01RAD

doc. Ing. Tomáš Hobza, Ph.D., Martin Kovanda, Máša Mašková, Filip Bár

24. září 2020

Obsah

Předmluva

Materiál byl sestaven na základě poznámek doc. Ing. Tomáše Hobzy, Ph.D., kterému bychom tímto chtěli poděkovat za rozsáhlou korekci vzniklého materiálu. Zmíněné přednášky proběhly v zimním semestru akademického roku 2020/2021 na Fakultě jaderné a fyzikálně inženýrské ČVUT v Praze. Přednášky nebyly uskutečněny prezenční formou vzhledem k probíhající pandemii Covid-19.

Tento učební text je určen posluchačům 1. ročníku navazujícího magisterského studia navštěvujícím kurs 01RAD *Regresní analýza dat*, který je zařazen mezi předměty oborů AMSM. Při sestavování textu se předpokládaly znalosti základů matematiky na úrovni absolvování kurzů 01MAB2-4, 01LAB1-2 a 01MIP.

Doporučená literatura:

(1) ...

1 Regresní analýza

1.1 Jednorozměrná lineární regrese

Předpokládejme, že se sledují dvě fyzikální veličiny X a Y mezi kterými existuje lineární závislost

$$Y = \beta_0 + \beta_1 X.$$

β_0 a β_1 nejsou známy, a proto se provádí experiment, při němž se zjišťují hodnoty dvojic (X, Y). Často se stává, že měření hodnot X probíhá prakticky zcela přesně (například X se nastavuje na předem dané úrovně), zatímco Y se měří s určitou chybou. Zavádí se tedy model

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad \forall\ i = 1, ..., n,$$

kde e_i je náhodný šum a $e_1, ..., e_n$ jsou $iid\ \mathcal{N}(0, \sigma^2)$ a dvojice $(x_1, y_1), ..., (x_n, y_n)$ získáme měřením. Neznáme parametry jsou $\beta_0, \beta_1, \sigma^2$, chtěli bychom je odhadnout na základě výběru (MLE odhady).

Rozdělení Y_i je $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$, a tedy věrohodnostní funkce výběru $y_1, ..., y_n$ je

$$L = \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x)^2}.$$

$$l = \ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2.$$

Je zřejmé, že pro libovolné σ^2 potřebujeme minimalizovat

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$$

přes β_0, β_1, na což použijeme metodu nejmenších čtverců (poznámka?).

$$\frac{\partial l}{\partial \beta_0} = 2\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i) = 0,$$

$$\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i) x_i = 0.$$

Z toho pak

$$\sum_{i=1}^n Y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i = 0,$$

$$\beta_0 = \overline{Y_n} - \beta_1 \overline{x_n} = \frac{1}{n}\sum_{i=1}^n Y_i - \beta_1 \frac{1}{n}\sum_{i=1}^n x_i.$$

Po vynásobení poslední rovnice n úpravou dostaneme vztah

$$\sum_{i=1}^{n}(Y_i - \overline{Y_n} + \beta_1\overline{x_n} - \beta_1 x_i)x_i = 0$$

a následně i vztah

$$\sum_{i=1}^{n} Y_i x_i - \overline{Y_n}\sum_{i=1}^{n} + \beta_1\overline{x_n}\sum_{i=1}^{n} x_i - \beta_1\sum_{i=1}^{n} x_i^2 = 0.$$

Z toho už následně vyjádříme

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{Y_n}\,\overline{x_n}}{\sum_{i=1}^{n} x_i^2 - n\overline{x_n}^2} \qquad \text{a} \qquad \widehat{\beta}_0 = \overline{Y_n} - \widehat{\beta}_1\overline{x_n}.$$

Nyní již spočítáme logaritmickou věrohodnostní funkci

$$\frac{\partial l}{\partial(\sigma^2)} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2 = 0,$$

odkud

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

Pokud dále označíme

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$

pak rozdíly

$$r_i = Y_i - \widehat{Y}_i$$

nazýváme **rezidua** (která by měla mít normální rozdělení, aby byly splněny předpoklady modelu) a

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = S_e$$

nazveme **reziduální součet čtverců**.

R^2 statistika

Tuto statistiku definujeme vztahem

$$R^2 = 1 - \frac{\sum_{i=1}^{n} r_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y_n})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y_n})^2},$$

který se dá chápat jako podíl součtu reziduálních čtverců a rozptylu Y. R^2 se interpretuje jako poměr variability v datech vysvětlené lineárním modelem. Čím větší je R^2, tím lépe vysvětluje náš model data, v ideálním případě pak $R^2 = 1$. Dále bychom chtěli:

1. sestrojit IS pro parametry modelu $\beta_0, \beta_1, \sigma^2$,
2. intervaly pro predikci hodnoty y v daném bodě x a

3. testovat hypotézy na parametrech modelu, například F-stat. v MATLABu testuje H_0 : $\beta_0 = 0$ a $\beta_1 = 0$, že vysvětlující proměnná y není korelovaná s vysvětlovanou proměnnou x.

Vše je podobné testům o parametrech $N(\mu, \sigma^2)$ (t-test, F-test), potřebujeme rozdělení odhadů $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma^2}$. Sdružené rozdělení $\widehat{\beta}_0, \widehat{\beta}_1$ se najde snadno, protože to jsou lineární funkce Y_i takže budou mít normální rozdělení, stačí tedy určit střední hodnoty, rozptyly, kovariance,... Označme výběrový rozptyl x jako

$$\sigma_x{}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x_n}^2.$$

Platí, že

1.
$$\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{n\sigma_x{}^2}\right),$$
$$\widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\Big(\frac{1}{n} + \frac{(\overline{x_n})^2}{n\sigma_x{}^2}\Big)\right) = \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n\sigma_x{}^2}\frac{1}{n}\sum_{i=1}^{n} x_i^2\right),$$
$$\mathbb{C}\mathrm{ov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{\overline{x_n}\sigma^2}{n\sigma_x{}^2},$$

2. $\widehat{\sigma}^2$ je nezávislé na $\widehat{\beta}_0$ a $\widehat{\beta}_1$,

3.
$$\frac{n\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Poznámka 1.1. První bod znamená, že $(\beta_0, \beta_1) \sim \mathcal{N}(\mu, \sum)$, kde

$$\boldsymbol{\mu} = (\beta_0, \beta_1) \quad \text{a} \quad \sum = \frac{\sigma^2}{n\sigma_x^2}\begin{pmatrix} \overline{x_n}^2 & -\overline{x_n} \\ -\overline{x_n} & 1 \end{pmatrix}.$$

Konfidenční intervaly

1. σ^2, a protože $\frac{n\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$, víme, že s pravděpodobností $\mathbb{P} = 1-\alpha$ bude

$$\chi^2_{\frac{\alpha}{2}}(n-2) \leqslant \frac{n\widehat{\sigma}^2}{\sigma^2} \leqslant \chi^2_{1-\frac{\alpha}{2}}(n-2),$$

a tedy $(1-\alpha)\%$ IS (interval spolehlivosti) pro σ^2 je

$$\frac{n\widehat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}(n-2)} \leqslant \sigma^2 \leqslant \frac{n\widehat{\sigma}^2}{\chi^2_{\frac{\alpha}{2}}(n-2)}.$$

2. β_1

Veličiny $\frac{\widehat{\beta}_1-\beta_1}{\sqrt{\frac{\sigma^2}{n\sigma_x^2}}} \sim \mathcal{N}(0,1)$ a $\frac{n\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$ jsou nezávislé. Z toho vyplývá, že

$$\frac{(\widehat{\beta}_1-\beta_1)\Big/\sqrt{\frac{\sigma^2}{n\sigma_x^2}}}{\sqrt{\frac{n\widehat{\sigma}^2}{\sigma^2}\frac{1}{n-2}}} \sim t(n-2).$$

Z toho potom

$$\frac{\widehat{\beta}_1-\beta_1}{\sqrt{\frac{\widehat{\sigma}^2}{(n-2)\sigma_x^2}}} = (\widehat{\beta}_1-\beta_1)\sqrt{\frac{(n-2)\sigma_x^2}{\widehat{\sigma}^2}} \sim t(n-2), \tag{1.1}$$

což znamená, že

$$-t_{1-\frac{\alpha}{2}}(n-2) \leqslant (\widehat{\beta}_1-\beta_1)\sqrt{\frac{(n-2)\sigma_x^2}{\widehat{\sigma}^2}} \leqslant t_{1-\frac{\alpha}{2}}(n-2)$$

s pravděpodobností $\mathbb{P} = 1-\alpha$, a tedy

$$\widehat{\beta}_1 - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{\widehat{\sigma}^2}{(n-2)\sigma_x^2}} \leqslant \beta_1 \leqslant \widehat{\beta}_1 + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{\widehat{\sigma}^2}{(n-2)\sigma_x^2}}$$

je $100(1-\alpha)\%$ IS pro β_1. Podobně pro β_0 dostaneme, že

$$\frac{\widehat{\beta}_0-\beta_0}{\sqrt{\sigma^2(\frac{1}{n}+\frac{\overline{x}_n^2}{\sigma_x^2})}}\frac{1}{\sqrt{\frac{n\widehat{\sigma}^2}{\sigma^2}\frac{1}{n-2}}} \sim t(n-2),$$

$$\frac{\widehat{\beta}_0-\beta_0}{\sqrt{(1+\frac{\overline{x}_n^2}{\sigma_x^2})\widehat{\sigma}^2\frac{1}{n-2}}} \sim t(n-2), \tag{1.2}$$

a tedy

$$\widehat{\beta}_0 - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{(1+\frac{\overline{x_n}^2}{\sigma_x^2})\widehat{\sigma}^2\frac{1}{n-2}} \leqslant \beta_0 \leqslant \widehat{\beta}_0 + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{(1+\frac{\overline{x_n}^2}{\sigma_x^2})\widehat{\sigma}^2\frac{1}{n-2}}$$

je $100(1-\alpha)\%$ IS pro β_0.

Statistiky (1.1) a (1.2) se dají použít i pro konstrukci testů například $H_0: \beta_1 = 0$. Za platnosti H_0 totiž

$$T_1 = \widehat{\beta}_1\sqrt{\frac{(n-2)\sigma_x^2}{\widehat{\sigma 2}}} \sim t(n-2),$$

a tedy H_0 zamítáme, pokud

$$|T_1| > t_{1-\frac{\alpha}{2}}(n-2).$$

TEST: H_0 zamítáme, pokud $|T_1| > t_{1-\frac{\alpha}{2}}(n-2)$.

PŘÍKLAD 1.2 (Měření rychlosti zvuku v závislosti na teplotě).

teplota	-20	0	20	50	100
rychlost (m/s)	323	327	340	364	386

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i = 30, \quad \overline{Y_n} = 348, \quad \sum_{i=1}^{n} X_i Y_i = 57140, \quad \sum_{i=1}^{n} X_i^2 = 13300,$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X_n^2} = \frac{1}{5}13300 - 900 = 1760,$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - 5\overline{x_n}\,\overline{Y_n}}{\sum_{i=1}^{n} X_i^2 - 5\overline{X_n^2}} = 0.561,$$

$$\widehat{\beta}_0 = \overline{Y_n} - \widehat{\beta}_1 \overline{x_n} = 331.16,$$

$$\widehat{\sigma}^2 = \frac{1}{5}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 = 11.37 \text{ a nestranný}$$

$$s^2 = \frac{1}{5-2}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 = 18.95.$$

Spočítáme IS například pro β_1. Dostaneme tedy $t_{0.975}(5-2) = 3.18$, který dosadíme do vzorečku na výpočet IS pro β_1, kde $\beta_1 \in (0.414, 0.709)$.
$\beta_1 = 0$, $T_1 = 12.097$, $|T_1| \geqslant t_{0.975}(3) = 3.18$, a proto nezamítáme H_0.

1.2 Intervaly predikce

Předpokládejme, že máme nové pozorování X, pro které je Y neznámé a my bychom chtěli predikovat hodnoty Y, případně najít intervaly spolehlivosti pro Y. Vzhledem k lineárnímu regresnímu modelu $Y = \beta_0 + \beta_1 X + e$ je přirozené vzít za predikci

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X.$$

Najdeme rozdělení rozdílu $Y - \widehat{Y}$. Zřejmě se jedná o normální rozdělení ($\beta_0 \sim \mathcal{N}(...)$, $\beta_1 \sim \mathcal{N}(...), e_1 \sim \mathcal{N}(...), Y \sim \mathcal{N}(...)$) stačí tedy určit střední hodnotu a rozptyl.

$$\mathbb{E}(\widehat{Y} - Y) = \mathbb{E}(\widehat{\beta}_0) + \mathbb{E}(\widehat{\beta}_1 X) - \beta_0 - \beta_1 X - \mathbb{E}(e) = \beta_0 + \beta_1 X - \beta_0 - \beta_1 X - 0 = 0.$$

Protože nový pár (X, Y) je nezávislý na předchozích datech, platí, že Y je nezávislé na $\widehat{Y}$ (β_0, β_1 jsou spočteny pouze pomocí $Y_1, ..., Y_n$). Pak tedy

$$\mathrm{D}(\widehat{Y} - Y) = \mathrm{D}(\widehat{Y}) + \mathrm{D}(Y) = \mathrm{D}(\widehat{Y}) + \sigma^2,$$

protože $\mathrm{D}(Y) = \mathrm{D}(e) = \sigma^2$.

$$\begin{aligned}
\mathrm{D}(\widehat{Y}) = \mathrm{D}(\widehat{\beta}_0 + \widehat{\beta}_1 X) &= \mathbb{E}(\widehat{\beta}_0 + \widehat{\beta}_1 X - \beta_0 - \beta_1 X)^2 = \mathbb{E}\left[\widehat{\beta}_0 - \beta_0 + X(\widehat{\beta}_1 - \beta_1)\right]^2 = \\
&= \underbrace{\mathbb{E}(\widehat{\beta}_0 - \beta_0)^2}_{\mathrm{D}\widehat{\beta}_0} + \underbrace{X^2\mathbb{E}(\widehat{\beta}_1 - \beta_1)}_{\mathrm{D}\widehat{\beta}_0} + 2X\underbrace{\mathbb{E}(\widehat{\beta}_0 - \beta_0)(\widehat{\beta}_1 - \beta_1)}_{\mathrm{D}(\widehat{\beta}_0, \widehat{\beta}_1)} = \\
&= \Big(\frac{1}{n} + \frac{(\overline{x_n})^2}{x\sigma_X^2}\Big)\sigma^2 + X^2\frac{\sigma^2}{n\sigma_X^2} - 2X\frac{\overline{x_n}\sigma^2}{n\sigma_X^2} = \sigma^2\Big(\frac{1}{n} + \frac{(\overline{x_n} - X)^2}{n\sigma_x^2}\Big)
\end{aligned}$$

Máme tedy

$$\widehat{Y} - Y \sim \mathcal{N}\left(0, \sigma^2\Big(1 + \frac{1}{n} + \frac{(\overline{x_n} - X)^2}{n\sigma_X^2}\Big)\right),$$

a proto

$$\frac{(\widehat{Y} - Y)\Big/\sqrt{\sigma^2(1 + \frac{1}{n} + \frac{(\overline{x_n} - X)^2}{n\sigma_x^2})}}{\sqrt{\frac{1}{n-2}\frac{n\widehat{\sigma}^2}{\sigma^2}}}$$

a tedy $100(1-\alpha)\%$ interval prediktu??? je

$$\widehat{Y} - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{\widehat{\sigma}^2}{n-2}\Big(n + 1 + \frac{(\overline{x_n} - X)^2}{\sigma_x^2}\Big)} \leqslant Y \leqslant \widehat{Y} + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\frac{\widehat{\sigma}^2}{n-2}\Big(n + 1 + \frac{(\overline{x_n} - X)^2}{\sigma_x^2}\Big)}.$$

Tohle kreslí MATLAB (polytool)

Příklad 1.3 (Rychlost zvuku). Mějme $\overline{x_n} = 30,\ \sigma_X^2 = 1760,\ \widehat{\beta}_1 = 0.561,\ \widehat{\beta}_0 = 331.16,$ $\sigma^2 = 11.37,$ nestraný, $\widehat{s}^2 = 18.95$. Nové $X = 35°C$ a $\widehat{Y} = 331.16 + 0.561 \cdot 35 = 350.8.$

$$\sqrt{\frac{\widehat{\sigma}^2}{n-2}\Big(n + 1 + \frac{(\overline{x_n} - X)^2}{\sigma_x^2}\Big)} = \sqrt{\frac{11.37}{3}\Big(6 + \frac{(30-35)^2}{1760}\Big)} = 4.77$$

$$t_{0.975}(3) = 3.1824 \text{ a tedy } IP = (335.6, 366.0)$$

Poznámka 1.4. Někdy dopředu známe kandidáta b_1 jako hodnotu parametru β_1 a chtěli bychom testovat $H_0 : \beta_1 = b_1$ vs. $H_1 : \beta_1 \neq b_1$. Test bude zamítnut H_0, pokud

$$|\beta_1 - b_1| \cdot \frac{\sqrt{S_{xx}}}{s_n} > t_{1-\frac{\alpha}{2}}(n-2).$$

1.2.1 Test významnosti interceptu

Otázka je, zda přímka prochází počátkem $(0,0)$, tedy $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 \neq 0$. Nezamítnutí H_0 znamená, že jednodušší model $y = \beta_1 x + e$ lépe popisuje datta, než $y = \beta_0 + \beta_1 x + e$. H_0 potom zamítneme, pokud

$$T_n = \frac{|\widehat{\beta}_0|}{\widehat{\sigma}(\widehat{\beta}_0)} = |\widehat{\beta}_0| \frac{1}{s_n\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}} > t_{1-\frac{\alpha}{2}}(n-2).$$

1.2.2 ANOVA přístup pro testování

Odvodili jsme t-test významnosti koeficientů a nyní odvodíme ekvivalentní F-test, který může být zobecněn na test celkové významnosti vícerozměrného regresního modelu (testy významnosti jednotlivých koeficientů mohou být totiž zavádějící).

Myšlenkou metody (analýza rozptylu ANOVA) je určit, kolik variability v pozorováních $(y_1, y_2, ..., y_n)$ je "vysvětleno"regresním modelem (přímkou). Míru variability v datech pak spočítáme jako podíl součtu sum od regrese a celkového počtu čtverců, tedy

$$\text{SST} = \sum_{i=1}^{n}(y_i - \overline{y}_n)^2,$$

pokud regresní přímka $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ dobře prokládá data, tedy $\widehat{y}_i \approx y_i$. Dále bude platit, že

$$\sum_{i=1}^{n} (\widehat{y}_i - \overline{\widehat{y}}_n)^2 \approx \sum_{i=1}^{n} (y_i - \overline{y}_n)^2.$$

Ukážeme, že $\overline{\widehat{y}} = \overline{y}_n$ a tak

$$\sum_{i=1}^{n} (\widehat{y}_i - \overline{\widehat{y}}_n)^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2 = \mathrm{SSR}$$

regresi sum ob squares, regresní součet čtverců. Podíl

$$\mathrm{R}^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\sum\limits_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2}{\sum\limits_{i=1}^{n} (y_i - \overline{y}_n)^2}$$

tak vyjadřuje variabilitu v $(y_1, ..., y_n)$ vysvětlené regresním modelem.

R^2 - *koeficient determinace (coefficient of determination)* (pro každý model by měl mít hodnotu $\mathrm{R}^2 \approx 1$). Ukážeme, že R^2 je kvadrát výběrového korelačního koeficientu mezi $\mathbf{x}$ a $\mathbf{y}$, což dává statistice ^{2} význam míry "dobré shody".

Pokud bychom znali rozdělení pravděpodobnostní statistiky R^2, nabízí se její použití pro test $H_0 : \beta_1 = 0$, kterou bychom zamítli, pokud bude $\mathrm{R}^2 \approx 1$. Protože každá monotonní funkce R^2 vede na ekvivalentní test, budeme uvažovat statistiku

$$\mathrm{F} = \frac{(n-2)\mathrm{R}}{1 - \mathrm{R}^2}.$$

Lemma 1.5. *Nechť $\widehat{e}_i = y_i - \widehat{y}_i$ značí rezidua, kde $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ a $\widehat{\beta}_0, \widehat{\beta}_1$ jsou LSE. Potom*

1. $\sum\limits_{i=1}^{n} \widehat{e}_i = 0$,
2. $\overline{\widehat{y}}_n = \overline{y}_n$,
3. $\sum\limits_{i=1}^{n} \widehat{e}_i \widehat{y}_i = 0$.

Důkaz. 1. Z rovnice $\frac{\partial S}{\partial \beta_0} = 0$ dostaneme

$$0 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = \sum_{i=1}^{n} (y_i - \widehat{y}_i) = \sum_{i=1}^{n} \widehat{e}_i.$$

2. Z bodu 1) plyne, že $\sum\limits_{i=1}^{n} \widehat{y}_i = \sum\limits_{i=1}^{n} y_i$, podělením n dostaneme dokazované tvrzení.

3. Z rovnice $\frac{\partial S}{\partial \beta_1} = 0$ dostaneme

$$0 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) x_i = \sum_{i=1}^{n} \widehat{e}_i x_i$$

a tedy

$$\sum_{i=1}^{n} \widehat{e}_i \widehat{y}_i = \sum_{i=1}^{n} \widehat{e}_i(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = \sum_{i=1}^{n} \widehat{e}_i \widehat{\beta}_0 + \sum_{i=1}^{n} x_i \widehat{e}_i \widehat{\beta}_1 = \widehat{\beta}_0 \underbrace{\sum_{i=1}^{n} \widehat{e}_i}_{=0} + \widehat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i \widehat{e}_i}_{=0} = 0.$$

□

Věta 1.6. *Předpokládejme, že* $\mathrm{SST} \neq 0$. *Potom platí*

1. $0 \leqslant \mathrm{R}^2 \leqslant 1$,
2. $\mathrm{R}^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$, *kde* $\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$ *jako reziduální součet čtverců,*
3. $\mathrm{R}^2 = 1 \Leftrightarrow (\forall i \in \hat{n})(\widehat{y}_i = y_i)$ *(všechna data leží na přímce),*
4. *pokud označíme* $\boldsymbol{x} = (x_1, ..., x_n)$ *a* $\boldsymbol{y} = (y_1, ..., y_n)$, *potom* $\mathrm{R}^2 = \varrho^2(\boldsymbol{x}, \boldsymbol{y})$, *kde*

$$\varrho(\boldsymbol{x}, \boldsymbol{y}) = \frac{\Big(\sum_{i=1}^{n} (x_i - \overline{x}_n)(y_i - \overline{y}_n) \Big)^2}{S_{xx} S_{yy}}$$

   je druhá mocnina výběrového korelačního koeficientu vektorů $\boldsymbol{x}, \boldsymbol{y}$,

5. $\mathrm{F} = \frac{\mathrm{SSR}}{s_n^2} = \mathrm{T}^2$,
6. *pokud jsou chyby* $e_1, ..., e_n$ *iid* $\mathcal{N}(0, \sigma^2)$ *a* $\beta_1 = 0$ *(platí* $H_0 : \ \beta_1 = 0$*) v modelu, potom* $\mathrm{F} \sim \mathrm{F}(1, n-2)$.

Důkaz. Důkaz věty bude založen na rozkladu

$$\sum_{i=1}^{n} (y_i - \overline{y}_n)^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

neboli $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$. Z lemmatu 1.5 vyplývá, že

$$\begin{aligned} \mathrm{SST} &= \sum_{i=1}^{n} (y_i - \overline{y}_n)^2 = \sum_{i=1}^{n} \big[(y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y}_n) \big]^2 = \\ &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2 + 2 \sum_{i=1}^{n} (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}_n) = \mathrm{SSE} + \mathrm{SSR} + 0, \end{aligned}$$

neboť

$$\sum_{i=1}^{n} (\underbrace{(y_i - \widehat{y}_i)}_{=\widehat{e}_i} (\widehat{y}_i - \overline{y}_n)) = \underbrace{\sum_{i=1}^{n} \widehat{e}_i \widehat{y}_i}_{=0} - \overline{y}_n \underbrace{\sum_{i=1}^{n} \widehat{e}_i}_{=0} = 0.$$

Z toho potom dokazujeme jednotlivé body věty.

1. Protože $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$, pak $0 \leqslant \mathrm{R}^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} \leqslant \frac{\mathrm{SST}}{\mathrm{SST}} = 1$.

2. $\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE} \;\Rightarrow\; \mathrm{R}^2 = \frac{\mathrm{SST}-\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$.

3. Z bodu 2 plyne, že $\mathrm{R}^2 = 1 \;\Leftrightarrow\; \mathrm{SSE} = 0$ a $\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = 0 \;\Leftrightarrow\; y_i = \widehat{y}_i \;\forall i \;\in \hat{n}$.

4. $\widehat{y}_i = \underbrace{\widehat{\beta}_0}_{=\overline{y}_n - \widehat{\beta}_1 x_n} + \widehat{\beta}_1 x_i = \overline{y}_n - \widehat{\beta}_1(\overline{x}_n - x_i)$. Proto pak

$$\mathrm{SSR} = \sum_{i=1}^{n}(\widehat{y}_i - \widehat{y}_n)^2 + \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \widehat{\beta}_1^2 S_{xx},$$

a protože $\widehat{\beta}_1 = \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)$, dostaneme

$$\varrho^2(\mathbf{x}, \mathbf{y}) = \frac{\Big[\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)\Big]^2}{S_{xx} S_{yy}} = \frac{\widehat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{\mathrm{SSR}}{\mathrm{SST}} = \mathrm{R}^2,$$

neboť $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y}_n)^2 = \mathrm{SST}$.

5. Z definice F plyne, že

$$\mathrm{F} = \frac{(n-2)\mathrm{R}^2}{1-\mathrm{R}^2} = \frac{(n-2)\frac{\mathrm{SSR}}{\mathrm{SST}}}{\frac{\mathrm{SSE}}{\mathrm{SST}}} = \frac{\mathrm{SSR}}{\frac{\mathrm{SSE}}{n-2}} = \frac{\mathrm{SSR}}{s_n^2}.$$

Protože $T_n = \widehat{\beta}_1 \frac{\sqrt{S_{xx}}}{s_n}$, pak

$$\mathrm{T}^2 = \frac{\widehat{\beta}_1^2 S_{xx}}{s_n^2} = \frac{\mathrm{SSR}}{s_n^2} = \mathrm{F}.$$

6. $T \sim t(n-2) \;\Rightarrow\; \mathrm{F} = \mathrm{T}^2 \sim \mathrm{F}(1, n-2)$.

□

Poznámka 1.7. 1. Z bodů 5 a 6 vyplývá, že použití libovolné statistiky T_n, R^2 nebo F vede na ekvivalentní test významnosti regrese.

2. R^2 poskytuje hrubou představu o kvalitě modelu, čím je blíže 1, tím lépe přímka prokládá data (nicméně je třeba jisté obezřetnosti, jak uvidíme později).

3. F lze chápat jako statistiku pro test významnosti velkých hodnot R^2.

Výsledky se většinou uvádí v tabulce ANOVA:

Source	df	SS	MS	F
Regression	1	SSR	MSR=SSR	$\frac{\mathrm{MSR}}{\mathrm{MSE}}$
Residual	$n-2$	SSE	$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-2} = s_n^2$	
Total	$n-1$	SST		

$$\mathrm{R}^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}$$

Kde **source** je zdroj součtu čtverců, **df** počet stupňů volnosti příslušný danému součtu čtverců, **SS** počet čtverců a **MS** ($\mathrm{MS} = \frac{\mathrm{SS}}{\mathrm{df}}$) "mean squares".

Poznámka 1.8. $H_0: \beta_1 = 0$ je zamítnul, pokud $\mathrm{F} > \mathrm{F}_{1-\alpha}(1, n-2)$. V tomto jednorozměrném případě je to ekvivalentní t-testu, neboť $\mathrm{F} = \mathrm{T}^2$.

Věta 1.9. *Mějme $e_1, ..., e_n$ iid $\mathcal{N}(0, \sigma^2)$. Za platnosti $H_0: \beta_1 = 0$ je splněno, že*

$$\frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2(1), \qquad \frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2), \qquad \frac{\mathrm{SST}}{\sigma^2} \sim \chi^2(n-1).$$

Poznámka 1.10. Proto v tabulce ANOVA 1.2.2 uvádí df po řadě $1, n-2, n-1$. Používají se však i v případě jiného rozdělení chyb. Představit si je lze takto:

1. $\mathrm{SSE} = \sum_{i=1}^{n} \widehat{e}_i^2$, na n-rezidní $\widehat{e}_1, ..., \widehat{e}_n$ máme 2 podmínky $\sum_{i=1}^{n} \widehat{e}_i = 0$ a $\sum_{i=1}^{n} x_i \widehat{e}_i = 0$. Z toho vyplývá, že mají $n-2$ stupňů volnosti.

2. $\mathrm{SST} = \sum_{i=1}^{n} (y_i - \overline{y}_n)^2 ... \; y_i - \overline{y}_n$ musí splňovat $\sum_{i=1}^{n} (y_i - \overline{y}_n) = 0$, a proto má $n-1$ stupňů volnosti.

3. $\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE}$, a počet stupňů volnosti je roven $(n-1) - (n-2) = 1$.

Důkaz. V důkazu věty ?? jsme ukázali, že $\mathrm{SSR} = \widehat{\beta}_1^2 S_{xx}$, takže $\frac{\mathrm{SSR}}{\sigma^2} = \left(\frac{\widehat{\beta}_1 \sqrt{S_{xx}}}{\sigma}\right)^2$, víme, že $\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$ a tedy $(\widehat{\beta}_1 - \beta_1)\frac{S_{xx}}{\sigma} \sim \mathcal{N}(0,1)$. Pro $\beta_1 = 0$ tedy

$$\widehat{\beta}_1 \frac{\sqrt{S_{xx}}}{\sigma} \sim \mathcal{N}(0,1) \;\Rightarrow\; \frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2(1).$$

Zároveň také $\frac{\mathrm{SSE}}{\sigma^2} = \frac{(n-2)s_n^2}{\sigma^2} \sim \chi^2(n-2)$ (viz dříve) a nezávisí na $\widehat{\beta}_1$. Z toho vyplývá, že $\frac{\mathrm{SSR}}{\sigma^2}$ a $\frac{\mathrm{SSE}}{\sigma^2}$ jsou nezávislé. Dále platí, že

$$\frac{\mathrm{SST}}{\sigma^2} = \frac{\mathrm{SSR}}{\sigma^2} + \frac{\mathrm{SSE}}{\sigma^2} \;\Rightarrow\; \frac{\mathrm{SST}}{\sigma^2} \sim \chi^2(n-1).$$

□

Poznámka 1.11. R^2 statistika - pozor na zjednodušení kvality modelu.

1. Nízké hodnoty R^2 nemusí znamenat, že regresní model není významný. V datech jen může být velké množství nevysvětlitelné náhodné variability. Například opakování hodnoty regresoru x snižují hodnotu R^2 oproti modelům s různými x.

2. Velké hodnoty R^2 mohou být způsobeny velkým měřítkem dat (S_{xx} je velká). Platí totiž, že

   $$\mathbb{E}(\mathrm{R}^2) \approx \frac{\beta_1^2 S_{xx}}{\beta_1^2 S_{xx} + \sigma^2},$$

   což je rostoucí funkce S_{xx}.

   Velký rozptyl $(x_1, ..., x_n)$ může mít za následek velké R^2 a přitom nic neříká o kvalitě modelu.

   $\mathbb{E}(\mathrm{R}^2)$ je také rostoucí funkcí β_1^2. Modely s *velkou* směrnicí tedy budou mít obecně větší $yRMR^2$, než modely s "malou"směrnicí.

Při hodnocení kvality modelu potřebujeme více kritérií. Mezi ně patří například

1. "velké" R^2,
2. "velké" F nebo |T| hodnoty,
3. "malé" hodnoty s_n^2 vzhledem k $\overline{y}_n$.

Další kritéria budeme probírat později.

PŘÍKLAD 1.12. Velká hodnota R^2 indikuje přibližně lineární vztah mezi x a y, ale vysoký stupeň korelace nemusí znamenat příčinný vztah. data: 1924-1937

y_i - počet mentálních onemocnění na 100000 obyvatel Anglie.
x_i - počet rádií v populaci.
model - $y_i = \beta_0 + \beta_1 x_i + e_i$.

$$\widehat{\beta}_0 = 4.5822, \qquad \widehat{\beta}_1 = 2.2042, \qquad R^2 = 0.984,$$

tzv. velmi významný lineární vztah mezi x a y. Závěr by mohl být, že rádia způsobují mentální onemocnění. I když by to mohla být pravda, nabízí se věrohodnější vysvětlení, a to takové, že x i y rostou lineárně s časem, tzn. y roste lineárně s x.

Rádia byla s časem dostupnější, lepší diagnostické procedury umožňovaly identifikovat více lidí s mentálními problémy.

2 Vícerozměrná lineární regrese

Předpokládejme model

$$Y_i = \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,$$

kde $\varepsilon_1, ..., \varepsilon_n$ *iid* $\mathcal{N}(0, \sigma^2)$. V maticové formě

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

kde $\mathbf{Y} = \mathbf{Y}_{n\times 1}$, $\varepsilon = \varepsilon_{n\times 1}$, $\beta = \beta_{p\times 1}$ a $\mathbf{X} = \mathbf{X}_{n\times p}$. Sloupce matice $\mathbf{X}$ označíme $X_1, ..., X_p$, tedy $\mathbf{X} = (X_1, ..., X_p)$ a předpokládejme, že jsou nezávislé. Pokud by nebyly nezávislé, nebylo by možné získat (rekonstruovat) parametr β z $\mathbf{X}$ a $\mathbf{Y}$ ani kdyby nebyl přítomný šum ε. (Vlastně bychom měli soustavu $\mathbf{X}\beta = \mathbf{Y}$.)

Poznámka 2.1. V jednorozměrné regresi by to odpovídalo případu, kdy jsou všechny X_i stejné, tzn. že by nebylo možné odhadnout přímku přímo z pozorování pouze v jednom bodě.

Dále předpokládejme, že

$$n > p, \quad h(\mathbf{X}) = p.$$

Zkusíme následně vypočítat MLE parametrů β, σ^2.

Věta 2.2. *Pro MLE parametrů β a σ^2 platí, že*

$$\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$$

a

$$\widehat{\sigma}^2 = \frac{1}{n}(\boldsymbol{Y} - X\widehat{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}) = \frac{1}{n}\|Y - \boldsymbol{X}\widehat{\beta}\|^2 = \frac{1}{n}\left\|Y - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\right\|^2.$$

Důkaz. zřejmě $Y_i \sim \mathcal{N}(\beta_1 X_{i1} + \ldots + \beta_p X_{ip}, \sigma^2)$ a její hustota tedy je

$$f_i(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(y - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2}{2\sigma^2}$$

a věrohodnostní funkce

$$\begin{aligned} L &= \prod_{i=1}^{n} f_i(Y_i) = \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp - \frac{\sum_{i=1}^n (Y_i - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2}{2\sigma^2} \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp - \frac{1}{2\sigma^2}\|Y - X\beta\|^2 \\ l &= \ln L = C - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\|Y - X\beta\|^2 \end{aligned}$$

Je třeba minimalizovat

$$\|Y - X\beta\|^2 = (Y - X\beta)^T(Y - X\beta) = (Y - \sum_{i=1}^{p} \beta_i X_i)^T(Y - \sum_{i=1}^{p} \beta_i X_I)$$
$$= Y^TY - 2\sum_{i=1}^{p} \beta_i Y X_i + \sum_{j=1}^{p}\sum_{i=1}^{p} \beta_i\beta_j X_i^T X_j.$$

Derivujeme podle β_i. Potom

$$-2Y^TX_i + 2\sum_{j=1}^{p} \beta_j X_i^T X_j = 0, \quad \text{a tedy} \quad Y^TX_i = \sum_{j=1}^{p} \beta_j X_i^T X_j, \quad \forall i \leqslant p.$$

V maticovém zápisu se $\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\beta$ nazývá **soustava normálních rovnic**. Matice $\mathbf{X}^T\mathbf{X}$ má rozměr $p \times p$ a je invertibilní, protože $h(\mathbf{X}) = p$ a $h(\mathbf{X}^T\mathbf{X}) = h(\mathbf{X})$ pro libovolnou matici $\mathbf{X}$. Proto tedy

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Derivujeme podle σ^2. Potom

$$-\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\|Y - X\beta\|^2 = 0,$$
$$\widehat{\sigma}^2 = \frac{1}{n}\left\|Y - X\widehat{\beta}\right\|^2 = \frac{1}{n}\underbrace{(Y - X\widehat{\beta})^T(Y - X\widehat{\beta})}_{R} = \frac{1}{n}R,$$

kde R je reziduální součet čtverců. □

Pro statistickou analýzu potřebujeme rozdělení odhadů $\widehat{\beta}, \widehat{\sigma}^2$.

Věta 2.3. *Platí, že*

$$\widehat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \quad a \quad \frac{n\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Odhady $\widehat{\beta}, \widehat{\sigma}^2$ jsou nezávislé.

Důkaz. $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, a proto

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \varepsilon) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon = \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon.$$

Z toho vyplývá, že $\mathbb{E}\widehat{\beta} = \beta$, protože $\mathbb{E}\varepsilon = 0$. Kovarianční matici můžeme napsat ve tvaru

$$\mathbb{E}(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)^T = \mathbb{E}((X^TX)^{-1}X^T\varepsilon\varepsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}(\varepsilon\varepsilon^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$
$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

□