

01RAD

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

26. září 2020

Obsah

1	Regresní analýza	1
1.1	Jednorozměrná lineární regrese	1
1.2	Intervaly predikce	5
2	Vícerozměrná lineární regrese	7
3	1-10	9
3.1	Jednorozměrná lineární regrese	9
3.1.1	Data s předpokladem normality dat	10
3.1.2	Data bez předpokladu normality	11
3.1.3	Vlastnosti odhadů	13
4	11-20	15
4.1	Gauss - Markov theorem	17
4.2	IS pro β_0, β_1	18
4.3	TH pro β_0, β_1	19
5	21-30	21
5.0.1	Test významnosti interceptu	21
5.0.2	ANOVA přístup pro testování	21

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 úrovni), zatímco Y se měří s určitou chybou. Zavádí se tedy model

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

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

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

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

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

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

$$\begin{aligned} \sum_{i=1}^n Y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i &= 0, \\ \beta_0 = \bar{Y}_n - \beta_1 \bar{x}_n &= \frac{1}{n} \sum_{i=1}^n Y_i - \beta_1 \frac{1}{n} \sum_{i=1}^n x_i. \end{aligned}$$

1 Regresní analýza

Po vynásobení poslední rovnice n úpravou dostaneme vztah

$$\sum_{i=1}^n (Y_i - \bar{Y}_n + \beta_1 \bar{x}_n - \beta_1 x_i) x_i = 0$$

a následně i vztah

$$\sum_{i=1}^n Y_i x_i - \bar{Y}_n \sum_{i=1}^n x_i + \beta_1 \bar{x}_n \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 = 0.$$

Z toho už následně vyjádříme

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i Y_i - n \bar{Y}_n \bar{x}_n}{\sum_{i=1}^n x_i^2 - n \bar{x}_n^2} \quad \text{a} \quad \hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n.$$

Nyní již spočítáme logaritmicke 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

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

Pokud dále označíme

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

pak rozdíl

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

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

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

nazveme **reziduální součet čtverců**.

R^2 statistika

Tuto statistiku definujeme vztahem

$$R^2 = 1 - \frac{\sum_{i=1}^n r_i^2}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2},$$

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

1 Regresní analýza

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 $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$. Sdružené rozdělení $\hat{\beta}_0, \hat{\beta}_1$ 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

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}_n^2.$$

Platí, že

1.

$$\begin{aligned} \hat{\beta}_1 &\sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{n\sigma_x^2}\right), \\ \hat{\beta}_0 &\sim \mathcal{N}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{(\bar{x}_n)^2}{n\sigma_x^2}\right)\right) = \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n\sigma_x^2} \frac{1}{n} \sum_{i=1}^n x_i^2\right), \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= -\frac{\bar{x}_n \sigma^2}{n\sigma_x^2}, \end{aligned}$$

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

3.

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

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

$$\mu = (\beta_0, \beta_1) \quad \text{a} \quad \Sigma = \frac{\sigma^2}{n\sigma_x^2} \begin{pmatrix} \bar{x}_n^2 & -\bar{x}_n \\ -\bar{x}_n & 1 \end{pmatrix}.$$

Konfidenční intervaly

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

$$\chi_{\frac{\alpha}{2}}^2(n-2) \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq \chi_{1-\frac{\alpha}{2}}^2(n-2),$$

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

$$\frac{n\hat{\sigma}^2}{\chi_{1-\frac{\alpha}{2}}^2(n-2)} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{\chi_{\frac{\alpha}{2}}^2(n-2)}.$$

1 Regresní analýza

2. β_1

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

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

Z toho potom

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

což znamená, že

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

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

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

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

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

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

a tedy

$$\hat{\beta}_0 - t_{1-\frac{\alpha}{2}}(n-2) \sqrt{(1 + \frac{\bar{x}_n^2}{\sigma_x^2})\hat{\sigma}^2 \frac{1}{n-2}} \leq \beta_0 \leq \hat{\beta}_0 + t_{1-\frac{\alpha}{2}}(n-2) \sqrt{(1 + \frac{\bar{x}_n^2}{\sigma_x^2})\hat{\sigma}^2 \frac{1}{n-2}}$$

je 100(1 - α)% 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

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

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

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

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

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

$$\hat{\beta}_0 = \overline{Y_n} - \hat{\beta}_1 \overline{X_n} = 331.16,$$

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

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

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 $\beta_1 \in (0.414, 0.709)$.

$\beta_1 = 0$, $T_1 = 12.097$, $|T_1| \geq t_{0.975}(3) = 3.18$, 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

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

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

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

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

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

protože $D(Y) = D(e) = \sigma^2$.

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

1 Regresní analýza

Máme tedy

$$\hat{Y} - Y \sim \mathcal{N}\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(\bar{x}_n - X)^2}{n\sigma_x^2}\right)\right),$$

a proto

$$\frac{(\hat{Y} - Y) / \sqrt{\sigma^2 \left(1 + \frac{1}{n} + \frac{(\bar{x}_n - X)^2}{n\sigma_x^2}\right)}}{\sqrt{\frac{1}{n-2} \frac{n\hat{\sigma}^2}{\sigma^2}}}$$

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

$$\hat{Y} - t_{1-\frac{\alpha}{2}}(n-2) \sqrt{\frac{\hat{\sigma}^2}{n-2} \left(n + 1 + \frac{(\bar{x}_n - X)^2}{\sigma_x^2}\right)} \leq Y \leq \hat{Y} + t_{1-\frac{\alpha}{2}}(n-2) \sqrt{\frac{\hat{\sigma}^2}{n-2} \left(n + 1 + \frac{(\bar{x}_n - X)^2}{\sigma_x^2}\right)}.$$

Tohle kreslí MATLAB (polytool)

PŘÍKLAD 1.3 (Rychlost zvuku). Mějme $\bar{x}_n = 30$, $\sigma_X^2 = 1760$, $\hat{\beta}_1 = 0.561$, $\hat{\beta}_0 = 331.16$, $\sigma^2 = 11.37$, nestraný, $\hat{s}^2 = 18.95$. Nové $X = 35^\circ\text{C}$ a $\hat{Y} = 331.16 + 0.561 \cdot 35 = 350.8$.

$$\sqrt{\frac{\hat{\sigma}^2}{n-2} \left(n + 1 + \frac{(\bar{x}_n - X)^2}{\sigma_x^2}\right)} = \sqrt{\frac{11.37}{3} \left(6 + \frac{(30 - 35)^2}{1760}\right)} = 4.77$$

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

2 Vícerozměrná lineární regrese

Předpokládejme model

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

kde $\varepsilon_1, \dots, \varepsilon_n \text{ iid } \mathcal{N}(0, \sigma^2)$. 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, \dots, X_p , tedy $\mathbf{X} = (X_1, \dots, X_p)$ 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

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

a

$$\hat{\sigma}^2 = \frac{1}{n} (\mathbf{Y} - \mathbf{X}\hat{\beta})^T (\mathbf{Y} - \mathbf{X}\hat{\beta}) = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}\|^2.$$

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

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

a věrohodnostní funkce

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

2 Vícerozměrná lineární regrese

Je třeba minimalizovat

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

Derivujeme podle β_i . Potom

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

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

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

Derivujeme podle σ^2 . Potom

$$\begin{aligned}-\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \|Y - X\beta\|^2 &= 0, \\ \hat{\sigma}^2 = \frac{1}{n} \|Y - X\hat{\beta}\|^2 &= \frac{1}{n} \underbrace{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}_R = \frac{1}{n} R,\end{aligned}$$

kde R je reziduální součet čtverců. □

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

Věta 2.3. Platí, že

$$\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}) \quad \text{a} \quad \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2.$$

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

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

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

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

$$\begin{aligned}\mathbb{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T &= \mathbb{E}((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \varepsilon \varepsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}(\varepsilon \varepsilon^T) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}\end{aligned}$$

□

3 1-10

3.1 Jednorozměrná lineární regrese

Předpokládejme, že sledujeme dvě veličiny x a y mezi kterými existuje lineární závislost

$$y = \beta_0 + \beta_1 x, \quad \text{kde } \beta_0, \beta_1 \text{ neznáme.}$$

Provede se experiment a zjistí se hodnoty dvojic (x, y) . Často se stává, že x je změřeno prakticky zcela přesně.

POZNÁMKA 3.1. To nastává například v případě, kdy se x nastavuje na předem dané úrovni a následně se k němu změří odpovídající y .

Oproti tomu u y obvykle předpokládáme měření s chybou. Chyba může být náhodná a proto i y budeme chápat jako náhodnou veličinu, kterou budeme značit Y . Pro dvojice $(x_1, Y_1), \dots, (x_n, Y_n)$ se zavádí model

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{kde } i = 1, \dots, n.$$

Jednotlivé proměnné se pak nazývají následovně

- Y_i – vysvětlovaná (závislá) proměnná
- x_i – vysvětlující (nezávislá) proměnná, *popřípadě prediktor nebo regresor*
- β_0, β_1 – neznámé regresní parametry
- e_i – náhodný šum, (náhodná chyba)

Budeme předpokládat, že e_i jsou nezávislé (někdy bude dokonce stačit, aby byly nekorelované) a $e_i \sim (0, \sigma^2)$. A tedy splňuje $\mathbb{E}[e_i] = 0$, $D[e_i] = \sigma^2$ pro $\forall i$ (homoskedasticita).

Měřením získáme data $(x_1, y_1), \dots, (x_n, y_n)$ a cílem statistické analýzy je určit, zda model ?? schopen popsat pozorovanou variabilitu u y .

První krok

Odhadneme neznámé parametry $\beta_0, \beta_1, \sigma^2$. Proložíme data přímkou ve tvaru

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

a porovnáme y_i – naměřená data a $\hat{y}(x_i)$ – predikovaná hodnota lineární regrese pro $\forall i$. To nám umožňuje posoudit adekvátnost modelu.

Pro proložení dat přímkou existuje několik způsobů. Zásadní ovšem bude znalost rozdělení e_i a tedy i Y_i i když apriori není zřejmé proč znát rozdělení a ne β_0, β_1 .

Zde máme následující možnosti:

1. Odhadnout β_0, β_1 pomocí metody nezávislé na rozdělení chyb

2. Udělat věrohodnostní předpoklad o rozdělení chyb, odhadnout β_0, β_1 a následně ověřit předpoklad

POZNÁMKA 3.2. Speciální důležitý případ je $e_i \sim N(0, \sigma^2)$ který při MLE odhadu β_0, β_1 vede na metodu nejmenších čtverců, která může být použita bez ohledu na rozdělení chyb.

Odhady parametrů

3.1.1 Data s předpokladem normality dat

Předpokládáme, že e_1, \dots, e_n iid $N(0, \sigma^2)$. To znamená, že $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ a jednotlivé Y_1, \dots, Y_n jsou nezávislé.

MLE odhady

Věrohodnostní funkce je ve tvaru

$$L = L(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right)$$

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

pro pevné $\sigma^2 > 0$ je maximalizace l ekvivalentní s minimalizováním S , kde

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

Proto tuto metodu někdy nazýváme metodou nejmenších čtverců.

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

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

Z první rovnice pak dostaneme

$$\beta_0 = \frac{1}{n} \sum_{i=1}^n y_i - \beta_1 - \frac{1}{n} \sum_{i=1}^n x_i = \bar{y}_n - \beta_1 \bar{x}_n$$

a dosazením do druhé dostaneme výraz

$$\sum_{i=1}^n y_i x_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 = 0,$$

$$\sum_{i=1}^n y_i x_i - \bar{y}_n \sum_{i=1}^n x_i - \beta_1 \bar{x}_n \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 = 0.$$

Jednotlivé MLE odhady parametrů pak mají následující tvar

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n \quad a \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - n \bar{x}_n \bar{y}_n}{\sum_{i=1}^n x_i^2 - n \bar{x}_n^2}.$$

Nyní najdeme odhad parametru σ^2

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

vyjádřením σ^2 z rovnice dostaneme výraz

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \text{SSE},$$

kde $\hat{y}_i = \beta_0 + \beta_1 x_i$ je predikce modelu (odhad $\mathbb{E}[Y_i]$) a zkratka SSE je odvozena z anglického *sum of the squares of errors*. Rozdíl $\hat{e}_i = y_i - \hat{y}_i$ nazýváme i -té reziduum. Velikost reziduí indikuje, jak dobře odhadnutá přímka odpovídá datům. Rezidua jsou vlastně odhady chyb e_i , jejich analýza hraje významnou roli v ověření předpokladů rozdělení chyb.

POZNÁMKA 3.3. Pro odhad σ^2 se používá častěji statistika $s_n^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \text{SSE}$, která je nestranným odhadem parametru σ^2 (pro libovolné rozdělení e_i), zatímco σ_{MLE}^2 je vychýlený odhad i pro normální rozdělení chyb.

Odhad σ

pro odhad parametru σ využíváme statistiku nazývanou standardní chyba regrese (standard error), která má tvar

$$s_n = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

Tento odhad není nestranný.

3.1.2 Data bez předpokladu normality

Bez předpokladu normality chyb. Tedy, že e_1, \dots, e_n jsou nekorelované, $e_1, \dots, e_n \sim (0, \sigma^2)$. Pro odhad β_0, β_1 lze použít minimalizaci S (nejmenší čtverce), což je rozumné provedení, když si uvědomíme ?????? interpret??? (strana 5).

Nechť $y = \beta_0 + \beta_1 x$ je rovnice nějaké přímky, potom $y_i - (\beta_0 + \beta_1 x_i)$ je vertikální vzdálenost bodu (x_i, y_i) od přímky a

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

je míra udávající, jak dobře přímka prokládá data. Dává smysl vybrat takovou přímku, která minimalizuje S. Minimalizací S získáme stejné odhady $\hat{\beta}_0, \hat{\beta}_1$ jako u MLE odhadů pro normální data. Teď se ale nazývají odhad metodou nejmenších čtverců LSE (least squares estimators). Existuje více měr vhodnosti přímky. Použití LSE pro libovolné rozdělení chyb má dvě zdůvodnění.

1. pro normální rozdělení chyby LSE splývá s MLE.
2. LSE odhad je navíc BLUE (best linear unbiased estimator) jak ukážeme v Gauss–Markov theorem

PŘÍKLAD 3.4. Necht' e_1, \dots, e_n jsou *iid* s hustotou

$$f(\varepsilon) = \frac{1}{2}e^{-|\varepsilon|} \quad \text{Laplaceovo rozdělení}$$

potom hustota Y_i je

$$f_{Y_i}(y_i) = \frac{1}{2}e^{-|y_i - \beta_0 - \beta_1 x_i|}$$

a věrohodnostní funkce L a l mají tvar

$$L = \frac{1}{2^n} e^{-\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|}$$

$$l = -n \ln 2 - \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i|$$

MLE odhady parametrů β_0, β_1 získáme minimalizací

$$A = \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i| \quad \dots \quad \text{MAD (minimum absolute deviation).}$$

Zde budou odhady jiné než u LSE.

Uvažujme 3 body: $(0, 0), (1, 0), (\frac{1}{2}, \frac{1}{2})$.

$$\begin{aligned} \text{MLE:} \quad & \beta_0 = \beta_1 = 0 \quad , \quad A = 0.5 \quad , \quad \hat{y} = 0 \\ \text{LSE:} \quad & \bar{x} = \frac{1}{2}, \bar{y} = \frac{1}{6} \quad , \quad \sum_{i=1}^n x_i^2 = \frac{5}{4}, \sum_{i=1}^n x_i y_i = \frac{1}{4} \quad , \quad \beta_1 = 0, \beta_0 = \frac{1}{6} \end{aligned}$$

POZNÁMKA 3.5. I když s_n^2 je nestranný odhad σ^2 , s_n je vychýlený odhad σ ! Je to obecná vlastnost odhadů (nestranných) rozptylů, neboť s^2 nestranný odhad $\sigma^2 \Rightarrow \mathbb{E}[s] \leq \sigma$

Uvažujme náhodnou veličinu X pro kterou platí, že $D[X] < +\infty$

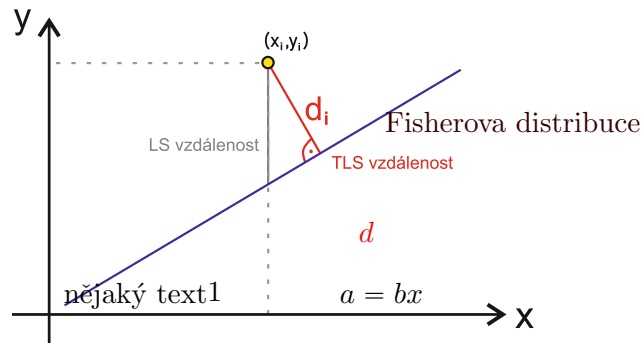
$$\begin{aligned} \mathbb{E}[X^2] &= D[X] + \mathbb{E}[X]^2 \quad \text{dosazením } X = s \quad \text{dostaneme} \\ \mathbb{E}[s^2] &= D[s] + \mathbb{E}[s]^2 \end{aligned}$$

$$\mathbb{E}[s]^2 \leq \sigma^2 \quad \mathbb{E}[s] \leq \sigma \quad (3.1)$$

a rovnost nastává pokud $D[s] = 0$.

Například pro normální chyby je $s_n^2 \propto \chi^2 \Rightarrow \mathbb{E}[s_n] < \sigma$

POZNÁMKA 3.6. předpokládali jsme, že hodnoty x_i jsou dány přesně, což nemusí být vždy pravda. Často obě veličiny (x, y) jsou měřeny nepřesně. EIV models "error in variable" v těchto modelech jsou často preferovány jiné odhady než LSE. Populární metoda: total least squares (ortogonal least squares). Zde minimalizujeme $\sum_{i=1}^n d_i^2$, kde d_i je minimální vzdálenost bodu a přímky (kolmice na přímku protínající bod). To znamená, že neupřednostňujeme veličinu x , ale přistupujeme k x a y rovnoměrně.



POZNÁMKA 3.7. v literatuře se někdy x uvažují jako realizace náhodné veličiny (ne vždy se x nastavuje předem, nebo je jasně dané (třeba pohlaví – ??? (8 strana))

Model má potom tvar

$$\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i \quad D[Y_i|X_i] = \sigma^2$$

pro většinu výsledků prezentovaných v této přednášce ale není podstatné, zde je x chápáno jako pevné nebo náhodné. Důkazy většinou fungují s podmíněnými výrazy (\mathbb{E}, D, \dots) při dané hodnotě x místo nepodmíněných. Nicméně větší pozornost je třeba u odvození asymptotických rozdělení odhadů.

3.1.3 Vlastnosti odhadů

Vlastnosti odhadů $\hat{\beta}_0, \hat{\beta}_1, s_n^2$.

Věta 3.8. Nechtě $\hat{\beta}_0, \hat{\beta}_1$ jsou LSE odhady parametrů β_0, β_1 v lineárním modelu

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, \dots, n,$$

kde e_i jsou nezávislé náhodné veličiny (postačí i nekorelovanost) se stejným rozptylem σ^2 . Potom platí:

1. $\mathbb{E}[\hat{\beta}_0] = \beta_0$, $\mathbb{E}[\hat{\beta}_1] = \beta_1$, (nestranné odhady)
2. $D[\hat{\beta}_0] = \frac{\sigma^2}{S_{xx}}$, kde $S_{xx} = \sum_{i=1}^n (x_i - \bar{x}_n)^2$
3. $D[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}} \right)$
4. Pokud navíc platí, že $e_i \sim \mathcal{N}(0, \sigma^2)$ $i = 1, \dots, n$ potom $\hat{\beta}_j \sim \mathcal{N}(\beta_j, D[\hat{\beta}_j])$ $j = 0, 1$

Důkaz

1. upravíme $\hat{\beta}_1$

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i x_i - n \bar{x}_n \bar{y}_n}{\sum_{i=1}^n x_i^2 - n \bar{x}_n^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = \\ &= \frac{1}{S_{xx}} \left(\sum_{i=1}^n (x_i - \bar{x}_n) y_i - \bar{y}_n \sum_{i=1}^n (x_i - \bar{x}_n) \right) = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) y_i \end{aligned}$$

potom má střední hodnota $\hat{\beta}_1$ tvar

$$\begin{aligned}\mathbb{E}[\hat{\beta}_1] &= \mathbb{E}\left[\frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) Y_i\right] = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) \mathbb{E}[Y_i] = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) (\beta_0 + \beta_1 x_i) = \\ &= \frac{\beta_0}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) + \frac{\beta_1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) x_i = 0 + \frac{\beta_1}{S_{xx}} S_{xx} = \beta_1\end{aligned}$$

a střední hodnota pro $\hat{\beta}_0$ má tvar

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{Y}_n - \hat{\beta}_1 \bar{x}_n] = \mathbb{E}[\bar{Y}_n] - \bar{x}_n \mathbb{E}[\hat{\beta}_1] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i] - \bar{x}_n \beta_1 = \beta_0 + \frac{\beta_1}{n} \sum_{i=1}^n x_i - \bar{x}_n \beta_1 = \beta_0$$

2.

$$D[\hat{\beta}_1] = D\left[\frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) Y_i\right] = \frac{1}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 D[Y_i] = \frac{\sigma^2 S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$$

3.

$$\begin{aligned}D[\hat{\beta}_0] &= D[\bar{Y}_n - \hat{\beta}_1 \bar{x}_n] = D[\bar{Y}_n] + \bar{x}_n^2 D[\hat{\beta}_1] - 2\bar{x}_n \text{cov}(\bar{Y}_n, \hat{\beta}_1) = \\ &= \frac{\sigma^2}{n} + \frac{\bar{x}_n^2 \sigma^2}{S_{xx}} - 2\bar{x}_n \text{cov}(\bar{Y}_n, \hat{\beta}_1) \\ \text{cov}(\bar{Y}_n, \hat{\beta}_1) &= \text{cov}\left(\bar{Y}_n, \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) Y_i\right) = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) \text{cov}(\bar{Y}_n, Y_i) \\ \text{cov}(\bar{Y}_n, Y_i) &= \text{cov}\left(\frac{1}{n} \sum_{j=1}^n Y_j, Y_i\right) = \frac{1}{n} \sum_{j=1}^n \text{cov}(Y_j, Y_i) = \frac{1}{n} \text{cov}(Y_i, Y_i) = \frac{1}{n} D Y_i = \frac{\sigma^2}{n} \\ &\Rightarrow \text{cov}(\bar{Y}_n, \hat{\beta}_1) = 0 = \frac{\sigma^2}{n S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n)\end{aligned}$$

4 11-20

Věta 4.1. *Za předpokladu předchozí věty platí*

$$\mathbb{E}(s_n^2) = \sigma^2$$

(s_n^2 je nestranný odhad σ^2).

Důkaz.

$$\mathbb{E}(s_n^2) = \frac{1}{n-2} \mathbb{E} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n-2} \underbrace{\sum_{i=1}^n \mathbb{E}(Y_i - \hat{Y}_i)^2}_{\text{ozn. } A}$$

Protože $\mathbb{E}(\hat{Y}_i) = \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i = \mathbb{E}Y_i$, platí, že:

$$\mathbb{E}(Y_i - \hat{Y}_i)^2 = D(Y_i - \hat{Y}_i) = \mathbb{E}(Y_i - \hat{Y}_i)^2 - \underbrace{(\mathbb{E}(Y_i - \hat{Y}_i))^2}_{=0}$$

Dostáváme tak

$$\begin{aligned} A &= \sum_{i=1}^n D(Y_i - \hat{Y}_i) = \sum_{i=1}^n [D(Y_i) + D(\hat{Y}_i) - 2\text{Cov}(Y_i, \hat{Y}_i)] = \\ &= n\sigma^2 + \sum_{i=1}^n D(\hat{Y}_i) - 2 \sum_{i=1}^n \text{Cov}(Y_i, \hat{Y}_i) \end{aligned} \quad (\#)$$

Rozepíšeme

$$D\hat{Y}_i = D(\hat{\beta}_0 + \hat{\beta}_1 x_i) = D\hat{\beta}_0 + x_i^2 D\hat{\beta}_1 + 2x_i,$$

kde

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}(\hat{Y}_n - \hat{\beta}_1 \bar{x}_n, \hat{\beta}_1) = \underbrace{\text{Cov}(\hat{Y}_n, \hat{\beta}_1)}_{=0 \text{ (viz. dříve)}} - \bar{x}_n \underbrace{D(\hat{\beta}_1)}_{\frac{\sigma^2}{s_{xx}}} = -\frac{\sigma^2 \bar{x}_n}{s_{xx}}$$

a tedy

$$\begin{aligned} D\hat{Y}_i &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}} + x_i^2 \frac{1}{s_{xx}} - \frac{2x_i \bar{x}_n}{s_{xx}} \right] = \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{s_{xx}} \right] \\ \sum_{i=1}^n D\hat{Y}_i &= \sigma^2 + \frac{\sigma^2}{s_{xx}} \underbrace{\sum_{i=1}^n (x_i - \bar{x}_n)^2}_{=s_{xx}} = 2\sigma^2 \end{aligned}$$

Následně máme

$$\mathbb{Cov}(Y_i, \hat{Y}_i) = \mathbb{Cov}(Y_i, \hat{\beta}_0 + \hat{\beta}_1 x_i) = \mathbb{Cov}(Y_i, \hat{\beta}_0) + x_i \mathbb{Cov}(Y_i, \hat{\beta}_1)$$

$$\mathbb{Cov}(Y_i, \hat{\beta}_1) = \frac{1}{s_{xx}} \sum_{j=1}^n (x_j - \bar{x}_n) \underbrace{\mathbb{Cov}(Y_i, Y_j)}_{=0 \text{ pro } i \neq j} = \frac{\sigma^2(x_i - \bar{x}_n)}{s_{xx}}$$

$$\mathbb{Cov}(Y_i, \hat{\beta}_0) = \mathbb{Cov}(Y_i, \bar{Y}_n - \bar{x}_n \hat{\beta}_1) = \mathbb{Cov}(Y_i, \bar{Y}) - \bar{x}_n \mathbb{Cov}(Y_i, \hat{\beta}_1) = \frac{\sigma^2}{n} - \frac{\bar{x}_n \sigma^2(x_i - \bar{x}_n)}{s_{xx}}$$

a tedy

$$\begin{aligned} \mathbb{Cov}(Y_i, \hat{Y}_i) &= \frac{\sigma^2}{n} - \frac{\bar{x}_n \sigma^2(x_i - \bar{x}_n)}{s_{xx}} + \frac{x_i \sigma^2(x_i - \bar{x}_n)}{s_{xx}} = \frac{\sigma^2}{n} + \frac{\sigma^2}{s_{xx}}(x_i - \bar{x}_n)^2 \\ \sum_{i=1}^n \mathbb{Cov}(Y_i, \hat{Y}_i) &= \sigma^2 + \frac{\sigma^2}{s_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = 2\sigma^2 \end{aligned}$$

Dosazením do (#) dostaneme

$$A = n\sigma^2 + 2\sigma^2 - 4\sigma^2$$

a celkem máme

$$\mathbb{E}(s_n^2) = \frac{1}{n-2} A = \sigma^2.$$

□

Tvrzení 4.2. *Nechť platí předpoklady věty 1 a necht' e_1, \dots, e_n iid $\mathcal{N}(0, \sigma^2)$. Potom platí:*

a) $\frac{(n-2)s_n^2}{\sigma^2} \sim \chi(n-2)$

b) s_n^2 je nezávislé na $\hat{\beta}_0$ a $\hat{\beta}_1$.

Důkaz. Vyplýne z obecnějších tvrzení pro vícerozměrnou regresi.

□

POZNÁMKA 4.3. Spočetli jsme

$$\underbrace{D(\hat{\beta}_0)}_{\text{ozn. } \sigma^2(\hat{\beta}_0)} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}} \right] \quad \text{a} \quad \underbrace{D(\hat{\beta}_1)}_{\text{ozn. } \sigma^2(\hat{\beta}_1)} = \frac{\sigma^2}{s_{xx}}$$

Nestranné odhady jsou:

$$\begin{aligned} \sigma^2(\hat{\beta}_0) &= s_n^2 \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}} \right] = s_n^2 \delta_0 \\ \sigma^2(\hat{\beta}_1) &= \frac{s_n^2}{s_{xx}} = s_n^2 \delta_1, \end{aligned}$$

kde δ_0 a δ_1 jsou tzv. variance multiplication factors.

Odhady směrodatné odchylky veličin $\hat{\beta}_0$ a $\hat{\beta}_1$ pak jsou

$$\hat{\sigma}(\hat{\beta}_0) = s_n \sqrt{\delta_0} \quad \text{a} \quad \hat{\sigma}(\hat{\beta}_1) = s_n \sqrt{\delta_1},$$

kterým se pak říká standardní chyby odhadů $\hat{\beta}_0$ a $\hat{\beta}_1$. Hrají zásadní roli při konstrukci IS a TH.

4.1 Gauss - Markov theorem

- Chyby normální \Rightarrow LSE pro $\hat{\beta}_0, \hat{\beta}_1$ je MLE ... parametrů (eficientní odhad)
- Pokud nejsou chyby normální, jaké je opodstatnění použít LSE?
Ukážeme, že LSE jsou BLUE (best linear unbiased estimators), tedy lineární nestranné odhady s minimálním rozptylem
- Je ale třeba poznamenat, že můžou existovat nelineární nebo vychýlené odhady parametrů β_0, β_1 , které jsou eficientnější než LSE, pokud se rozdělení chyb liší výrazně od normálního (tím se zabývá robustní regresní analýza).

Uvažujme model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \dots, n \quad (*)$$

Definice 4.4. Lineární odhad parametru β je statistika tvaru

$$\hat{\beta} = \sum_{i=1}^n c_i Y_i,$$

kde c_i jsou dané reálné konstanty a $i = 1, \dots, n$.

Věta 4.5. Necht' e_1, \dots, e_n v modelu $(*)$ jsou nekorelované a mají stejný rozptyl $D(e_i) = \sigma^2, i = 1, \dots, n$. Potom LSE $\hat{\beta}_j, j = 0, 1$ je BLUE parametru β_j .

Důkaz. Ukážeme pro β_1 , pro β_0 je důkaz podobný.

Necht' $\hat{\beta}_1 = \sum_{i=1}^n c_i Y_i$, pak

$$D\hat{\beta}_1 = \sum_{i=1}^n c_i^2 D Y_i = \sigma^2 \sum_{i=1}^n c_i^2$$

Aby byl $\hat{\beta}_1$ nestranný, musí platit $\mathbb{E}\hat{\beta}_1 = \beta_1$, tedy

$$\mathbb{E}\hat{\beta}_1 = \sum_{i=1}^n c_i \mathbb{E}Y_i = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i \stackrel{!}{=} \beta_1$$

protože to musí platit pro lib. β_0, β_1 , dostáváme

$$\sum_{i=1}^n c_i = 0 \quad \text{a} \quad \sum_{i=1}^n c_i x_i = 1.$$

Hledání lineárního, nestranného odhadu β_1 je tedy redukováno na minimalizaci $\sum_{i=1}^n c_i^2$ za vazebných podmínek $\sum_{i=1}^n c_i = 0$ a $\sum_{i=1}^n c_i x_i = 1$.

Lagrangeova funkce: $L = \sum_{i=1}^n c_i^2 - 2\lambda_1 \left(\sum_{i=1}^n c_i \right) - 2\lambda_2 \left(\sum_{i=1}^n c_i x_i - 1 \right)$.

$$\frac{\partial L}{\partial c_i} = 2c_i - 2\lambda_1 - 2\lambda_2 x_i = 0, \quad i = 1, \dots, n$$

$$\frac{\partial L}{\partial \lambda_1} = -2 \left(\sum_{i=1}^n c_i \right) = 0$$

$$\frac{\partial L}{\partial \lambda_2} = -2 \left(\sum_{i=1}^n c_i x_i - 1 \right) = 0$$

Sečteme prvních n rovnic

$$\underbrace{\sum_{i=1}^n c_i}_{=0} - n\lambda_1 - \lambda_2 \sum_{i=1}^n x_i = 0 \Rightarrow n\lambda_1 + \lambda_2 \sum_{i=1}^n x_i = 0 \Rightarrow \lambda_1 = -\lambda_2 \bar{x}_n$$

Sečteme dále prvních n rovnic vynásobených x_i :

$$\begin{aligned} \sum_{i=1}^n c_i x_i - \lambda_1 \sum_{i=1}^n x_i - \lambda_2 \sum_{i=1}^n x_i^2 &= 0 \\ \Rightarrow \lambda_1 \sum_{i=1}^n x_i + \lambda_2 \sum_{i=1}^n x_i^2 &= 1 \\ -\lambda_2 \bar{x}_n \cdot n\bar{x}_n + \lambda_2 \sum_{i=1}^n x_i^2 &= 1 \\ \lambda_2 \left(\sum_{i=1}^n x_i^2 - n\bar{x}_n^2 \right) &= 1 \Rightarrow \lambda_2 = \frac{1}{s_{xx}} \quad \text{a} \quad \lambda_1 = -\frac{\bar{x}_n}{s_{xx}} \end{aligned}$$

Dosadíme za λ_1, λ_2 :

$$c_i + \frac{\bar{x}_n}{s_{xx}} - \frac{x_i}{s_{xx}} = 0 \Rightarrow c_i = \frac{x_i - \bar{x}_n}{s_{xx}}$$

a $\hat{\beta}_1 = \frac{1}{s_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) Y_i$, což je LSE.

□

POZNÁMKA 4.6. Ukázali jsme pouze, že to je stacionární bod, že je tam i minimum ukážeme v obecnější větě ve vícerozměrné regresi.

4.2 IS pro β_0, β_1

- IS poskytují jistou "míru přesnosti" bodových odhadů
- pro jejich konstrukci potřebujeme znát rozdělení pravděpodobnosti bodového odhadu
- budeme tedy uvažovat normalitu chyb
- spočtené IS se ale často používají, i když rozdělení chyb není normální, jejich použití se zdůvodňuje tím, že LSE odhady par. β jsou lineární funkcí $Y_i, i = 1, \dots, n$, což umožňuje aplikovat CLT a dostat asymptotickou normalitu odhadů $\hat{\beta}_0, \hat{\beta}_1$

Uvažujme model $Y_i = \beta_0 + \beta_1 x_i + e_i$, e_i i.i.d $\mathcal{N}(0, \sigma^2)$. Víme:

$$\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2(\hat{\beta}_i)), \quad \frac{(n-2)s_n^2}{\sigma^2} \sim \chi^2(n-1) \quad \text{a nezávisí na } \hat{\beta}_0, \hat{\beta}_1.$$

POZNÁMKA 4.7.

$$X \sim \mathcal{N}(0, 1), Y \sim \chi^2(n), X, Y \text{ nezávislé} \Rightarrow \frac{X}{\sqrt{Y/n}} \sim t(n)$$

Tedy

$$T_i = \frac{\frac{\hat{\beta}_i - \beta_i}{\sigma(\hat{\beta}_i)}}{\frac{s_n}{\sigma}} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}}(\hat{\beta}_i) \sim t(n-2, i=0, 1)$$

neboť $\sigma(\hat{\beta}_i) = \sigma\sqrt{\delta_i}$ a $\hat{\sigma}(\hat{\beta}_i) = s_n\sqrt{\delta_i}$.

Tzn. $P\left[-t_{1-\alpha/2}(n-2) \leq \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}}(\hat{\beta}_i) \leq t_{1-\alpha/2}(n-2)\right]$ a vyjádřením β_i dostaneme

$$P\left[\hat{\beta}_i - t_{1-\alpha/2}(n-2)\hat{\sigma}(\hat{\beta}_i) \leq \beta_i \leq \hat{\beta}_i + t_{1-\alpha/2}(n-2)\hat{\sigma}(\hat{\beta}_i)\right] = 1 - \alpha$$

a tedy $(\hat{\beta}_i \pm t_{1-\alpha/2}(n-2)\hat{\sigma}(\hat{\beta}_i))$ je $100(1-\alpha)\%$ IS pro $\beta_i, i=0, 1$.

Dosazením za $\hat{\sigma}(\hat{\beta}_i)$ dostaneme

- $100(1-\alpha)\%$ IS pro β_0 : $\hat{\beta}_0 \pm t_{1-\alpha/2}(n-2) \cdot s_n \sqrt{\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}}$
- $100(1-\alpha)\%$ IS pro β_1 : $\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2) \cdot s_n \frac{1}{\sqrt{s_{xx}}}$

POZNÁMKA 4.8. Z tvarů IS lze pozorovat, že IS pro β_0 bude ve většině praktických případů širší než IS pro β_1 , tzn. směrnice je obecně odhadnuta s větší přesností než absolutní člen (intercept).

POZNÁMKA 4.9. Někdy se konstruují simultánní IS pro oba parametry. (Obr?) Zmíníme podrobněji u vícerozměrné regrese.

4.3 TH pro β_0, β_1

Chtěli bychom ověřit platnost předpokladu lineárního vztahu mezi x a y .

Předpokládejme nyní, že model je lineární a že x je jediná dostupná vysvětlující proměnná. Otázkou je, zda je x užitečná ve vysvětlení variability v y , chceme tedy rozhodnout mezi dvěma modely:

$$Y_i = \beta_0 + e_i \quad \text{a} \quad Y_i = \beta_0 + \beta_1 x_i + e_i$$

tzn. otestovat hypotézu $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

Pokud nezamítneme H_0 , závěr bude, že x nevysvětluje nic z variability y a není v modelu významné. Pokud zamítneme H_0 , znamená to, že x je významné.

POZNÁMKA 4.10. Tyto závěry jsou správné pouze za předpokladu, že model je lineární!

- nezamítnutí H_0 nemusí znamenat, že x není užitečná, může to pouze indikovat, že vztah mezi y a x není lineární
- zamítnutí H_0 naopak říká, že existuje lineární trend mezi x a y , ale mohou tam být i jiné typy závislosti

Pro konstrukci testů využijeme odvozené IS.

POZNÁMKA 4.11. Opakování: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0 \Rightarrow (\underline{\theta}, \bar{\theta})$ je $100(1-\alpha)\%$ IS pro θ . Pak $W = \{x | \theta_0 \notin (\underline{\theta}, \bar{\theta})\}$ je kritický obor test na hladině α .

$H_0 : \beta_1 = 0$ zamítneme, pokud $0 \notin \left(\hat{\beta}_1 \pm t_{1-\alpha/2}(n-2) \cdot \frac{s_n}{\sqrt{s_{xx}}} \right)$, tzn.

$$\text{buď } \hat{\beta}_1 + t_{1-\alpha/2}(n-2) \cdot \frac{s_n}{\sqrt{s_{xx}}} < 0 \iff \hat{\beta}_1 \frac{\sqrt{s_{xx}}}{s_n} < -t_{1-\alpha/2}(n-2)$$

$$\text{nebo } \hat{\beta}_1 - t_{1-\alpha/2}(n-2) \cdot \frac{s_n}{\sqrt{s_{xx}}} > 0 \iff \hat{\beta}_1 \frac{\sqrt{s_{xx}}}{s_n} > t_{1-\alpha/2}(n-2)$$

A zapsáno dohromady

$$|T_n| = |\hat{\beta}_1| \frac{\sqrt{s_{xx}}}{s_n} > t_{1-\alpha/2}(n-2).$$

POZNÁMKA 4.12. Intuitivní interpretace: $|T_n| = |\hat{\beta}_1| \frac{\sqrt{s_{xx}}}{s_n} = \frac{|\hat{\beta}_1|}{\hat{\sigma}(\hat{\beta}_1)}$ je převrácená hodnota relativní chyby.

Pokud je β_1 dobře odhadnuto, očekáváme malý rozptyl $\hat{\sigma}(\hat{\beta}_1)$, tedy T bude velké. t-test tedy říká, že zamítneme H_0 , pokud je relativní chyba odhadu malá.

5 21-30

POZNÁMKA 5.1. 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

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

5.0.1 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 data, než $y = \beta_0 + \beta_1 x + e$. H_0 potom zamítneme, pokud

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

5.0.2 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, \dots, y_n) je "vysvětleno" regresním modelem (přímkou). Míru variability v datech pak spočítáme jako podíl součtu sum od regrese a celkového počtu čtverců, tedy

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

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

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

Ukážeme, že $\bar{\hat{y}} = \bar{y}_n$ a tak

$$\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}}_n)^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 = \text{SSR}$$

regresi sum of squares, regresní součet čtverců. Podíl

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2}$$

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

R^2 - *koeficient determinace (coefficient of determination)* (pro každý model by měl mít hodnotu $R^2 \approx 1$). Ukážeme, že R^2 je kvadrát výběrového korelačního koeficientu mezi \mathbf{x} a \mathbf{y} , což dává statistice R^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 $R^2 \approx 1$. Protože každá monotónní funkce R^2 vede na ekvivalentní test, budeme uvažovat statistiku

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

Lemma 5.2. *Nechť $\hat{e}_i = y_i - \hat{y}_i$ značí rezidua, kde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ a $\hat{\beta}_0, \hat{\beta}_1$ jsou LSE. Potom*

1. $\sum_{i=1}^n \hat{e}_i = 0$,
2. $\bar{\hat{y}}_n = \bar{y}_n$,
3. $\sum_{i=1}^n \hat{e}_i \hat{y}_i = 0$.

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

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

2. Z bodu 1) plyne, že $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$, dělením n dostaneme dokazované tvrzení.

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

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

a tedy

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

□

Věta 5.3. *Předpokládejme, že $SST \neq 0$. Potom platí*

1. $0 \leq R^2 \leq 1$,
2. $R^2 = 1 - \frac{SSE}{SST}$, kde $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ jako reziduální součet čtverců,
3. $R^2 = 1 \Leftrightarrow (\forall i \in \hat{n})(\hat{y}_i = y_i)$ (všechna data leží na přímce),

4. pokud označíme $\mathbf{x} = (x_1, \dots, x_n)$ a $\mathbf{y} = (y_1, \dots, y_n)$, potom $R^2 = \varrho^2(\mathbf{x}, \mathbf{y})$, kde

$$\varrho(\mathbf{x}, \mathbf{y}) = \frac{\left(\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n) \right)^2}{S_{xx}S_{yy}}$$

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

5. $F = \frac{SSR}{s_n^2} = T^2$,

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

Důkaz. Důkaz věty bude založen na rozkladu

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

neboli $SST = SSR + SSE$. Z lemmatu 5.2 vyplývá, že

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

neboť

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

Z toho potom dokazujeme jednotlivé body věty.

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

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

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

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

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

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

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

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

5. Z definice F plyne, že

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

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

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

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

□

POZNÁMKA 5.4. 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	$\frac{MSR}{MSE}$
Residual	$n-2$	SSE	$MSE = \frac{SSE}{n-2} = s_n^2$	
Total	$n-1$	SST		

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

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

POZNÁMKA 5.5. $H_0 : \beta_1 = 0$ je zamítnul, pokud $F > F_{1-\alpha}(1, n-2)$. V tomto jednorozměrném případě je to ekvivalentní t -testu, neboť $F = T^2$.

Věta 5.6. Mějme e_1, \dots, e_n iid $\mathcal{N}(0, \sigma^2)$. Za platnosti $H_0 : \beta_1 = 0$ je splněno, že

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

POZNÁMKA 5.7. Proto v tabulce ANOVA 5.0.2 uvádí df po řadě 1, $n-2$, $n-1$. Používají se však i v případě jiného rozdělení chyb. Představit si je lze takto:

1. $SSE = \sum_{i=1}^n \hat{e}_i^2$, na n -rezidní $\hat{e}_1, \dots, \hat{e}_n$ máme 2 podmínky $\sum_{i=1}^n \hat{e}_i = 0$ a $\sum_{i=1}^n x_i \hat{e}_i = 0$. Z toho vyplývá, že mají $n-2$ stupňů volnosti.
2. $SST = \sum_{i=1}^n (y_i - \bar{y}_n)^2$... $y_i - \bar{y}_n$ musí splňovat $\sum_{i=1}^n (y_i - \bar{y}_n) = 0$, a proto má $n-1$ stupňů volnosti.

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

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

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

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

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

□

POZNÁMKA 5.8. 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

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

což je rostoucí funkce S_{xx} .

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

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

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 \bar{y}_n .

Další kritéria budeme probírat později.

PŘÍKLAD 5.9. 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 - $y_i = \beta_0 + \beta_1 x_i + e_i$.

$$\hat{\beta}_0 = 4.5822, \quad \hat{\beta}_1 = 2.2042, \quad R^2 = 0.984,$$

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.