_j (x_j-\bar x)^2 = \frac{n-1}{n}\, s^2_{\boldsymbol{x}} = \mathrm{D}\,\mathrm{Emp}(\boldsymbol{x})\,,$$
$$\hat\sigma^2_{\boldsymbol{y}} := \frac{1}{n}\sum_j (y_j-\bar y)^2 = \frac{n-1}{n}\, s^2_{\boldsymbol{y}} = \mathrm{D}\,\mathrm{Emp}(\boldsymbol{y})\,,$$
$$c_{\boldsymbol{x},\boldsymbol{y}} := \frac{1}{n}\sum_j (x_j-\bar x)\,(y_j-\bar y) = \mathrm{cov}(\mathrm{Emp}(\boldsymbol{x},\boldsymbol{y}))\,.$$

Odhad korelace = realizace výběrového koeficientu korelace =

$$r_{\boldsymbol{x},\boldsymbol{y}} := \frac{\sum_j (x_j-\bar x)\,(y_j-\bar y)}{\sqrt{\sum_j (x_j-\bar x)^2\,\sum_j (y_j-\bar y)^2}} = \frac{c_{\boldsymbol{x},\boldsymbol{y}}}{\hat\sigma_{\boldsymbol{x}}\,\hat\sigma_{\boldsymbol{y}}} = \varrho(\mathrm{Emp}(\boldsymbol{x},\boldsymbol{y}))\,.$$

Lze psát

$$\hat\vartheta_1 = \frac{c_{\boldsymbol{x},\boldsymbol{y}}}{\hat\sigma^2_{\boldsymbol{x}}}\,.$$

Rovnice regresní přímky pro závislost Y na X:

$$y-\bar y = \hat\vartheta_1\,(x-\bar x)\,,$$
$$y-\bar y = \frac{c_{\boldsymbol{x},\boldsymbol{y}}}{\hat\sigma^2_{\boldsymbol{x}}}\,(x-\bar x)\,,$$
$$\frac{y-\bar y}{\hat\sigma_{\boldsymbol{y}}} = r_{\boldsymbol{x},\boldsymbol{y}}\,\frac{x-\bar x}{\hat\sigma_{\boldsymbol{x}}}\,,$$

rovnice regresní přímky pro závislost X na Y (to je **jiná** přímka!):

$$x-\bar x = \hat\vartheta_1^*\,(y-\bar y)\,,$$
$$x-\bar x = \frac{c_{\boldsymbol{x},\boldsymbol{y}}}{\hat\sigma^2_{\boldsymbol{y}}}\,(y-\bar y)\,,$$
$$\frac{x-\bar x}{\hat\sigma_{\boldsymbol{x}}} = r_{\boldsymbol{x},\boldsymbol{y}}\,\frac{y-\bar y}{\hat\sigma_{\boldsymbol{y}}}\,.$$

Směrnice obou regresních přímek mají stejná znaménka a součin $\hat\vartheta_1\,\hat\vartheta_1^* = \dfrac{c^2_{\boldsymbol{x},\boldsymbol{y}}}{\hat\sigma^2_{\boldsymbol{x}}\,\hat\sigma^2_{\boldsymbol{y}}} = r^2_{\boldsymbol{x},\boldsymbol{y}}$, takže $r_{\boldsymbol{x},\boldsymbol{y}} = \sqrt{\hat\vartheta_1\,\hat\vartheta_1^*}\,\mathrm{sign}(\hat\vartheta_1)\,.$

2.10 Chyba lineární regrese

Odhadli jsme lineární regresní funkci $\hat g$,

$$\hat g(x) := \hat\vartheta_0 + \hat\vartheta_1\,x\,,$$

pomocí ní hodnoty nezávisle proměnné v jednotlivých realizacích

$$\hat y_j := \hat g(x_j) = \hat\vartheta_0 + \hat\vartheta_1\,x_j\,,$$
$$\hat{\boldsymbol{y}} := (\hat y_1,\ldots,\hat y_n)$$

a chyby (**rezidua**)

$$\hat{e}_j := y_j - \hat{y}_j = y_j - \hat{\vartheta}_0 - \hat{\vartheta}_1\, x_j = y_j - \bar{y} - \hat{\vartheta}_1\, (x_j - \bar{x})\,,$$
$$\hat{\boldsymbol{e}} := (\hat{e}_1, \ldots, \hat{e}_n)\,.$$

Věta: $\dfrac{1}{n}\sum_j \hat{y}_j = \bar{y}\,.$

Důkaz: $\dfrac{1}{n}\sum_j \hat{y}_j = \dfrac{1}{n}\sum_j \left(\hat{\vartheta}_0 + \hat{\vartheta}_1\, x_j\right) = \hat{\vartheta}_0 + \hat{\vartheta}_1\, \bar{x} = \bar{y}\,.$

2.11 Složky rozptylu regresního odhadu

(Někdy se nedělí n.)

Celkový rozptyl (angl. *total variation*) $:= \hat{\sigma}^2_{\boldsymbol{y}} = \dfrac{1}{n}\sum_j (y_j - \bar{y})^2\,.$

Rozptyl modelu (angl. *explained variation*) $:= \hat{\sigma}^2_{\hat{\boldsymbol{y}}} = \dfrac{1}{n}\sum_j (\hat{y}_j - \bar{y})^2\,.$

Reziduální rozptyl (angl. *unexplained variation*) :=

$$\hat{\sigma}^2_{\hat{\boldsymbol{e}}} = \frac{1}{n}\sum_j \hat{e}_j^2 = \frac{1}{n}\sum_j (y_j - \hat{y}_j)^2\,.$$

Věta: $\hat{\sigma}^2_{\boldsymbol{y}} = \hat{\sigma}^2_{\hat{\boldsymbol{y}}} + \hat{\sigma}^2_{\hat{\boldsymbol{e}}}\,.$

Věta: $r^2_{\boldsymbol{x},\boldsymbol{y}} = \dfrac{\hat{\sigma}^2_{\hat{\boldsymbol{y}}}}{\hat{\sigma}^2_{\boldsymbol{y}}} = 1 - \dfrac{\hat{\sigma}^2_{\hat{\boldsymbol{e}}}}{\hat{\sigma}^2_{\boldsymbol{y}}}\,.$

2.12 Odhad rozptylu

Max. věrohodný odhad rozptylu *původního* rozdělení:

$$\lambda(\vartheta_0, \vartheta_1, \sigma) = -n\,\ln\sigma\,\sqrt{2\,\pi} - \frac{1}{2\,\sigma^2}\sum_j (y_j - \vartheta_0 - \vartheta_1\, x_j)^2 =$$

$$= \underbrace{-n\,\ln\sqrt{2\,\pi}}_{\text{konst.}} - n\,\ln\sigma - \frac{1}{2\,\sigma^2}\sum_j e_j^2\,,$$

$$0 = \frac{\partial\lambda(\vartheta_0, \vartheta_1, \sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_j e_j^2\,,$$

řešením je

$$\hat{\sigma}^2_{\hat{\boldsymbol{e}}} = \frac{1}{n}\sum_j \hat{e}_j^2 = \mathrm{D}\,\mathrm{Emp}(\hat{\boldsymbol{e}})\,.$$

Tento odhad je vychýlený; nestranný odhad je *(odvození viz [Likeš, Machek])*

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_j \hat{e}_j^2 = \frac{1}{n-2}\sum_j \left(y_j - \hat{\vartheta}_0 - \hat{\vartheta}_1\, x_j\right)^2\,.$$

2.13 Rozdělení odhadů a testy hypotéz o nich

Věta: Odhady $\hat{\vartheta}_0, \hat{\vartheta}_1, \hat{\sigma}^2$ skutečných parametrů $\vartheta_0, \vartheta_1, \sigma^2$ jsou nestranné, konzistentní, asymptoticky normální.
Odhady rozptylů regresních koeficientů

$$\hat{\sigma}^2_{\hat{\vartheta}_0} = \frac{\hat{\sigma}^2}{n^2\,\hat{\sigma}^2_{\boldsymbol{x}}} \sum_j x_j^2 = \frac{\hat{\sigma}^2_{\hat{\boldsymbol{e}}}}{n\,(n-2)\,\hat{\sigma}^2_{\boldsymbol{x}}} \sum_j x_j^2\,,$$

$$\hat{\sigma}^2_{\hat{\vartheta}_1} = \frac{\hat{\sigma}^2}{n\,\hat{\sigma}^2_{\boldsymbol{x}}} = \frac{\hat{\sigma}^2_{\hat{\boldsymbol{e}}}}{(n-2)\,\hat{\sigma}^2_{\boldsymbol{x}}}$$

nejsou nezávislé.
Test abolutního členu, tj. nulové hypotézy $H_0\colon \vartheta_0 = c$:

$$\frac{\hat{\vartheta}_0 - c}{\frac{\hat{\sigma}_{\hat{\boldsymbol{e}}}}{\hat{\sigma}_{\boldsymbol{x}}} \sqrt{\frac{1}{n}\sum_j x_j^2}} \sqrt{n-2}$$

testujeme na rozdělení $\mathrm{t}(n-2)$.
Test směrnice, tj. nulové hypotézy $H_0\colon \vartheta_1 = c$:

$$\frac{\hat{\vartheta}_1 - c}{\frac{\hat{\sigma}_{\hat{\boldsymbol{e}}}}{\hat{\sigma}_{\boldsymbol{x}}}} \sqrt{n-2}$$

testujeme na rozdělení $\mathrm{t}(n-2)$.
Test chyby regrese $\hat{\vartheta}_0 + \hat{\vartheta}_1\,x$ pro dané x, tj. nulové hypotézy $H_0\colon \vartheta_0 + \vartheta_1\,x = c$:

$$\frac{\hat{\vartheta}_0 + \hat{\vartheta}_1\,x - c}{\hat{\sigma}_{\hat{\boldsymbol{e}}} \sqrt{n+1+\frac{n\,(x-\bar{\boldsymbol{x}})^2}{\hat{\sigma}^2_{\boldsymbol{x}}}}} \sqrt{n-2}$$

testujeme na rozdělení $\mathrm{t}(n-2)$.

2.14 Testy korelace

Test nekorelovanosti (nulovosti korelace): Za předpokladu $H_0\colon \varrho(X,Y) = 0$ testujeme

$$\frac{r_{\boldsymbol{x},\boldsymbol{y}}}{\sqrt{1-r^2_{\boldsymbol{x},\boldsymbol{y}}}} \sqrt{n-2} = \frac{\hat{\sigma}_{\hat{\boldsymbol{y}}}}{\hat{\sigma}_{\hat{\boldsymbol{e}}}} \sqrt{n-2}$$

na rozdělení $\mathrm{t}(n-2)$. (Zde $\frac{\hat{\sigma}_{\hat{\boldsymbol{y}}}}{\hat{\sigma}_{\hat{\boldsymbol{e}}}}$ je poměr směrodatných odchylek odpovídajících vysvětlené a nevysvětlené části celkového rozptylu.)
Test (nenulové) hodnoty korelace: Ozančme funkci

$$h(t) := \frac{1}{2}\,\ln\frac{1+t}{1-t}\,.$$

Za předpokladu $H_0\colon \varrho(X,Y) = c$, kde $c \neq 0$, pochází

$$z_{\boldsymbol{x},\boldsymbol{y}} := h(r_{\boldsymbol{x},\boldsymbol{y}}) = \frac{1}{2}\,\ln\frac{1+r_{\boldsymbol{x},\boldsymbol{y}}}{1-r_{\boldsymbol{x},\boldsymbol{y}}}$$

z rozdělení přibližně $\mathrm{N}\left(\mu, \sigma^2\right)$, kde

$$\mu := h(c) = \frac{1}{2} \ln \frac{1+c}{1-c}\,,$$
$$\sigma^2 := \frac{1}{n-3}\,,$$
$$\sigma = \sqrt{\frac{1}{n-3}}\,.$$

Normované kritérium

$$\frac{z_{\boldsymbol{x},\boldsymbol{y}} - \mu}{\sigma} = \left(h(r_{\boldsymbol{x},\boldsymbol{y}}) - h(c)\right) \sqrt{n-3}$$

testujeme na rozdělení $\mathrm{N}\,(0,1)$.

Test rovnosti dvou korelací: Z nezávislých výběrů rozsahu n, resp. m, vypočítáme výběrové koeficienty korelace $r_{\boldsymbol{x},\boldsymbol{y}}$, resp. $r_{\boldsymbol{u},\boldsymbol{v}}$, z nich

$$z_{\boldsymbol{x},\boldsymbol{y}} := h(r_{\boldsymbol{x},\boldsymbol{y}}) = \frac{1}{2} \ln \frac{1+r_{\boldsymbol{x},\boldsymbol{y}}}{1-r_{\boldsymbol{x},\boldsymbol{y}}}\,,$$
$$z_{\boldsymbol{u},\boldsymbol{v}} := h(r_{\boldsymbol{u},\boldsymbol{v}}) = \frac{1}{2} \ln \frac{1+r_{\boldsymbol{u},\boldsymbol{v}}}{1-r_{\boldsymbol{u},\boldsymbol{v}}}\,.$$

Za předpokladu $H_0\colon \varrho(X,Y) = \varrho(U,V)$ pochází $z := z_{\boldsymbol{x},\boldsymbol{y}} - z_{\boldsymbol{u},\boldsymbol{v}} = h(r_{\boldsymbol{x},\boldsymbol{y}}) - h(r_{\boldsymbol{u},\boldsymbol{v}})$ z rozdělení přibližně $\mathrm{N}\left(0, \sigma^2\right)$, kde

$$\sigma^2 := \frac{1}{n-3} + \frac{1}{m-3}\,,$$
$$\sigma = \sqrt{\frac{1}{n-3} + \frac{1}{m-3}}\,.$$

Normované kritérium

$$\frac{z_{\boldsymbol{x},\boldsymbol{y}} - z_{\boldsymbol{u},\boldsymbol{v}}}{\sigma} = \frac{h(r_{\boldsymbol{x},\boldsymbol{y}}) - h(r_{\boldsymbol{u},\boldsymbol{v}})}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}}$$

testujeme na rozdělení $\mathrm{N}\,(0,1)$.

2.15 Příklad použití

Příklad 1 [Markechová, Tirpáková, Stehlíková, př. 9.3, 9, 13]: Měřili jsme výšku X [cm] a hmotnost Y [kg] $n = 10$ žáků, výsledky jsou v tabulce:

x_j [cm]	151	154	165	146	158	172	142	169	176	138
y_j [kg]	43	48	56	42	55	54	44	49	55	38

Odhadneme parametry jednotlivých rozdělení (která považujeme přibližně za normální):
$\bar{x} = \frac{1}{n} \sum_j x_j = 157.1$ cm
$\hat{\sigma}^2_{\boldsymbol{x}} = \frac{1}{n} \sum_j (x_j - \bar{x})^2 \doteq 154.7$ cm^2
$\hat{\sigma}_{\boldsymbol{x}} = \sqrt{\hat{\sigma}^2_{\boldsymbol{x}}} \doteq 12.4$ cm
$\bar{y} = \frac{1}{n} \sum_j y_j = 48.4$ kg
$\hat{\sigma}^2_{\boldsymbol{y}} = \frac{1}{n} \sum_j (y_j - \bar{y})^2 \doteq 37.44$ kg^2
$\hat{\sigma}_{\boldsymbol{y}} = \sqrt{\hat{\sigma}^2_{\boldsymbol{y}}} \doteq 6.12$ kg

2.15.1 Odhady koeficientů 1. regresní přímky

$\hat{\vartheta}_1 = \frac{\sum_j (x_j - \bar{x})\,(y_j - \bar{y})}{\sum_j (x_j - \bar{x})^2} = \frac{\sum_j x_j\, y_j - n\,\bar{x}\,\bar{y}}{\sum_j x_j^2 - n\,\bar{x}^2} \doteq 0.421 \text{ kg} / \text{cm}$
$\hat{\vartheta}_0 = \bar{y} - \hat{\vartheta}_1\,\bar{x} \doteq -17.78 \text{ kg}$
Rovnice 1. regresní přímky *(ve stejných jednotkách jako výše)*:

$$y = \hat{\vartheta}_0 + \hat{\vartheta}_1\, x \doteq -17.78 + 0.421\, x\,,$$

nebo

$$y - \bar{y} = \hat{\vartheta}_1\ (x - \bar{x})\ ,$$
$$y - 48.4 = 0.421\ (x - 157.1)\ ,$$

např. pro výšku $x = 160$ cm dostáváme odhad hmotnosti $y = \hat{\vartheta}_0 + \hat{\vartheta}_1\, x \doteq -17.78 + 0.421 \cdot 160 = 49.6$ kg.
Kontrola: Bod $(\bar{x}, \bar{y})$ leží na 1. regresní přímce, $\hat{\vartheta}_0 + \hat{\vartheta}_1\,\bar{x} \doteq -17.78 + 0.421 \cdot 157.1 = 48.4$ kg.

2.15.2 Odhady koeficientů 2. regresní přímky

$\hat{\vartheta}_1^* = \frac{\sum_j (x_j - \bar{x})\,(y_j - \bar{y})}{\sum_j (y_j - \bar{y})^2} \doteq 1.74 \text{ cm} / \text{kg}$
$\hat{\vartheta}_0^* = \bar{x} - \hat{\vartheta}_1^*\,\bar{y} \doteq 72.9 \text{ cm}$
Rovnice 2. regresní přímky:

$$x = \hat{\vartheta}_0^* + \hat{\vartheta}_1^*\, y \doteq 72.9 + 1.74\, y\,,$$

nebo

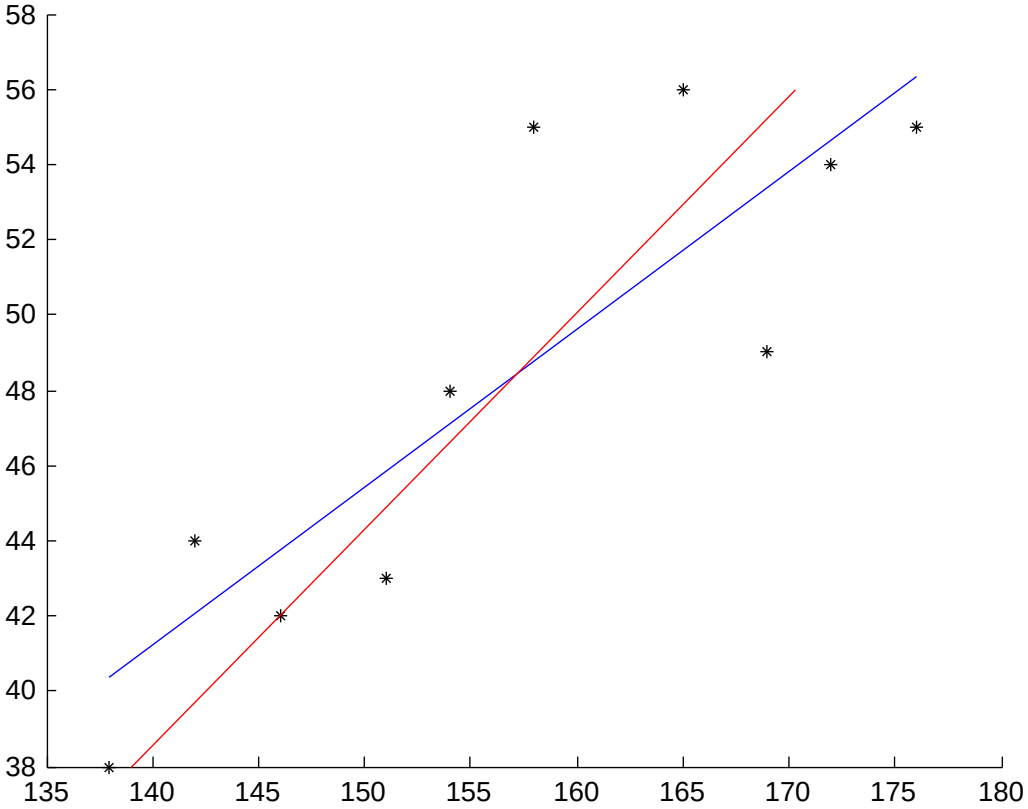
$$x - \bar{x} = \hat{\vartheta}_1^*\ (y - \bar{y})\ ,$$
$$x - 157.1 = 1.74\ (y - 48.4)\ ,$$

$$y - \bar{y} = \frac{1}{\hat{\vartheta}_1^*}\,(x - \bar{x})\,,$$
$$y - 48.4 = 0.575\,(x - 157.1)\,,$$

např. pro hmotnost $y = 50$ kg dostáváme odhad výšky $x = \hat{\vartheta}_0 + \hat{\vartheta}_1\, x \doteq 72.9 + 1.74 \cdot 50 = 160$ cm. *(To není totéž jako u 1. regresní přímky!)*
Kontrola: Bod $(\bar{x}, \bar{y})$ leží na 2. regresní přímce, $\hat{\vartheta}_0^* + \hat{\vartheta}_1^*\,\bar{y} \doteq 72.9 + 1.74 \cdot 48.4 = 157$ kg. *(Jen přibližná shoda kvůli zaokrouhlovacím chybám.)*
Součin směrnic obou regresních přímek:

$$\hat{\vartheta}_1\,\hat{\vartheta}_1^* = 0.421 \cdot 1.74 = 0.733 = r^2_{\boldsymbol{x},\boldsymbol{y}}\,,$$
$$r_{\boldsymbol{x},\boldsymbol{y}} = \sqrt{0.733} = 0.856\,.$$

(Kladné znaménko vyjadřuje, že větším hodnotám výšky odpovídá v průměru větší hmotnost.)
`http://lib.stat.cmu.edu/DASL/Stories/USCrime.html`



2.15.3 Odhady rozptylů, kovariance a korelace

$$\hat{\sigma}_{\boldsymbol{x}}^2 := \frac{1}{n}\sum_j (x_j - \bar{x})^2 \doteq 154.7\ \mathrm{cm}^2\,,$$

$$\hat{\sigma}_{\boldsymbol{y}}^2 := \frac{1}{n}\sum_j (y_j - \bar{y})^2 \doteq 37.44\ \mathrm{kg}^2\,,$$

$$c_{\boldsymbol{x},\boldsymbol{y}} := \frac{1}{n}\sum_j (x_j - \bar{x})\,(y_j - \bar{y}) \doteq 65.2\ \mathrm{kg\ cm}\,,$$

$$r_{\boldsymbol{x},\boldsymbol{y}} := \frac{\sum_j (x_j - \bar{x})\,(y_j - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2 \sum_j (y_j - \bar{y})^2}} = \frac{c_{\boldsymbol{x},\boldsymbol{y}}}{\hat{\sigma}_{\boldsymbol{x}}\,\hat{\sigma}_{\boldsymbol{y}}} \doteq \frac{65.2}{\sqrt{154.7 \cdot 37.44}} \doteq 0.856\,.$$

2.15.4 Chyba lineární regrese

Odhady hodnot nezávisle proměnné a chyby (**rezidua**)

$$\hat{y}_j = \hat{\vartheta}_0 + \hat{\vartheta}_1\, x_j\,,$$
$$\hat{e}_j = y_j - \hat{\vartheta}_0 - \hat{\vartheta}_1\, x_j\,,$$

$\hat{y}_j$ [kg]	45.8	47.1	51.7	43.7	48.8	54.7	42.0	53.4	56.4	40.4
$\hat{e}_j$ [kg]	−2.8	0.9	4.3	−1.7	6.2	−0.7	2.0	−4.4	−1.4	−2.4

Kontrola:

$$\frac{1}{n}\sum_j \hat{y}_j \doteq 48.4 \doteq \bar{y}\,,$$
$$\frac{1}{n}\sum_j \hat{e}_j = 0\,.$$

2.15.5 Složky rozptylu

Rozptyl modelu: $\hat{\sigma}^2_{\hat{\boldsymbol{y}}} = \frac{1}{n}\sum_j (\hat{y}_j - \bar{y})^2 \doteq 27.44\ \mathrm{kg}^2$.

Reziduální rozptyl: $\hat{\sigma}^2_{\hat{\boldsymbol{e}}} = \frac{1}{n}\sum_j \hat{e}_j^2 = \frac{1}{n}\sum_j (y_j - \hat{y}_j)^2 \doteq 10\ \mathrm{kg}^2$.

Celkový rozptyl: $\hat{\sigma}^2_{\boldsymbol{y}} = \frac{1}{n}\sum_j (y_j - \bar{y})^2 \doteq 37.44\ \mathrm{kg}^2 \doteq \hat{\sigma}^2_{\hat{\boldsymbol{y}}} + \hat{\sigma}^2_{\hat{\boldsymbol{e}}}$.

Kontrola:

$$r^2_{\boldsymbol{x},\boldsymbol{y}} = \frac{\hat{\sigma}^2_{\hat{\boldsymbol{y}}}}{\hat{\sigma}^2_{\boldsymbol{y}}} = 1 - \frac{\hat{\sigma}^2_{\hat{\boldsymbol{e}}}}{\hat{\sigma}^2_{\boldsymbol{y}}}\,,$$
$$0.733 \doteq \frac{27.44}{37.44} \doteq 1 - \frac{10}{37.44}\,.$$

2.15.6 Odhady rozptylu původního rozdělení

Max. věrohodný:

$$\hat{\sigma}^2_{\hat{\boldsymbol{e}}} = \frac{1}{n}\sum_j \hat{e}_j^2 \doteq 10\ \mathrm{kg}^2\,,$$

nestranný:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_j \hat{e}_j^2 \doteq 12.5\ \mathrm{kg}^2\ .$$

2.15.7 Testy korelace

Test nekorelovanosti (nulovosti korelace): Nulovou hypotézu $H_0 : \varrho(X, Y) = 0$ posoudíme podle (realizace) statistiky

$$\frac{r_{\boldsymbol{x},\boldsymbol{y}}}{\sqrt{1 - r^2_{\boldsymbol{x},\boldsymbol{y}}}}\sqrt{n-2} \doteq \frac{0.856}{\sqrt{1 - 0.856^2}}\sqrt{10-2} \doteq 4.69\,,$$

kterou lze určit i z poměru rozpylu modelu a reziduálního modelu,

$$\sqrt{\frac{\hat{\sigma}^2_{\hat{\boldsymbol{y}}}}{\hat{\sigma}^2_{\hat{\boldsymbol{e}}}}\,(n-2)} \doteq \sqrt{\frac{27.44}{10}\,(10-2)} \doteq 4.69\,.$$

Testujeme ji na Studentovo rozdělení t(8). Pro oboustranný test (alternativní hypotéza $H_1 : \varrho(X,Y) \neq 0$) vychází dosažená významnost 0.00078, pro jednostranný test (alternativní hypotéza $H_1 : \varrho(X,Y) > 0$) 0.0016, což dovoluje obě hypotézy zamítnout na vysoké hladině významnosti.
(Samozřejmě nelze zamítnout nulovou hypotézu $H_0 : \varrho(X,Y) \geq 0$ proti alternativní hypotéze $H_1 : \varrho(X,Y) < 0$.)
Test (nenulové) hodnoty korelace: Testujeme nulovou hypotézu $H_0 : \varrho(X,Y) = c$ proti *alternativní hypotéze* $H_1 : \varrho(X,Y) \neq c$ např. pro $c = 0.6$. Provedeme nelineární transformaci obou porovnávaných hodnot:

$$z_{\boldsymbol{x},\boldsymbol{y}} = \frac{1}{2} \ln \frac{1 + r_{\boldsymbol{x},\boldsymbol{y}}}{1 - r_{\boldsymbol{x},\boldsymbol{y}}} = \frac{1}{2} \ln \frac{1 + 0.856}{1 - 0.856} \doteq 1.28\,,$$
$$\frac{1}{2} \ln \frac{1 + c}{1 - c} \doteq 0.693\,,$$

jejich rozdíl poměřujeme směrodatnou odchylkou

$$\sigma_{Z_{\boldsymbol{X},\boldsymbol{Y}}} = \sqrt{\frac{1}{n-3}} \doteq 0.378\,,$$

testujeme

$$\frac{1.28 - 0.693}{0.378} \doteq 1.55$$

na rozdělení $\mathrm{N}(0,1)$, což nestačí k zamítnutí nulové hypotézy na hladině významnosti 5 % (ani pro jednostranné hypotézy $H_0 : \varrho(X,Y) \leq c$, $H_1 : \varrho(X,Y) > c$).
Test rovnosti dvou korelací: Pokud v jiném průzkumu rozsahu $m = 20$ vyšel odhad korelace stejných náhodných veličin např. $d = 0.4$, transformujeme i tuto hodnotu:

$$\frac{1}{2} \ln \frac{1 + d}{1 - d} = \frac{1}{2} \ln \frac{1 + 0.4}{1 - 0.4} \doteq 0.424$$

rozdíl poměřujeme směrodatnou odchylkou

$$\sqrt{\frac{1}{n-3} + \frac{1}{m-3}} \doteq 0.449\,,$$

testujeme

$$\frac{1.28 - 0.424}{0.449} \doteq 1.91$$

na rozdělení $\mathrm{N}(0,1)$, což nestačí k zamítnutí nulové hypotézy o rovnosti obou korelací na hladině významnosti 5 % (ale můžeme na této hladině zamítnout jednostrannou hypotézu $H_0 : \varrho(X,Y) \leq d$ proti $H_1 : \varrho(X,Y) > d$).

2.16 Odhad náhodné veličiny: lineární model dimenze k

Úloha 4: Odhadujeme spojitou náhodnou veličinu Y pomocí k **vysvětlujících** náhodných veličin $X_1, \ldots, X_k$ na základě realizace náhodného výběru

$$((y_1, x_{11}, \ldots, x_{1k}), \ldots, (y_n, x_{n1}, \ldots, x_{nk}))\,,$$

kde $n > k$, obvykle $n \gg k$. Předpokládáme lineární model

$$Y = \sum_{i=1}^{k} \vartheta_i \, X_i + \mathcal{E} \,,$$

kde $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_k)^T \in \mathbb{R}^k$ je (sloupcový) vektor neznámých parametů (**regresních koeficientů**),
$\mathcal{E}$ je náhodná veličina s rozdělením $\mathrm{N}(0, \sigma^2)$,
σ^2 je (konstantní, známý nebo neznámý) rozptyl,
Otázka: Kam se poděl absolutní člen ϑ_0 z modelu dimenze 1?
Odpověď 1: Je-li potřebný, lze jej zahrnout tak, že jedna z vysvětlujících proměnných je konstanta, BÚNO 1.
Odpověď 2: Od všech veličin $Y, X_1, \ldots, X_k$ lze odečíst jejich výběrové průměry.
Pro realizace dostáváme

$$y_j = \sum_{i=1}^{k} \vartheta_i \, x_{ji} + e_j \,,$$

maticový zápis:

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\vartheta} + \boldsymbol{e} \,,$$

kde $\boldsymbol{y} = (y_1, \ldots, y_n)^T, \boldsymbol{e} = (e_1, \ldots, e_n)^T \in \mathbb{R}^n$ jsou (sloupcové) vektory realizací *(nezávislých)* náhodných veličin,

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} .$$

Max. věrohodný odhad (za předpokladu normálního rozdělení chyb) = odhad metodou nejmenších čtverců $\hat{\boldsymbol{\vartheta}}$; který minimalizuje **reziduální součet čtverců** (angl. **residual sum of squares**, **RSS**)

$$R_{SS} = \sum_{j=1}^{n} \hat{e}_j^2 = \sum_{j=1}^{n} \left(y_j - \sum_{i=1}^{k} \hat{\vartheta}_i \, x_{ji} \right)^2 = \|\boldsymbol{e}\|^2 = \left\| \boldsymbol{y} - \boldsymbol{X}\,\hat{\boldsymbol{\vartheta}} \right\|^2 .$$

Geometrické řešení: $\boldsymbol{X}\,\hat{\boldsymbol{\vartheta}}$ je lineární kombinace sloupců matice $\boldsymbol{X}$, která má nejmenší euklidovskou vzdálenost od $\boldsymbol{y}$, tj. kolmý průmět $\boldsymbol{y}$ do lineárního podprostoru generovaného sloupci matice $\boldsymbol{X}$.
Ten je charakterizován tím, že každý jeho vektor $\boldsymbol{z}$ je kolmý na sloupce matice $\boldsymbol{X}$, tj.

$$\boldsymbol{X}^{\boldsymbol{T}}\,\boldsymbol{z} = \boldsymbol{0} \,.$$

$\Rightarrow$ **soustava normálních rovnic**

$$\boldsymbol{X}^{\boldsymbol{T}}\,(\boldsymbol{y} - \boldsymbol{X}\,\hat{\boldsymbol{\vartheta}}) = \boldsymbol{0} \,,$$

$$\boldsymbol{X}^{\boldsymbol{T}}\,\boldsymbol{X}\,\hat{\boldsymbol{\vartheta}} = \boldsymbol{X}^{\boldsymbol{T}}\,\boldsymbol{y} \,.$$

Předp.: $\boldsymbol{X}$ má plnou hodnost, k.

Pozn.: To je obvyklý případ, ale pro reálná data nelze zaručit. Pokud to není splněno, lze nějakou vysvětlující proměnnou vyjádřit jako lineární kombinaci ostatních a v modelu ji vynechat.
Důsledek: (Symetrická) matice soustavy normálních rovnic $\boldsymbol{X^T X}$ je regulární, řešení je

$$\hat{\boldsymbol{\vartheta}} = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T y}\,.$$

Pozn.: Matice $\left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T}$ je **pseudoinverze** obdélníkové matice $\boldsymbol{X}$. Vynásobíme-li jí matici $\boldsymbol{X}$ zleva, dostaneme jednotkovou matici.
Věta: $\forall i : \hat{\boldsymbol{\vartheta}}_i$ (přesněji $\hat{\Theta}_i$) je nejlepší nestranný odhad $\boldsymbol{\vartheta}_i$.
Kovarianční matice vektoru odhadů $\hat{\Theta}$ je

$$\boldsymbol{\Sigma}_{\hat{\Theta}} = \sigma^2 \left(\boldsymbol{X^T X}\right)^{-1}\,.$$

Věta: Hodnota regresní funkce v bodě $\boldsymbol{x} = (x_1, \ldots, x_k)^T \in \mathbb{R}^k$,

$$\boldsymbol{x^T}\,\hat{\boldsymbol{\vartheta}} = \sum_{i=1}^{k} \hat{\vartheta}_i\, x_i\,,$$

je nejlepší nestranný odhad vysvětlované náhodné veličiny Y v bodě $\boldsymbol{x}$.

2.17 Intervalové odhady regresních koeficientů při známém rozptylu

Věta: (Symetrický) $(1-\alpha)$-konfidenční interval pro odhad parametru $\boldsymbol{\vartheta}_i$ je

$$\hat{\boldsymbol{\vartheta}}_i \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\left(\boldsymbol{\Sigma}_{\hat{\Theta}}\right)_{ii}} = \hat{\boldsymbol{\vartheta}}_i \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sigma\sqrt{c_{ii}}\,,$$

kde $\left(\boldsymbol{\Sigma}_{\hat{\Theta}}\right)_{ii}$, resp. c_{ii}, je i-tý prvek na diagonále matice $\boldsymbol{\Sigma}_{\hat{\Theta}}$, resp. $\boldsymbol{C} := \left(\boldsymbol{X^T X}\right)^{-1}$.
Pozn.: I pro intervalový odhad jednoho regresního koeficientu potřebujeme vypočítat matici $\left(\boldsymbol{X^T X}\right)^{-1}$.
Pokud rozptyl neznáme, je potřeba jej nahradit odhadem *(viz dále)*.

2.18 Odhady rozptylu σ^2 původního rozdělení

Maximálně věrohodný:

$$\hat{\sigma}^2_{\hat{\boldsymbol{e}}} = \frac{1}{n}\,R_{SS} = \frac{1}{n}\sum_{j=1}^{n}\hat{e}_j^2 = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \sum_{i=1}^{k}\hat{\vartheta}_i\, x_{ji}\right)^2,$$

nestranný:

$$\hat{\sigma}^2 = \frac{1}{n-k}\,R_{SS} = \frac{1}{n-k}\sum_{j=1}^{n}\hat{e}_j^2 = \frac{1}{n-k}\sum_{j=1}^{n}\left(y_j - \sum_{i=1}^{k}\hat{\vartheta}_i\, x_{ji}\right)^2.$$

$\dfrac{R_{SS}}{\sigma^2}$ pochází z rozdělení $\chi^2(n-k)$.
Odhady R_{SS} a $\hat{\boldsymbol{\vartheta}}$ jsou nezávislé. (Nikoli však jednotlivé složky vektoru $\hat{\boldsymbol{\vartheta}}$ mezi sebou!)

2.19 Intervalové odhady regresních koeficientů při neznámém rozptylu

Věta: (Symetrický) $(1-\alpha)$-konfidenční interval pro odhad parametru ϑ_i je

$$\hat{\boldsymbol{\vartheta}}_i \pm q_{\mathrm{t}(n-k)}\left(1-\frac{\alpha}{2}\right)\,\hat{\sigma}\,\sqrt{c_{ii}} = \hat{\boldsymbol{\vartheta}}_i \pm q_{\mathrm{t}(n-k)}\left(1-\frac{\alpha}{2}\right)\cdot\sqrt{\frac{R_{SS}\,c_{ii}}{n-k}}\,,$$

kde c_{ii} je i-tý prvek na diagonále matice $\boldsymbol{C} = \left(\boldsymbol{X}^{\boldsymbol{T}}\,\boldsymbol{X}\right)^{-1}$.

3 Volba vysvětlujících proměnných

Můžeme vyhodnotit, zda se jednotlivé regresní koeficienty statisticky významně liší od 0, např. pro závislost kriminality na různých sociálních faktorech ve státech USA v r. 1960 (převzato z [Wasserman] a
`http://lib.stat.cmu.edu/DASL/Stories/USCrime.html`):
Popis proměnných v originále (Intercept=absolutní člen):
Crime rate: # of offenses reported to police per million population
The number of males of age 14-24 per 1000 population
Indicator variable for Southern states (0 = No, 1 = Yes)
Mean # of years of schooling $\times 10$ for persons of age 25 or older
1960 per capita expenditure on police by state and local government
1959 per capita expenditure on police by state and local government
Labor force participation rate per 1000 civilian urban males age 14–24
The number of males per 1000 females
State population size in hundred thousands
Unemployment rate of urban males per 1000 of age 14–24
Unemployment rate of urban males per 1000 of age 35–39
Median value of transferable goods and assets or family income in tens of $

Covariate	$\widehat{\beta}_j$	$\widehat{\mathrm{se}}(\widehat{\beta}_j)$	t value	p-value	
(Intercept)	-589.39	167.59	-3.51	0.001	**
Age	1.04	0.45	2.33	0.025	*
Southern State	11.29	13.24	0.85	0.399	
Education	1.18	0.68	1.7	0.093	
Expenditures	0.96	0.25	3.86	0.000	***
Labor	0.11	0.15	0.69	0.493	
Number of Males	0.30	0.22	1.36	0.181	
Population	0.09	0.14	0.65	0.518	
Unemployment (14–24)	-0.68	0.48	-1.4	0.165	
Unemployment (25–39)	2.15	0.95	2.26	0.030	*
Wealth	-0.08	0.09	-0.91	0.367	

Může se zdát, že např. vyšší výdaje na vzdělání a prevenci kriminality souvisí s vyšší kriminalitou. To ovšem může být z mnoha jiných důvodů než je kauzální závislost (v kterémkoli směru).
Naopak některé vysvětlující proměnné, např. „jižní státy“, jsou „zbytečné“. Jejich vynecháním se model téměř nezhorší a přitom zjednoduší.

3.1 Kritéria pro výběr modelu

underfitting: příliš jednoduchý model, nepřesný
overfitting: zbytečně složitý model
Kritéria vyvažují počet vysvětlujících proměnných a chybu modelu (obvykle R_{SS}), např.

$$R_{SS}(\mathcal{S}) + 2\,\hat{\sigma}^2\ |\mathcal{S}|\ ,$$

kde $\mathcal{S}$ je množina vybraných proměnných, $|\mathcal{S}|$ jejich počet a $R_{SS}(\mathcal{S})$ je reziduální součet čtverců modelu používajícího pouze proměnné z $\mathcal{S}$.

$$\lambda_{\mathcal{S}} - c\ |\mathcal{S}|\ ,$$

kde $\lambda_{\mathcal{S}}$ je logaritmus věrohodnosti modelu používajícího pouze proměnné z $\mathcal{S}$. Doporučené hodnoty c např. 1 nebo $\frac{1}{n}\ln n$.
Zde nezáleží na použitých jednotkách, neboť porovnáváme pouze relativní přesnost jednotlivých modelů.

3.2 Volba vysvětlujících proměnných

A. Postupně přidáváme vždy tu proměnnou, která nejvíc vylepší kritérium.
B. Použijeme všechny proměnné a postupně ubíráme vždy tu, která nejvíc vylepší kritérium.
C. Začneme od jakékoli podmnožiny kritérií a náhodně ji upravujeme (přidání/ubrání/výměna), větší pravděpodobnost dáme změnám, které víc vylepšují kritérium.

3.3 Rozdělení na trénovací a testovací data, cross-validation

Místo skutečné chyby používáme její odhad na základě trénovacích dat.
Testovací data by měla být s trénovacími disjunktní.
Úspornější postupy:
cross-validation: rozdělíme data na m disjunktních množin, vždy použijeme jednu jako testovací a ostatní jako trénovací,
leave-one-out cross-validation: extrémní případ předchozího pro $m = n$.