

ANALÝZA DAT V R

6. LINEÁRNÍ REGRESE

Mgr. Markéta Pavlíková

Katedra pravděpodobnosti a matematické statistiky MFF UK

www.biostatisticka.cz

PŘEHLED TESTŮ

rozdělení	normální	spojité	alternativní / diskrétní
populační parametr (o čem je hypotéza)	populační průměr	populační medián (distribuční funkce)	pravděpodobnost jevu / (ne)závislost / poměr šancí
jeden výběr	jednovýběrový t-test	jednovýběrový Wilcoxonův test znaménkový test	test proporcí
výběr dvojic	párový t-test	párový Wilcoxonův test znaménkový test	McNemarův test
dva nezávislé výběry (třídění na dvě kategorie / závislost na binární proměnné)	dvouvýběrový t-test	Mann-Whitney (dvouvýběrový Wilcoxon) Kolmogorov-Smirnov	Fisherův exaktní test χ^2 -test
k nezávislých výběrů (závislost na kategoriální proměnné)	analýza rozptylu (jednoduché třídění, F-test)	Kruskal-Wallis	Fisherův exaktní test χ^2 -test
závislost na spojité proměnné	lineární regrese	zobecněná lineární regrese (speciální případy) jádrová regrese (kernel smoothing) ...	logistická regrese multinomiální logistická regrese
opakovaná měření	smíšená lineární regrese (mixed models)	smíšená zobecněná lineární regrese	

PŘEHLED TESTŮ

rozdělení	normální	spojité	alternativní / diskrétní
populační parametr (o čem je hypotéza)	populační průměr	populační medián (distribuční funkce)	pravděpodobnost jevu / (ne)závislost / poměr šancí
jeden výběr	jednovýběrový t-test	jednovýběrový Wilcoxonův test znaménkový test	test proporcí
výběr dvojic	párový t-test	párový Wilcoxonův test znaménkový test	McNemarův test
dva nezávislé výběry (třídění na dvě kategorie / závislost na binární proměnné)	dvouvýběrový t-test	Mann-Whitney (dvouvýběrový Wilcoxon) Kolmogorov-Smirnov	Fisherův exaktní test χ^2 -test
k nezávislých výběrů (závislost na kategoriální proměnné)	analýza rozptylu (jednoduché třídění, F-test)	Kruskal-Wallis	Fisherův exaktní test χ^2 -test
závislost na spojité proměnné	lineární regrese	zobecněná lineární regrese (speciální případy) jádrová regrese (kernel smoothing) ...	logistická regrese multinomiální logistická regrese
opakovaná měření	smíšená lineární regrese (mixed models)	smíšená zobecněná lineární regrese	

PŘEHLED TESTŮ

rozdělení	normální	spojité	alternativní / diskrétní
populační parametr (o čem je hypotéza)	populační průměr	populační medián (distribuční funkce)	pravděpodobnost jevu / (ne)závislost / poměr šancí
jeden výběr	jednovýběrový t-test	jednovýběrový Wilcoxonův test znaménkový test	test proporcí
výběr dvojic	párový t-test	párový Wilcoxonův test znaménkový test	McNemarův test
dva nezávislé výběry (třídění na dvě kategorie / závislost na binární proměnné)	dvouvýběrový t-test	Mann-Whitney (dvouvýběrový Wilcoxon) Kolmogorov-Smirnov	Fisherův exaktní test χ^2 -test
k nezávislých výběrů (závislost na kategoriální proměnné)	analýza rozptylu (jednoduché třídění, F-test)	Kruskal-Wallis	Fisherův exaktní test χ^2 -test
závislost na spojité proměnné	lineární regrese	zobecněná lineární regrese (speciální případy) jádrová regrese (kernel smoothing) ...	logistická regrese multinomiální logistická regrese
opakovaná měření	smíšená lineární regrese (mixed models)	smíšená zobecněná lineární regrese	

ANALÝZA ROZPTYLU

analýza rozptylu jednoduchého třídění (ANOVA)

- ▶ $Y_{11}, \dots, Y_{1n_1} \sim N(\mu_1, \sigma^2)$ (první výběr, průměr $\bar{Y}_{1\bullet}$)
 $Y_{21}, \dots, Y_{2n_2} \sim N(\mu_2, \sigma^2)$ (druhý výběr, průměr $\bar{Y}_{2\bullet}$)

...

- ▶ $Y_{k1}, \dots, Y_{kn_k} \sim N(\mu_k, \sigma^2)$ (k -tý výběr, průměr $\bar{Y}_{k\bullet}$)

- ▶ **nezávislé** výběry (shodné rozptyly, normální rozdělení)
- ▶ $H_0 : \mu_1 = \mu_2 = \dots = \mu_k \quad (= \mu) \quad H_1 : \text{neplatí } H_0$
- ▶ celkový průměr $\bar{Y}_{\bullet\bullet}$, celkový rozsah $n = \sum_{i=1}^k n_i$
- ▶ rozklad součtu čtverců

$$\sum_{i=1}^k \sum_{t=1}^{n_i} (Y_{it} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^k n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^k \sum_{t=1}^{n_i} (Y_{it} - \bar{Y}_{i\bullet})^2$$

(celková variabilita) = (variabilita mezi) + (variabilita uvnitř)

$$S_T = S_A + S_e$$

$$f_T = f_A + f_e$$

$$(n - 1) = (k - 1) + (n - k)$$

ANALÝZA ROZPTYLU

(celková variabilita) = (variabilita mezi) + (variabilita uvnitř)

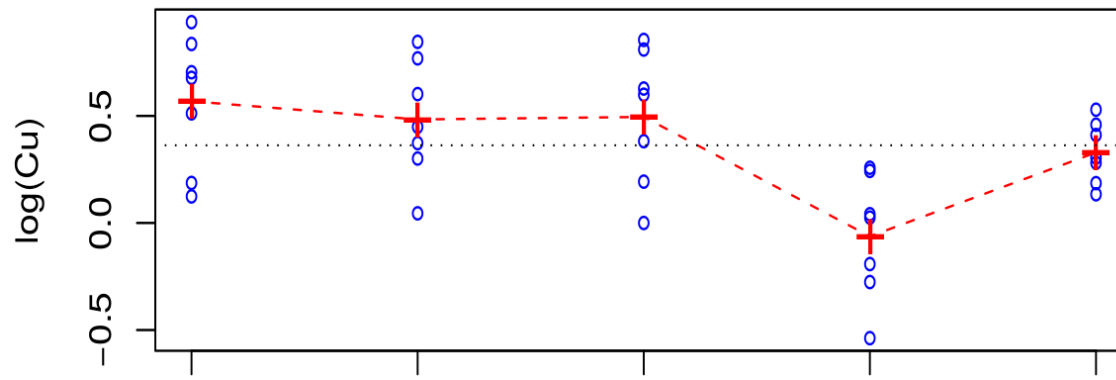
$$\sum_{i=1}^k \sum_{t=1}^{n_i} (Y_{it} - \bar{Y}_{\bullet\bullet})^2 = \sum_{i=1}^k n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 + \sum_{i=1}^k \sum_{t=1}^{n_i} (Y_{it} - \bar{Y}_{i\bullet})^2$$

S_T
vysvětlováno

S_A
vysvětleno závislostí

S_E
nevysvětleno

H_0 zamítnout, je-li $F_A = \frac{S_A/f_A}{S_E/f_E} \geq F_{f_A, f_E}(1 - \alpha)$



ANALÝZA ROZPTYLU

varianty zápisu modelu AR jednoduchého třídění

- ▶ **model** – idealizovaná představa o vzniku pozorované hodnoty
- ▶ měření = úroveň + náhodná „chyba“
měření = **systematická složka** + **náhodná složka**

$$\begin{aligned} Y_{it} &= \mu_i + E_{it} & 1 \leq t \leq n_i, \quad 1 \leq i \leq k \\ &= \mu + (\mu_i - \mu) + E_{it} & E_{it} \text{ nezávislé} \\ &= \mu + \alpha_i + E_{it} & E_{it} \sim N(0, \sigma^2) \end{aligned}$$

- ▶ **reparametrizace** (α_i – efekty faktoru A):

$$\sum_{i=1}^k \alpha_i = 0 \quad \text{podmínka navíc = stupeň volnosti dolů}$$

- ▶ $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k$ (totéž jako $\mu_1 = \mu_2 = \dots = \mu_k$)
- ▶ pro $k = 2$ je $F_A = T^2$ (vztah s dvouvýběrovým t -testem)

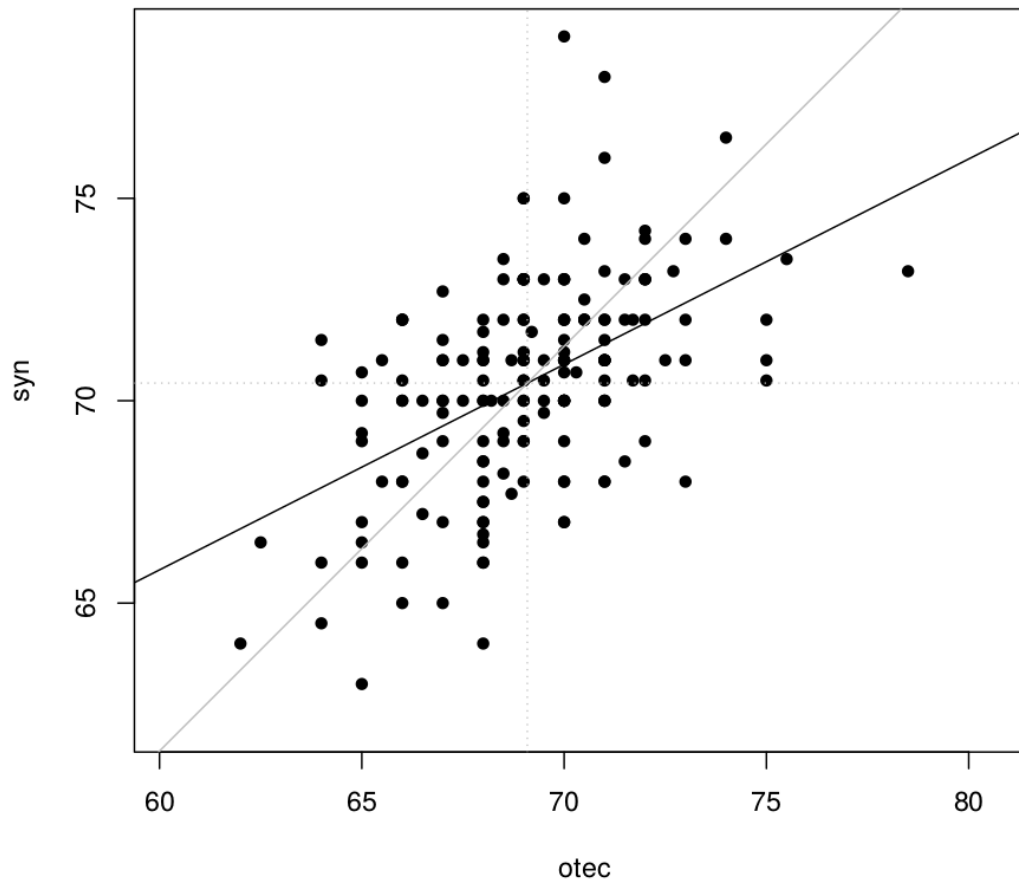
REGRESE – HISTORICKÝ EXKURZ

- ▶ tendence (návrat) k průměrnosti
F. Galton (1886) vyšetřoval dědičnost výšky postavy
- ▶ uvažujme otce, jejichž výška je rovna průměrné výšce generace **všech** otců; průměrná výška synů otců této výšky bude rovna průměrné výšce **všech** synů
- ▶ uvažujme otce o 10 cm **vyšší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen asi o 5 cm **vyšší**, než průměrná výška generace synů
- ▶ uvažujme otce o 10 cm **nižší**, než je průměrná výška generace otců: průměrná výška synů těchto otců bude jen o asi 5 cm **nižší**, než průměrná výška generace synů
- ▶ průměrné výšky synů nereprodukuje celou odchylku výšky otce od průměru, je tu návrat k průměru (regrese)

REGRESE – HISTORICKÝ EXKURZ

Závislost výšky syna na výšce otce

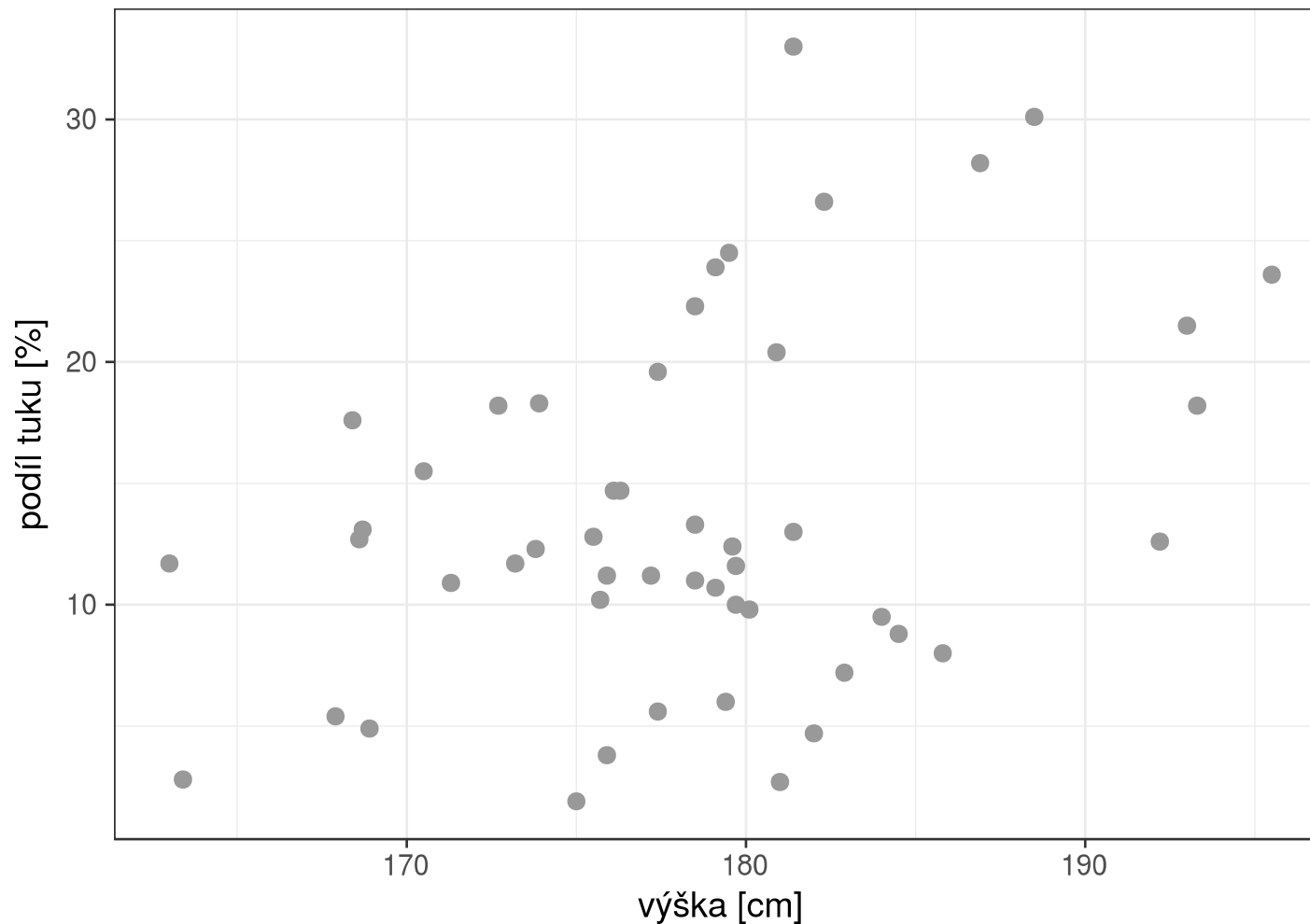
výšky v ang. palcích, Galtonova data, ke každému otci náhodně vybrán jeden z jeho synů
proložena přímka: $y = 35,4 + 0,5x$, pro porovnání šedivě přímka s jednotkovou směrnici



LINEÁRNÍ REGRESE – PŘÍKLAD

data Policie, zkoumáme závislost procenta tuku na výšce postavy

Vztah % tuku na výšce



LINEÁRNÍ REGRESE – MODEL

regresní přímka

- ▶ **předpokládaná** závislost střední hodnoty Y na nenáhodné x :

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad E_i \sim N(0, \sigma^2)$$

- ▶ $Y_i =$ **systematická složka** $(\beta_0 + \beta_1 x_i)$ + **náhodná složka** (E_i)
- ▶ k daným $x_1, \dots, x_n$ zjistíme $Y_1, \dots, Y_n$
- ▶ předpoklady:
 - ▶ **nezávislá** pozorování $Y_1, \dots, Y_n$ (tedy také $E_1, \dots, E_n$)
 - ▶ **stejný** rozptyl σ^2
 - ▶ **normální** rozdělení $E_1, \dots, E_n$ (potřebné až pro testy, normalitu nelze ověřovat testováním přímo $Y_1, \dots, Y_n$!)
- ▶ neznámé populační parametry β_0, β_1 odhadujeme metodou **nejmenších čtverců**:

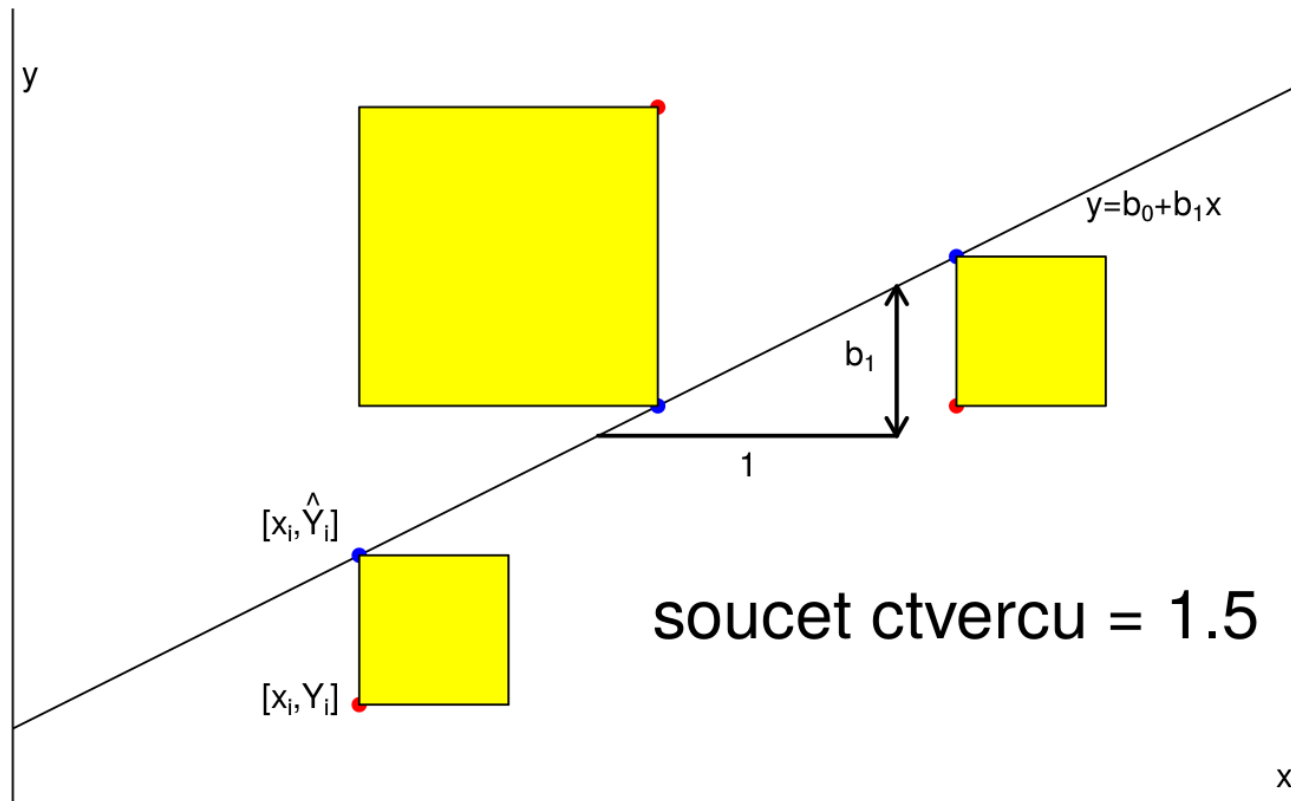
minimalizovat přes β_0, β_1 výraz
$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$$

- ▶ odhady označíme b_0, b_1

LINEÁRNÍ REGRESE – ČTVERCE

Metoda nejmenších čtverců

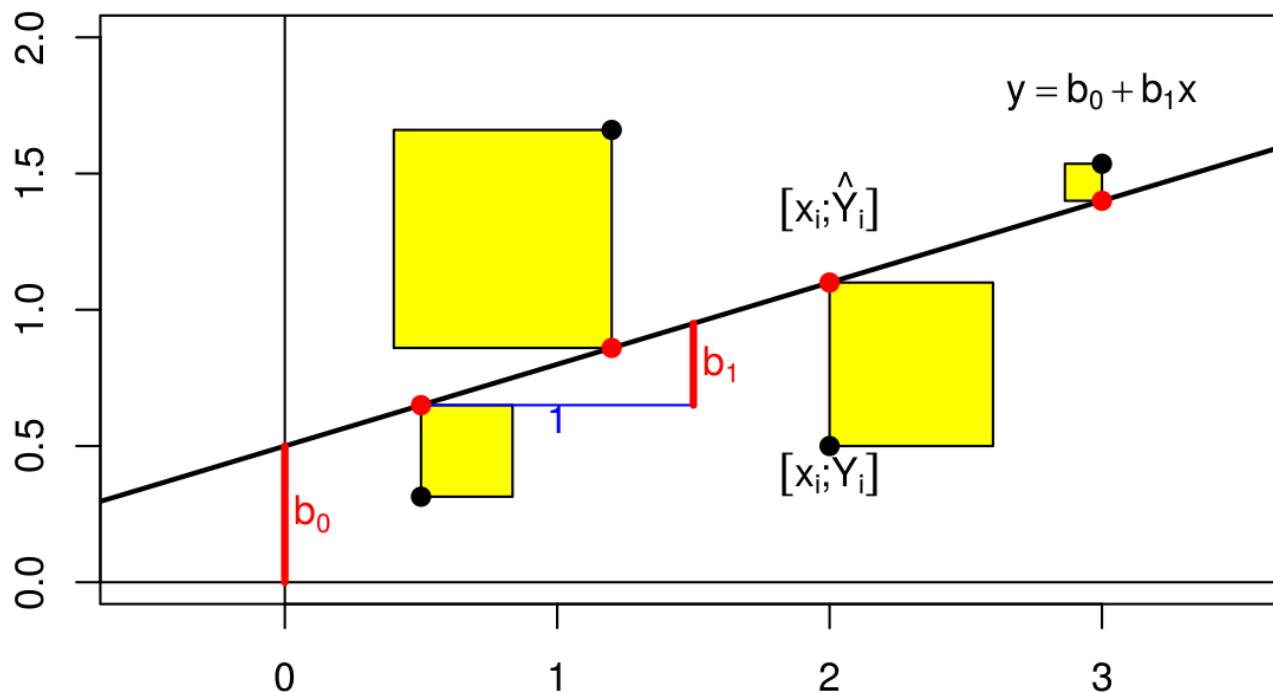
přímka vedená třemi červenými body $[x_i, Y_i]$



LINEÁRNÍ REGRESE - ČTVERCE

metoda nejmenších čtverců

odhadovaná závislost:	$y = \beta_0 + \beta_1 \cdot x$	(populace)
odhad závislosti:	$y = b_0 + b_1 \cdot x$	(výběr)
i -tá vyrovnaná hodnota:	$\hat{Y}_i = b_0 + b_1 \cdot x_i$	(výběr)
i -té reziduum:	$U_i = Y_i - \hat{Y}_i$	(výběr)
celková plocha čtverců:	$S_e = \sum_{i=1}^n U_i^2$	(výběr)



LINEÁRNÍ REGRESE – ODHADY

- ▶ b_1 – odhad směrnice β_1
- ▶ $b_1 = (b_0 + b_1(x + 1)) - (b_0 + b_1x)$
– odhad změny střední hodnoty závisle proměnné Y při **jednotkové změně** nezávisle proměnné x
- ▶ i -té reziduum $U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1x_i)$
- ▶ $Y_i = \hat{Y}_i + U_i$
- ▶ (vysvětlováno) = (vysvětleno závislostí) + (nevysvětleno)
- ▶ reziduální součet čtverců (nevysvětlená variabilita):

$$S_e = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - b_0 - b_1x_i)^2 = \sum_{i=1}^n U_i^2$$

- ▶ reziduální rozptyl (odhad rozptylu σ^2)

$$S^2 = \frac{S_e}{n - 2} \text{ stupně volnosti}$$

LINEÁRNÍ REGRESE – ODHADY

prokazování závislosti

- ▶ modelujeme závislost $E Y$ na x pomocí $E Y = \beta_0 + \beta_1 x$
- ▶ nezávislost $y = \beta_0 + \beta_1 x$ na x znamená $\beta_1 = 0$
- ▶ hypotézu $H_0 : \beta_1 = 0$ testujeme pomocí statistiky

$$T = \frac{b_1}{\text{S.E.}(b_1)}$$

- ▶ hypotézu zamítáme, je-li $|T| \geq t_{n-2}(1 - \alpha/2)$
tj. je-li příslušná p -hodnota $\leq \alpha$
- ▶ pokud H_0 zamítneme, říkáme, na hladině α je **závislost průkazná**

LINEÁRNÍ REGRESE - INTERPRETACE

koeficient determinace

[coefficient of determination]

- ▶ podíl variability Y vysvětlené uvažovanou závislostí (jakou část variability Y se podařilo závislostí na x vysvětlit)
- ▶

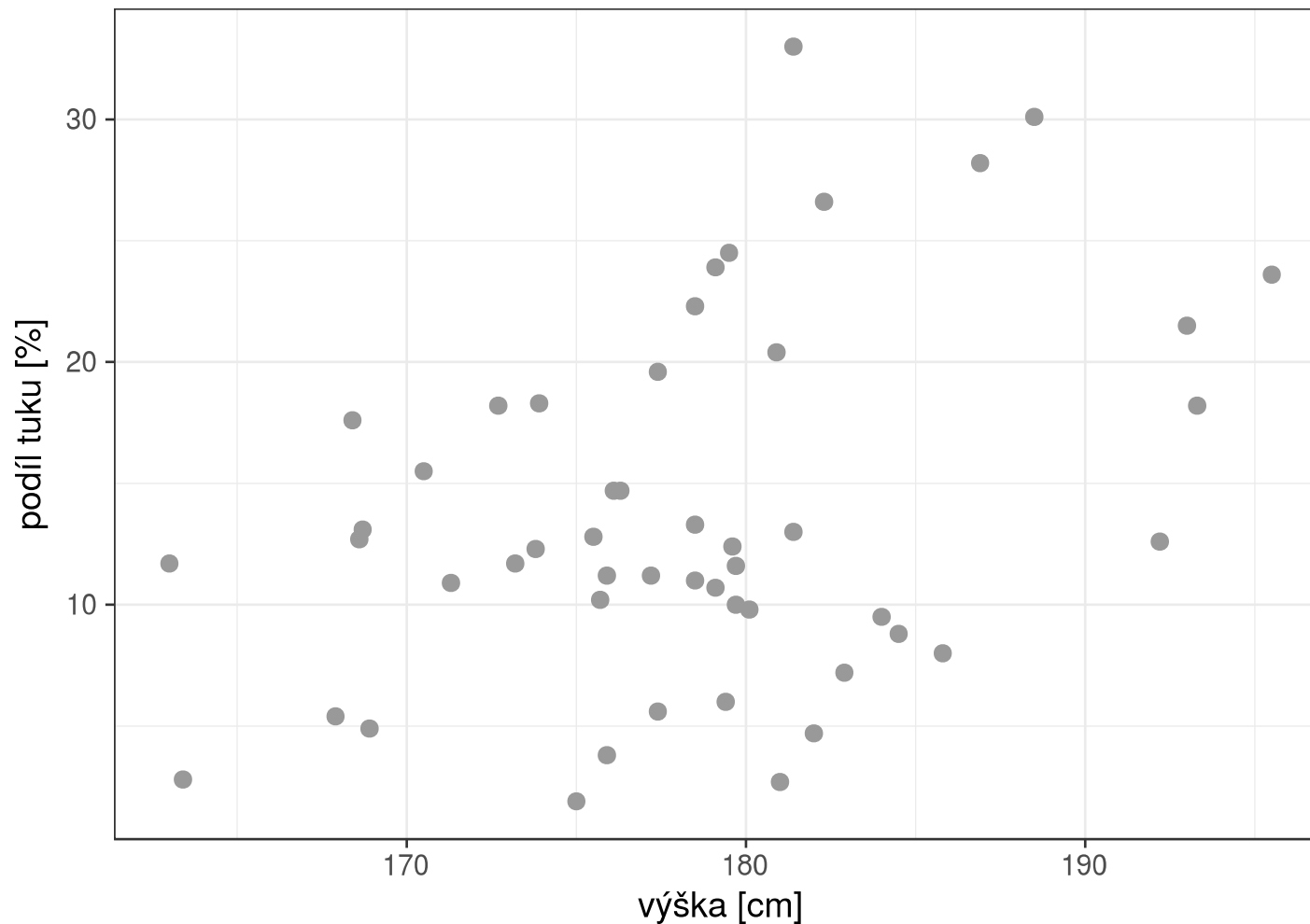
$$\begin{aligned} R^2 &= \frac{\text{variabilita vysvětlená}}{\text{variabilita vysvětlovaná}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{\text{variabilita nevysvětlená}}{\text{variabilita vysvětlovaná}} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \\ &= 1 - \frac{S_e}{\sum(Y_i - \bar{Y})^2} \end{aligned}$$

- ▶ R^2 je bezrozměrné číslo, často vyjádřeno v procentech
- ▶ R^2 ukazuje, zda má smysl předpovídat pomocí regrese
- ▶ v případě regresní přímky je $R^2 = r_{XY}^2$

LINEÁRNÍ REGRESE – PŘÍKLAD

data Policie, zkoumáme závislost procenta tuku na výšce postavy

Vztah % tuku na výšce



LINEÁRNÍ REGRESE – PŘÍKLAD

příklad závislost procenta tuku na výšce

data: Policie

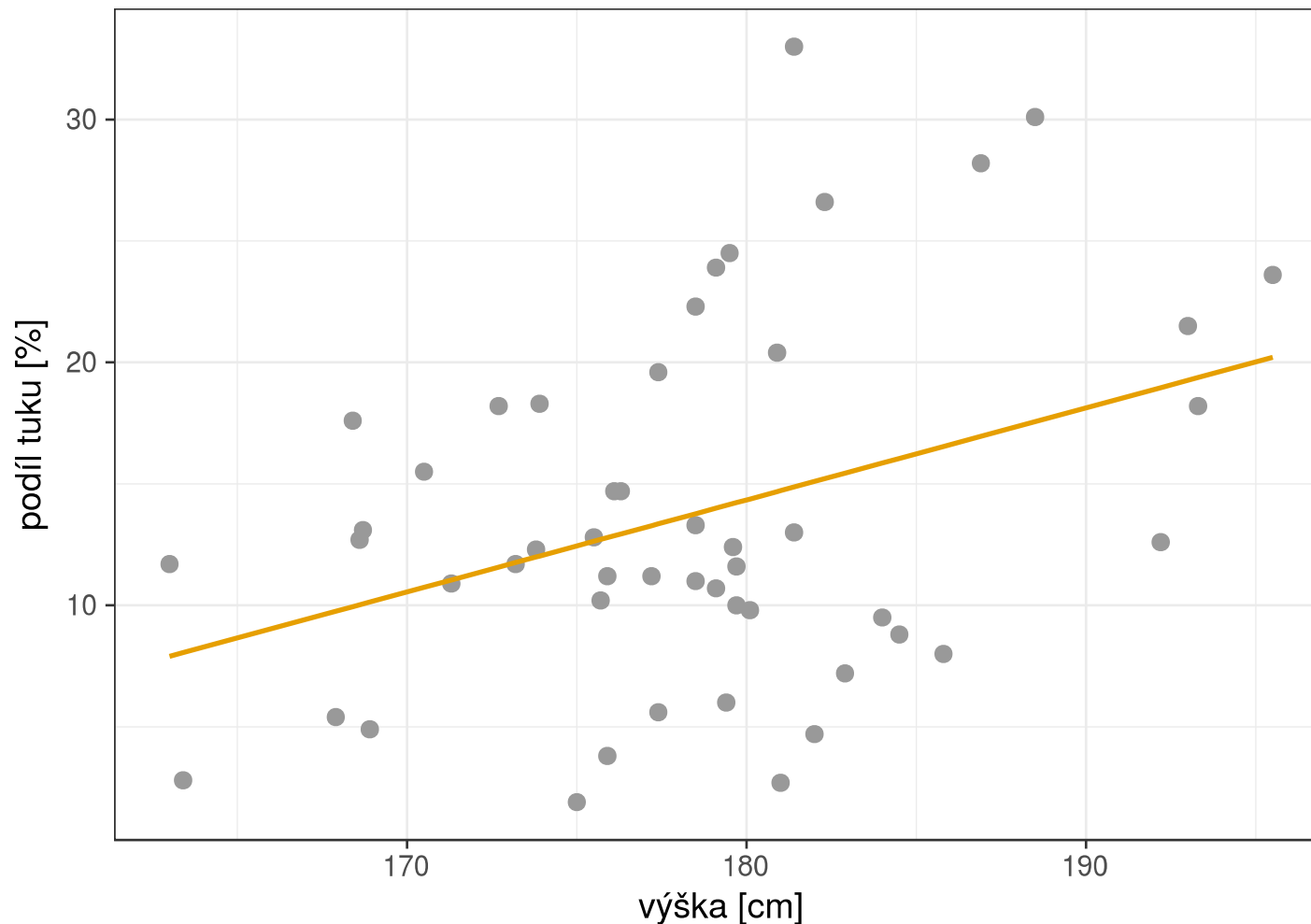
regresor	b_j	S.E. (b_j)	t	p
abs. člen	-53,870	24,657	-2,185	0,0338
height	0,379	0,138	2,742	0,0086

- ▶ předpověď: $\hat{Y}_i = -53,870 + 0,379 \cdot x_i$
- ▶ $\widehat{\text{fat}} = -53,870 + 0,379 \cdot \text{height}$
- ▶ závislost procenta tuku na výšce je na 5% hladině průkazná, neboť $p = 0,86 \%$
- ▶ na každý centimetr výšky *v průměru* přibude 0,379 procentního bodu tuku
- ▶ `[summary(lm(fat~height))]`

LINEÁRNÍ REGRESE – PŘÍKLAD

data Policie, zkoumáme závislost procenta tuku na výšce postavy

Vztah % tuku na výšce



LINEÁRNÍ REGRESE – INTERPRETACE

tabulka analýzy rozptylu

varia- bilita	součet čtverců	st. vol.	prům. čtverec	F	p
regrese	362,54	1	362,54	7,519	0,0086
rezid.	2314,41	48	48,22		
celk.	2676,95	49	(54,63)		



$$s^2 = 48,22$$



$$R^2 = \frac{362,54}{2676,95} = 1 - \frac{2314,41}{2676,95} = 0,135$$

- závislostí na výšce jsme vysvětlili jen 13,5 % variability procenta tuku
- `[anova(lm(fat~height))]`

LINEÁRNÍ REGRESE – PŘÍKLAD V R

```
fit<-(lm(fat~height,data=Policie))
```

```
> summary(fit)
```

Call:

```
lm(formula = fat ~ height, data = Policie)
```

Residuals:

Min	1Q	Median	3Q	Max
-12.0187	-5.0696	-0.8255	4.5207	18.1297

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-53.8700	24.6571	-2.185	0.03382 *
height	0.3789	0.1382	2.742	0.00855 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.944 on 48 degrees of freedom

Multiple R-squared: 0.1354, Adjusted R-squared: 0.1174

F-statistic: 7.519 on 1 and 48 DF, p-value: 0.008554

```
> anova(fit)
```

Analysis of Variance Table

Response: fat

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
height	1	362.54	362.54	7.519	0.008554 **
Residuals	48	2314.41	48.22		

MNOHONÁSOBNÁ LINEÁRNÍ REGRESE

- ▶ závislost na dvou (nebo více) nezávisle proměnných
- ▶ pozorování $(x_1, v_1, Y_1), \dots, (x_n, v_n, Y_n)$
- ▶ představa (model)

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 v_i}_{E Y_i} + E_i$$

- ▶ střední hodnota Y_i (tj. systematická, nenáhodná složka Y_i) vysvětlena pomocí x_i, v_i jako $\beta_0 + \beta_1 x_i + \beta_2 v_i$
- ▶ $E_1, \dots, E_n$ (také $Y_1, \dots, Y_n$) jsou **nezávislé** náhodné veličiny
- ▶ $E_i \sim N(0, \sigma^2)$ (normální rozdělení se stejným rozptylem)
- ▶ b_0, b_1, b_2 – odhady parametrů $\beta_0, \beta_1, \beta_2$

MNOHONÁSOBNÁ LINEÁRNÍ REGRESE

interpretace

- ▶ b_1 – odhad změny střední hodnoty Y při **jednotkové** změně x a **nezměněné** hodnotě v
- ▶ b_2 – odhad změny střední hodnoty Y při **jednotkové** změně v a **nezměněné** hodnotě x
- ▶ U_i – **reziduum**

$$U_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 x_i + b_2 v_i)$$

- ▶ **rozklad variability** $S_T = S_R + S_e$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

MNOHONÁSOBNÁ LINEÁRNÍ REGRESE

- ▶ **koeficient determinace R^2**

podíl celkové variability, který se podařilo vysvětlit závislostí Y na x a v (jakou část variability Y se podařilo vysvětlit)

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T}$$

- ▶ $H_0 : \beta_1 = \beta_2 = 0$ (chování Y nezávisí ani na x ani na v)

$$F = \frac{S_R/2}{S_e/(n-3)} \geq F_{2,n-3}(1-\alpha)$$

- ▶ p -hodnota tohoto testu bývá uváděna spolu s R^2

MNOHONÁSOBNÁ LINEÁRNÍ REGRESE

testy o přínosu jednotlivých regresorů

- ▶ model $y = \beta_0 + \beta_1 x + \beta_2 v$

- ▶ $H_0 : \beta_2 = 0$

k vysvětlení chování Y stačí x , tj. $y = \beta_0 + \beta_1 x$

$$T_2 = \frac{b_2}{\text{S.E.}(b_2)}, \quad \text{zamítat pro } |T_2| \geq t_{n-3}(1 - \alpha/2)$$

- ▶ $H_0 : \beta_1 = 0$

k vysvětlení chování Y stačí v , tj. $y = \beta_0 + \beta_2 v$

$$T_1 = \frac{b_1}{\text{S.E.}(b_1)}, \quad \text{zamítat pro } |T_1| \geq t_{n-3}(1 - \alpha/2)$$

- ▶ $H_0 : \beta_0 = 0$ zpravidla nemá reálný smysl

LINEÁRNÍ REGRESE – PŘÍKLAD

příklad: závislost procenta tuku na výšce a váze

data: Policie

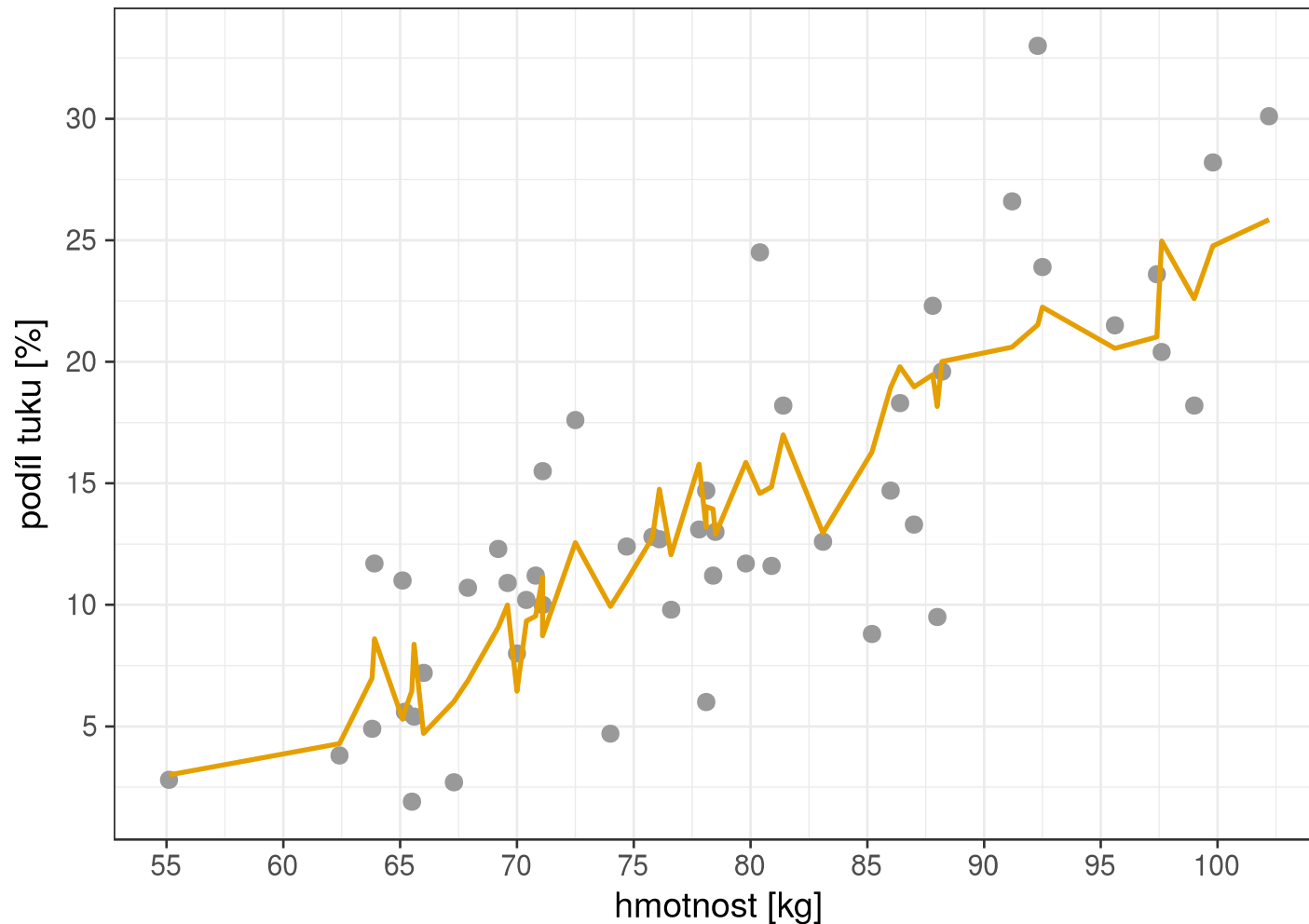
regresor	b_j	S.E. (b_j)	t	p
abs. člen	11,327	16,682	0,679	0,5005
height	-0,262	0,110	-2,376	0,0216
weight	0,624	0,0690	9,050	<0,0001

- ▶ `[summary(lm(fat~height+weight))]`
- ▶ při **stejné výšce** očekáváme na každý kg hmotnosti o 0,6 proc. bodu více tuku
- ▶ u mužů, kteří se liší výškou o 10 cm a **mají stejnou hmotnost** očekáváme, že ti vyšší mají v průměru o 2,6 proc. bodu **méně** tuku
- ▶ na 5% hladině nelze vyloučit výšku, průkazně přispívá k vysvětlení pomocí váhy
- ▶ na 1% hladině nelze vyloučit váhu, průkazně přispívá k vysvětlení pomocí výšky

LINEÁRNÍ REGRESE – PŘÍKLAD

data Policie, závislost procenta tuku na **výšce a hmotnosti**

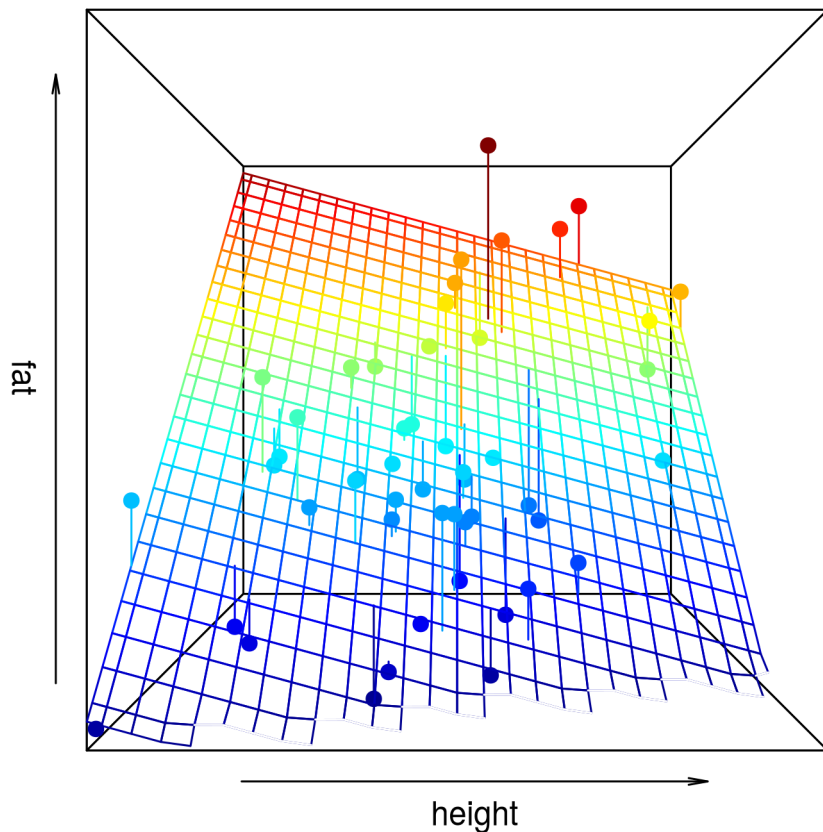
Vztah % tuku na hmotnosti



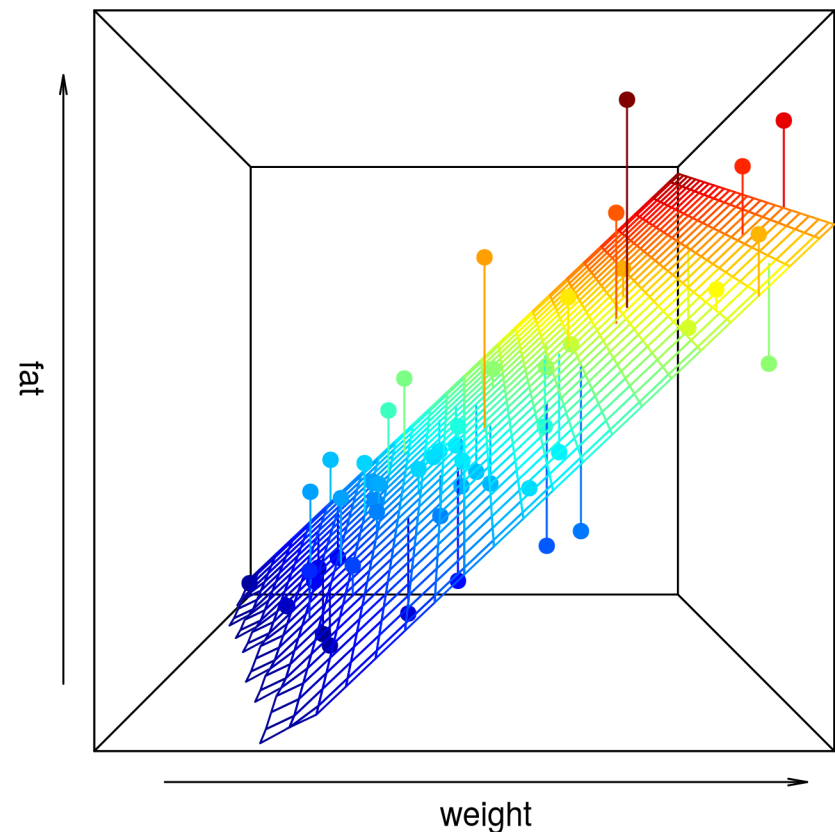
LINEÁRNÍ REGRESE – PŘÍKLAD

data Policie, závislost procenta tuku na **výšce a hmotnosti**

Tuk v závislosti na výšce a hmotnosti



Tuk v závislosti na výšce a hmotnosti



LINEÁRNÍ REGRESE – PŘÍKLAD

tabulka analýzy rozptylu

(F -statistika je v summary(), v commanderu nutno zvolit typ I a přínosy regresorů sečíst)

variabilita	souč. čtv,	st. vol.	prům. čtv.	F	p
regrese	1833,11	2	916,55	51,050	<0,001
rezid.	843,85	47	17,95		
celk.	2676,95	49	(54,63)		

- ▶ $R^2 = 1833,11/2676,95 = 1 - 843,85/2676,95 = 0,685$
- ▶ závislostí na výšce a váze jsme vysvětlili 68,5 % variability procenta tuku
- ▶ $s^2 = 17,95$
- ▶ na každé rozumné hladině zamítáme hypotézu, podle které procento tuku nezávisí ani na výšce ani na váze

LINEÁRNÍ REGRESE – PŘÍKLAD V R

```
Call:
lm(formula = fat ~ height + weight, data = Policie)

Residuals:
    Min       1Q   Median       3Q      Max
-8.657 -2.920  0.067  2.559 11.478

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.3273    16.6818   0.679   0.5005
height       -0.2615     0.1101  -2.376   0.0216 *
weight        0.6244     0.0690   9.050  7.24e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.237 on 47 degrees of freedom
Multiple R-squared:  0.6848, Adjusted R-squared:  0.6714
F-statistic: 51.05 on 2 and 47 DF, p-value: 1.651e-12
```

```
> anova(fit)

Analysis of Variance Table

Response: fat
          Df Sum Sq Mean Sq F value    Pr(>F)
height     1  362.54   362.54   20.193 4.552e-05 ***
weight     1 1470.56  1470.56   81.906 7.243e-12 ***
Residuals 47  843.85    17.95

```

LINEÁRNÍ REGRESE – DIAGNOSTIKA

regresní diagnostika

zda byly splněny předpoklady

- a) zvolili jsme správně **tvar závislosti**?
 - b) je **rozptyl** všude **stejný**?
 - c) je přiměřeně splněn předpoklad o **normálním rozdělení**?
 - d) jsou opravdu pozorování **nezávislá**?
problém často tam, kde působí čas
-
- ▶ k odstranění problémů s body a), b), c) často pomůže transformace, např. logaritmování závisle proměnné
 - ▶ `[plot(lm(fat~height+weight))]`

DĚKUJI ZA POZORNOST

www.biostatisticka.cz