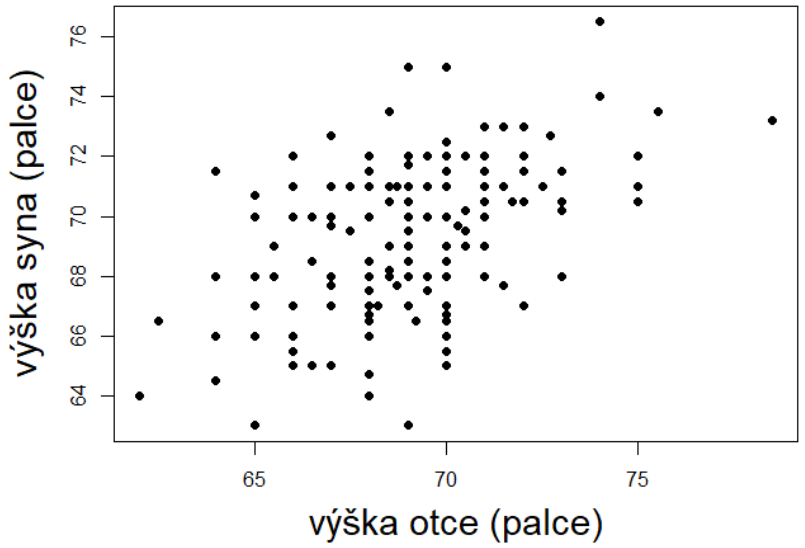
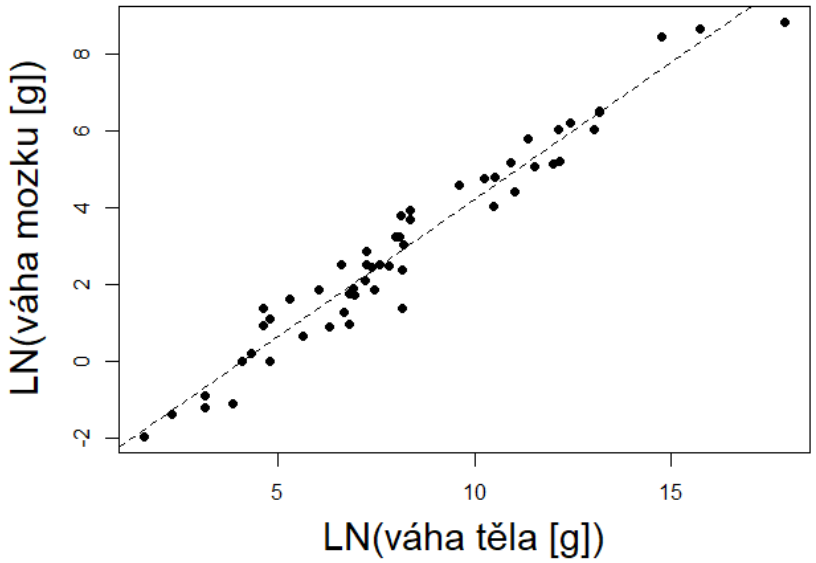
Příkladová data

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

Váha těla [g]	váha mozku [g]
10	0.25
60	1
101	4
200	5
1000	6.6
14830	98.2





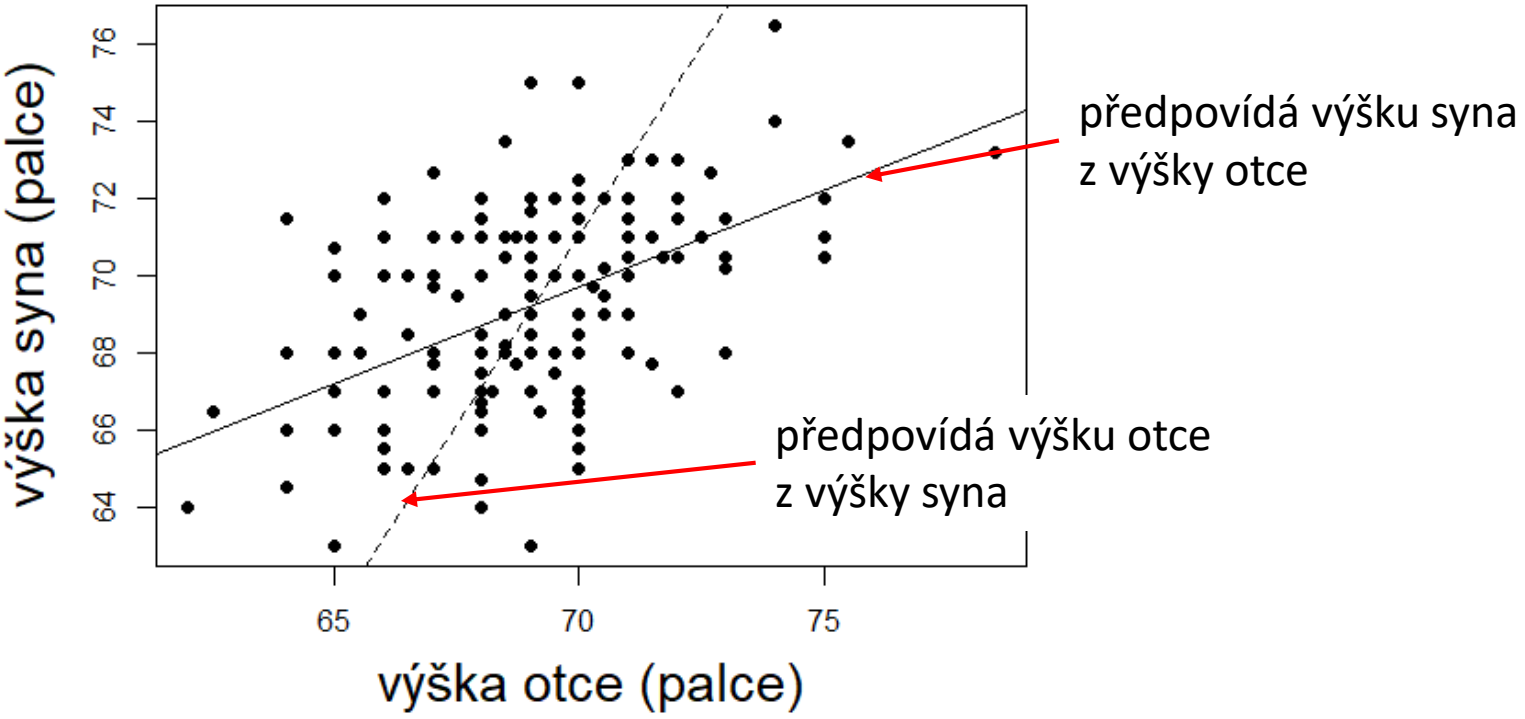
Analýza vztahu dvou kvantitativních proměnných

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

KORELACE popisuje sílu vzájemné závislosti.

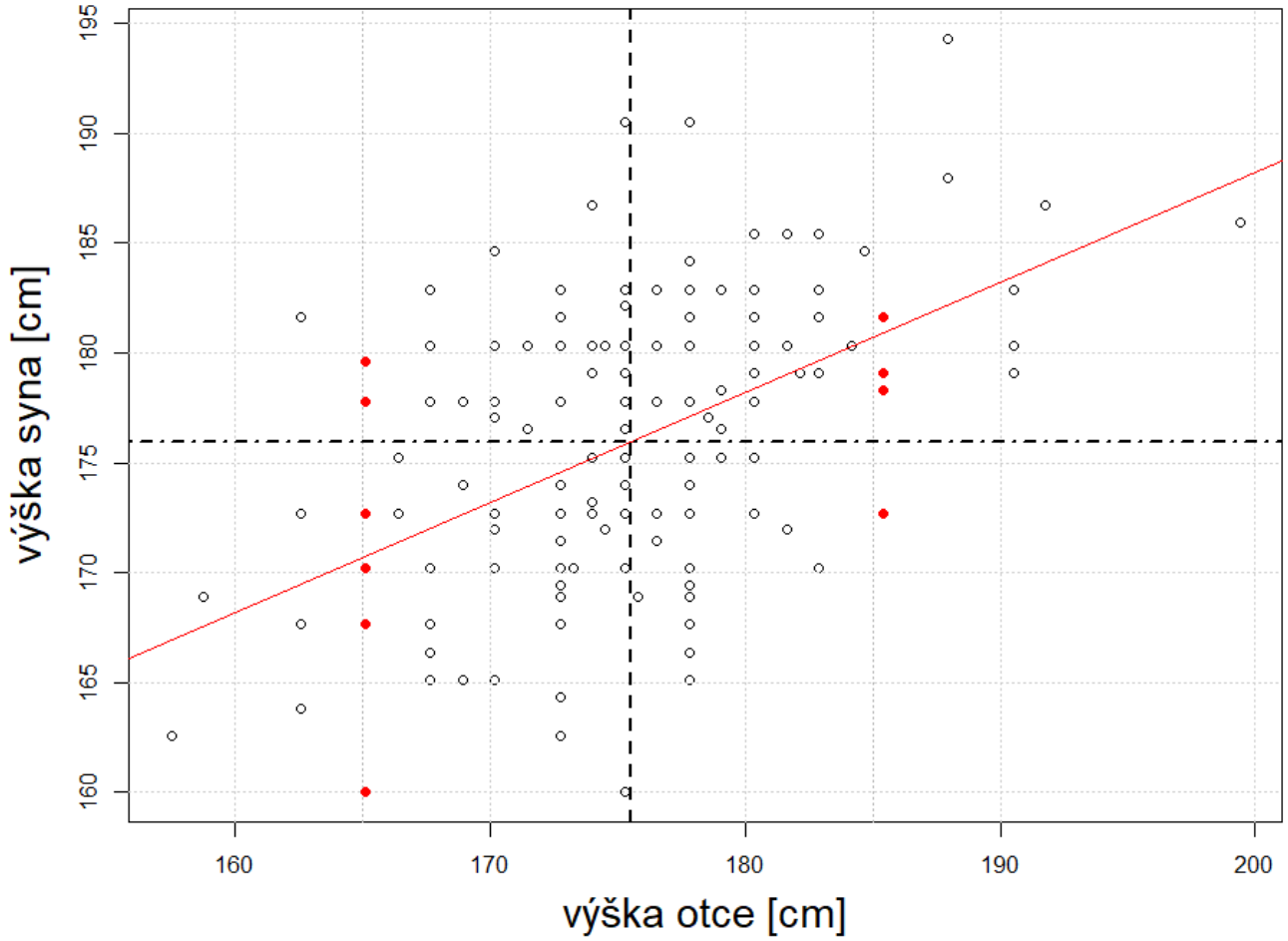
REGRESE pomocí jedné proměnné popisuje hodnoty druhé proměnné

Příklad: výšky otce a syna (data GaltonSyn)



Regrese – původ názvu

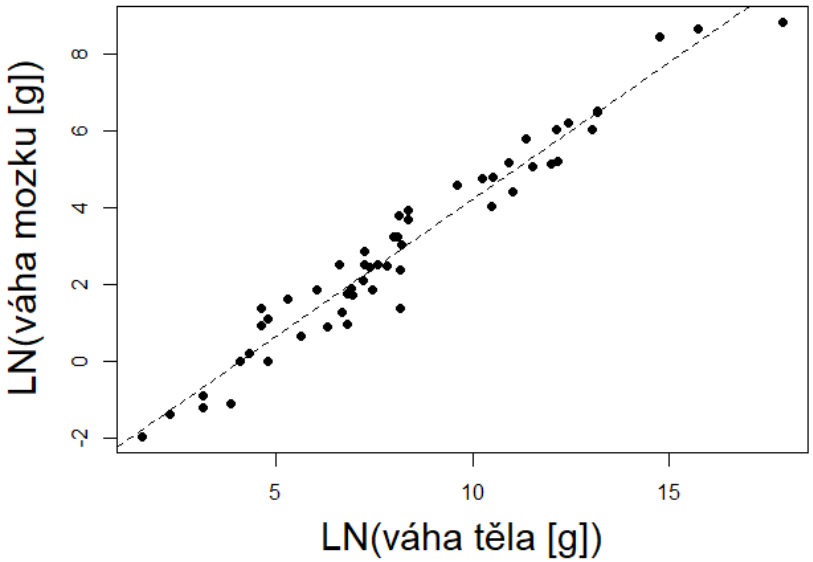
Sir F. Galton (1886): dědičnost výšky postavy

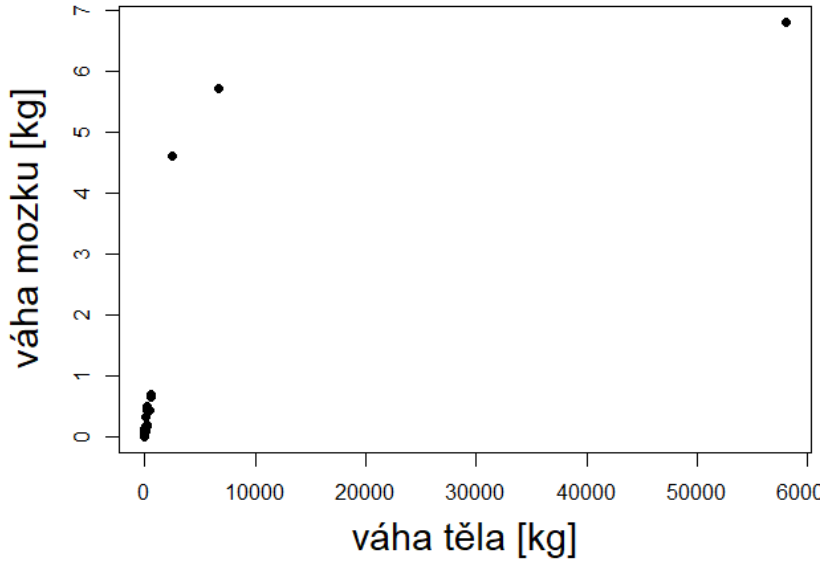


Regrese – vysvětlení variability Y pomocí X

- 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]

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

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]
- $E_i \sim N(0, \sigma^2)$ náhodná chyba, přirozená variabilita [error term, disturbance term]
- β_0 a β_1 jsou parametry platné pro celou populaci, tedy neznámé
  → hledáme odhady b_0 a b_1 a testujeme jejich nenulovost
- Parametry β_0 a β_1 určují přímku závislosti:
  β_0 je průsečík s osou y [intercept],
  když $X = 0$, potom $Y = \beta_0$
  β_1 je sklon přímky [slope];
  když X zvětším o 1 jednotku
  potom Y naroste (v průměru) o β_1.

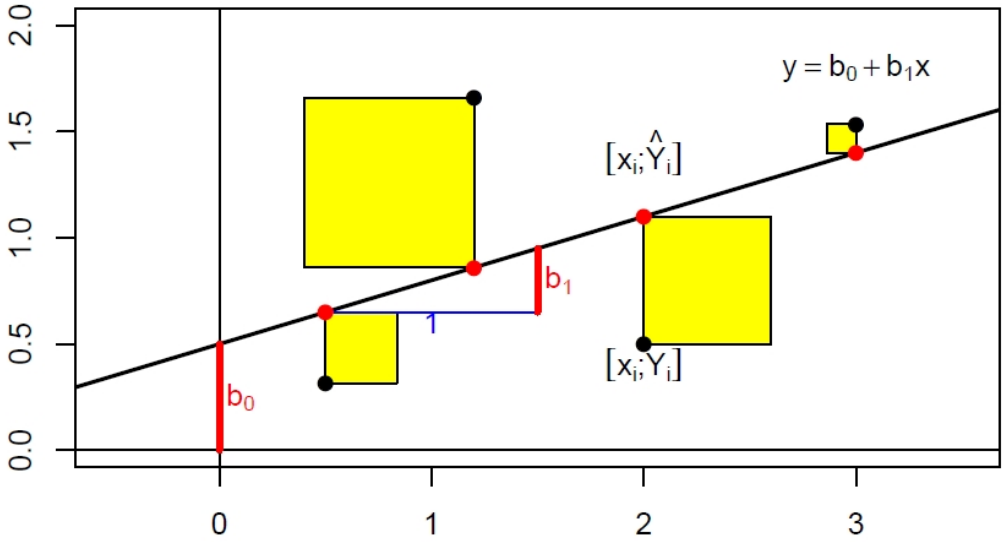
Odhad regresních koeficientů: β_0, β_1, σ^2

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

1) Odhady b_0 a b_1 hledáme **metodou nejmenších čtverců**
[method of the least squares]

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

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



→ 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: σ^2

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

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 „Y-ových“ datech (vysvětlovaná proměnná)

$DF_{TOT} = n - 1$

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

$DF_{REG} = k$

k … počet vysvětlujících proměnných

$SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ … reziduální variabilita, variabilita modelem nevysvětlená

$DF_E = n - k - 1$

Platí: $SS_{TOT} = SS_{REG} + SS_E$

Regresní model – příklad Galton:

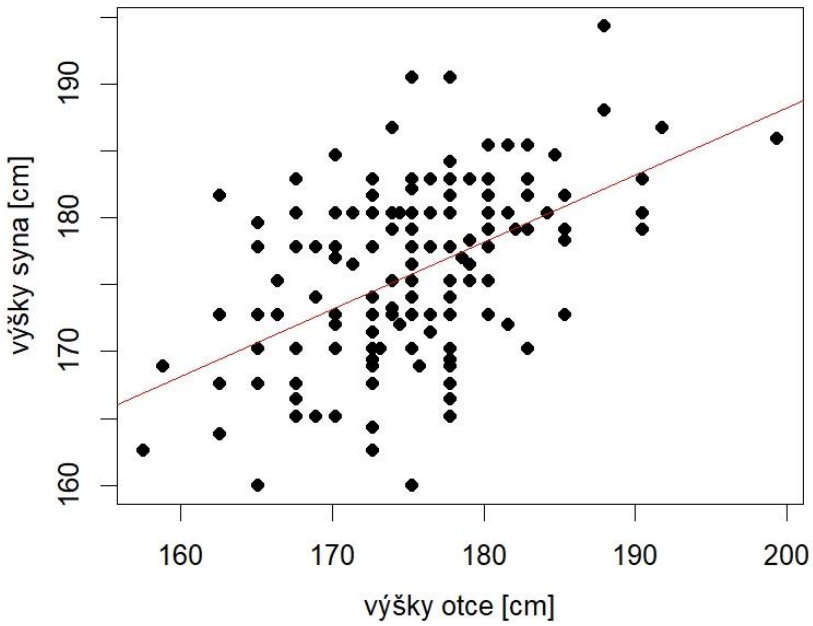
```
> summary(lm(syn~otec))

Call:
lm(formula = syn ~ otec)

Residuals:
     Min       1Q   Median       3Q
-15.8326  -3.4311  -0.5134   3.9341  1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 87.85290   11.49732   7.641 1.47e-12 ***
otec         0.50188    0.06547   7.666 1.28e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.545 on 171 degrees of freedom
Multiple R-squared:  0.2558,       Adjusted R-squared:  0.2514
F-statistic: 58.76 on 1 and 171 DF,  p-value: 1.275e-12
```



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 $(Y_i - \hat{Y}_i)$. 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:

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

- ***EY*** je lineární funkcí 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“ ve Zvárovi) či proměnné transformovat (příklad „mozky“). Odhalení druhého případu je složitější, zvlášť když nemám další proměnné k dispozici. Popisuje ho příklad „tuk“ (Zvára).
- Špatně zvolený model dává vychýlený odhad 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.

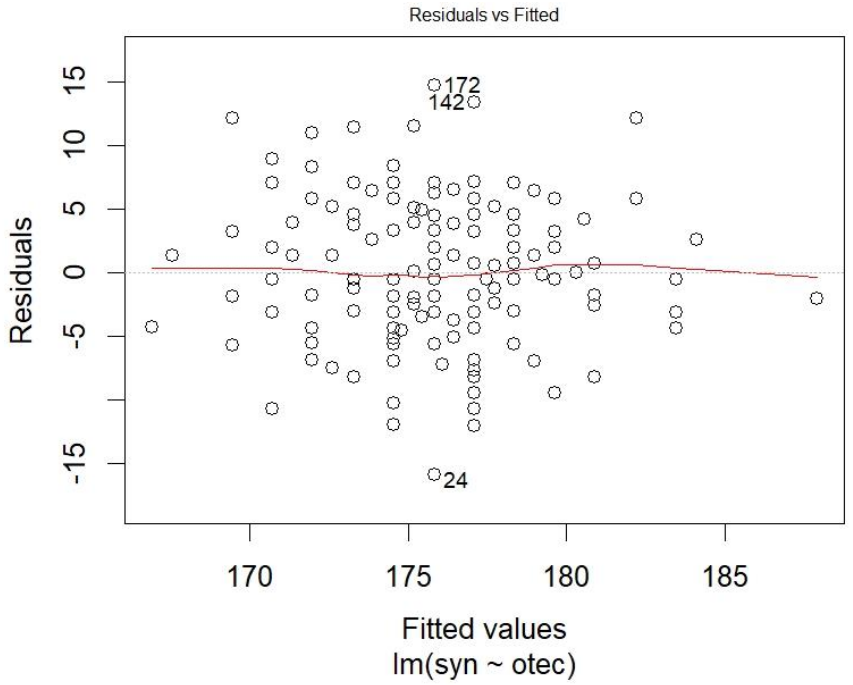
Předpoklady – příklad Galton:

```
> plot(lm(syn~otec))  # zadání formulí
> plot(gal.lm1)       # zadání modelem

Hit <Return> to see next plot:
```

Musím stisknout ENTER, potom naskočí graf.

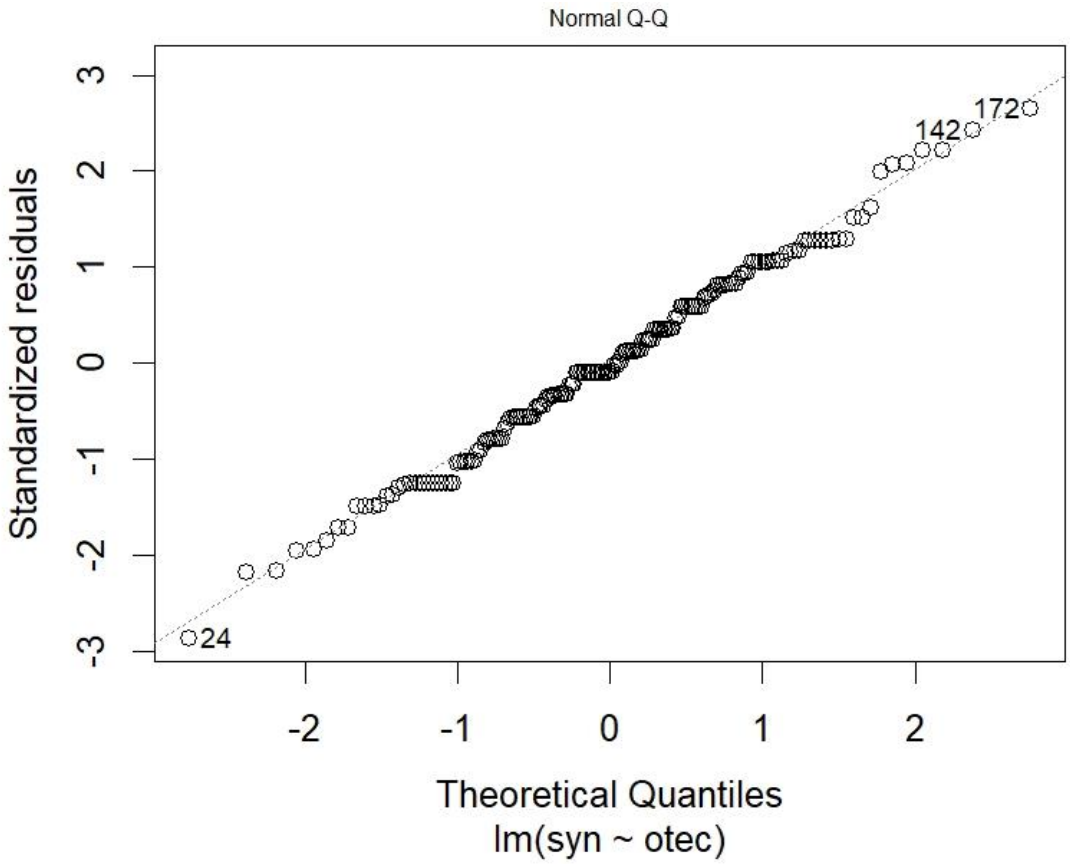
Rezidua vs. Předpovědi (stejnost rozptylu):
Tento graf má odhalit závislost rozptylu σ^2 na (předpovídané) střední hodnotě Y.
Správně mají být body rozložené stejnoměrně podle vodorovné osy.



Předpoklady – příklad Galton:

R: `plot(model)`

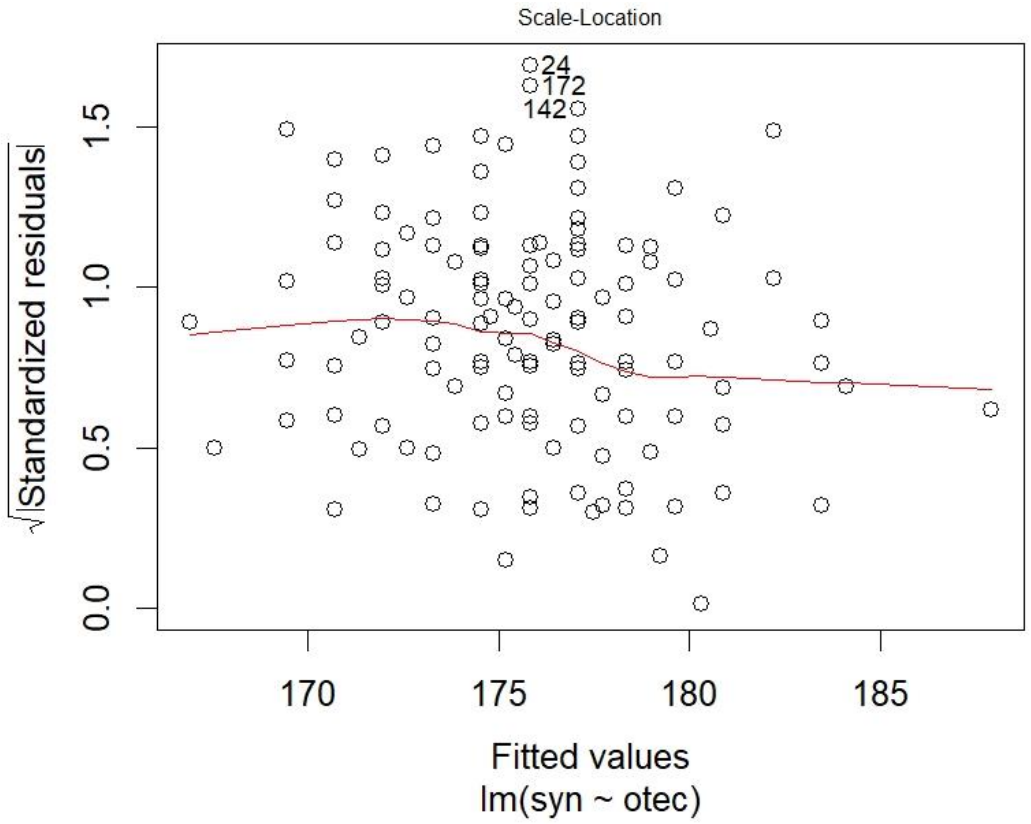
Q-Q plot reziduí (normální rozdělení reziduí E_i):



Předpoklady – příklad Galton:

R: `plot(model)`

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



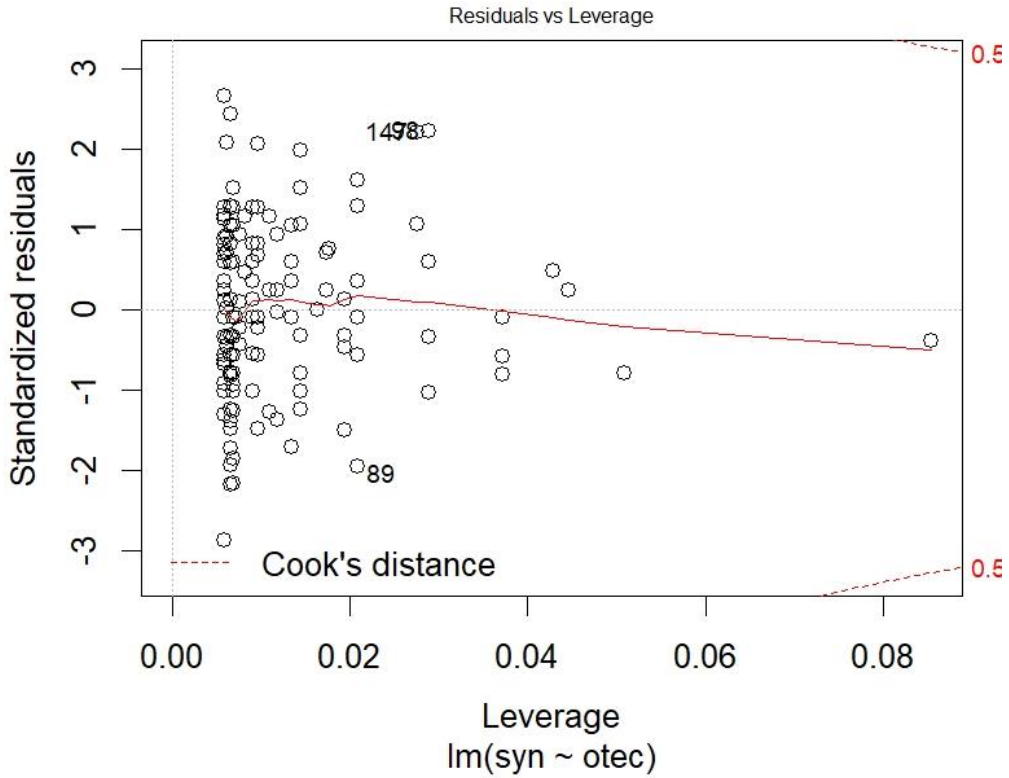
Předpoklady – příklad Galton:

R: `plot(model)`

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]



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

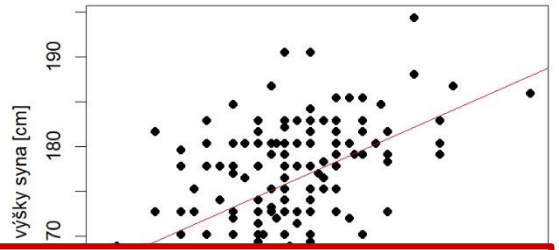
$$Y_i = \beta_0 + \beta_1 \cdot X_i + E_i \qquad E_i \sim N(0, \sigma^2)$$

- Modelujeme závislost ***EY*** na ***X*** jako $EY = \beta_0 + \beta_1 \cdot X$
- Hodnotu β_0 (průsečík s osou Y) testujeme zřídka, protože hypotéza většinou nemá biologicky rozumnou interpretaci.
- Nezávislost ***EY*** na ***X*** znamená, že $\beta_1 = 0$.
- Hypotézu H_0: $\beta_1 = 0$ 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.

Regresní model – příklad Galton:




```
> summary(lm(syn~ot
```

S.E.(odhadu)
- rozptýlenost kolem skutečné hodnoty
- „přesnost“ odhadu

odhad b_0 (průsečík)
odhad b_1 * otec

test. statistika a p-hodnota
H_0: $b_1 = 0$ → výška otce nemá vliv
H_1: $b_1 \neq 0$ → výška otce má vliv

```
Coefficients:
             Estimate  Std. Error   t-value   Pr(>|t|)
(Intercept) 87.85290    11.49732     7.641   1.47e-12 ***
otec         0.50188     0.06547     7.666   1.28e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

β_0 (průsečík s Y)
β_1 * otec

Oba koeficienty průkazné
→ syn = 87.9 + 0.50*otec

Test podmodelu (jednoduché lineární) regrese

Nejjednodušší model je „průměr“: $Y_i = EY$ odhadujeme jako $Y_i = \overline{Y}$.

Máme tedy **model**, zde rostoucí či klesající přímku: $Y_i = \beta_0 + \beta_1 \cdot X_i + E_i$

a **podmodel**, zde vodorovnou přímku, průměr: $Y_i = EY + E_i$.

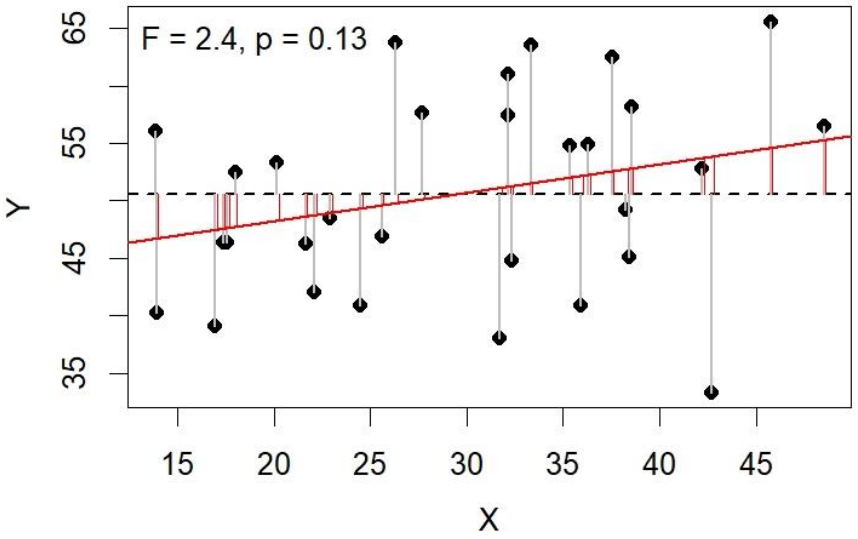
Variabilitu v datech $var(Y)$ odhadujeme jako $MS_{TOT} = \frac{\sum(Y_i - \overline{Y})^2}{n-1}$, tj. $\frac{SS_{TOT}}{df_{TOT}}$.

MS znamená: *mean square*, „průměrný čtverec“

Složitější model je dobrý tehdy, když vysvětlí nějaký podstatný díl této variability.

Složitější model je zbytečný tehdy, když množství „nově“ vysvětlené variability je nevýznamné, neprůkazně odlišné od nuly.

Slabý model:



Test podmodelu jednoduché lineární regrese

model: $\boldsymbol{Y_i = \beta_0 + \beta_1 \cdot X_i + E_i}$ *versus* podmodel: $\boldsymbol{Y_i = EY + E_i}$

Konstrukce testu vypadá takto:

H_0: Rozdíl ve variabilitě vysvětlené modelem a podmodelem je malý, tj. složitější model vysvětlí navíc jen nevýznamné množství variability Y.

H_1: Rozdíl ve variabilitě vysvětlené modelem a podmodelem je významný.

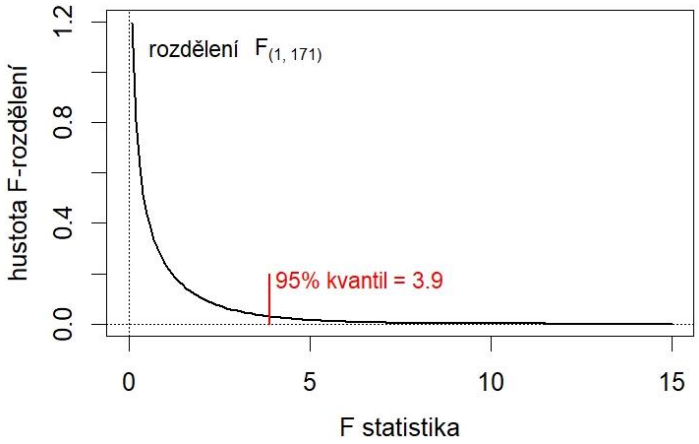
V jednoduchém lineárním modelu je rozdíl mezi modelem a podmodelem spočtený právě jako $\boldsymbol{SS_{REG} = \sum(\hat{Y}_i - \bar{Y})^2}$. Ve složitějších modelech je to složitější výpočet.

$\boldsymbol{SS_{REG}}$ porovnáváme se $\boldsymbol{SS_{Error}}$ modelu pomocí ***F*** statistiky.

Součty čtverců použijeme ve tvaru průměrných součtů čtverců, $\boldsymbol{MS_{REG}}$ a $\boldsymbol{MS_{Error}}$.

V tomto případě jsou $\boldsymbol{df_{REG} = 1}$ a $\boldsymbol{df_{Error} = n - 2}$.

$$\boldsymbol{F = \frac{MS_{REG}}{MS_{Error}} = \frac{\frac{SS_{REG}}{1}}{\frac{SS_{Error}}{n-2}} \sim_{H_0} F_{1,n-2}}$$



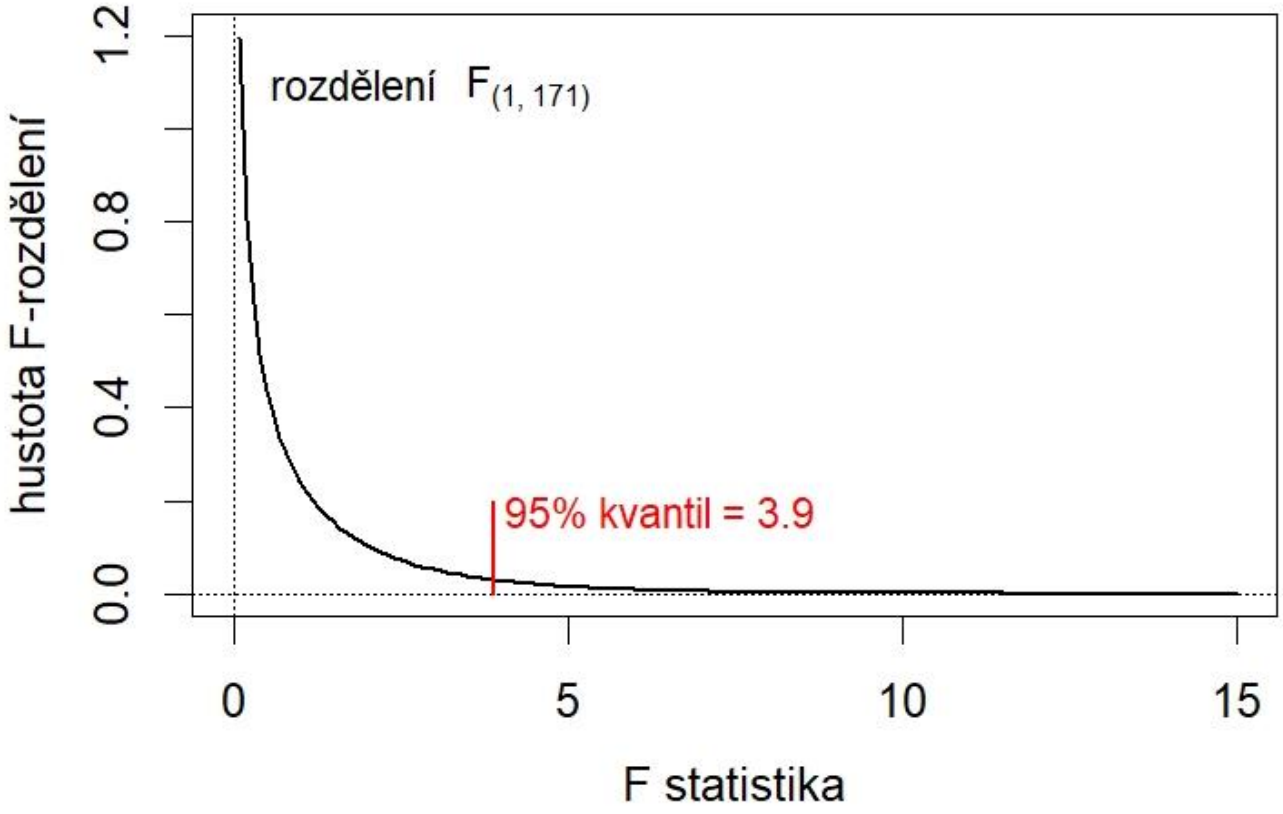
Test podmodelu jednoduché lineární regrese

***H_0**: Rozdíl ve variabilitě vysvětlené modelem a podmodelem je malý,*
tj. složitější model vysvětlí navíc jen nevýznamné množství variability Y.

***H_1**: Rozdíl ve variabilitě vysvětlené modelem a podmodelem je významný.*

$$F = \frac{MS_{REG}}{MS_{Error}} \sim_{H_0}$$

$$\sim F_{1,\,n-2}$$



Regresní model – příklad Galton:

```
> anova(lm(syn~otec))
Analysis of Variance Table
Response: syn
            Df  Sum Sq  Mean Sq  F value    Pr(>F)
otec         1  1807.1  1807.08   58.763 1.275e-12 ***
Residuals  171  5268.6    30.75
```

$$F = \frac{\frac{SS_{REG}}{df_{REG}}}{\frac{SS_{Error}}{df_{Error}}} = \frac{\frac{1807.1}{1}}{\frac{5268.6}{171}} = \frac{MS_{REG}}{MS_{Error}} = \frac{1807.08}{30.75} = 58.763$$

```
> summary(lm(syn~otec))
...
Residual standard error: 5.545 on 171 degrees of freedom
Multiple R-squared:  0.2558,    Adjusted R-squared:  0.2514
F-statistic: 58.76 on 1 and 171 DF,  p-value: 1.275e-12
```

Regresní model – příklad Galton:

$$F = \frac{\frac{SS_{REG}}{df_{REG}}}{\frac{SS_{Error}}{df_{Error}}} = \frac{MS_{REG}}{MS_{Error}}$$

```
> anova(lm(syn ~ otec + matka))
Analysis of Variance Table
Response: syn
           Df Sum Sq Mean Sq F value    Pr(>F)
otec        1 1807.1 1807.08  63.056 2.615e-13 ***
matka       1  386.7  386.70  13.493 0.0003205 ***
Residuals 170 4871.9   28.66
```

```
> anova(lm(syn ~ matka + otec))
Analysis of Variance Table
Response: syn
           Df Sum Sq Mean Sq F value    Pr(>F)
matka       1  599.7  599.71  20.926 9.167e-06 ***
otec        1 1594.1 1594.07  55.623 4.299e-12 ***
Residuals 170 4871.9   28.66
```

Na pořadí záleží!

Test podmodelu (jednoduché lineární) regrese

***H_0**: Rozdíl ve variabilitě vysvětlené modelem a podmodelem je malý, tj. složitější model vysvětlí navíc jen nevýznamné množství variability Y.*

***H_1**: Rozdíl ve variabilitě vysvětlené modelem a podmodelem je významný.*

$$F = \frac{MS_{REG}}{MS_{Error}} = \frac{\frac{SS_{REG}}{df_{REG}}}{\frac{SS_{Error}}{df_{Error}}} \sim_{H_0} F_{k,n-k-1}$$

Porovnáváme s kvantilem $F_{1,n-k-1}(1-\alpha)$

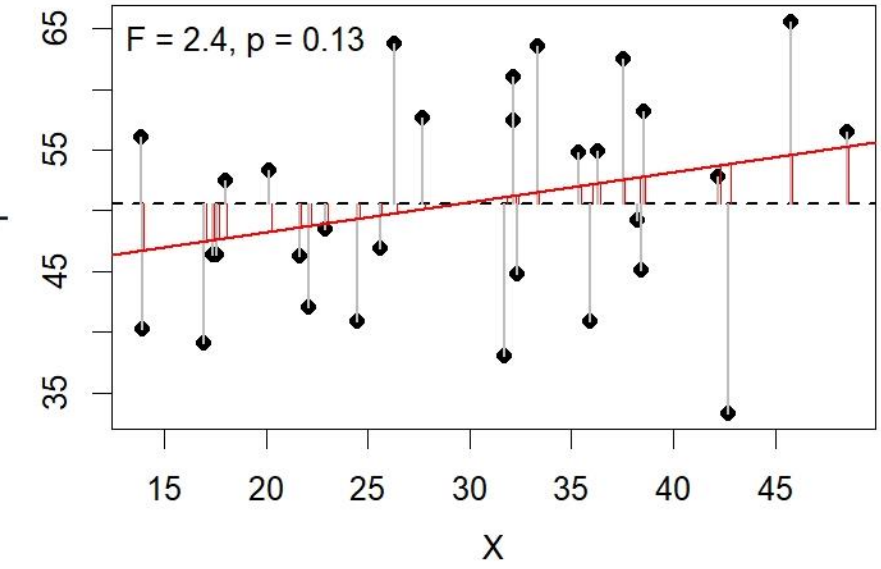
- F-statistika vypovídá o významnosti té části variability ***Y***, kterou lze vysvětlit přidáním (další) vysvětlující proměnné.
- V případě jednoduché lineární regrese s jednou nezávislou proměnnou je *p*-hodnota F-testu analýzy rozptylu shodná s *p*-hodnotou *t*-testu nenulovosti koeficientu b_1. To je proto, že v tomto nejjednodušším případě platí $F = T^2 \sim F_{1,\,n-2}$

Test podmodelu jednoduché lineární regrese

OPRAVA textu v učebnici Lepš & Šmilauer (2016), str. 257 dole.
Věta „*Lze ukázat, že pokud platí nulová hypotéza o nezávislosti, jsou všechny tři hodnoty **MS** odhadem variance **Y**.*“ není pravdivá.

Pokud jsou ***X*** a ***Y*** nezávislé, potom hodnoty ***X*** nedokáží vysvětlovat/předpovídat výsledné hodnoty ***Y*** a regresní přímka nemá statistický smysl. Graficky by taková přímka měla být velmi blízko obyčejné průměrné hodnotě ***Y***.

Potom rozdíly mezi napočítanou (regresní) hodnotou $\widehat{\boldsymbol{Y}}_{\boldsymbol{i}}$ a průměrem $\overline{\boldsymbol{Y}}$ budou velmi malé. Proto také součet čtverců $\boldsymbol{SS_{REG}} = \sum\left(\widehat{\boldsymbol{Y}}_{\boldsymbol{i}} - \overline{\boldsymbol{Y}}\right)^{\boldsymbol{2}}$ bude spíše malé číslo blízko nuly. Tento součet čtverců ale nepopisuje variabilitu v datech ***Y***.



Ovšem $\boldsymbol{SS_{Error}} = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$ slabého modelu může popisovat variabilitu ***Y***.

Koeficient determinace R^2

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

Součty čtverců ≈ rozklad variability

$$\boldsymbol{SS_{TOT}} = \sum (\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})^2$$

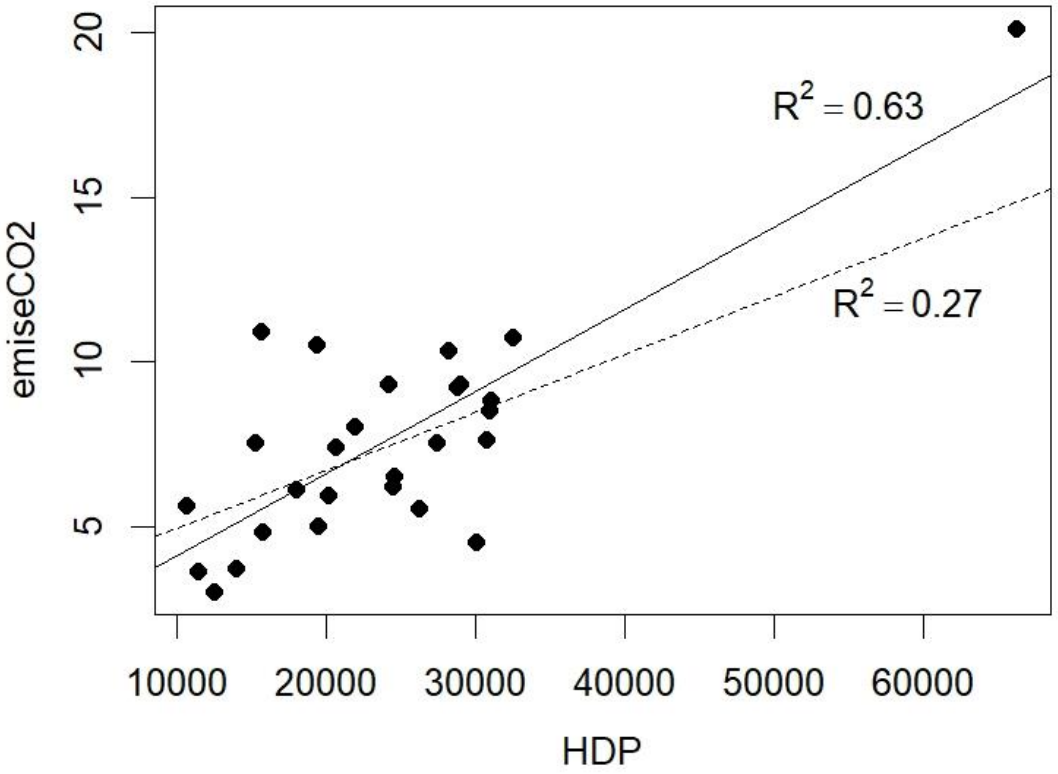
$$\boldsymbol{SS_{REG}} = \sum (\widehat{\boldsymbol{Y}}_i - \overline{\boldsymbol{Y}})^2$$

$$\boldsymbol{SS_{Error}} = \sum (\boldsymbol{Y_i} - \widehat{\boldsymbol{Y}}_i)^2$$

- $R^2 \in \langle 0,1 \rangle$
- Interpretujeme jako podíl vysvětlené variability **(REG)** vzhledem k celkové variabilitě v datech *Y* **(TOT)**
- 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)
- **Adjustovaný** (upravený, korigovaný) $\boldsymbol{R^2}$: když mám více vysvětlujících proměnných ($Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot V_i + \beta_3 \cdot W_i + E_i$), tak je obyčejný R^2 nadhodnocený, vychýlený. Proto se používá tato úprava.

Koeficient determinace R^2

Poznámka: R^2 se může velmi měnit s množinou zahrnutých pozorování. Odlehlé pozorování může hodnotu 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 …



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.

Mnohonásobná lineární regrese [multiple linear regression]

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

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

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

Konfidenční interval $\boldsymbol{\beta_j}$:

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

Porovnání vlivu regresorů na ***Y*** mezi sebou (*viz Zvára, str. 199 dole a str. 200*):

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

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

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

Regrese – další čtení:

- **Standardizované beta koeficienty**: *Zvára, str. 199 dole a str. 200.*
- **Transformace dat v regresi, kontrola předpokladů**:
  - Lepš & Šmilauer, str.259
  - Zvára, str. 203: Breuschův-Paganův test
- **Regresní analýza v R**: Lepš & Šmilauer, str.274
- **Vícenásobná regrese**: Zvára, str. 197
  Lepš & Šmilauer, str. 294.
- **Zobecněné lineární modely**: Lepš & Šmilauer, str. 316
- **Nelineární závislost**: Lepš & Šmilauer, str. 338