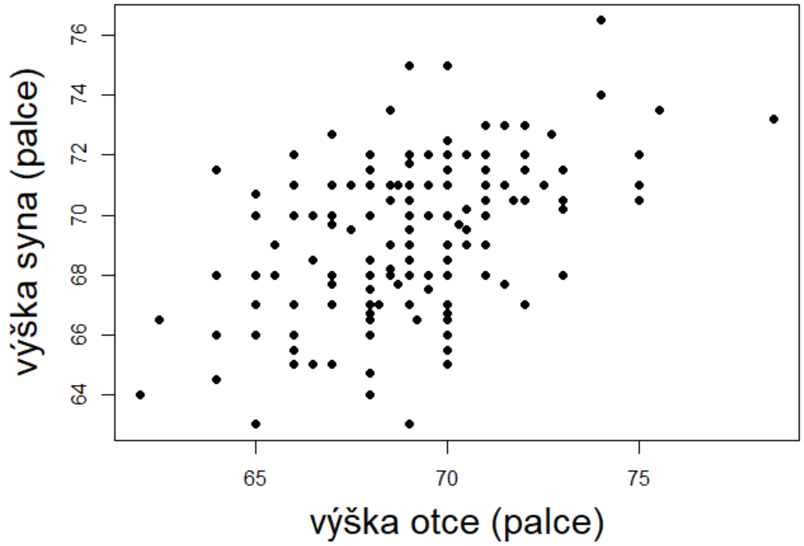
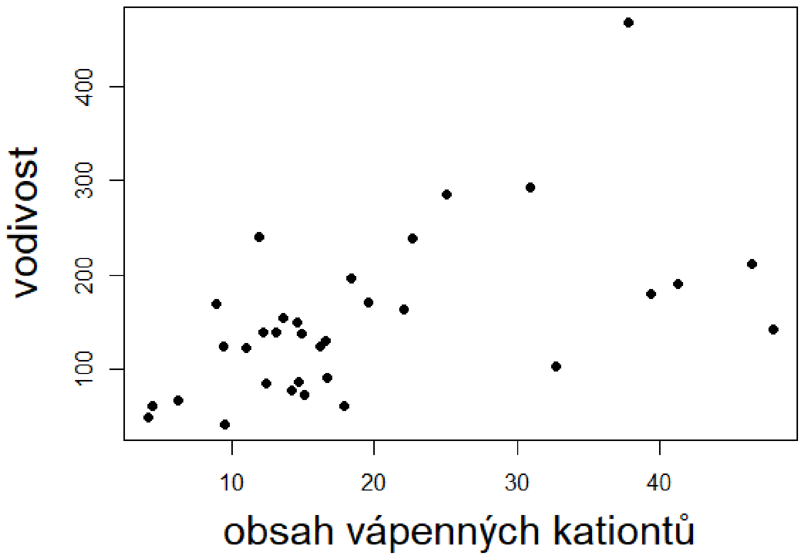
Příkladová data

Výška otce	Výška syna
175	178
177	173
188	188
173	173
163	164
163	168
178	169
…	…



Vodivost vody	Ca ionty
164	22.081
155	13.600
467	37.800
171	19.600
67	6.280
78	14.237



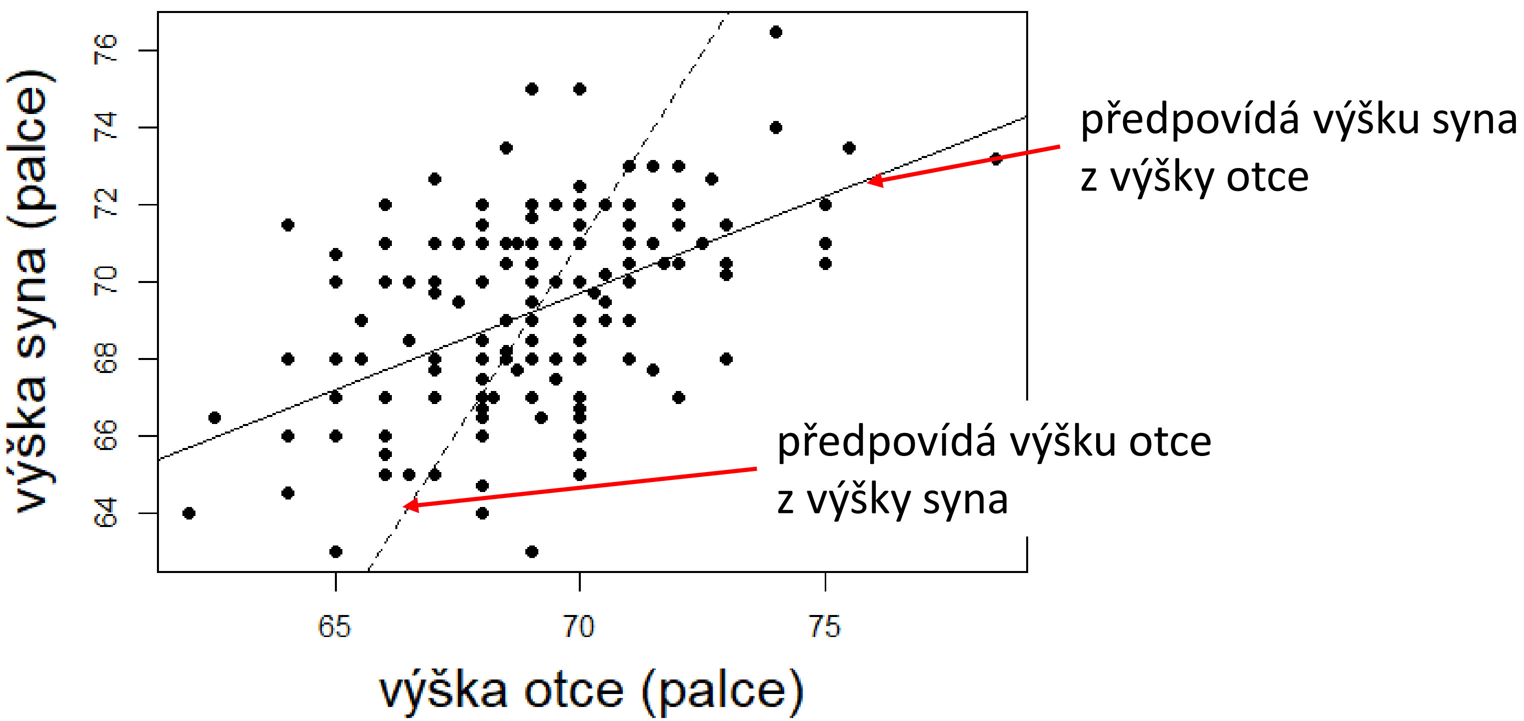
Analýza vztahu dvou kvantitativních proměnných

Dva přístupy, pohledy: **korelace a regrese**.

KORELACE popisuje sílu vzájemné závislosti.

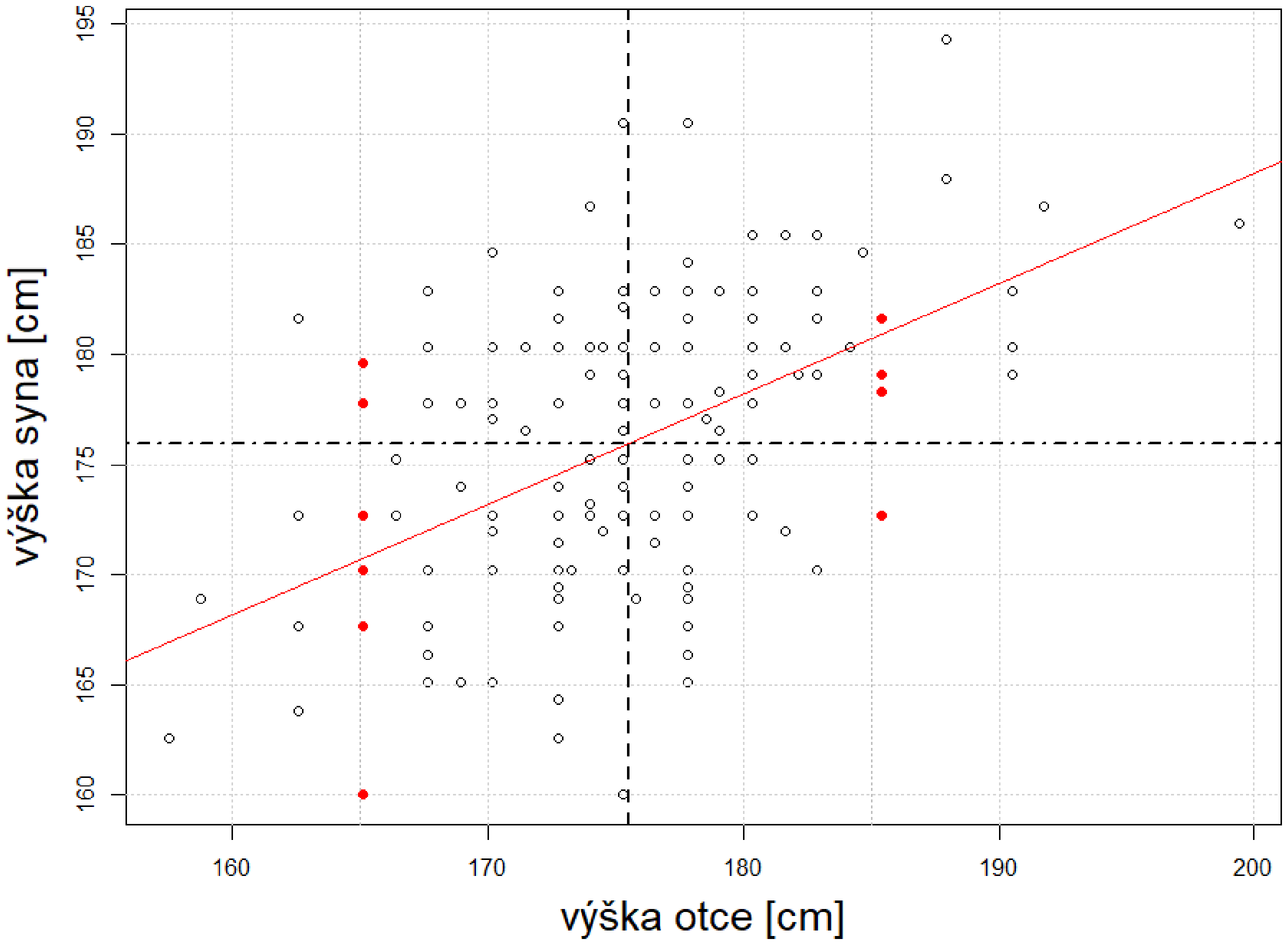
REGRESE pomocí jedné proměnné popisuje hodnoty druhé proměnné

<u>Příklad</u>: výšky otce a syna (data GaltonSyn)



Regrese – původ názvu

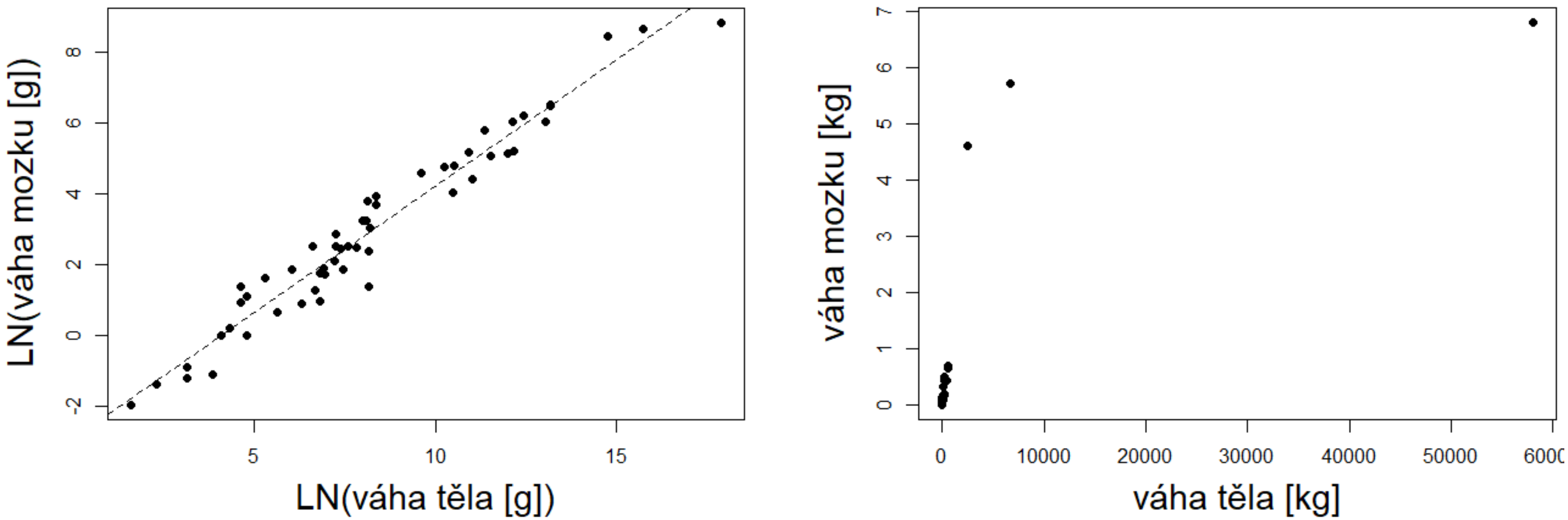
Sir F. Galton (1886): dědičnost výšky postavy



Regrese – vysvětlení variability Y pomocí X

- Opět spojitá, kvantitativní data
- Hodnoty proměnné ***Y*** modelujeme pomocí hodnot proměnné ***X***
- Lineární regresní model: $Y = \beta_0 + \beta_1 \cdot X + E$ … rovnice přímky
- Modelem vysvětlujeme variabilitu v hodnotách ***Y***, prokazujeme závislost ***Y*** na ***X*** nebo předpovídáme střední hodnotu ***Y*** pro nové hodnoty ***X***.
- V interpretaci zohledňujeme logickou závislost proměnných, „co ovlivňuje co".

Příklad: váha mozku vysvětlovaná váhou celého těla u 54 vybraných savců



Lineární regresní model [simple linear regression, bivariate regression]

$$Y_i = \underbrace{\beta_0 + \beta_1 \cdot X_i}_{} + \underbrace{E_i}_{} \qquad E_i \sim N(0, \sigma^2)$$

systematická složka + náhodná složka modelu [deterministic + stochastic component]

- ***Y*** nazýváme vysvětlovaná proměnná, závislá proměnná, odpověď, odezva
  [explained variable, dependent variable, response]
- ***X*** nazýváme vysvětlující proměnná, nezávislá proměnná, prediktor, regresor
  [explanatory variable, independent variable, predictor]
- $\boldsymbol{E_i \sim N(0, \sigma^2)}$ náhodná chyba, přirozená variabilita
- $\boldsymbol{\beta_0}$ a $\boldsymbol{\beta_1}$ jsou parametry platné pro celou populaci, tedy neznámé
  → hledáme odhady $\boldsymbol{b_0}$ a $\boldsymbol{b_1}$ a testujeme jejich nenulovost
- Parametry $\boldsymbol{\beta_0}$ a $\boldsymbol{\beta_1}$ určují přímku závislosti:
  $\boldsymbol{\beta_0}$ je průsečík s osou y [intercept],
  když *X* = 0, potom $Y = \beta_0$
  $\boldsymbol{\beta_1}$ je sklon přímky [slope];
  když *X* zvětším o 1 jednotku
  potom *Y* naroste (v průměru) o β_1.

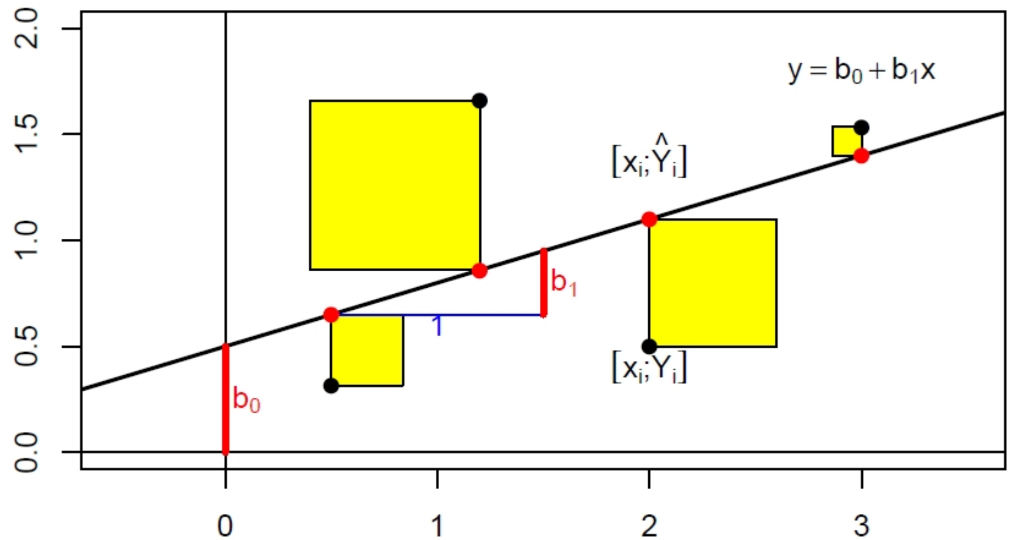
Odhad regresních koeficientů: β_0, β_1, σ^2

$$\boldsymbol{Y_i = \beta_0 + \beta_1 \cdot X_i + E_i} \qquad \boldsymbol{E_i \sim N(0, \sigma^2)}$$

1) Odhady $\boldsymbol{b_0}$ a $\boldsymbol{b_1}$ hledáme **metodou nejmenších čtverců**

[method of the least squares]

→ „nafitovaná“ hodnota:
$\boldsymbol{\widehat{Y}_i = b_0 + b_1 \cdot X_i}$
[fitted value], česky lépe modelovaná, vyhlazená hodnota



→ Reziduum $\boldsymbol{U_i}$:
$\boldsymbol{U_i = Y_i - \widehat{Y}_i = Y_i - b_0 + b_1 \cdot X_i}$

→ Součet čtverců (reziduální):

$SS_E = \sum_{i=1}^{n} U_i^2 = \sum\left(Y_i - \widehat{Y}_i\right)^2 = \sum(Y_i - b_0 + \beta_1 \cdot X_i)^2$ … **aby byl minimální**

→ $b_1 = \frac{s_{XY}}{s_X^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

→ $b_0 = \bar{Y} - b_1 \cdot \bar{X}$

Odhad regresních koeficientů: σ²

$$\boldsymbol{Y_i = \beta_0 + \beta_1 \cdot X_i + E_i} \qquad \boldsymbol{E_i \sim N(0, \sigma^2)}$$

2) Variabilitu náhodné odchylky $\boldsymbol{\sigma^2}$ odhadujeme jako reziduální rozptyl, tj.

$$S^2 = \frac{SS_E}{n-2}$$

Rozklad variability modelu (podobně jako v analýze rozptylu)

$SS_{TOT} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ … celková variabilita v datech

$DF_{TOT} = n - 1$

$SS_{REG} = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ … regresní, modelová variabilita, variabilita vysvětlená modelem

$DF_{REG} = k$

k … počet vysvětlujících proměnných

$SS_E = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ … reziduální variabilita, variabilita modelem <u>ne</u>vysvětlená

$DF_E = n - k - 1$

Platí: $SS_{TOT} = SS_{REG} + SS_E$

Předpoklady regresního lineárního modelu:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + E_i \qquad E_i \sim N(0, \sigma^2) \qquad Y_i \sim N(\beta_0 + \beta_1 \cdot X_i, \sigma^2)$$

- Y_i jsou vzájemně nezávislé hodnoty, pozorování.
- Y_i jsou zatíženy náhodnou variabilitou, pro kterou předpokládáme normální rozdělení: nelze ověřit předem, protože se střední hodnota ***EY*** mění a my teprve hledáme funkci, která tuto změnu popisuje. Proto nejprve modelujeme a potom ověřujeme. Normalitu zkontrolujeme na reziduálech $\left(Y_i - \hat{Y}_i\right)$. Předobrazem reziduálů v modelu jsou členy E_i.
- Pro E_i předpokládáme $N(0, \sigma^2)$ a že σ^2 se nemění.
- X_i naopak považujeme za přesné hodnoty bez náhodné chyby (variability). To splňují např. laboratorní teploty v různých pokusných boxech. Naopak váha těla savců z příkladu má jistě svoji variabilitu, předpoklad není dodržen.
- ***EY*** je lineární funkcí hodnot X_i (viz dále)

Předpoklady regresního lineárního modelu:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + E_i \qquad E_i \sim N(0, \sigma^2) \qquad Y_i \sim N\left(\beta_0 + \beta_1 \cdot X_i, \sigma^2\right)$$

- ***EY*** je <u>lineární funkcí</u> hodnot X_i. Nesplnění tohoto předpokladu znamená, že buď závislost není čistě lineární nebo ***EY*** závisí ještě na další proměnné, např. ***V***.

  Výrazně zakřivené vztahy vidím většinou hned na bodovém grafu. Potom mohu zvolit např. kvadratickou regresi (viz příklad „kořeny“) či proměnné transformovat (příklad „mozky“). Odhalení druhého případu je složitější, zvlášť když nemám další proměnné k dispozici. Popisuje ho příklad „tuk“.
- Špatně zvolený model <u>dává vychýlený odhad</u> středních hodnot ***EY***. Projevilo by se to například v používání modelu v praxi, kdy by předpovídané průměry a naměřené průměry byly systematicky vzájemně posunuté, vychýlené.
- Předpokládaný lineární vztah dobře funguje, když ***X*** i ***Y***, respektive jejich reziduály, mají normální rozdělení. Pokud normalita chybí, pomůžeme si transformací. Normalita ***X*** a ***Y*** ale není předpokladem regresního modelu.

Testy regresních koeficientů, prokazování závislosti *Y* na *X*

$$Y_i = \beta_0 + \beta_1 \cdot X_i + E_i \qquad E_i \sim N(0, \sigma^2)$$

- Modelujeme závislost ***EY*** na ***X*** jako $EY = \beta_0 + \beta_1 \cdot X$
- Hodnotu β_0 testujeme zřídka, protože hypotéza většinou nemá biologicky rozumnou interpretaci.
- Nezávislost ***EY*** na ***X*** znamená, že $\beta_1 = 0$.
- Hypotézu $H_0\colon\ \beta_1 = 0$ testujeme pomocí statistiky

$$T = \frac{b_1 - 0}{S.E.(b_1)} \sim_{H_0} t_{n-2}$$

Toto je jeden z hlavních výsledků regresní analýzy. Pokud *p-hodnota* < *α*, zamítám hypotézu o nezávislosti, tedy závislost ***Y*** na ***X*** je průkazná.

Např.: `> summary(lm(syn~otec))`

→ syn = 88.02 + 0.50*otec

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 88.01687   11.49887   7.654 1.36e-12 ***
otec         0.50096    0.06548   7.651 1.39e-12 ***
```

Koeficient determinace R^2

$$\boldsymbol{R^2} = \frac{\boldsymbol{SS_{REG}}}{\boldsymbol{SS_{TOT}}} = 1 - \frac{SS_E}{SS_{TOT}}$$

- $R^2 \in \langle 0,1 \rangle$
- Interpretujeme jako podíl vysvětlené variability vzhledem k celkové variabilitě v datech *Y*
- Bezrozměrný koeficient, často vyjádřený v procentech
- Koeficient ukazuje, jestli má model smysl, jestli vysvětlí nějaký podstatný díl variability.
- Pro lineární regresi platí $R^2 = r_{XY}^2$ (Pearsonův korelační koeficient ^2)

Poznámka: $\boldsymbol{R^2}$ se může velmi měnit s množinou zahrnutých pozorování. Odlehlé pozorování může hodnotu $\boldsymbol{R^2}$ i zdvojnásobit prostě proto, že má velký reziduální čtverec, kterým zvětší jak reziduální průměrný čtverec, tak regresní (modelový) průměrný čtverec. Naše radost nad množstvím vysvětlené variability pak může být vratká a krátká …

Test celého modelu jednoduché lineární regrese

Např.: `> anova(lm(syn~otec))`

```
Analysis of Variance Table
Response: syn
           Df Sum Sq Mean Sq F value    Pr(>F)
otec        1 1800.7 1800.69  58.532 1.392e-12 ***
Residuals 171 5260.7   30.76
```

Efekt	Analýza rozptylu (Galton v CV_10_data_KoreRegre) Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	1800,689	1	1800,689	58,53171	0,000000
Rezid.	5260,702	171	30,764		
Celk.	7061,391				

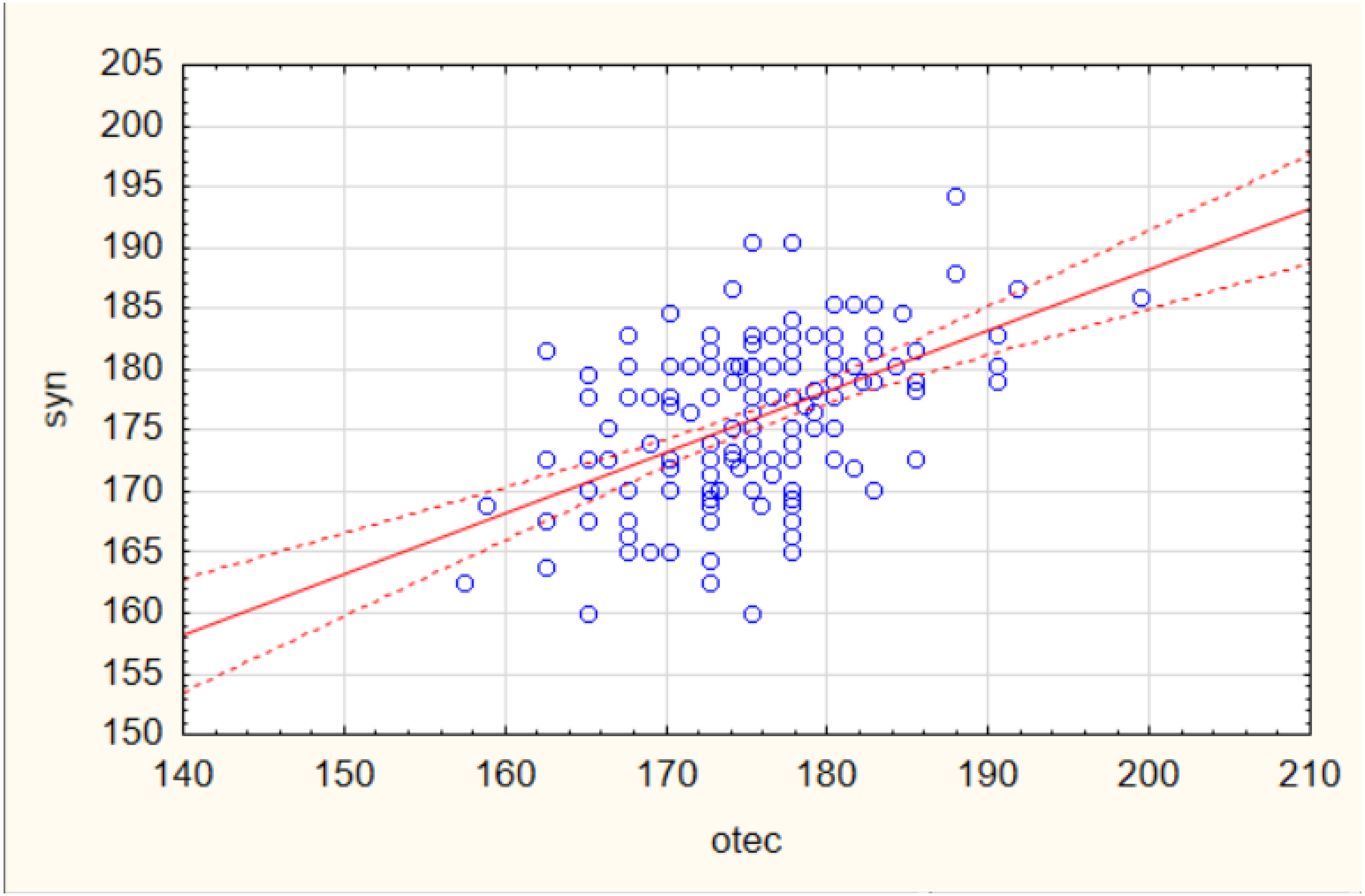
H_0: model vysvětlí jen nevýznamný díl variability

$$F = \frac{\frac{SS_{REG}}{1}}{\frac{SS_E}{n-2}} \sim F_{1,n-2}$$

Porovnáváme s kvantilem $F^{-1}_{1,n-2}\left(1-\frac{\alpha}{2}\right)$, zde = 5.11

- Tabulka analýzy rozptylu: porovnávám variabilitu vysvětlenou pomocí proměnné ***X*** (výška otce) s variabilitou reziduální, která zbyde po aplikaci modelu.
- F-statistika vypovídá o významnosti té části variability ***Y***, kterou lze vysvětlit modelem (STAT) nebo přidáním další vysvětlující proměnné (R, rozdíl později).
- V případě jednoduché lineární regrese s jednou nezávislou proměnnou je *p*-hodnota F-testu analýzy rozptylu shodná s *p*-hodnotou *t*-testu nenulovosti koeficientu b_1. To je proto, že v tomto nejjednodušším případě platí $F = T^2 \sim F_{1,\,n-2}$

Konfidenční interval pro celou regresní přímku



Model lineární regrese a příčinná závislost

Ideálně ***Y*** logicky závisí na ***X***.

Je-li vztah závislosti nejasný a obě proměnné jsou zatíženy náhodnou chybou, studujeme spíše korelaci proměnných.

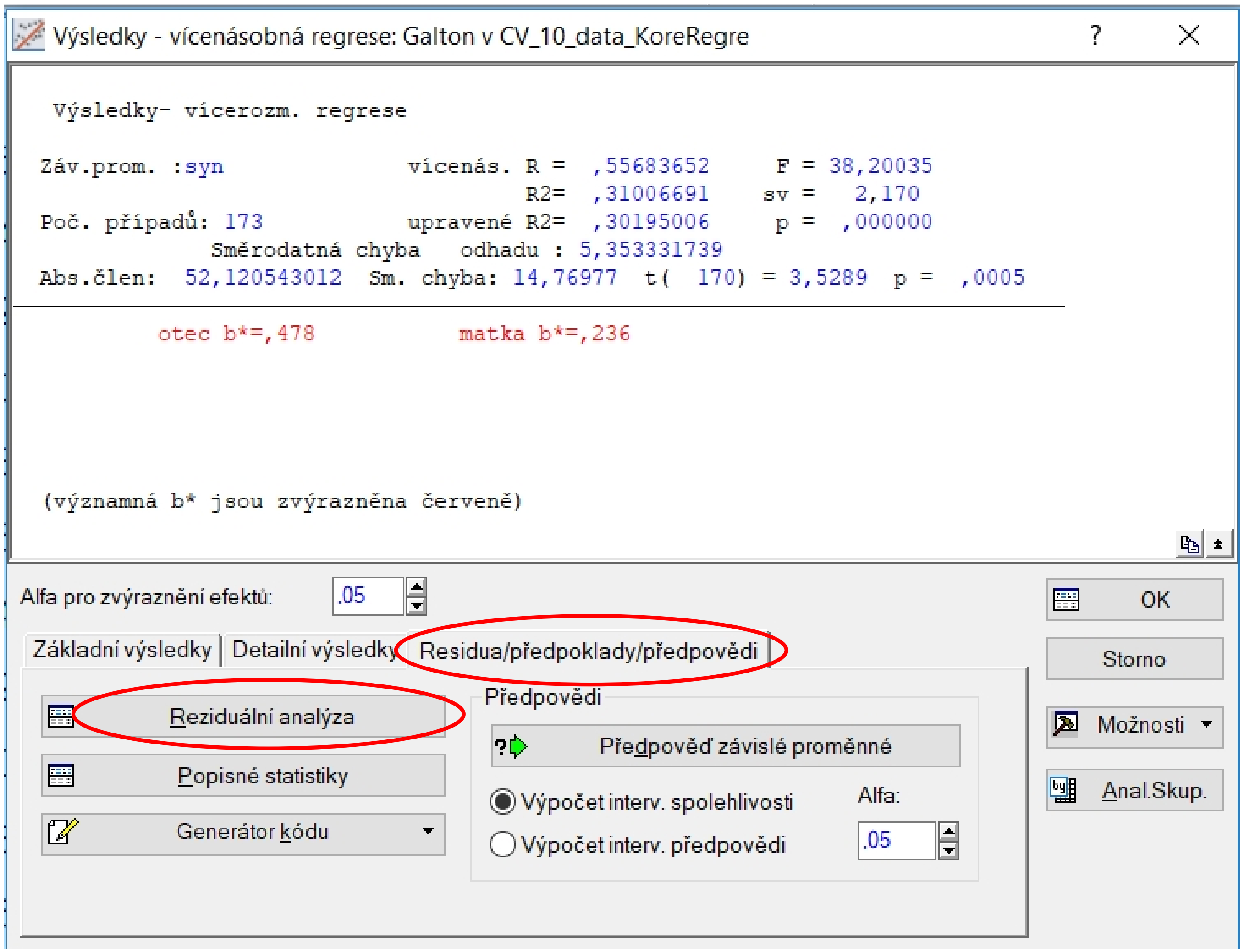
V praxi používáme regresi i ve sporných případech, kdy kauzální vztah není jasný. Přesto nás zajímá rovnice, která vztah obou proměnných (v daném uspořádání) popisuje. Mluvíme pak spíše o vysvětlované a vysvětlující proměnné a signifikantní model považujeme jen za nepřímý „důkaz“ příčinné závislosti ***Y*** na ***X***.

Statistickými prostředky nelze dokazovat příčinné závislosti (kauzalitu)! To umíme dělat jen manipulativními experimenty, kdy jsme schopni měnit hodnoty jen jedné proměnné, zatímco ostatní uvažované proměnné udržujeme na stálé úrovni.

Interpretace i predikce modelu je založena především na zkoumaném rozsahu hodnot vysvětlující proměnné. Se změnou rozsahu často narazíme na nelinearitu (v přírodě spíše běžnou) a náš model přestává platit.

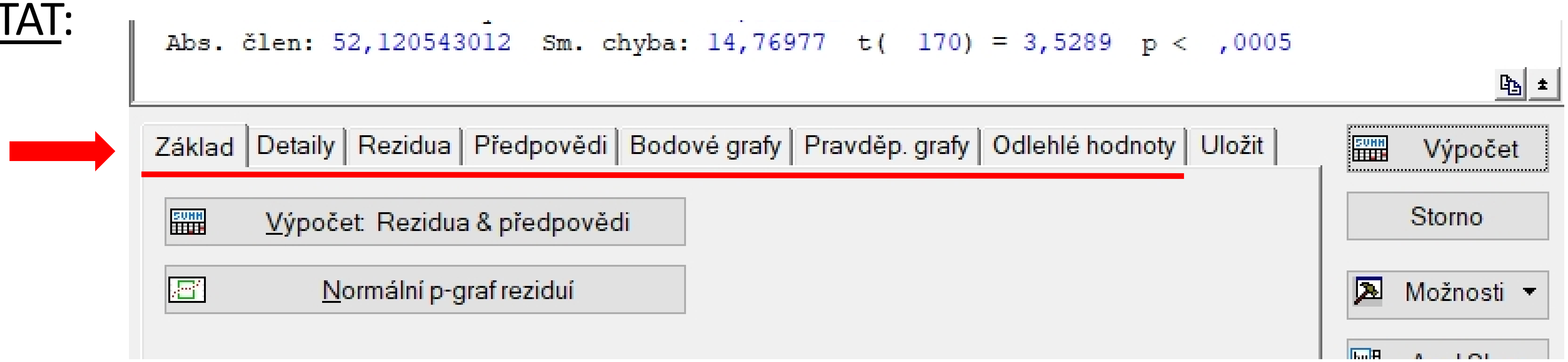
Ověřování předpokladů – regresní diagnostika [regression diagnostics]

STAT:



Ověřování předpokladů – diagnostické grafy

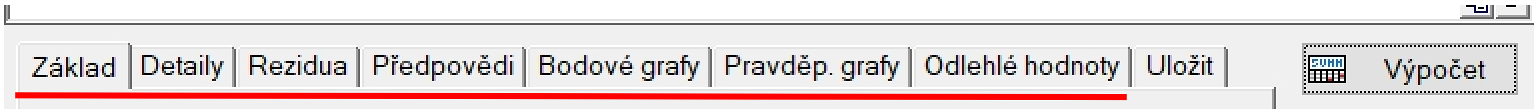
STAT:



- **Histogram reziduí (normalita):** záložka *Rezidua*
- **Q-Q plot reziduí (normalita):** *Základ* nebo *Pravděpodobnostní grafy*
- **Rezidua vs. Předpovědi (stejnost rozptylu):** *Bodové grafy*. Tento graf má odhalit závislost rozptylu $\boldsymbol{\sigma^2}$ na (předpovídané) střední hodnotě $\boldsymbol{Y}$. Správně mají být body rozložené stejnoměrně podle vodorovné osy.
- **Rezidua vs. Nezávislé proměnné (stejnost rozptylu):** *Rezidua*. V této kombinaci zkoumáme případnou závislost rozptylu $\boldsymbol{\sigma^2}$ na jednotlivých vysvětlujících proměnných. Správně jsou body rozložené stejnoměrně podle vodorovné osy.

Ověřování předpokladů – diagnostické grafy

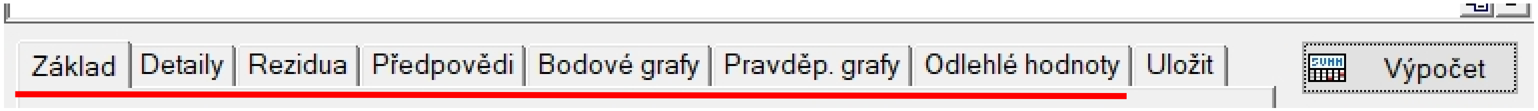
STAT:



- **Korelace mezi po sobě jdoucími reziduály** [autocorrelation]**(nezávislost mezi Y_i):** *Detaily. Durbin-Watsonova statistika*. Výsledek nazvaný *Sériové korelace* udává korelaci bodů daných souřadnicemi [U_i, U_{i+1}]. Vychází z úvahy, že závislá pozorování Y_i a Y_{i+1} budou mít podobnou odchylku od průměru. Například sourozenci budou mít podobně vychýlenou výšku. Nebo dotazník vyplněný stejným člověkem bude mít podobné odpovědi. To může fungovat za předpokladu, že pozorování v tabulce jsou zapsána tak, jak byla získána v „terénu“. Pokud by takto závislých pozorování bylo v datech hodně, byla by *Sériová korelace* výrazně odlišná od nuly.

Ověřování předpokladů – diagnostické grafy a statistiky

STAT:



- **Mahalanobisova vzdálenost (odlehlá pozorování [outlier]):** *Odlehlé hodnoty*. *Typ odlehlých hodnot*: zvolit. Počítá prostorovou vzdálenost pozorování od centroidu (těžiště) vysvětlujících proměnných, upravenou pro korelované proměné. Výstupní tabulka je uspořádána od největších odchylek po nejmenší, takže potenciálně problematická, odlehlá pozorování jsou na prvních řádcích.
- **Cookova vzdálenost (příliš vlivná pozorování [leverage case]):** *Odlehlé hodnoty*. *Typ odlehlých hodnot*: zvolit. Pro každé pozorování spočte rozdíl v odhadu regresních koeficientů v modelu ***s*** a ***bez*** daného řádku (pozorování). Pokud je rozdíl velký, je jasné, že dané pozorování podstatně ovlivňuje směr regresní přímky, tedy celého modelu. Opět, nejvlivnější pozorování jsou na prvních řádcích.

Ověřování předpokladů – regresní diagnostika

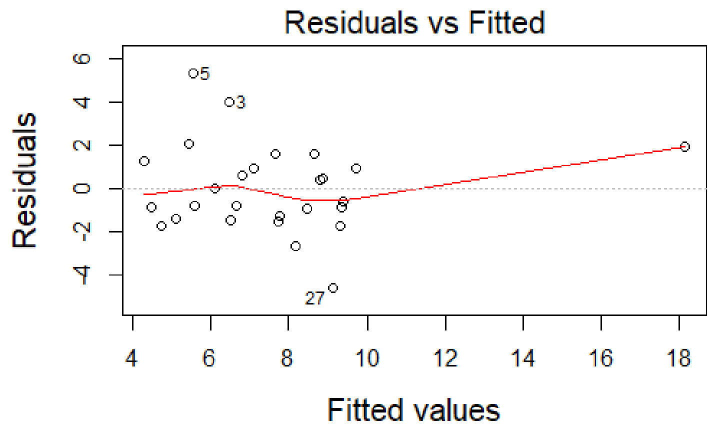
R: `model <- lm(Y~X)`

`model$reziduals` … s tímto vektorem pak tvořím histogramy a Q-Q diagramy

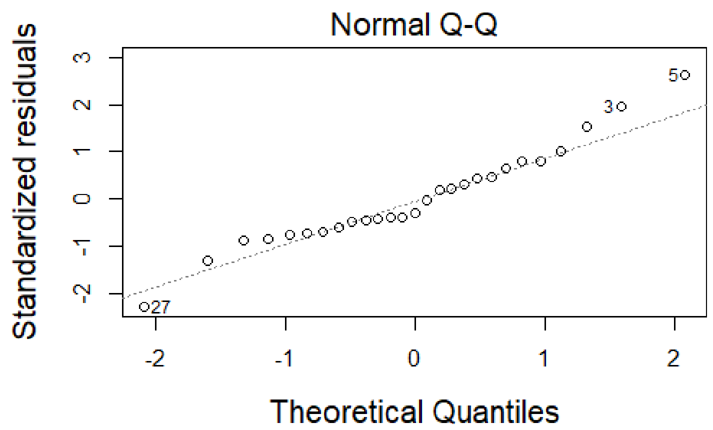
`plot(model)` … 6 předchystaných grafů, předvolba tiskne 1.,2.,3. a 5. graf.

Rezidua vs. Předpovědi (stejnost rozptylu):

Tento graf má odhalit závislost rozptylu σ^2 na (předpovídané) střední hodnotě Y. Správně mají být body rozložené stejnoměrně podle vodorovné osy.



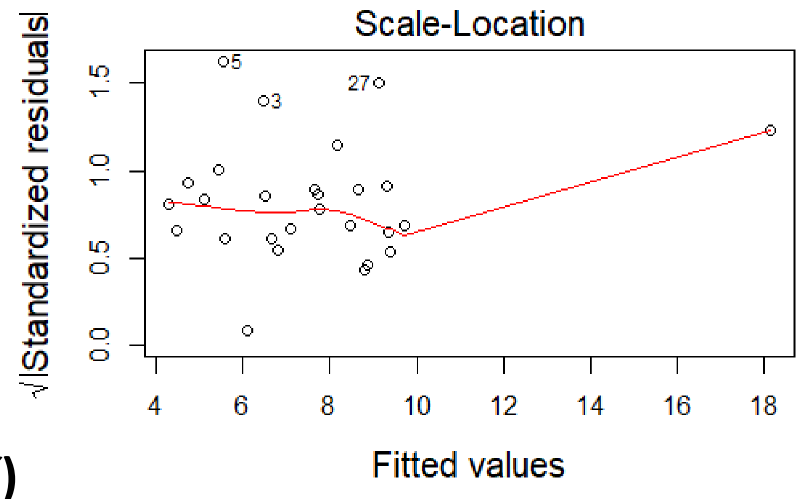
Q-Q plot reziduí (normalita):



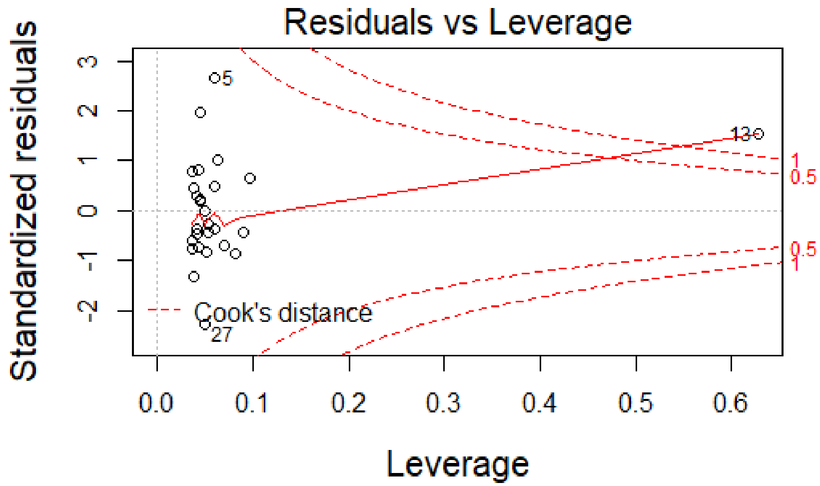
Ověřování předpokladů – regresní diagnostika

R: `plot(model)`

Odmocněná Rezidua vs. Předpovědi (stejnost rozptylu, normalita).
Při porušení předpokladu vykazují body
Nějaký druh závislosti (lineární či nelineární).



Cookova vzdálenost (příliš vlivná pozorování)
Pro každé pozorování spočte rozdíl v odhadu regresních koeficientů v modelu ***s*** a ***bez*** daného řádku (pozorování). Pokud je rozdíl velký, je jasné, že dané pozorování podstatně ovlivňuje směr regresní přímky, tedy celého modelu.
[lever = páka; leverage = vliv páky, páčení]



Ověřování předpokladů – regresní diagnostika

R: **Rezidua vs. Nezávislé proměnné:** ilustruje případnou závislost rozptylu na hodnotách vysvětlujících, nezávislých proměnných.
V Rku možnost **Breuschova-Paganova testu** (knihovna **`lmtest`**)

```
> bptest(model, varformula=~ …, data = …)
```

Mnohonásobná lineární regrese [multiple linear regression]

Poznámka: Něco jiného je mnohorozměrná regrese [multidimensional regression], ve které modeluji více závislých proměnných pomocí více nezávislých proměnných.

Model: $\boldsymbol{Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot V_i + \beta_3 \cdot W_i + E_i}$ $\quad \mathbf{k = 3}$, $\boldsymbol{E_i \sim N(0, \sigma^2)}$

Jiný zápis: $\boldsymbol{Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3 \cdot X_{3i} + E_i}$

Hodnoty vysvětlujících proměnných se pak dají zapsat jako matice:

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & X_{21} & X_{31} \\ X_{12} & X_{22} & X_{32} \\ \vdots & \vdots & \vdots \\ X_{1n} & X_{2n} & X_{3n} \end{pmatrix}$$

Odhad rozptylu $\boldsymbol{\sigma^2}$: $S^2 = \frac{SS_E}{n-k-1}$

Počet pozorování – počet regresorů – 1

Výsledky: odhady regresních koeficientů b_0, b_1, b_2, b_3; R^2; F – test modelu

Interpretace $\boldsymbol{b_j}$: o kolik vzroste (klesne) hodnota $\boldsymbol{Y}$, když $\boldsymbol{X_j}$ vzroste o jednotku a ostatní vysvětlující proměnné se nezmění.

Mnohonásobná lineární regrese

Model: $\boldsymbol{Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3 \cdot X_{3i} + E_i} \quad \mathbf{k = 3}, \quad \boldsymbol{E_i \sim N(0, \sigma^2)}$

Hodnocení regresních koeficientů:

Hypotéza H_0: $\boldsymbol{b_j = 0}$ → $T = \frac{\boldsymbol{b_j} - 0}{\boldsymbol{S.E.(b_j)}} \sim_{\boldsymbol{H_0}} \boldsymbol{t_{n-k-1}}$ → znamená to, že proměnná $\boldsymbol{X_j}$ nepřidá do modelu novou informaci o střední hodnotě ***Y***, nic významně nového nevysvětlí.

Konfidenční interval $\boldsymbol{\beta_j}$:

$$\left(\boldsymbol{b_j} - S.E.(\boldsymbol{b_j}) \cdot \boldsymbol{t_{n-k-1}}(1 - {}^{\alpha}/_{2}), \boldsymbol{b_j} + S.E.(\boldsymbol{b_j}) \cdot \boldsymbol{t_{n-k-1}}(1 - {}^{\alpha}/_{2})\right)$$

Porovnání vlivu regresorů na ***Y*** mezi sebou

→ přepočítám na standardizovaný tvar: $b_j^* = b_j \cdot \frac{sd(X_j)}{sd(Y)}$

Příklad: % tuku ~ výška + váha. $b_{VYSKA}^* = -0.254, b_{VAHA}^* = 0.968$

Mohu říci, že váha má zhruba 4-krát větší vliv na výsledné % tuku než výška.

Příklady