

# **01RAD**

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

26. listopadu 2020

# Obsah

<b>1 SME – regresní analýza</b>	<b>1</b>
1.1 Jednorozměrná lineární regrese . . . . .	1
1.2 Intervaly predikce . . . . .	5
1.3 Vícerozměrná lineární regrese . . . . .	6
<b>2 Jednorozměrná lineární regrese</b>	<b>9</b>
2.1 Data s předpokladem normality dat . . . . .	10
2.2 Data bez předpokladu normality . . . . .	11
2.3 Vlastnosti odhadů . . . . .	13
2.4 Gauss - Markov theorem . . . . .	17
2.5 IS pro $\beta_0, \beta_1$ . . . . .	18
2.6 TH pro $\beta_0, \beta_1$ . . . . .	19
2.6.1 Test významnosti interceptu . . . . .	20
2.7 ANOVA přístup pro testování . . . . .	20
2.8 Regrese skrz počátek . . . . .	25
2.8.1 Odhadování a testy v případě $\beta_0 = 0$ . . . . .	26
2.8.2 Ad 1 . . . . .	28
2.8.3 Ad 2 . . . . .	29
2.8.4 Ad 3 - Analýza reziduí . . . . .	31
2.9 Grafy reziduí . . . . .	32
2.9.1 Vlastnosti vektoru reziduí $\hat{e}$ . . . . .	39
2.9.2 Gauss - Markov theorem . . . . .	40
2.9.3 Testování modelu - tabulka ANOVA . . . . .	41
<b>3 IS a t-testy pro parametry</b>	<b>46</b>
3.0.1 Obecná lineární hypotéza . . . . .	47
3.0.2 Predikce . . . . .	49
3.1 Rezidua, diagnostika a transformace . . . . .	51
3.1.1 Rezidua . . . . .	51
3.1.2 Grafy reziduí . . . . .	53
3.1.3 Partial residual plot . . . . .	54
3.2 Míry influence . . . . .	60
3.3 Transformace . . . . .	62
3.3.1 Transformace vysvětlované proměnné $y$ . . . . .	62
3.4 Korelované chyby . . . . .	64
3.4.1 Durbin-Watson statistika . . . . .	65

# 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 SME – regresní analýza

## 1.1 Jednorozměrná lineární regrese

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

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

$\beta_0$  a  $\beta_1$  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

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

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  $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$ , 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  $\sigma^2$  potřebujeme minimalizovat

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

přes  $\beta_0, \beta_1$ , 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 SME – 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

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

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 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

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

Pokud dále označíme

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

pak rozdíly

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

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

## 1 SME – 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

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

Konfidenční intervaly

1.  $\sigma^2$ , 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  $\sigma^2$  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 SME – regresní analýza

2.  $\beta_1$

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  $100(1 - \alpha)\%$  IS pro  $\beta_1$ . Podobně pro  $\beta_0$  dostaneme, že

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

a tedy

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

je  $100(1 - \alpha)\%$  IS pro  $\beta_0$ .

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

## 1 SME – regresní analýza

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

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

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i = 30, \quad \bar{Y}_n = 348, \quad \sum_{i=1}^n X_i Y_i = 57140, \quad \sum_{i=1}^n X_i^2 = 13300,$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - 5\bar{X}_n \bar{Y}_n}{\sum_{i=1}^n X_i^2 - 5\bar{X}_n^2} = 0.561,$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n = 331.16,$$

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

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

Spočítáme IS například pro  $\beta_1$ . Dostaneme tedy  $t_{0.975}(5-2) = 3.18$ , který dosadíme do vzorečku na výpočet IS pro  $\beta_1$ , 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 ( $\beta_0 \sim \mathcal{N}(\dots)$ ,  $\beta_1 \sim \mathcal{N}(\dots)$ ,  $e \sim \mathcal{N}(\dots)$ ,  $Y \sim \mathcal{N}(\dots)$ ) stačí tedy určit střední hodnotu a rozptyl.

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

Protože nový pár  $(X, Y)$  je nezávislý na předchozích datech, platí, že  $Y$  je nezávislé na  $\hat{Y}$  ( $\beta_0, \beta_1$  jsou spočteny pouze pomocí  $Y_1, \dots, Y_n$ ). Pak tedy

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

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

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

## 1 SME – 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  $X = 35^\circ C$  a  $\hat{Y} = 331.16 + 0.561 \cdot 35 = 350.8$ .

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

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

### 1.3 Vícerozměrná lineární regrese

Předpokládejme model

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

kde  $\varepsilon_1, \dots, \varepsilon_n$  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  $\beta$  z  $\mathbf{X}$  a  $\mathbf{Y}$  ani kdyby nebyl přítomný šum  $\varepsilon$ . (Vlastně bychom měli soustavu  $\mathbf{X}\beta = \mathbf{Y}$ .)

**Poznámka 1.4.** 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  $\beta, \sigma^2$ .

**Věta 1.5.** Pro MLE parametrů  $\beta$  a  $\sigma^2$  platí, že

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

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

## 1 SME – regresní analýza

*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

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

Je třeba minimalizovat

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

Derivujeme podle  $\beta_i$ . 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  $\sigma^2$ . 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.  $\square$

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

**Věta 1.6.** Platí, že

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

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

## 1 SME – regresní analýza

*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

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

□

## 2 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 2.1.** 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

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

Jednotlivé proměnné se pak nazývají následovně

- $Y_i$  – vysvětlovaná (závislá) proměnná
- $x_i$  – vysvětlující (nezávislá) proměnná, *popřípadě prediktor nebo regresor*
- $\beta_0, \beta_1$  – 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 tady i  $Y_i$  i když apriori není zřejmé proč znát rozdělení a ne  $\beta_0, \beta_1$ .

Zde máme následující možnosti:

1. Odhadnout  $\beta_0, \beta_1$  pomocí metody nezáviselých na rozdělení chyb
2. Udělat věrohodnostní předpoklad o rozdělení chyb, odhadnout  $\beta_0, \beta_1$  a následně ověřit předpoklad

## 2 Jednorozměrná lineární regrese

**Poznámka 2.2.** Speciální důležitý případ je  $e_i \sim N(0, \sigma^2)$  který při MLE odhadu  $\beta_0, \beta_1$  vede na metodu nejmenších čtverců, která může být použita bez ohledu na rozdělení chyb.

### Odhady parametrů

#### 2.1 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

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

## 2 Jednorozměrná lineární regrese

Nyní najdeme odhad parametru  $\sigma^2$

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

vyjádřením  $\sigma^2$  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 reziduů 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 2.3.** Pro odhad  $\sigma^2$  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  $\sigma^2$  (pro libovolné rozdělení  $e_i$ ), zatímco  $\sigma_{\text{MLE}}^2$  je vychýlený odhad i pro normální rozdělení chyb.

### Odhad $\sigma$

pro odhad parametru  $\sigma$  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.

## 2.2 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  $\beta_0, \beta_1$  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. Ted' 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

## 2 Jednorozměrná lineární regrese

PŘÍKLAD 2.4. Nechť  $e_1, \dots, e_n$  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 odhad parametrů  $\beta_0, \beta_1$  získáme minimalizací

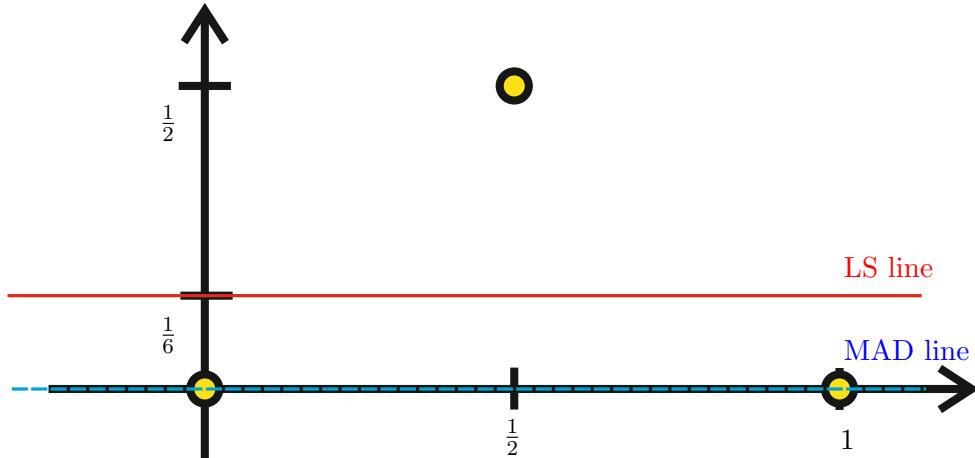
$$A = \sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i| \quad \dots \quad \text{MAD (minimum absolute deviation).}$$

Zde budou odhady jiné než u LSE.

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

$$\text{MLE: } \beta_0 = \beta_1 = 0, \quad A = 0.5, \quad \hat{y} = 0$$

$$\text{LSE: } \bar{x} = \frac{1}{2}, \bar{y} = \frac{1}{6}, \quad \sum_{i=1}^n x_i^2 = \frac{5}{4}, \quad \sum_{i=1}^n x_i y_i = \frac{1}{4}, \quad \beta_1 = 0, \beta_0 = \frac{1}{6}$$



POZNÁMKA 2.5. I když  $s_n^2$  je nestranný odhad  $\sigma^2$ ,  $s_n$  je vychýlený odhad  $\sigma$ ! 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  $D[X] < +\infty$

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

$$\mathbb{E}[s^2] = D[s] + \mathbb{E}[s]^2$$

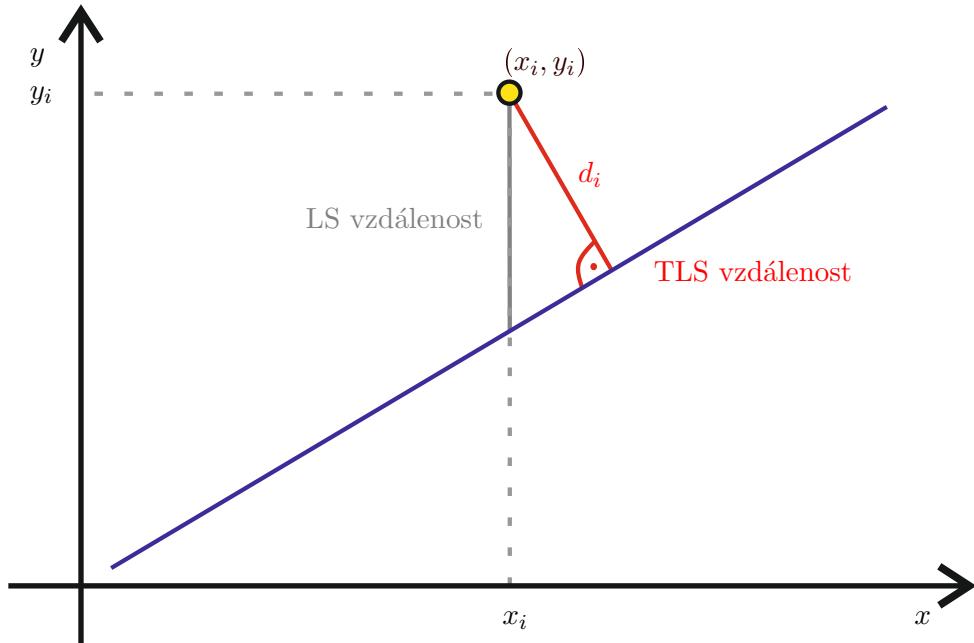
$$\mathbb{E}[s]^2 \leq \sigma^2 \quad \mathbb{E}[s] \leq \sigma \tag{2.1}$$

a rovnost nastává pokud  $D[s] = 0$ .

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

## 2 Jednorozměrná lineární regrese

Poznámka 2.6. 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 2.7. 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

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

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

### 2.3 Vlastnosti odhadů

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

**Věta 2.8.** Nechť  $\hat{\beta}_0, \hat{\beta}_1$  jsou LSE odhady parametrů  $\beta_0, \beta_1$  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  $\sigma^2$ . Potom platí:

## 2 Jednorozměrná lineární regrese

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.

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

## 2 Jednorozměrná lineární regrese

**Věta 2.9.** Za předpokladu předchozí věty platí

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

tedy  $s_n^2$  je nestranný odhad  $\sigma^2$ .

Důkaz.

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

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

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

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

$$\begin{aligned} \text{Cov}(Y_i, \hat{Y}_i) &= \text{Cov}(Y_i, \hat{\beta}_0 + \hat{\beta}_1 x_0) = \text{Cov}(Y_i, \hat{\beta}_0) + x_0 \text{Cov}(Y_i, \hat{\beta}_1) \\ \text{Cov}(Y_i, \hat{\beta}_1) &= \frac{1}{s_{xx}} \sum_{j=1}^n (x_j - \bar{x}_n) \underbrace{\text{Cov}(Y_i, Y_j)}_{=0 \text{ pro } i \neq j} = \frac{\sigma^2(x_i - \bar{x}_n)}{s_{xx}} \\ \text{Cov}(Y_i, \hat{\beta}_0) &= \text{Cov}(Y_i, \bar{Y}_n - \bar{x}_n \hat{\beta}_1) = \text{Cov}(Y_i, \bar{Y}) - \bar{x}_n \text{Cov}(Y_i, \hat{\beta}_1) = \frac{\sigma^2}{n} - \frac{\bar{x}_n \sigma^2 (x_i - \bar{x}_n)}{s_{xx}} \end{aligned}$$

## 2 Jednorozměrná lineární regrese

a tedy

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

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 2.10.** Nechť platí předpoklady věty 1 a nechť  $e_1, \dots, e_n$  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 2.11. 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  $\delta_0$  a  $\delta_1$  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.

## 2.4 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  $\beta_0, \beta_1$ , 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 2.12.** Lineární odhad parametru  $\beta$  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 2.13.** Nechť  $e_1, \dots, e_n$  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  $\beta_j$ .

*Důkaz.* Ukážeme pro  $\beta_1$ , pro  $\beta_0$  je důkaz podobný.

Nechť  $\hat{\beta}_1 = \sum_{i=1}^n c_i Y_i$ , 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  $E\hat{\beta}_1 = \beta_1$ , tedy

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

protože to musí platit pro lib.  $\beta_0, \beta_1$ , 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  $\beta_1$  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:  $L = \sum_{i=1}^n c_i^2 - 2\lambda_1 (\sum_{i=1}^n c_i) - 2\lambda_2 (\sum_{i=1}^n c_i x_i - 1)$ .

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

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

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

## 2 Jednorozměrná lineární regrese

Sečteme prvních  $n$  rovnic

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

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  $\lambda_1, \lambda_2$ :

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

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

□

**Poznámka 2.14.** Ukázali jsme pouze, že to je stacionární bod, že je tam i minimum ukážeme v obecnější větě ve vícerozměrné regresi.

### 2.5 IS pro $\beta_0, \beta_1$

- IS poskytuje 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.  $\beta$  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 2.15.**

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

## 2 Jednorozměrná lineární regrese

Tedy

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

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  $\beta_i$  dostaneme

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

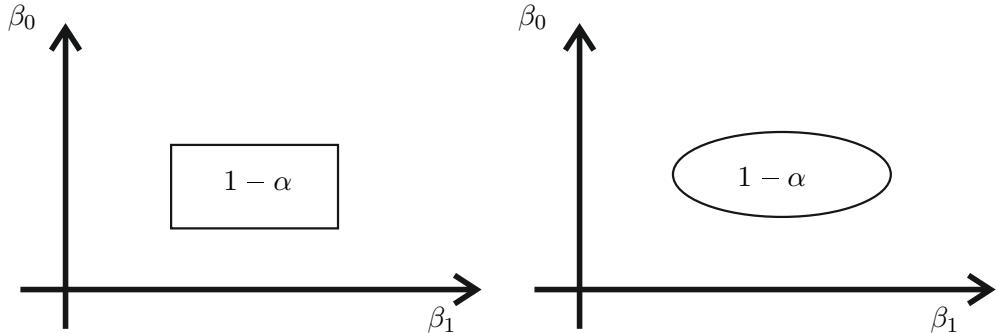
a tedy  $(\hat{\beta}_i \pm t_{1-\alpha/2}(n-2)\hat{\sigma}(\hat{\beta}_i))$  je  $100(1-\alpha)\%$  IS pro  $\beta_i, i=0, 1$ .

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

- $100(1-\alpha)\%$  IS pro  $\beta_0: \hat{\beta}_0 \pm t_{1-\alpha/2}(n-2) \cdot s_n \sqrt{\frac{1}{n} + \frac{\bar{x}_n^2}{s_{xx}}}$
- $100(1-\alpha)\%$  IS pro  $\beta_1: \hat{\beta}_1 \pm t_{1-\alpha/2}(n-2) \cdot s_n \frac{1}{\sqrt{s_{xx}}}$

**POZNÁMKA 2.16.** Z tvaru IS lze pozorovat, že IS pro  $\beta_0$  bude ve většině praktických případů širší než IS pro  $\beta_1$ , tzn. směrnice je obecně odhadnuta s větší přesností než absolutní člen (intercept).

**POZNÁMKA 2.17.** Někdy se konstruují simultánní IS pro oba parametry.



Zmíníme podrobněji u vícerozměrné regrese.

### 2.6 TH pro $\beta_0, \beta_1$

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. Otázku 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 nezamítнемe  $H_0$ , závěr bude, že  $x$  nevysvětuje nic z variability  $y$  a není v modelu významné. Pokud zamítнемe  $H_0$ , znamená to, že  $x$  je významné.

**POZNÁMKA 2.18.** Tyto závěry jsou správné pouze za předpokladu, že model je lineární!

## 2 Jednorozměrná lineární regrese

- 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 ří, že existuje lineární trend mezi  $x$  a  $y$ , ale mohou tam být i jiné typy závislosti

Pro konstrukci testů využijeme odvozené IS.

Poznámka 2.19. Opakování:  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0 \Rightarrow (\underline{\theta}, \bar{\theta})$  je  $100(1 - \alpha)\%$  IS pro  $\theta$ . Pak  $W = \{x | \theta_0 \notin (\underline{\theta}, \bar{\theta})\}$  je kritický obor test na hladině  $\alpha$ .

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

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

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

A zapsáno dohromady

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

Poznámka 2.20. 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  $\beta_1$  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.

Poznámka 2.21. Někdy dopředu známe kandidáta  $b_1$  jako hodnotu parametru  $\beta_1$  a chtěli bychom testovat  $H_0 : \beta_1 = b_1$  vs.  $H_1 : \beta_1 \neq b_1$ . Test bude zamítnut  $H_0$ , pokud

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

### 2.6.1 Test významnosti interceptu

Otzáka 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

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

## 2.7 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).

## 2 Jednorozměrná lineární regrese

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

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

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

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

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

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

regresi sum of squares, regresní součet čtverců. Podíl

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

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 zamítlí, pokud bude  $R^2 \approx 1$ . Protože každá monotonní funkce  $R^2$  vede na ekvivalentní test, budeme uvažovat statistiku

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

**Lemma 2.22.** 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 Jednorozměrná lineární regrese

2. Z bodu 1) plyne, že  $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$ , podělením  $n$  dostaneme dokazované tvrzení.
3. Z rovnice  $\frac{\partial S}{\partial \beta_1} = 0$  dostaneme

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

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 2.23.** *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  $R^2 = \rho^2(\mathbf{x}, \mathbf{y})$ , kde

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

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 2.22 vyplývá, že

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

## 2 Jednorozměrná lineární regrese

neboť

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

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.  $\hat{y}_i = \underbrace{\hat{\beta}_0}_{=\bar{y}_n} + \hat{\beta}_1 x_i = \bar{y}_n + \hat{\beta}_1 (\bar{x}_n - x_i)$ . 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 2.24.** 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  $R^2$ .

## 2 Jednorozměrná lineární regrese

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

Výsledky se většinou uvádí v tabulce ANOVA:

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

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

POZNÁMKA 2.25.  $H_0 : \beta_1 = 0$  je zamítelný, pokud  $F > F_{1-\alpha}(1, n - 2)$ . V tomto jednorozměrném případě je to ekvivalentní  $t$ -testu, neboť  $F = T^2$ .

**Věta 2.26.** Mějme  $e_1, \dots, e_n$  iid  $\mathcal{N}(0, \sigma^2)$ . Za platnosti  $H_0 : \beta_1 = 0$  je splněno, že

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

POZNÁMKA 2.27. Proto v tabulce ANOVA 2.7 uvádí df po řadě  $1, n - 2, n - 1$ . Používají se však i v případě jiného rozdělení chyb. Představit si je lze takto:

1.  $\text{SSE} = \sum_{i=1}^n \hat{e}_i^2$ , na  $n$ -reziďní  $\hat{e}_1, \dots, \hat{e}_n$  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.  $\text{SST} = \sum_{i=1}^n (y_i - \bar{y}_n)^2 \dots y_i - \bar{y}_n$  musí splňovat  $\sum_{i=1}^n (y_i - \bar{y}_n) = 0$ , a proto má  $n - 1$  stupňů volnosti.
3.  $\text{SSR} = \text{SST} - \text{SSE}$ , a počet stupňů volnosti je roven  $(n - 1) - (n - 2) = 1$ .

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

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

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

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

□

POZNÁMKA 2.28.  $R^2$  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  $x$  snižuje hodnotu  $R^2$  oproti modelům s různými  $x$ .

## 2 Jednorozměrná lineární regrese

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  $\beta_1^2$ . Modely s *velkou* směrnicí tedy budou mít obecně větší  $yRM R^2$ , 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  $\bar{y}_n$ .

Další kritéria budeme probírat později.

**PŘÍKLAD 2.29.** 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.

**POZNÁMKA 2.30.** korelace VS příčinnost

- **Příčinná spojitost** - i když je příčinná spojitost mezi  $x$  a  $y$  korelace samotná nám neřekne, zda  $x$  ovlivňuje  $y$  nebo naopak.
- **Skrytá příčinnost** - skrytá veličina  $z$  ovlivňuje  $x$  i  $y$ , což způsobuje jejich korelovanost.
- **Confounding factor** - skryté proměnné  $z$  i  $x$  ovlivňují  $y$ , výsledek tedy závisí i na  $z$ .
- **Coincidence** - korelace je náhodná.

## 2.8 Regrese skrz počátek

Existují případy, kdy přípustný model vyžaduje  $\beta_0 = 0$ , tj.

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

## 2 Jednorozměrná lineární regrese

PŘÍKLAD 2.31.

- Je to předem známo na základě nějakých fyzikálních úvah

$$\mathbb{E}[Y_0] = \beta_0 = 0$$

potom nemá smysl odhadnout  $\beta_0$ , protože to obecně sníží přesnost odhadu  $\sigma^2$  a tedy i  $\beta_1$

- Na začátek předpokládáme, že  $\beta_0 \neq 0$  a t-test nezamítne hypotézu  $H_0 : \beta_0 = 0$ , potom  $\beta_0$  může být z modelu odstraněn.

POZNÁMKA 2.32. V praktických situacích si často nemůžeme být jisti, že model platí i blízko počátku. Část statistiků trvá na přitomnosti interceptu v modelu, i když je nevýznamný.

Položit  $\beta_0$  apriorně, může nýt chybné i když  $\mathbb{E}[Y_0] = 0$ . Pokud totiž nevíme jistě, že model je lineární na okolí 0, volba  $\beta_0 = 0$  může vést k vychýleným odhadům  $\beta_1$ , pokud jsou nezávislé proměnné daleko od  $x = 0$ .

—————PICTURE—————

### 2.8.1 Odhad a testy v případě $\beta_0 = 0$

LSE parametru  $\beta_1$  dostaneme minimalizací  $S = \sum_{i=1}^n (y_i - \beta_1 x_i)^2$  ve tvaru:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2},$$

pokud  $e_1, \dots, e_n$  i.i.d.  $N(0, \sigma^2)$ , potom  $\mathbb{E}[\hat{\beta}_1] = \beta_1$  a  $D[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$ . Takže  $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n x_i^2})$

a  $s_n^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-1} = \frac{SSE}{n-1}$  je nestranný odhad  $\sigma^2$ . Dále  $\frac{SSE}{\sigma^2} \sim \chi^2(n-1)$  a nezávisí na  $\hat{\beta}_1$ .  $H_0 : \beta_1 = 0$  lze otestovat za pomoci statistiky:

$$T = \frac{\frac{\hat{\beta}_1}{s_n}}{\sqrt{\sum x_i^2}} \sim t(n-1)$$

$$100(1-\alpha)\% \text{ IS pro } \beta_1 \text{ je } (\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-1) \frac{s_n}{\sqrt{\sum x_i^2}})$$

- Zatím je vše podobné jako pro případ  $\beta_1 \neq 0$ .
- Rozdíl je ale v tabulce ANOVA a v míře dobré shody, problém je, že neplatí rozklad  $SST = SSR + SSE$  neboť součet reziduí  $\sum_{i=1}^n (y_i - \hat{y})$  nemusí být 0 a tedy  $\bar{\hat{y}}_n \neq \bar{y}_n$ . Odvodíme nový rozklad, který platí v obou případech, dokážeme ho ale jen pro  $\beta_0 = 0$

**Věta 2.33.** *V modelu s  $\beta_0 = 0$  platí*

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

## 2 Jednorozměrná lineární regrese

Důkaz.

$$\begin{aligned} \sum_{i=1}^n y_i^2 &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n \hat{y}_i^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i) \hat{y}_i \\ \text{z rovnice } \frac{dS}{d\beta_1} &= 0 \quad \text{dostaneme} \quad \sum_{i=1}^n (y_i - \hat{\beta}_1 x_i) x_i = 0 \\ &\sum_{i=1}^n (y_i - \hat{y}_i) \hat{y}_i = 0 \quad \text{Q. E. D.} \end{aligned}$$

□

Pokud vezmeme  $\sum y_i^2$  jako míru variability v datech, analogie  $R^2$  statistiky bude

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} \quad \Leftrightarrow \quad 1 - R^2 = \frac{\sum_{i=1}^n y_i^2 - \sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} = \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n y_i^2} \\ &= \text{definujeme } F = \frac{(n-1)R^2}{1-R^2} \text{ potom} \\ F &= \frac{\sum_{i=1}^n \hat{y}_i^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \frac{\hat{\beta}_1 \sum_{i=1}^n x_i^2}{s_n^2} = T^2 \end{aligned}$$

vztah mezi  $R^2$ ,  $F$  a  $T^2$  je tedy stejný jako pro  $\beta_0 \neq 0$ .

**POZNÁMKA 2.34.** Tato definice  $R^2$  se ale v praxi moc nepoužívá, protože neumožňuje přímé srovnání modelů bez a s interceptem.

$$\begin{aligned} \beta_0 = 0 \quad : \quad R^2 &= 1 - \frac{\text{SSE}}{\sum_{i=1}^n y_i^2} & \beta_0 \neq 0 \quad : \quad R^2 &= 1 - \frac{\text{SSE}}{\sum_{i=1}^n (y_i - \bar{y}_n)^2} \\ \text{obecně ale } \sum_{i=1}^n (y_i - \bar{y}_n)^2 &< \sum_{i=1}^n y_i^2, \text{ R}^2 \text{ v modelu s } \beta_0 = 0 \text{ tedy bude větší než R}^2 \text{ modelu s } \beta_0 \neq 0 \text{ i když jsou jejich SSE srovnatelné.} \end{aligned}$$

- Definice vhodné  $R^2$  pro  $\beta_0 = 0$  vyvolává jistou kontroverzi a existuje několik verzí.
- Možná volba je  $R^2 = (\rho(y_I, \bar{y}_I))^2$ , kde  $\bar{y}_I = (\bar{y}_1, \dots, \bar{y}_n)$  protože tato vlastnost platí i pro případ  $\beta_0 = 0$ .
- Další možnost je srovnat modely pomocí hodnot  $s_n^2$ . (preferuje se model s nejnižší hodnotou  $s_n^2$ )

## 2 Jednorozměrná lineární regrese

Source	df	SS	MS	F
Regression	1	$\sum_{i=1}^n \hat{y}_i^2$	$\text{MSR} = \frac{\text{SSR}}{1}$	$\frac{\text{SSR}}{s_n^2}$
Residual	$n - 1$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\text{MSE} = \frac{\text{SSE}}{n-1}$	
Total	$n$	$\sum_{i=1}^n y_i^2$		
		$R^2 = \rho^2(\mathbf{y}, \hat{\mathbf{y}})$		

Tabulka 2.1: Tabulka ANOVA pro  $\beta_0 = 0$ .

### Predikce

Jakmile máme model, často bývá cílem odhadnout hodnoty veličiny  $Y_0$  pro nové  $x_0$ , které není v původních datech. Budeme uvažovat dva typy predikce:

1. predikce střední hodnoty  $\mu_0 = \mathbb{E}[Y_0]$  v bodě  $x_0$ ,
2. predikce hodnoty nového pozorování  $Y_0$  v bodě  $x_0$ .

Pro oba typy použijeme bodový odhad

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Intervalové odhady se ale budou lišit.

#### 2.8.2 Ad 1

Protože je  $\mu_0 = \beta_0 + \beta_1 x_0$  vlastně parametr, lze pro něj odvodit IS (za předpokladu normality chyb).

Spočteme tedy  $D(\hat{Y}_0)$ . Dosazením odhadů  $\hat{\beta}_0$  a  $\hat{\beta}_1$  dostaneme  $\hat{Y}_0 = \bar{y} + \hat{\beta}_1(x_0 - \bar{x})$  a

$$D\hat{Y}_0 = D(\bar{Y}) + (x_0 - \bar{x})^2 D(\hat{\beta}_1) + 2(x_0 - \bar{x}) \underbrace{\text{Cov}(\bar{Y}, \hat{\beta}_1)}_{=0} = \frac{\sigma^2}{n} + \frac{\sigma^2(x_0 - \bar{x})^2}{S_{xx}} = \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right].$$

Nahrazením  $\sigma^2$  statistika  $s_n^2$  dostaneme odhad  $D(\hat{Y}_0)$  ve tvaru

$$\hat{\sigma}^2(\hat{Y}_0) = s_n^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right].$$

$\hat{\sigma}(\hat{Y}_0)$  se obvykle nazývá **standardní chyba predikce v bodě**  $x_0$ . Jsou-li  $e_1, \dots, e_m$  iid  $\mathbb{N}(0, \sigma^2)$ , platí, že

$$\hat{Y}_0 \sim \mathbb{N}\left(\mu_0, \underbrace{\sigma^2 \left[ \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}_{\sigma^2(\hat{Y}_0)}\right)$$

a tedy

$$\frac{\hat{Y}_0 - \mu_0}{\sigma(\hat{Y}_0)} \sim \mathbb{N}(0, 1).$$

## 2 Jednorozměrná lineární regrese

Celkem tedy

$$T = \frac{\frac{\hat{Y}_0 - \mu_0}{\sqrt{\sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}}}{\sqrt{\frac{(n-2)s_n^2}{\sigma^2} \frac{1}{n-2}}} = \frac{\hat{Y}_0 - \mu_0}{\sqrt{s_n^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}} = \frac{\hat{Y}_0 - \mu_0}{\hat{\sigma}^2(\hat{Y}_0)} \sim t(n-2).$$

Vyjádřením získáme  $100(1-\alpha)\%$  IS pro  $\mu_0$  ve tvaru

$$\hat{Y}_0 \pm t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}^2(hY_0).$$

**POZNÁMKA 2.35.** Z tvaru IS je vidět, že bude nejkratší pro  $x_0 = \bar{x}$  a s rostoucí vzdáleností  $|x_0 - \bar{x}|$  se prodlužuje.

- Speciálně potom čím dále jsme od oblasti, kde jsou naše data  $x$ , tím méně spolehlivé jsou naše predikce.
- Je třeba opatrnosti při predikci hodnot  $Y$  mimo interval  $(\min x_i, \max x_i)$ .

### 2.8.3 Ad 2

Intervalové odhady pro  $Y_0$  nejsou IS, protože  $Y_0$  není parametr. Říká se jim **intervaly predikce**. Potřebujeme rozptyl  $Y_0 - \hat{Y}_0$ , pokud je nené pozorování  $Y_0$  nezávislé na  $Y_i, i \in \hat{n}$ , potom

$$D(Y_0 - \hat{Y}_0) = \underbrace{DY_0}_{\sigma^2} + D\hat{Y}_0 + 0 = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right].$$

Odhad tohto rozptylu bude  $s_p^2$ , kde

$$s_p = s_n \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

Za předpokladu normality chyb pak

$$T = \frac{Y_0 - \hat{Y}_0}{s_n \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} = \frac{Y_0 - \hat{Y}_0}{s_p} \sim t(n-2).$$

Vyjádřením získáme  $100(1-\alpha)\%$  interval predikce pro  $Y_0$  ve tvaru

$$\hat{Y}_0 \pm t_{1-\frac{\alpha}{2}}(n-2)s_p.$$

**POZNÁMKA 2.36.** Přesnost predikce

- roste s rostoucím  $n$  a rostoucím rozsahem  $x$  naměřeným pomocí  $S_{xx}$ ,
- klesá s rostoucím  $|x_0 - \bar{x}|$ .

Pokud můžeme předem zvolit  $x_1, \dots, x_n$ , lze přesnost predikce zvýšit volbou dostatečně rozptýlených hodnot  $x$ . To ale může zvyšovat  $R^2$  a někdy vést k horšímu modelu.

To je **základní rozpor v regresní analýze**:

## 2 Jednorozměrná lineární regrese

- dobrý model nemusí poskytovat dobré predikce,
- dobré predikce mohou vycházet z méně přesných modelů.

**Poznámka 2.37.** Odvozené výsledky platí za předpokladu normality chyb. Protože jsou ale za podmínek regularity odhady  $\hat{\beta}_0, \hat{\beta}_1$  asymptoticky normální, IS pro  $\mathbb{E}[Y_0]$  budou fungovat (jsou použitelné i pro velká  $n$ ). IP pro  $Y_0$  ale závisí na normalitě chyb i pro velká  $n$ , mohou tedy být nepřesné pro nenormální chyby.

**Příklad 2.38** (Ověření adekvátnosti modelu). Ověření adekvátnosti modelu je důležitá součást analýzy. Měla by být provedena dříve, než budeme interpretovat parametry modelu nebo přijímat nějaké závěry založené na modelu.

Všechny výsledky týkající se  $\beta_0, \beta_1$  byly odvozeny za předpokladu **linearity modelu** a některé za předpokladu **normality chyb**.

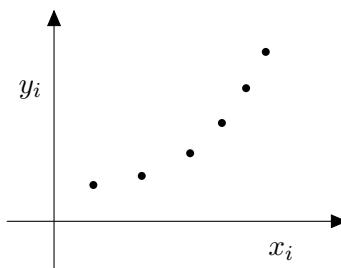
Bylo by tedy dobré mít testy ověřující linearitu.

Základní procedury jsou následující:

- 1) Prozkoumání **scatter plotu** dvojic  $(x_i, y_i)$ . Příklad lze vidět na obrázku 2.1. Takový scatter plot může indikovat, že lepší model bude

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i.$$

Scatter plot ale může být zavádějící, pokud je odklon od linearity způsoben spíše chybějící



Obrázek 2.1: Scatter plot naměřených dat.

proměnnou než polynomiální závislostí na  $x$ .

### 2) Analýza hodnot testovacích statistik.

- Např. malá hodnota  $R^2$  společně s významem hodnot???  $t$ -statistiky pro parametry  $\beta_1$  obecně naznačuje, že skutečný model obsahuje i jiné proměnné  $x$ ,
  - velká hodnota  $R^2$  a významná  $t$ -statistika ale samo o sobě neznamená, že je model lineární.
- 3) **Obrázky reziduí.** Je to efektivní diagnostický nástroj. Rezidua odhadují, kolik variability v datech zůstne po odstranění lineární části v  $x$ . Dá se také očekávat, že jejich hodnoty budou užitečné pro detekci odchylek od normality.

**Příklad 2.39.** Analýza scatter plotů a obrázků reziduí je dost subjektivní. Bylo by dobré mít nějaký objektivní analytický nástroj pro ověření linearity modelu. Bohužel nejsou k dispozici skoro žádné takové nástroje. Pro většinu dat jsou v praxi nejvíce využívány metody 1) - 3).

Jinak je tomu u navržených experimentů typu industriálních nebo klinických studií, kde existuje doporučený analytický test, tzv. *lack of fit* test (LOFT). Ten předpokládá, že máme více pozorování pro jednu  $x_i$ .

### 2.8.4 Ad 3 - Analýza reziduí

Intuitivně, pokud je náš model správný, měla by se rezidua chovat jako náhodný výběr z  $\mathbb{N}(0, \sigma^2)$ . Pokud se bude zdát, že se tak nechovají, bude to znamenat neadekvátnost modelu. Později ukážeme grafický nástroj. Nejprve ale začneme vlastnostmi reziduů.

**Věta 2.40.** *Nechť  $\hat{e}_i$  jsou rezidua modelu (\*) odhadnutého metodou nejmenších čtverců. Potom platí:*

1.  $\mathbb{E}\hat{e}_i = 0, \quad i = 1, \dots, n$
2.  $D\hat{e}_i = \sigma^2 = \sigma^2 \left[ 1 - \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right] \approx \sigma^2 \text{ pro velká } n$
3.  $\text{Cov}(\hat{e}_i, \hat{e}_j) = -\sigma^2 \left[ 1 - \left( \frac{1}{n} + \frac{(\bar{x} - x_i)(\bar{x} - x_j)}{S_{xx}} \right) \right]$
4.  $\text{Cov}(\hat{e}_i, \hat{Y}_i) = 0 = 0, \quad i = 1, \dots, n$
5. Pokud jsou  $e_1, \dots, e_n$  iid  $\mathcal{N}(0, \sigma^2)$ , potom platí:

$$\hat{Z}_i = \frac{\hat{e}_i}{\sigma_{\hat{e}_i}} \sim \mathcal{N}(0, 1).$$

*Důkaz.* 1.  $\hat{e}_i = Y_i - \hat{Y}_i$ , takže  $\mathbb{E}(\hat{e}_i) = \mathbb{E}Y_i - \mathbb{E}\hat{Y}_i$ , ale  $\mathbb{E}\hat{Y}_i = \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i = \mathbb{E}Y_i$

2.

$$D\hat{e}_i = D(Y_i - \hat{Y}_i) = DY_i + \underbrace{D\hat{Y}_i}_{\sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right]} - 2 \underbrace{\text{Cov}(Y_i, \hat{Y}_i)}_{\sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right]} = \sigma^2 \left[ 1 - \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right]$$

3.

$$\begin{aligned} \text{Cov}(\hat{e}_i, \hat{e}_j) &= \text{Cov}(Y_i - \hat{Y}_i, Y_j - \hat{Y}_j) = \underbrace{\text{Cov}(Y_i, Y_j)}_{=0} - \text{Cov}(Y_i, \hat{Y}_j) - \text{Cov}(Y_i, \hat{Y}_j) + \text{Cov}(\hat{Y}_i, \hat{Y}_j) \\ \text{Cov}(\hat{Y}_i, \hat{Y}_j) &= \text{Cov}(\hat{\beta}_0 + \hat{\beta}_1 x_i, \hat{\beta}_0 + x_j \hat{\beta}_1) = \underbrace{D(\hat{\beta}_0)}_{\sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} + (x_i + x_j) \underbrace{\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)}_{-\frac{\sigma^2 \bar{x}}{S_{xx}}} + x_i x_j \underbrace{D(\hat{\beta}_1)}_{\frac{\sigma^2}{S_{xx}}} = \\ &= \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} - \frac{(x_i + x_j)\bar{x}}{S_{xx}} + \frac{x_i x_j}{S_{xx}} \right] = \sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \right] \end{aligned}$$

Podobně bychom dostali

$$\text{Cov}(Y_i, \hat{Y}_j) + \text{Cov}(\hat{Y}_i, Y_j) = 2\sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \right]$$

$$\text{takže } \text{Cov}(\hat{e}_i, \hat{e}_j) = -\sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \right].$$

4.

$$\begin{aligned}\text{Cov}(\hat{e}_i, \hat{Y}_i) &= \text{Cov}(Y - i - \hat{Y}_i, \hat{Y}_i) = \underbrace{\text{Cov}(Y_i, \hat{Y}_i)}_{=\sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right]} - \underbrace{\text{D}(\hat{Y}_i)}_{=\text{Cov}(\hat{Y}_i, \hat{Y}_i) = \sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right]} = 0\end{aligned}$$

5.  $e_i \sim \mathcal{N}(0, \sigma^2) \implies \hat{e}_i \sim \mathcal{N}(\cdot, \cdot)$ , protože  $\hat{e}_i$  je LK  $Y_1, \dots, Y_n$

- 1)  $\implies \mathbb{E}\hat{e}_i = 0$
  - 2)  $\implies \text{D}\hat{e}_i = \sigma_{\hat{e}_i}^2$
- $$\implies \frac{\hat{e}_i}{\sigma_{\hat{e}_i}} \sim \mathcal{N}(0, 1)$$

□

POZNÁMKA 2.41. Z bodu 3) věty plyne, že  $\text{Cov}(\hat{e}_i, \hat{e}_j) \approx 0$  pro velké  $n$ . Pokud jsou testy  $e_i$  iid  $\mathcal{N}(0, \sigma^2)$ , měla by se standardizovaná rezidua  $\hat{Z}_i = \frac{\hat{e}_i}{\sigma_{\hat{e}_i}}$  chovat pro velké  $n$  jako náhodný výběr z  $\mathcal{N}(0, 1)$  rozdělení. V praxi ale budeme potřebovat odhad  $\sigma^2$  pro výpočet  $\hat{Z}_i$ .

Nejznámější procedura: odhadnout  $\sigma^2$  pomocí  $s_n^2$ , potom

$$\hat{z}_i = \frac{\hat{e}_i}{s_n \sqrt{1 - \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)}} \quad \text{standardizovaná rezidua}$$

by se opět pro velká  $n$  měla chovat jako NV z  $\mathcal{N}(0, 1)$ .

POZNÁMKA 2.42.  $\hat{e}_i$  se užívají pro grafickou analýzu.

Jiná třída reziduí – PRESS rezidua (? metody zkoumání reziduí):

ozn.  $\hat{\beta}_{0(-i)}, \hat{\beta}_{1(-i)}$  odhadu parametrů  $\beta_0, \beta_1$ , pokud je vynecháno  $i$ -té pozorování. Pak  $i$ -té PRESS reziduum je definováno jako

$$\hat{e}_{(-i)} = \hat{Y}_i - \hat{Y}_{(-i)}, \quad \text{kde } \hat{Y}_{(-i)} = \hat{\beta}_{0(-i)} + x_i \hat{\beta}_{1(-i)}.$$

Podrobněji se jim budeme věnovat později.

## 2.9 Grafy reziduí

- Histogram reziduí (náhled normality reziduí).
- Kvantilový graf (QQ plot) standardizovaných reziduí – seřadíme dle velikosti:  $\hat{r}_{(1)} \leq \hat{r}_{(2)} \leq \dots \leq \hat{r}_n$  a vyneseme oproti  $\Phi^{-1}((i - \frac{1}{2})\frac{1}{n})$ ,  $i = 1, \dots, n$ . Body by měly ležet přibližně na přímce ( $\mathbb{E}(e_i) \approx \Phi^{-1}((i - \frac{1}{2})\frac{1}{n})$  pro normální chyby).
- Použití: ověření normality, detekce odlehlych pozorování (obr. 3.6 str. 1077 GLM).
- Standardizovaná rezidua  $\times$  jednotlivým vysvětlujícím proměnným  $x - \hat{r}_i$  nezávisí na  $\sigma$ , graf  $\hat{r}_i \times x_i$  lze použít pro detekci nelinearity nebo nekonstantního rozptylu.
- Standardizovaná rezidua  $\hat{r}_i \times$  predikovaným hodnotám  $\hat{y}_i - \text{Cov}(\hat{e}_i, \hat{Y}_i) = 0$ , tedy  $\hat{e}_i(\hat{r}_i) \times \hat{Y}_i$  by měly být nekorelované, pokud platí model (\*). Tzn. graf  $\hat{r}_i \times \hat{y}_i$  by měl být náhodně rozptýlený kolem osy  $x$ , navíc  $\hat{r}_i$  by měla ležet v  $(-3, 3)$  ( $\hat{r}_i \approx \mathcal{N}(0, 1)$ ).

Obrázky....

## 2 Jednorozměrná lineární regrese

- Standardizovaná rezidua  $\times$  pořadí pozorování – možná detekce řadové korelace mezi pozorováními.

Obrázek....

Předpokládejme, že kromě  $y_i$  máme pro každé  $i \in \hat{n}$  k dispozici také  $m$  nezávislých proměnných  $x_{i1}, x_{i2}, \dots, x_{im}$ . Pak získáme model

$$Y_i = \beta_0 + \sum_{j=1}^m \beta_j x_{ij} + e_i, \quad i \in \hat{n},$$

kde  $e_1, \dots, e_n$  jsou **nezávislé (nekorelované)** chyby a  $e_i \sim \mathcal{N}(0, \sigma^2)$ . Na základě pozorování  $(x_{i1}, \dots, x_{im}, y_i)$ ,  $i \in \hat{n}$  chceme odhadnout parametr  $\beta = (\beta_0, \beta_1, \dots, \beta_m)^T$  (proložení dat  $m+1$  dimenzionální nadrovinou). Předpokládejme, že  $n > m+1$ , tj. že máme více dat, než parametrů. Maticově můžeme tento stav zapsat jako

$$\mathbf{Y} = (Y_1, \dots, Y_n)^T, \quad \mathbf{y} = (y_1, \dots, y_n)^T, \quad \mathbf{e} = (e_1, \dots, e_n)^T.$$

Označme

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ 1 & \vdots & \dots & \vdots \\ \vdots & x_{n1} & \dots & x_{nm} \end{bmatrix}$$

jako **matici modelu** (regresní matici, !!něco matrix, nepřečtu to???). Dostaneme tak model ve tvaru (důležitému)

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (m+1)} \beta_{(m+1) \times 1} + \mathbf{e}_{bn \times 1}. \quad (2.2)$$

Nyní budeme předpokládat, že  $e_1, \dots, e_n$  jsou nezávislé a  $e_i \sim \mathcal{N}(0, \sigma^2)$ , tzn.  $\mathbf{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$  a  $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 I_n)$ .

Věrohodnostní funkce je potom ve tvaru

$$\begin{aligned} L(\beta, \sigma^2) &= f_\pi(\mathbf{y}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2} = \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2} (\mathbf{y} - \boldsymbol{\mu})^T (\mathbf{y} - \boldsymbol{\mu})} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)}, \end{aligned}$$

kde  $\mu_i = \beta_0 + \sum_{j=1}^m \beta_j x_{ij}$  a  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T = \mathbf{X}\beta$ .

Pro pevné  $\sigma^2$  je

$$\max_{\beta} L(\beta, \sigma^2) \Leftrightarrow \min_{\beta} \underbrace{(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)}_{g(\beta)}$$

je opět pomocí derivací, ukážeme algebraický přístup.

**Věta 2.43.** Uvažujme model 2.2 a nechť  $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ . Potom  $\hat{\beta}$  je MLE  $\beta$  právě tehdy, když  $\hat{\beta}$  je řešením soustavy rovnic

$$\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y} \quad (\text{soustava normálních rovnic}).$$

Je-li matice  $\mathbf{X}^T \mathbf{X}$  singulární, má tato soustava jednoznačné řešení ve tvaru

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

## 2 Jednorozměrná lineární regrese

*Důkaz.*  $\Leftarrow$  Ukážeme, že každé řešení  $\hat{\beta}$  soustavy  $\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$  minimalizuje  $g(\beta)$  a pro každé  $\beta$  platí, že

$$g(\beta) = (\mathbf{y} - \mathbf{X} \beta)^T (\mathbf{y} - \mathbf{X} \beta) = \mathbf{y}^T \mathbf{y} - 2 \underbrace{\mathbf{y}^T \mathbf{X} \beta}_{\hat{\beta}^T \mathbf{X}^T \mathbf{X}} + \beta^T \mathbf{X}^T \mathbf{X} \beta = \mathbf{y}^T \mathbf{y} - 2 \hat{\beta}^T \mathbf{X}^T \mathbf{X} \beta + \beta^T \mathbf{X}^T \mathbf{X} \beta$$

má platit i pro  $\hat{\beta}$ :

$$g(\hat{\beta}) = \mathbf{y}^T \mathbf{y} - 2 \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta} + \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta} = \mathbf{y}^T \mathbf{y} - \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta}$$

a tedy

$$g(\beta) - g(\hat{\beta}) = \beta^T \mathbf{X}^T \mathbf{X} \beta - 2 \hat{\beta}^T \mathbf{X}^T \mathbf{X} \beta + \hat{\beta}^T \mathbf{X}^T \mathbf{X} \hat{\beta} = (\mathbf{X} \beta - \mathbf{X} \hat{\beta})^T (\mathbf{X} \beta - \mathbf{X} \hat{\beta}) = \quad (2.3)$$

$$= (\mathbf{X}(\beta - \hat{\beta}))^T (\mathbf{X}(\beta - \hat{\beta})) = \langle \mathbf{X}(\beta - \hat{\beta}), \mathbf{X}(\beta - \hat{\beta}) \rangle \geq 0, \quad \forall \beta, \quad (2.4)$$

tedy  $\hat{\beta}$  minimalizuje  $g(\beta)$  a je tedy MLE parametru  $\beta$ .

$\Rightarrow$  Předpokládejme, že  $\hat{\beta}_1$  minimalizuje  $g(\beta)$  (je tedy MLE). To potom znamená, že  $g(\hat{\beta}_1) \leq g(\beta)$ ,  $\forall \beta$ , speciálně  $g(\hat{\beta}_1) \leq g(\hat{\beta})$ , kde  $\hat{\beta}$  je řešení soustavy  $\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$ . Z rovnice 2.4 vyplývá, že  $g(\hat{\beta}_1) \geq g(\hat{\beta})$ . Celkem tedy  $g(\hat{\beta}_1) = g(\hat{\beta})$ . Dosazením do 2.4 dostaneme, že

$$0 = g(\hat{\beta}_1) - g(\hat{\beta}) = \langle \mathbf{X}(\hat{\beta}_1 - \hat{\beta}), \mathbf{X}(\hat{\beta}_1 - \hat{\beta}) \rangle$$

a tedy  $\mathbf{X}(\hat{\beta}_1 - \hat{\beta}) = \mathbf{0}$ . Potom ale vynásobením  $\mathbf{X}^T$  zleva dostaneme, že

$$\mathbf{X}^T \mathbf{X} \hat{\beta}_1 \underbrace{\mathbf{X}^T \mathbf{X} \hat{\beta}}_{\mathbf{X}^T \mathbf{y}} = 0 \Rightarrow \mathbf{X}^T \mathbf{X} \hat{\beta}_1 = \mathbf{X}^T \mathbf{y}$$

a  $\hat{\beta}_1$  splňuje soustavu  $\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$ .

Aby byl důkaz korektní, je třeba ukázat, že soustava  $\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$  má vždy alespoň 1 řešení. Pokud existuje  $(\mathbf{X}^T \mathbf{X})^{-1}$ , není co dokazovat, řešení máme přímo. Co když je ale  $\mathbf{X}^T \mathbf{X}$  singulární?

□

**Lemma 2.44.** Soustava lineárních rovnic  $\mathbb{A}x = \mathbf{y}$  má řešení právě tehdy, když  $\langle \mathbf{y}, \mathbf{z} \rangle = 0$  pro všechna  $\mathbf{z}$  splňující  $\mathbb{A}\mathbf{z} = \mathbf{0}$ .

**Věta 2.45.** Soustava normálních rovnic  $\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$  má vždy alespoň jedno řešení.

*Důkaz.* Musíme ukázat, že  $\langle \mathbf{X}^T \mathbf{y}, \mathbf{z} \rangle = 0$ ,  $\forall \mathbf{z}$  splňující  $\mathbf{X}^T \mathbf{X} \mathbf{z} = \mathbf{0}$ . Potom  $\mathbf{X}^T \mathbf{X} \mathbf{z} = \mathbf{0} \Rightarrow \langle \mathbf{X}^T \mathbf{X} \mathbf{z}, \mathbf{z} \rangle = \langle \mathbf{X} \mathbf{z}, \mathbf{X} \mathbf{z} \rangle = 0$  a tedy  $\mathbf{X} \mathbf{z} = \mathbf{0}$ . Celkem tedy  $\langle \mathbf{X}^T \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{y}, \mathbf{X} \mathbf{z} \rangle = 0$ . Obecně totiž platí, že  $\langle \mathbf{x}, \mathbb{A} \mathbf{y} \rangle = \langle \mathbb{A}^T \mathbf{x}, \mathbf{y} \rangle$ . □

**Poznámka 2.46.** Z vět vyplývá, že MLE  $\beta$  může být nalezeno řešením  $m+1$  lineárních rovnic o  $m+1$  neznámých. Málokdy existuje analytické řešení, je třeba použít numerické metody. Matice  $\mathbf{X}^T \mathbf{X}$  může být v praktických aplikacích špatně podmíněná, což ovlivňuje numerickou přesnost  $\hat{\beta}$ . Proto se často užívají metody jako Choleského rozklad, QR rozklad, singulární rozklad (SVD).

Odvodili jsme to pro normální chyby. Minimalizace  $g(\beta)$  lze ale použít i pro jiné druhy chyb, potom se  $\hat{\beta}$  nazývá **ordinary least squares estimate (OLS)** (obyčejné nejmenší čtverce). Asi nejužívanější metoda pro oblast  $\beta$ .

Jak poznat, že mají normální rovnice jednoznačné řešení bez nutnosti výpočtu  $\mathbf{X}^T \mathbf{X}$ ?

## 2 Jednorozměrná lineární regrese

**Věta 2.47.** Matice  $\mathbf{X}^T \mathbf{X}$  je nesignulární právě tehdy, když jsou sloupce matice  $\mathbf{X}$  LN.

*Důkaz.*  $\Leftarrow$  Sporem. Nechť jsou sloupce  $\mathbf{X}$  LN a matice  $\mathbf{X}^T \mathbf{X}$  singulární, tzn.  $\exists c \neq 0$  tak, že  $\mathbf{X}^T \mathbf{X} c = 0$ . Potom

$$0 = \langle c, \mathbf{X}^T \mathbf{X} c \rangle = \langle \mathbf{X} c, \mathbf{X} c \rangle \Rightarrow \mathbf{X} c = 0, \quad \sum c_i \mathbf{x}_i^c = 0,$$

kde  $c = (c_1, \dots, c_m)^T$  a  $\mathbf{x}_i^c$  je  $i$ -tý sloupec matice  $\mathbf{X}$ . Potom sloupce  $\mathbf{X}$  jsou LZ. Spor.

$\Rightarrow$  Sporem. Předpokládejme, že  $\mathbf{X}^T \mathbf{X}$  je regulární a sloupce  $\mathbf{X}$  LZ. Potom existuje  $c \neq 0$  takové, že  $\mathbf{X} c = 0$ ,  $\mathbf{X}^T \mathbf{X} c = 0$ . Z toho vyplývá, že  $\mathbf{X}^T \mathbf{X}$  je singulární. Spor.

□

**Poznámka 2.48.** Pokud  $\mathbf{X}_{nx(m-1)}$ ,  $n > m + 1$ ,  $h(\mathbf{X}) = m + 1$ ,  $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 I_m)$ . Potom existuje jednoznačné řešení normálních rovnic  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ .

**Poznámka 2.49.** Pokud jsou sloupce  $\mathbf{X}$  LZ, je  $\mathbf{X}^T \mathbf{X}$  sigulární, což je většinou detekováno numerickou metodou výpočtu  $\hat{\beta}$ .

**Poznámka 2.50.** • pokud jsou sloupce  $\mathbf{X}$  LZ, je  $\mathbf{x}^T \mathbf{x}$  singulární, což je většinou detekováno numerickou metodou výpočtu  $\hat{\beta}$

- horší situace je, pokud jsou sloupce  $\mathbf{X}$  ”téměř” LZ  $\rightarrow$  tzv. **multikolinearita** – způsobuje problémy při výpočtu  $\hat{\beta}$ , protože je  $\mathbf{x}^T \mathbf{x}$  ”téměř” singulární, jak ji detekovat probereme na konci přednášky

### Odhad parametru $\sigma^2$

Pro normální chybu získáme MLE  $\sigma^2$  derivací  $\ln L(\beta, \sigma^2)$ , z čehož plyne:

$$\hat{\sigma}_n^2 = \frac{1}{n} SSE = \frac{1}{n} (\mathbf{y} - \mathbf{X} \hat{\beta})^T (\mathbf{y} - \mathbf{X} \hat{\beta}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

kde  $\hat{y}_i = (\mathbf{X} \hat{\beta})_i = \mathbf{x}_i^T \hat{\beta}$ ,  $i = 1, \dots, n$

a  $\mathbf{x}_i^T$  značí  $i$ -tý řádek matice  $\mathbf{X}$ . Protože se jedná o vychýlený odhad, používá se obecně odhad

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

a  $s_n = \sqrt{s_n^2}$  jako odhad  $\sigma$  (už není nestranný).

Pro  $e_i \sim \mathcal{N}(0, \sigma^2)$  se také používají statistiky  $s_n^2, s_n$ .

*Př. Ex. 5.13, str. 158 (nebo 138? jinak?)*

*Ex. 5.15, str. 203*

**Vlastnosti odhadů**  $\hat{\beta}, s_n^2$

**Věta 2.51.** Nechť  $\hat{\beta}$  je OLS odhad parametru  $\beta$  v modelu (\*\*), kde  $h(\mathbf{X}) = m-1$  a  $e_1, \dots, e_n$  nezávislé,  $e_i \sim \mathcal{N}(0, \sigma^2)$ . Potom platí:

1.  $\mathbb{E}(\hat{\beta}) = \beta$  (tj.  $\hat{\beta}$  je nestranný)
2.  $\text{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$
3.  $\mathbb{E}(s_n^2) = \sigma^2$
4. Pokud navíc  $e_i \sim \mathcal{N}(0, \sigma^2), i = 1, \dots, n$ , potom  $\hat{\beta} \sim \mathcal{N}_{m-1}(\beta, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1})$ . Speciálně  $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2 \nu_i)$ , kde  $\nu_i$  je  $i$ -tý diagonální prvek matice  $(\mathbf{X}^T \mathbf{X})^{-1}$ .

Důkaz. 1.

$$\begin{aligned} h(\mathbf{X}) = m - 1 &\implies \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \\ \mathbb{E}\hat{\beta} &= \mathbb{E}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}\mathbf{Y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta \end{aligned}$$

2. Značení:  $\mathbf{Y}$  velikosti  $(n \times 1)$  nádoný vektor,  $\text{Cov}(\mathbf{Y}) = \Sigma, \mathbb{A}_{m,n}$  matice, potom  $\text{Cov}(\mathbb{A}\mathbf{X}) = \mathbb{A}\Sigma\mathbb{A}^T$ .

Protože  $\hat{\beta} = \mathbb{A}\mathbf{Y}$ , kde  $\mathbb{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ ,  $\hat{\beta}$  je LK  $Y_1, Y_m$  a  $\text{Cov}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$

$$\text{Cov}\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{I}_n \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}.$$

3. Nejdříve přepíšeme vektor reziduí  $\hat{e} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \hat{\mathbf{Y}}$  a  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbb{H}\mathbf{Y}$ , kde  $\mathbb{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$  je tzv. **projekční matici**. Pak  $\hat{e} = \mathbf{Y} - \mathbb{H}\mathbf{Y} = (\mathbf{I}_n - \mathbb{H})\mathbf{Y}$ .

Dále platí  $(\mathbf{I}_n - \mathbb{H})\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$ , takže

$$\hat{e} = (\mathbf{I}_n - \mathbb{H})\mathbf{Y} = (\mathbf{I}_n - \mathbb{H})(\mathbf{Y}\beta + \mathbf{e}) = \underbrace{(\mathbf{I}_n - \mathbb{H})\mathbf{X}\beta}_{=0} + (\mathbf{I}_n - \mathbb{H})\mathbf{e} = (\mathbf{I}_n - \mathbb{H})\mathbf{e}$$

Zřejmě  $\mathbb{H}^T = \mathbb{H}$  a  $\mathbb{H}^2 = [\mathbf{X} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] [\mathbf{X} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] = \mathbf{X} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbb{H}$  a  $(\mathbf{I}_n - \mathbb{H})^2 = \mathbf{I}_n - \mathbb{H}$  (neboli  $\mathbb{H}$  je symetrická a idempotentní?).

$$SSE = (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) = \hat{e}^T \hat{e} = \hat{e}^T (\mathbf{I}_n - \mathbb{H})(\mathbf{I}_n - \mathbb{H})\mathbf{e} = \mathbf{e}^T (\mathbf{I}_n - \mathbb{H})\mathbf{e} = \sum_{i=1}^n \sum_{j=1}^n g_{ij} e_i e_j,$$

kde  $g_{ij}$  je (i,j)-prvek matice  $(\mathbf{I}_n - \mathbb{H})$ .

$$\begin{aligned} \mathbb{E}(SSE) &= \sum_{i=1}^n \sum_{j=1}^n g_{ij} \underbrace{\mathbb{E}(e_i e_j)}_{\text{Cov}(e_i, e_j)} = [\text{nekorelované: } \mathbb{E}e_i = 0] = \sum_{i=1}^n g_{ii} \mathbb{E}e_i = \sigma^2 \sum_{i=1}^n g_{ii} \\ \sum_{i=1}^n g_{ii} &= \text{tr}(\mathbf{I}_n - \mathbb{H}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbb{H}) = n - \text{tr}(\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) = \\ &= n - \text{tr}(\mathbf{X}^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1}) = n - \text{tr}(\mathbb{I}_{m+1}) = n - (m+1) \end{aligned}$$

Celkem pak dostáváme  $\mathbb{E}s_n^2 = \frac{1}{n-(m+1)} \mathbb{E}(SSE) = \frac{1}{n-(m+1)} \sigma^2(n - (m+1)) = \sigma^2$ .

## 2 Jednorozměrná lineární regrese

4.  $\hat{\beta}$  e LK  $Y_1, \dots, Y_n$ , nezávislé, normálně rozdělené  $\rightarrow \hat{\beta} \sim \mathcal{N}_{m+1}(\beta, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1})$ .

□

**Poznámka 2.52.** Vlastnosti projekční matice:

- $\mathbb{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ ,  $\hat{\mathbf{Y}} = \mathbb{H}\mathbf{Y}$ ,  $\mathbb{H}^T = \mathbb{H}$ ,  $(\mathbb{I}_n - \mathbb{H})^T = (\mathbb{I}_n - \mathbb{H})$  – symetrie
- $\mathbb{H}^2 = \mathbb{H}$ ,  $(\mathbb{I}_n - \mathbb{H})^2 = \mathbb{I}_n - \mathbb{H}$  – idempotentnost
- $\mathbb{H}\mathbf{X} = \mathbf{X}$ ,  $\text{tr}(\mathbb{H}) = \sum_{i=1}^n h_{ii} = m + 1$
- $\mathbb{H}(\mathbb{I}_n - \mathbb{H}) = (\mathbb{I}_n - \mathbb{H})\mathbb{H} = \mathbf{0}$ .

**Věta 2.53.** Nechť  $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$  je LM (\*\*), kde  $h(\mathbf{X}) = m + 1$  a  $\mathbf{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbb{I}_n)$ . Potom

1.  $\hat{\beta}$  a  $s_n^2$  jsou nezávislé náhodné veličiny,

2.  $(n - m - 1) \frac{s_n^2}{\sigma^2} \sim \chi^2(n - m - 1)$ .

3. Jestliže  $v_i = (\mathbf{X}^T \mathbf{X})_{ii}^{-1}$ , potom  $T_i = \frac{\hat{\beta}_i - \beta_i}{s_n \sqrt{v_i}} \sim t(n - m - 1)$ .

4. Nechť  $\mathbb{C} \in \mathbb{R}^{r, m+1}$  takové, že  $h(\mathbf{e}) = r$ . Potom kvadratická forma

$$\frac{q}{\sigma^2} = \frac{(\hat{\beta} - \beta)^T \mathbb{C}^T [C(\mathbf{X}^T \mathbf{X})^{-1} \mathbb{C}^T]^{-1} \mathbb{C}(\hat{\beta} - \beta)}{\sigma^2} \sim \chi^2(r).$$

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

a tedy  $\hat{\beta} - \beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{e}$

dále víme, že  $\hat{\mathbf{e}} = (\mathbb{I}_n - \mathbf{H})\mathbf{e}$  a vektor  $(\hat{\beta} - \beta, \hat{\mathbf{e}})^T$  lze zapsat jako

$$\mathbf{Z} = \begin{pmatrix} \hat{\beta} - \beta \\ \hat{\mathbf{e}} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \\ \mathbb{I}_n - \mathbf{H} \end{pmatrix} \mathbf{e} = \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} \mathbf{e},$$

kde  $\mathbf{Z}$  je funkcií pouze  $(e_1, \dots, e_n) = \mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n) \Rightarrow \mathbf{Z}$  má vícerozměrné normální rozdělení ( i když degenerované, protože  $\text{Cov}(\mathbf{Z})$  je singulární ) abychom ukázali, že  $\hat{\beta}$  a  $\hat{\mathbf{e}}$  jsou nezávislé.

$(s_n^2 = \frac{1}{n-m-1} \hat{\mathbf{e}}^T \hat{\mathbf{e}})$ , tedy i  $\hat{\beta}$  a  $s_n^2$  jsou nezávislé )

**Poznámka 2.54.**  $\mathbf{B}\mathbf{B}^T = \mathbb{I}_n - \mathbf{H}$  je singulární, protože  $(\mathbb{I}_n - \mathbf{H})\mathbf{X} = 0$

Stačí nám tedy ukázat, že  $\text{Cov}(\hat{\beta}_i, \hat{e}_j) = 0$  pro  $i = 0, \dots, m$  a  $j = 1, \dots, n$  spočtěme  $\text{Cov}(\mathbf{Z})$ :

$$\text{Cov}(\mathbf{Z}) = \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} \text{Cov}(\mathbf{e}) (\mathbf{A}^T \mathbf{B}^T) = \sigma^2 \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} (\mathbf{A}^T \mathbf{B}^T) = \sigma^2 \begin{pmatrix} \mathbf{A}\mathbf{A}^T & \mathbf{A}\mathbf{B}^T \\ \mathbf{B}\mathbf{A}^T & \mathbf{B}\mathbf{B}^T \end{pmatrix}$$

$$\begin{aligned} \left( \text{Cov}(\hat{\beta}_i, \hat{e}_j) \right)_{i=0, \dots, m, j=1, \dots, n} &= \mathbf{A}\mathbf{B}^T = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbb{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) = \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = 0 \end{aligned}$$

## 2 Jednorozměrná lineární regrese

2. Výsledky z LA:

- $\mathbf{A}_{n \times n}$  symetrická matice  $\Rightarrow$  existuje ortogonální matice  $\mathbf{Q}$  a diagonální matice  $\Lambda$  tak, že  $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T$ , sloupce  $\mathbf{Q}$  jsou ON vlastní vektory matice  $\mathbf{A}$  a diagonální prvky matice  $\Lambda$  jsou ??? odpovídající vlastní čísla.
- $\mathbf{A}_{n \times n}$  idempotentní matice  $\Rightarrow$  vlastní čísla jsou pouze 0 nebo 1  $\Rightarrow h(\mathbf{A}) = \text{tr}(\mathbf{A})$

V důkazu předchozí věty  $(n - m - 1) \frac{s_n^2}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} = \frac{1}{\sigma^2} \mathbf{e}^T (\mathbf{I}_n - \mathbf{H}) \mathbf{e}$   
protože je  $\mathbf{I}_n - \mathbf{H}$  symetrické a idempotentní

$$\mathbf{I}_n - \mathbf{H} = \mathbf{Q}\Lambda\mathbf{Q}^T \quad \text{kde} \quad \begin{array}{l} \mathbf{Q} \dots \text{ortogonální matice} \\ \Lambda \dots \text{diagonální matice s vlastními čísly } \mathbf{I}_n - \mathbf{H} \end{array}$$

protože vlastní čísla  $\mathbf{I}_n - \mathbf{H}$  jsou 0 nebo 1 a  $\text{tr}(\mathbf{I}_n - \mathbf{H}) = h(\mathbf{I}_n - \mathbf{H}) = n - m - 1$   
 $\Lambda$  může být zapsána ve tvaru:

$$\Lambda = \begin{pmatrix} \mathbf{I}_{n-m-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

takže

$$\mathbf{e}(\mathbf{I}_n - \mathbf{H})\mathbf{e} = \mathbf{e}\mathbf{Q}\Lambda\mathbf{Q}^T\mathbf{e} = \mathbf{q}^T\Lambda\mathbf{q} \quad \text{kde} \quad \mathbf{q} = \mathbf{Q}^T\mathbf{e}$$

**Věta 2.55.**  $\mathbf{V} \sim \mathbf{N}_n(\mathbf{0}, \mathbf{I}_n)$  a  $\mathbf{Q}$  je ortogonální matice, potom  $\mathbf{Q}\mathbf{V} \sim \mathbf{N}_n(\mathbf{0}, \mathbf{I}_n)$

tzv.  $\mathbf{q}$  je vektor nezávislých  $N(0, \sigma^2)$  veličin ( $\mathbf{q} \sim \mathbf{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ )

a  $\frac{1}{\sigma^2} \mathbf{e}^T (\mathbf{I}_n - \mathbf{H}) \mathbf{e} = \frac{1}{\sigma^2} \mathbf{q}^T \Lambda \mathbf{q} = \sum_{i=1}^{n-m-1} \frac{q_i^2}{\sigma^2} \sim \chi^2(n - m - 1)$   
je suma druhých mocnin  $n - m - 1$  nezávislých  $N(0, 1)$  veličin.

3. Z předchozí věty:

$$\frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{v_i}} \sim N(0, 1) \quad \text{a} \quad \frac{s_n}{\sigma} = \sqrt{\frac{(n-m-1)s_n^2}{n-m-1}} = \sqrt{\frac{\chi^2(n-m-1)}{n-m-1}}$$

a z bodu 1) nezávislost

$$T_i = \frac{\hat{\beta}_i - \beta_i}{s_n \sqrt{v_i}} = \frac{\frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{v_i}}}{\frac{s_n}{\sigma}} \sim t(n - m - 1)$$

4.  $\mathbf{C}\hat{\beta} \sim \mathbf{N}_r(\mathbf{C}\beta, \sigma^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T)$  a tedy

$$\mathbf{C}(\hat{\beta} - \beta) = \mathbf{C}\hat{\beta} - \mathbf{C}\beta \sim \mathbf{N}_r(\mathbf{0}, \sigma^2 \mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T)$$

stačí tedy ukázat, že pokud  $\mathbf{Z} \sim \mathbf{N}_r(\mathbf{0}, \Sigma)$ , potom  $\mathbf{Z}^T \Sigma \mathbf{Z} \sim \chi^2(r)$

protože  $\Sigma$  je pozitivně definitní, existuje regulární matice  $\mathbf{R}$  taková, že  $\Sigma = \mathbf{R}\mathbf{R}^T$

protože  $\mathbf{U} = \mathbf{R}^{-1}\mathbf{Z}$ , potom  $\mathbb{E}\mathbf{U} = \mathbf{R}^{-1}\mathbb{E}[\mathbf{Z}] = \mathbf{0}$

$$\text{Cov}(\mathbf{U}) = \mathbf{R}^{-1}\Sigma(\mathbf{R}^{-1})^T = \mathbf{R}^{-1}\mathbf{R}\mathbf{R}^T(\mathbf{R}^T)^{-1} = \mathbf{I}_r \quad \text{tedy } \mathbf{U} \sim \mathbf{N}_r(\mathbf{0}, \mathbf{I}_r)$$

takže složky  $\mathbf{U}$  jsou nezávislé  $N(0, 1)$  rozdělené náhodné veličiny.

$$\mathbf{R}^T \Sigma^{-1} \mathbf{R} = \mathbf{U}^T \mathbf{R}^T (\mathbf{R}^T)^{-1} \mathbf{R} \mathbf{U} = \mathbf{U}^T \mathbf{U} = \sum_i^r U_i^2 \sim \chi^2(r)$$

$$\text{za } \mathbf{R} = \mathbf{C}(\hat{\beta} - \beta) \text{ a } \Sigma^{-1} = \frac{1}{\sigma^2} [\mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T]^{-1} \quad \text{Q. E. D.}$$

□

### 2.9.1 Vlastnosti vektoru reziduí $\hat{\mathbf{e}}$

**Věta 2.56.** Uvažujeme model  $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ , kde  $e_1, \dots, e_n$  jsou nekorelované a  $e_i \sim (0, \sigma^2)$ . Nechť  $\hat{\beta}$  je OLS  $\beta$  a  $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$  je vektor reziduí. Potom platí:

1.  $\mathbb{E}[\hat{\mathbf{e}}] = 0$
2.  $\text{Cov}(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$
3. pokud navíc  $\mathbf{e} \sim \mathbf{N}_n(0, \sigma^2 \mathbf{I}_n)$ , potom  $\hat{\mathbf{e}} \sim \mathbf{N}_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$
4. jestliže má model intercept, tj.  $\beta_0 \neq 0$ , potom  $\sum_{i=1}^n \hat{e}_i = 0$
5.  $\sum_{i=1}^n \hat{e}_i \hat{y}_i = 0$

Důkaz. Ukázali jsme, že  $\hat{\mathbf{e}} = (\mathbf{I}_n - \mathbf{H})\mathbf{e}$

1.  $\mathbb{E}[\hat{\mathbf{e}}] = (\mathbf{I}_n - \mathbf{H}) \cdot \mathbb{E}[\mathbf{e}] = (\mathbf{I}_n - \mathbf{H}) \cdot 0 = 0$
2.  $\text{Cov}(\hat{\mathbf{e}}) = (\mathbf{I}_n - \mathbf{H})\text{Cov}(\mathbf{e})(\mathbf{I}_n - \mathbf{H})^T = \sigma^2(\mathbf{I}_n - \mathbf{H})$
3.  $\hat{\mathbf{e}}$  je LK složek  $\mathbf{e} \Rightarrow \hat{\mathbf{e}} \sim \mathbf{N}_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$
4. Soustava normálních rovnic  $\mathbf{X}^T \mathbf{X}\beta = \mathbf{X}^T \mathbf{y}$  lze zapsat  $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$

První rovnice:

$$\sum_{i=1}^n x_{i1} \cdot (y_i - \mathbf{x}_i^T \beta) = 0$$

pro  $\hat{\beta}$  tedy platí

$$0 = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \hat{\beta}) = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n \hat{e}_i$$

5. Z předchozího bodu platí pro OLS  $\hat{\beta}$

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0 \quad \text{přenásobením zleva } \hat{\beta}^T$$

$$0 = \hat{\beta}^T \mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = \hat{y}^T(\mathbf{y} - \hat{\mathbf{y}}) = \hat{\mathbf{y}}^T \hat{\mathbf{e}} = \sum_{i=1}^n \hat{y}_i \hat{e}_i$$

□

Použitím bodů 4. a 5. dostaneme (stejně jako u jednorozměrné regrese)

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

tedy

$$\text{SST} = \text{SSR} + \text{SSE}$$

### 2.9.2 Gauss - Markov theorem

$e_i$  i.i.d.  $N(0, \sigma^2) \Rightarrow$  OLS  $\hat{\beta}$  je MLE, tzn. je eficientní MVVE parametr  $\beta$

Chybou nenormální:

- ukážeme, že OLS  $\hat{\beta}$  je BLUE (best linear unbiased estimation) parametru  $\beta$  (za jistých podmínek)
- mohou ale existovat lepší lineární vychýlené odhady nebo nelineární odhady.

**Definice 2.57.** Nechť  $\beta$  je vektor regresních parametrů v lineárním modelu (LM). Řekněme, že  $\hat{\beta}$  je lineární odhad  $\beta$ , jestliže každé  $\beta_i$  je LK pozorování  $Y_i$ ,  $i = 1, \dots, n$ , tedy

$$\hat{\beta}_i = \sum_{j=1}^n a_{ij} Y_j \quad i = 0, \dots, m$$

V maticovém zápisu

$$\hat{\beta} = \mathbf{AY} \quad \text{kde} \quad \mathbf{A} = (a_{ij})$$

pro  $i = 0, \dots, m$  a  $j = 1, \dots, n$

**Poznámka 2.58.** Pokud v modelu  $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$  platí  $h(\mathbf{X}) = m+1$ , potom OLS  $\hat{\beta}$  je lineární, neboť  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ , kde  $\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$

**Věta 2.59** (Gauss-Markov). *Uvažujeme model  $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ , kde matice  $\mathbf{X}$  má plnou hodnost,  $e_i$ ,  $i = 1, \dots, n$  jsou nekorelované a  $e_i \sim (0, \sigma^2)$ . Potom OLS odhad  $\hat{\beta}$  je BLUE parametru  $\beta$  (best linear unbiased estimation)*

**Důkaz.** Nechť  $\hat{\beta} = \mathbf{AY}$  je lineární odhad  $\beta$ , aby byl nestranný musí platit  $\mathbb{E}[\hat{\beta}] = \beta$ , tzn.  $\mathbb{E}[\mathbf{AY}] = \mathbf{A}\mathbb{E}[\mathbf{Y}] = \mathbf{AX}\beta = \beta$ , tedy  $(\mathbf{AX} - \mathbf{I}_{m+1})\beta = 0$  protože to musí platit  $\forall \beta \in \mathbb{R}^{m+1}$ , dostáváme  $\mathbf{AX} - \mathbf{I}_{m+1} = 0$ , nebo ekvivalentně  $\mathbf{AX} = \mathbf{I}_{m+1}$ .

Spočteme kovarianční matici  $\hat{\beta}$

$$\text{Cov}(\hat{\beta}) = \text{Cov}(\mathbf{AY}) = \mathbf{ACov}(\mathbf{Y})\mathbf{A}^T = \sigma^2 \mathbf{AA}^T = \sigma^2 \mathbf{I}_n$$

zapišme  $\mathbf{A}$  ve tvaru  $\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + \mathbf{D}$  kde  $\mathbf{D}$  je rozdíl mezi  $\mathbf{A}$  a maticí pro OLS odhad. Pokud ukážeme, že pro nestranný lineární odhad  $\hat{\beta} = \mathbf{AY}$ , který minimalizuje rozptyl, musí platit  $\mathbf{D} = 0$ , bude věta dokázána.

Dosazením dostaneme:

$$\begin{aligned} \text{Cov}(\hat{\beta}) &= \sigma^2 ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + \mathbf{D}) ((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + \mathbf{D})^T = \\ &= \sigma^2 [(\mathbf{X}^T \mathbf{X})^{-1} + \mathbf{DX}(\mathbf{X}^T \mathbf{X})^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D}^T + \mathbf{DD}^T] \end{aligned}$$

z podmínek nerovnosti

$$\mathbf{AX} = [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T + \mathbf{D}] \mathbf{X} = \mathbf{I}_{m+1} + \mathbf{DX} = \mathbf{I}_{m+1} \Rightarrow \mathbf{DX} = 0 \text{ a tedy i } \mathbf{D}^T \mathbf{X}^T = \mathbf{0}$$

tzn.

$$\text{Cov}(\hat{\beta}) = \sigma^2 [(\mathbf{X}^T \mathbf{X})^{-1} + \mathbf{DD}^T]$$

## 2 Jednorozměrná lineární regrese

pro diagonální prvky platí

$$D[\hat{\beta}_i] = \sigma^2[v_i + \sum_{j=1}^n d_{ij}^2] \quad i = 0, \dots, m$$

protože  $v_i \geq 0$  a  $\sum_{j=1}^n d_{ij}^2 \geq 0 \Rightarrow D[\hat{\beta}_i]$  je minimalizován volnou  $\sum_{j=1}^n d_{ij}^2 = 0$ , tj.  $d_{ij} = 0 \quad j = 1, \dots, n$  platí  $\forall i = 0, \dots, m \Rightarrow \mathbf{D} = \mathbf{0}$  tzn. lineárně nestranný odhad  $\hat{\beta}$ , který minimalizuje  $D[\hat{\beta}_i]$ ,  $i = 0, \dots, m$  je  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$   $\square$

### 2.9.3 Testování modelu - tabulka ANOVA

#### Celkový F-test ( overall F-test )

- Zajímá nás, zda je model statisticky signifikantní, tj. zda alespoň jeden z koeficientů  $\beta_1, \dots, \beta_m$  je nulový.
- Mohli bychom testovat jednotlivé koeficienty  $H_0 : \beta_j = 0$  pomocí alternativy t-testu.
- Celková chyba I. druhu by takto ale mohla být velká, pokud máme hodně proměnných. Museli bychom hodně snížit  $\alpha$  pro jednotlivé testy, což což zvýší pravděpodobnost chyby II. druhu (tzn. riziko akceptování nenulových koeficientů jako nulových a tedy vynechání významných proměnných z modelu)
- Navíc je zde problém multikolinearity (viz později) jejíž jedním efektem jsou velké stand. chyby dohadů. To může vést k akceptování všech koeficientů jeho 0, i když je model celkově významný (uvidíme na příkladu)

Bylo by dobré mít jednu statistiku pro test

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0 \quad \times \quad H_1 : (\exists i \in \hat{m}, \beta_i \neq 0)$$

ANOVA přístup pro jedn. regresi naznačuje, že statistika

$$F = \frac{\frac{\text{SSR}}{m}}{\frac{s^2}{\hat{m}}} \quad \text{by mohla být užitečná}$$

(vyplýne i z obecnějších přístupů k testování později) **Označení:**  $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$  – průměr j-tého sloupce matice  $\mathbf{X}$ ,

$$\bar{\mathbf{X}} = \begin{pmatrix} \bar{x}_0 & \bar{x}_1 & \cdots & \bar{x}_m \\ \vdots & \vdots & & \vdots \\ \bar{x}_0 & \bar{x}_1 & \cdots & \bar{x}_m \end{pmatrix}_{n \times m+1} \quad \underbrace{(\mathbf{X}_c)_{ij}}_{\text{centrované matice regresorů}} = x_{ij} - \bar{x}_j, \quad i = 1, \dots, n, j = 1, \dots, m$$

**Věta 2.60.** V modelu  $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ , kde  $e_i$  jsou nekorelované a  $e_i \sim \mathcal{N}(0, \sigma^2)$  pro  $i = 1, \dots, n$  platí

$$\mathbb{E} \left[ \frac{\text{SSR}}{n} \right] = \sigma^2 + \frac{\beta^T (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}}) \beta}{m} = \sigma^2 + \frac{\beta_s^T \mathbf{X}_c^T \mathbf{X}_c \beta_s}{m},$$

kde  $\beta_s = (\beta_1, \dots, \beta_m)$ .

## 2 Jednorozměrná lineární regrese

Důkaz.

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \sum_{j=1}^m x_{ij} \hat{\beta}_j \\ \frac{\partial SSE}{\partial \beta_0} &= \sum_{i=1}^n \left( y_i - (\beta_0 + \sum_{j=1}^m x_{ij} \hat{\beta}_j) \right) = 0 \implies \hat{\beta}_0 = \bar{y} - \sum_{j=1}^m \bar{x}_j \hat{\beta}_j\end{aligned}$$

Celkem pak  $\hat{y}_i - \bar{x} = \sum_{j=1}^m (x_{ij} - \bar{x}_j) \hat{\beta}_j$ ,  $i = 1, \dots, n$  a zapsáno maticově:

$$\hat{Y} = \bar{Y} = (\mathbf{X} - \bar{\mathbf{X}})\hat{\beta}, \quad \text{kde } \bar{Y} = (\bar{y}, \bar{y}, \dots, \bar{y})_{1 \times n}^T,$$

protože první sloupec matice  $\mathbf{X} - \bar{\mathbf{X}}$  je nulový. Potom

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (\hat{Y} - \bar{Y})^T (\hat{Y} - \bar{Y}) = \hat{\beta}^T \underbrace{(\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}})}_{\mathbb{A}} \hat{\beta} = \hat{\beta}^T \mathbb{A} \hat{\beta}$$

□

**Věta 2.61.** Nechť  $Z = \mathbf{Y}^T \mathbb{A} \mathbf{Y}$  je kvadratická forma a nechť  $\mathbb{E} \mathbf{Y} = \boldsymbol{\mu}$  a  $\text{Cov} \mathbf{Y} = \Sigma$ . Potom platí:

$$\mathbb{E} Z = \text{tr}(\mathbb{A} \Sigma) + \boldsymbol{\mu}^T \mathbb{A} \boldsymbol{\mu}.$$

Důkaz. Nejdříve zjednodušíme matici  $\mathbb{A}$ :

$$\begin{aligned}\bar{\mathbf{X}} &= \frac{1}{n} \mathbb{B}, \quad \text{kde } \mathbb{B} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \quad \text{a tedy} \\ \mathbf{X} - \bar{\mathbf{X}} &= \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) \mathbf{X} \quad \text{a} \quad (\mathbf{X} - \bar{\mathbf{X}})^T = \mathbf{X}^T \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) \\ \mathbb{A} &= (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}}) = \mathbf{X}^T \underbrace{\left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right)^2}_{\mathbb{I}_n - \frac{2}{n} \mathbb{B} + \frac{\mathbb{B}^2}{n^2}} \mathbf{X} = \mathbf{X}^T \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) \mathbf{X}\end{aligned}$$

Dále rozepíšeme  $\underbrace{\mathbb{A} \Sigma}_{=\mathbb{A} \text{Cov} \hat{\beta}} = \sigma^2 \mathbf{X}^T \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$  a spočítáme  $\text{tr}(\mathbb{A} \Sigma)$ :

$$\begin{aligned}\text{tr}(\mathbb{A} \Sigma) &= \sigma^2 \text{tr} \left[ \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) \right] = \sigma^2 \text{tr} \left[ \mathbb{H} - \frac{1}{n} \mathbb{H} \mathbb{B} \right] = \\ &= \sigma^2 \left[ \text{tr} \mathbb{H} - \frac{1}{n} \text{tr} (\mathbb{H} \mathbb{B}) \right] = \sigma^2 \left[ \underbrace{\text{tr} \mathbb{H}}_{=m+1} \frac{1}{n} \underbrace{\text{tr} \mathbb{B}}_{=n} \right] = \sigma^2 m,\end{aligned}$$

jelikož víme, že  $\mathbb{H} \mathbf{X} = \mathbf{X}$  a  $\mathbf{1} = (1, \dots, 1)^T$  je první sloupec  $\mathbf{X}$ , takže  $\mathbb{H} \mathbf{1} = \mathbf{1}$  a tedy  $\mathbb{H} \mathbb{B} = \mathbb{B}$ . Celkem tak dostáváme

$$\mathbb{E} \left( \frac{SSR}{m} \right) = \frac{1}{m} (\sigma^2 m + \boldsymbol{\beta}^T (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}}) \boldsymbol{\beta}) = \sigma^2 + \frac{1}{m} \boldsymbol{\beta}^T (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}}) \boldsymbol{\beta}$$

Navíc platí  $(\mathbf{X} - \bar{\mathbf{X}}) \boldsymbol{\beta} = \mathbf{X}_c \boldsymbol{\beta}_s$ , protože první sloupec matice  $\mathbf{X} - \bar{\mathbf{X}}$  je nulový vektor. □

## 2 Jednorozměrná lineární regrese

**Poznámka 2.62.** Pokud  $\beta_s = 0$ , potom  $\mathbb{E}\left(\frac{SSR}{m}\right) = \sigma^2 = \mathbb{E}s_n^2$ , takže  $\beta_s \neq 0$  implikuje, že  $\mathbb{E}\left(\frac{SSR}{m}\right) > \sigma^2$ , tedy velké hodnoty  $F = \frac{SSR/m}{s_n^2}$  budou znamenat zamítnutí  $H_0 : \beta_s = 0$ . Budeme proto potřebovat rozdělení  $F$  za platnosti  $H_0$ .

**Věta 2.63.** Nechť v modelu  $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$  (\*\*\*) jsou  $e_1, \dots, e_n$  iid  $\mathcal{N}(0, \sigma^2)$ . Pokud  $\beta_s = 0$ , tj.  $\beta_1 = \beta_2 = \dots = \beta_m = 0$ , potom

$$F \sim F(m, n - m - 1).$$

*Důkaz.* V důkazu minulé věty jsme ukázali

$$SSR = \hat{\beta}^T \mathbf{X}^T \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) \mathbf{X} \hat{\beta} = \hat{\mathbf{Y}}^T \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) \hat{\mathbf{Y}}$$

a potřebujeme rozepsat  $\hat{\mathbf{Y}}$ :

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbb{H}\mathbf{Y} = \mathbb{H}(\mathbf{X}\beta + \mathbf{e}) = \mathbb{H}(\mathbf{1}\beta_0 + \mathbf{X}_v\beta_s + \mathbf{e}) = \beta_0 \underbrace{\mathbb{H}\mathbf{1}}_{=\mathbf{1}} + \mathbb{H}\mathbf{X}_v \underbrace{\beta_s}_{=0} + \mathbb{H}\mathbf{e} = \beta_0 \mathbf{1} + \mathbb{H}\mathbf{e} \\ SSR &= (\beta_0 \mathbf{1}^T + \mathbf{e}^T \mathbb{H}) \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) (\beta_0 \mathbf{1} + \mathbb{H}\mathbf{e}) = \mathbf{e}^T \mathbb{H} \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) \mathbb{H}\mathbf{e}, \end{aligned}$$

protože  $(\mathbb{I}_n - \frac{1}{n} \mathbb{B}) \mathbf{1} = 0$  a  $(\mathbb{I}_n - \frac{1}{n} \mathbb{B})$  je symetrická.

Dále platí  $\mathbb{H} = \mathbb{H}^T$ ,  $\mathbb{H}^2 = \mathbb{H}$  a  $\mathbb{H}\mathbb{B} = \mathbb{B}\mathbb{H} = \mathbb{B}$  (protože  $\mathbb{H}\mathbf{1} = \mathbf{1}$ ) a celkem tedy dostáváme

$$SSR = \mathbf{e}^T \underbrace{\left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right)}_{\text{ozn. } \mathbb{C}} \mathbf{e} = \mathbf{e}^T \mathbb{C} \mathbf{e}.$$

Pro matici  $\mathbb{C}$  platí

$$\begin{aligned} \mathbb{C}^T &= \left( \mathbb{I}_n^T - \frac{1}{n} \mathbb{B}^T \right) = \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) = \mathbb{C} \\ \mathbb{C}^2 &= \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) = \mathbb{H}^2 - \frac{1}{n} \mathbb{H}\mathbb{B} - \frac{1}{n} \mathbb{B}\mathbb{H} + \frac{1}{n^2} \mathbb{B}^2 = \mathbb{H} - \frac{2}{n} \mathbb{B} + \frac{1}{n} \mathbb{B} = \mathbb{H} - \frac{1}{n} \mathbb{B} = \mathbb{C}, \end{aligned}$$

tedy  $\mathbb{C}$  je symetrická a idempotentní, a proto

$$h(\mathbb{C}) = \text{tr}(\mathbb{C}) = \text{tr} \left( \mathbb{I}_n - \frac{1}{n} \mathbb{B} \right) = m + 1 - 1 = m.$$

Z věty o spektrálním rozkladu plyne existence  $Q$  OG a ??  $\Lambda$  tak, že

$$\mathbb{C} = \mathbf{Q}^T \Lambda \mathbf{Q} = \mathbf{Q}^T \begin{pmatrix} I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Q},$$

která má vlastní čísla 0 a 1, protože se jedná o idempotentní matici. Dále potom

$$SSR = \mathbf{e}^T \mathbf{Q}^T \begin{pmatrix} I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \underbrace{\mathbf{Q}\mathbf{e}}_{\mathbf{Z} \sim \mathcal{N}(0, \sigma^2 I_m)} = \mathbf{Z}^T \begin{pmatrix} I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Z} = \sum_{i=1}^n Z_i^2,$$

## 2 Jednorozměrná lineární regrese

kde  $Z_i \sim \mathcal{N}(0, \sigma^2)$  jsou nezávislé. Z toho vyplývá, že

$$\frac{Z_i}{\sigma} \sim \mathcal{N}(0, 1) \quad \text{a} \quad \frac{\text{SSR}}{\sigma^2} \sim \chi^2(m).$$

To znamená, že

$$\frac{\frac{\text{SSR}}{\sigma^2 m}}{\frac{(n-m-1)s_n^2}{\sigma^2} \frac{1}{n-m-1}} = \frac{\frac{\text{SSR}}{\sigma^2 m}}{s_m^2} = F \sim F(m, n-m-1),$$

pokud ukážeme, že SSR a  $s_n^2$  jsou nezávislé. K tomu ale stačí dokázat, že SSR je nezávislé na reziduích  $\hat{e}_i$ ,  $i \in \hat{n}$ .

$$\text{SSR} = \mathbf{e}^T \mathbf{H} \left( I_n - \frac{1}{n} \mathbf{B} \right) \mathbf{H} \mathbf{e} = \mathbf{e}^T \mathbf{H} \underbrace{\left( I_n - \frac{1}{n} \mathbf{B} \right) \left( I_n - \frac{1}{n} \mathbf{B} \right)}_{=I_n - \frac{1}{n} \mathbf{B}} \mathbf{H} \mathbf{e} = \frac{T}{\mathbf{w}^T \mathbf{w}},$$

kde  $\mathbf{w} = (I_n - \frac{1}{n} \mathbf{B}) \mathbf{H} \mathbf{e} \equiv \mathbf{K} \mathbf{e}$ ,  $\hat{\mathbf{e}} = (I_n - \mathbf{H}) \mathbf{e} \equiv \mathbf{L} \mathbf{e}$ . Stačí tedy ukázat, že  $\mathbf{w}$  a  $\hat{\mathbf{e}}$  jsou nezávislé vektory. Víme, že

$$\begin{pmatrix} \mathbf{w} \\ \hat{\mathbf{e}} \end{pmatrix} = \begin{pmatrix} \mathbf{K} \\ \mathbf{L} \end{pmatrix} \mathbf{e},$$

tzn. má vícerozměrné normální rozdelení. Pokud je výraz  $\mathbf{KL}^T$  z rovnice 2.5 roven nule, pak jsou  $\mathbf{w}$  a  $\hat{\mathbf{e}}$  nezávislé.

$$\text{Cov} \begin{pmatrix} \mathbf{w} \\ \hat{\mathbf{e}} \end{pmatrix} = \begin{pmatrix} \mathbf{K} \\ \mathbf{L} \end{pmatrix} \text{Cov } \mathbf{e}(\mathbf{K}^T, \mathbf{L}^T) = \sim^2 \begin{pmatrix} \mathbf{KK}^T & \mathbf{KL}^T \\ \mathbf{LK}^T & \mathbf{LL}^T \end{pmatrix} \quad (2.5)$$

Pro  $\mathbf{KL}^T$  platí, že

$$\mathbf{KL}^T = \left( I_n - \frac{1}{n} \mathbf{B} \right) \underbrace{\mathbf{H} (I_n - \mathbf{H})}_{\mathbf{H} - \mathbf{H}^2 = \mathbf{0}} = 0.$$

□

**TEST:**  $H_0$  zamítáme, pokud  $F > F_{1-\alpha}(m, n-m-1)$ .

**Poznámka 2.64.** Odvozeno pro  $e_i \sim \mathcal{N}(0, \sigma^2)$ , obecně se používá, i když to nevíme, pro velké  $n$  může být často zdůvodněno pomocí CLV.

### Tabulka ANOVA

Source	df	SS	MS	F
Regression	$m$	SSR	$\text{MSR} = \frac{\text{SSR}}{m}$	$\frac{\text{MSR}}{\text{MSE}}$
Residual	$n - (m + 1)$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-m-1} = s_n^2$	
Total	$n - 1$	SST		
		$R^2$	$\bar{R}^2$	

### Koeficient (vícenásobná) determinace $R^2$

Podobně jako u jednorozměrné regrese, lze F-test chápát jako test významnosti  $R^2$ , definovaného jako

$$R^2 \equiv \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

protože

$$F = \frac{\frac{SSR}{m}}{\frac{SSE}{n-m-1}} = \frac{n-m-1}{m} \left( \frac{\frac{SSR}{SST}}{\frac{SSE}{SST}} \right) = \frac{n-m-1}{m} \frac{R^2}{1-R^2},$$

což je rostoucí funkce  $R^2$  (opět  $R^2 \in [0, 1]$ ).

**POZNÁMKA 2.65.**  $R^2$  je možno zvětšovat přidáváním nových proměnných  $x$ , i když jsou statisticky nevýznamné. (Pro  $n$  LN proměnných  $x$  a  $n$  pozorování dostaneme "perfect fit", tedy přeúčení.) Vysvětlení:

$$R^2 = 1 - \frac{SSE}{SST},$$

kde SST je pevně dán daty  $y$ , ale SSE může být snížena přidáním proměnných  $x$ . Minimizujeme totiž  $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$  přes větší množinu  $\boldsymbol{\beta}$ . To znamená, že  $\frac{SSE}{SST}$  je nerostoucí funkce počtu proměnných, a tedy  $R^2$  je neklesající funkce počtu proměnných. Z tohoto důvodu se někdy definuje **upravený koeficient determinace** (adjusted coefficient of determination)

$$\bar{R}^2 = R_{adj}^2 = 1 - \frac{\frac{SSE}{n-m-1}}{\frac{SST}{n-1}} = 1 - \frac{n-1}{n-m-1} \frac{SSE}{SST}.$$

(S rostoucím  $m$  klesá SSE, ale i  $n - m - 1$ .)

### 3 IS a t-testy pro parametry

- Pokud se model ukáže jako významný, bude nás zajímat, které koeficienty přispívají.
- Lze použít IS a TH stejně, jako u jednorozměrné regrese.
- Výsledky jsou odvozeny pro normální chyby.
- V praxi se používají i pro jiné typy chyb (za jistých předpokladů budou platit asymptoticky, lze je použít pro velká  $n$ ).

Pro konstrukci použijeme dokázanou vlastnost

$$T_j = \frac{\hat{\beta}_j - \beta_j}{s_n \sqrt{v_j}} \sim t(n - m - 1), \quad \text{kde} \quad v_j = (\mathbf{X}^T \mathbf{X})_{jj}^{-1}.$$

Standardním postupem získáme  $100(1 - \alpha)\%$ . IS pro  $\beta_j$  ve tvaru

$$(\hat{\beta}_j - t_{1-\frac{\alpha}{2}}(n - m - 1)s_n \sqrt{v_j}, \hat{\beta}_j + t_{1-\frac{\alpha}{2}}(n - m - 1)s_n \sqrt{v_j})$$

s jejich pomocí lze odvodit kritický obor pro test

$$H_0 : \beta_j = b_j \quad \text{vs.} \quad H_1 : \beta_j \neq b_j$$

ve tvaru

$$\frac{|\hat{\beta}_j - b_j|}{s_n \sqrt{v_j}} > t_{1-\frac{\alpha}{2}}(n - m - 1).$$

Pro  $b_j = 0$  dostaneme test významnosti  $\beta_j$ , tzn.  $H_0 : \beta_j = 0$  zamítneme, pokud

$$\frac{|\hat{\beta}_j|}{s_n \sqrt{v_j}} > t_{1-\frac{\alpha}{2}}(n - m - 1).$$

**Poznámka 3.1.** • Pokud nejsou porušeny předpoklady modelu nebo není přítomna kolinearita, lze zvážit odstranění všech nevýznamných proměnných (dle t-testu).

- V případě kolinearita, model může být významný (dle celkového F-testu), ale všechny nebo téměř všechny proměnné se mohou jevit jako nevýznamné (dle t-testů).
- Naopak, pokud má model velký počet možných proměnných, některé proměnné se mohou jevit významné, i když jsou náhodným šumem.
- Při použití t-testů je třeba být obezřetný.

**Příklad 3.2.** 5.26, str. 230 a 5.27, str. 231

**Poznámka 3.3.** Statistiky F,  $R^2$  a t jsou užitečné pro rozkrytí efektů jednoduchých proměnných, nemohou být ale používány úplně automaticky.

### 3.0.1 Obecná lineární hypotéza

F-test a t-testy jsou speciálním případem **obecné lineární hypotézy**

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b} \quad \text{vs.} \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{b},$$

kde  $\mathbf{C} \in \mathbb{R}^{r \times (m+1)}$  a  $h(\mathbf{C}) = r$ , tzn.  $r \leq m + 1$ . Rovnice  $\mathbf{C}\boldsymbol{\beta} = \mathbf{b}$  reprezentuje  $r$  lineárně nezávislých podmínek

$$\sum_{j=0}^m c_{ij}\beta_j = b_i, \quad i = 1, \dots, r.$$

**POZNÁMKA 3.4.** a) Volba  $\mathbf{b} = (0, \dots, 0)^T$  a  $\mathbf{C} = \left( \begin{array}{c|cccc} 0 & 1 & 0 & \dots & 0 \\ \hline 0 & 0 & 1 & & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & 1 \end{array} \right)_{m \times (m+1)}$  vede na test

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0} \quad \Leftrightarrow \quad H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0.$$

b) Volba  $\mathbf{b} = \mathbf{0}$  a  $\mathbf{C} = (0, \dots, 0, 1, 0, \dots, 0)$  vede na test

$$H_0 : \beta_j = 0.$$

c) V modelu  $Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + e$  chceme testovat zároveň, že  $\beta_2 = 0$  a  $\beta_3 = \beta_4$ . To lze udělat volbou  $\mathbf{C} = \left( \begin{array}{ccccc} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{array} \right)$ ,  $\mathbf{b} = (0, 0)^T$ .

Pro test  $H_0$  naladíme 2 modely:

**plný model (full model)** bez podmínek na  $\mathbf{C}\boldsymbol{\beta}$ ,

**redukovaný model (reduced model)** za předpokladu, že platí  $H_0 : \mathbf{C}\boldsymbol{\beta} = b$ .

Označme příslušné reziduální součty čtverců  $SSE_F$  a  $SSE_R$  (bude platit  $SSE_F \leq SSE_R$ ).

- Pokud neplatí  $H_0$ , dá se očekávat, že  $\Delta SSE = SSE_R - SSE_F$  bude významně větší, než náhodná chyba  $\sigma^2$ ,  $H_0$  tedy budeme zamítat, pokud  $\frac{\Delta SSE}{s_n^2}$  bude velké.
- Zobecnění F-testu, tj. za platnosti  $H_0$  ukázeme pro normální chyby vztah

$$F = \frac{\frac{\Delta SSE}{r}}{\frac{s_n^2}{s_n^2}} \sim F(r, n - m - 1).$$

**PŘÍKLAD 3.5.** Uvažujme F-test pro  $H_0 : \beta_1 = \beta_2 = \dots = \beta_m = 0$  v plném modelu. Redukovaný model bude  $Y_i = \beta_0 + e_i$ ,  $i = 1, \dots, n \Rightarrow \hat{\beta}_0 = \bar{y}$  a  $SSE_R = \sum_{i=1}^n (y_i - \bar{y})^2 = SST$ , tedy

$$\Delta SSE = SST - SSE_P = \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SSR$$

a statistiku  $F = \frac{\frac{SSR}{m}}{\frac{s_n^2}{s_n^2}} = F_{overall} \sim F(m, n - m - 1)$ , jak jsme již ukázali.

### 3 IS a t-testy pro parametry

**Věta 3.6.** Nechť v modelu (\*\*) platí, že  $e_1, \dots, e_n$  jsou nezávislé a  $e_i \sim \mathcal{N}(0, \sigma^2)$ . Označme  $SSE_F$  reziuální s.č. plného modelu a  $SSE_R$  reziduální s.č. modelu, kde platí  $H_0 : C\beta = b$ . Potom, za platnosti  $H_0$  je splněno

$$F = \frac{\frac{\Delta SSE}{r}}{s_n^2} \sim F(r, n - m - 1).$$

**Lemma 3.7.** Označme  $\hat{\beta}_F$  a  $\hat{\beta}_R$  LSE parametru  $\beta$  v plném a redukovaném modelu. Potom platí

$$1. \hat{\beta}_F = \hat{\beta}_R - (\mathbf{X}^T \mathbf{X})^{-1} \mathbb{C}^T \mathbb{A} (\mathbb{C} \hat{\beta}_F - b), \text{ kde } \mathbb{A} = (\mathbb{C} (\mathbf{X}^T \mathbf{X})^{-1} \mathbb{C}^T)^{-1}$$

$$2. \Delta SSE = SSE_R - SSE_F = (\mathbb{C} \hat{\beta}_F - b)^T \mathbb{A} (\mathbb{C} \hat{\beta}_F - b).$$

Důkaz. 1. Víme, že  $\hat{\beta}_F = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  a musíme najít  $\hat{\beta}_R$ . Budeme proto minimalizovat

$$g(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

za podmínky  $C\beta = b$ . Sestavíme Lagrangeovu funkci

$$\begin{aligned} L &= L(\beta) = g(\beta) - 2\lambda^T (\mathbb{C}\beta - b), \text{ kde } \lambda = (\lambda_1, \dots, \lambda_r) \\ L &= \mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}\beta + \beta^T \mathbf{X}^T \mathbf{X}\beta - 2\lambda^T \mathbb{C}\beta + 2\lambda^T b \end{aligned}$$

a tedy

$$\begin{aligned} \frac{\partial L}{\partial \beta} &= \left( \frac{\partial L}{\partial \beta_0}, \frac{\partial L}{\partial \beta_1}, \dots, \frac{\partial L}{\partial \beta_m} \right)^T = 2\mathbf{X}^T \mathbf{X}\beta - 2\mathbf{X}^T \mathbf{y} - 2\mathbb{C}^T \lambda = 0 \\ \frac{\partial L}{\partial \lambda} &= \left( \frac{\partial L}{\partial \lambda_1}, \dots, \frac{\partial L}{\partial \lambda_r} \right)^T = \mathbb{C}\beta - b = 0. \end{aligned}$$

Z první rovnice dostáváme

$$\hat{\beta}_R = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} + (\mathbf{X}^T \mathbf{X})^{-1} \mathbb{C}^T \lambda = \hat{\beta}_F + (\mathbf{X}^T \mathbf{X})^{-1} \mathbb{C}^T \lambda \quad (+)$$

a dosadíme do druhé

$$\mathbb{C}\hat{\beta}_R - b = \mathbb{C}\hat{\beta}_F - b + \mathbb{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbb{C}^T \lambda = 0.$$

Můžeme tak spočítat  $\lambda = -(\mathbb{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbb{C}^T)^{-1} (\mathbb{C}\hat{\beta}_F - b)$  a dosazením do rovnice (+) získáme

$$\hat{\beta}_R = \hat{\beta}_F - (\mathbf{X}^T \mathbf{X})^{-1} \mathbb{C}^T (\mathbb{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbb{C}^T)^{-1} (\mathbb{C}\hat{\beta}_F - b) = \hat{\beta}_F - (\mathbf{X}^T \mathbf{X})^{-1} \mathbb{C} \mathbb{A} (\mathbb{C}\hat{\beta}_F - b).$$

2. Z důkazu věty  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \beta$  víme, že

$$g(\beta) - g(\hat{\beta}_F) = (\beta - \hat{\beta}_F) \mathbf{X}^T \mathbf{X} (\beta - \hat{\beta}_F) \quad \forall \beta.$$

Dosadíme  $\beta = \hat{\beta}_R$ :

$$\begin{aligned} \Delta SSE &= g(\hat{\beta}_R) - g(\hat{\beta}_F) = (\hat{\beta}_R - \hat{\beta}_F) \mathbf{X}^T \mathbf{X} (\hat{\beta}_R - \hat{\beta}_F) = \\ &= (\mathbb{C}\hat{\beta}_F - b)^T \mathbb{A}^T \mathbb{C} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X} \mathbf{X})^{-1} \mathbb{C}^T \mathbb{A} (\mathbb{C}\hat{\beta}_F - b) = (\times) \end{aligned}$$

a protože  $\mathbb{A}^T = \mathbb{A}$ , platí  $\underbrace{\mathbb{A}^T \mathbb{C} (\mathbf{X} \mathbf{X})^{-1} \mathbb{C}^T \mathbb{A}}_{=\mathbb{A}^{-1}} = \mathbb{A} \implies (\times) = (\mathbb{C}\hat{\beta} - b)^T \mathbb{A} (\mathbb{C}\hat{\beta} - b)$ .

□

### 3 IS a t-testy pro parametry

*Důkaz.* Důkaz věty: Nejdříve ukážeme, že  $\frac{\Delta SSE}{\sigma^2} \sim \chi^2(r)$  za  $H_0 : \mathbb{C}\beta = b$ . Za  $H_0$ :  $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta_R, \sigma^2\mathbb{I}_n)$  a  $\hat{\beta}_F = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ , tzn.

$$\hat{\mathbf{r}} = \mathbb{C}\hat{\beta}_F - b \sim \mathcal{N}(\mathbb{E}(\hat{\mathbf{r}}), \text{Cov}(\hat{\mathbf{r}})).$$

$$\begin{aligned} \mathbb{E}\hat{\mathbf{r}} &= \mathbb{E}(\mathbb{C}\hat{\beta}_F - b) = \mathbb{E}(\mathbb{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}) - b = \mathbb{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}\mathbf{Y} - b = \\ &= \mathbb{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta_R - b = \mathbb{C}\beta_R - b = 0 \quad \text{za platnosti } H_0 \\ \text{Cov}(\hat{\mathbf{r}}) &= \mathbb{C}\text{Cov}(\hat{\beta}_F)\mathbb{C}^T = \sigma^2\mathbb{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbb{C}^T = \sigma^2\mathbb{A}^{-1} \\ \implies \hat{\mathbf{r}} &= \mathbb{C}\hat{\beta}_F - b \sim \mathcal{N}(0, \sigma^2\mathbb{A}^{-1}) \end{aligned}$$

Tedy

$$\frac{\Delta SSE}{\sigma^2} = \frac{\hat{\mathbf{r}}^T\mathbb{A}\hat{\mathbf{r}}}{\sigma^2} \sim \chi^2(r).$$

Navíc bod 4) věty na str (55), kde  $\mathbb{Z} \sim \mathcal{N}_r(0, \Sigma) \implies \mathbb{Z}^T\Sigma^{-1}\mathbb{Z} \sim \chi^2(r)$  a bod 1)  $\implies \hat{\beta}_F$  a  $s_n^2$  jsou nezávislé.

Tedy  $\Delta SSE$  je funkcií pouze  $\hat{\beta}_F$ , tzn. nezávisí na  $s_n^2$ , takže

$$F = \frac{\frac{\Delta SSE}{\sigma^2 r}}{\frac{(n-m-1)s_n^2}{\sigma^2(n-m-1)}} = \frac{\frac{\Delta SSE}{r}}{s_n^2} \sim F(r, n-m-1).$$

□

POZNÁMKA 3.8. Použitím rozkladu  $SST = SSE + SSR$  dostaneme

$$\Delta SSE = SSR_F - SSR_R.$$

Interpretace: nárůst regresního součtu čtverců díky neplatnosti  $H_0$ . Dále

$$SSR_F = SSR_R + \Delta SSE,$$

kde  $\Delta SSE$  je *extra sum of squares* přidaná k  $SSR$  díky neplatnosti  $H_0$ .

Např. pokud  $\beta = (\beta_0, \beta_1, \dots, \beta_{m-1}, 0)$ , tzn.  $\beta_m = 0$  a skutečný model má  $\beta = \beta_F$ , potom  $\Delta SSE$  je extra regresní součet čtverců získaný díky přidání  $\beta_m$  do modelu.

umožňuje rozklad  $SSR$  plného modelu na jednotlivé části  $(x_1, x_2|x_1, x_3|x_2x_1, \dots)$ .

Př. analogie k Př. 5.25 str. 238

POZNÁMKA 3.9. Joint confidence region viz. Ex 5.30 str. 239

#### 3.0.2 Predikce

Jakmile máme adekvátní model, můžeme ho použít pro bodové a intervalové predikce jako u jednorozměrné regrese

a) **predikce  $\mathbb{E}[\mathbf{Y}_x]$**

Nechť  $\mathbf{x}_0 = (1, x_{0,1}, \dots, x_{0,m})^T$  je nový bod proměnné  $\mathbf{x}$  bodový odhad  $\mathbb{E}[\mathbf{Y}_{\mathbf{x}_0}]$  je roven

$$\hat{y}_{\mathbf{x}_0} = \hat{\beta}_0 + \sum_{j=1}^m x_{0,j}\hat{\beta}_j = \mathbf{x}_0^T\hat{\beta}$$

### 3 IS a t-testy pro parametry

tzn.  $D[\hat{\mathbf{Y}}_{\mathbf{x}_0}] = \mathbf{x}_0^T \cdot D[\hat{\beta}] \cdot \mathbf{x}_0 = \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$  a může být odhadnut pomocí  $\hat{\sigma}^2(\hat{\mathbf{Y}}_{\mathbf{x}_0}) = s_n^2 [\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0]$  (rozptyl predikce). Speciálně pokud  $\mathbf{x}_0^T = \mathbf{x}_i^T$  (i-tý řádek matice  $\mathbf{X}$ )

$$\hat{\sigma}^2(\hat{\mathbf{Y}}_{\mathbf{x}_i}) = s_n^2 [\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i] = s_n^2 h_{ii} \quad \text{kde } h_{ii} = (\mathbf{H})_{ii} \quad \text{a } \mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T.$$

Pro normální chyby lze odvodit interval spolehlivosti pro  $\mathbb{E}[\mathbf{Y}_{\mathbf{x}_0}] = \gamma_{\mathbf{x}_0}$ , protože  $\hat{\mathbf{Y}}_{\mathbf{x}_0}$  je LK náhodné veličiny s vícerozměrným normálním rozdělením, má normální rozdělení se  $\mathbb{E}[\hat{\mathbf{Y}}_{\mathbf{x}_0}] = \gamma_{\mathbf{x}_0} = \mathbf{x}_0^T \beta$  a  $D[\hat{\mathbf{Y}}_{\mathbf{x}_0}] = \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$  tzn.

$$\frac{\hat{\mathbf{Y}}_{\mathbf{x}_0} - \gamma_{\mathbf{x}_0}}{\sigma \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim N(0, 1)$$

a díky nezávislosti  $\hat{\beta}$  a  $s_n^2$

$$\begin{aligned} \frac{\hat{\mathbf{Y}}_{\mathbf{x}_0} - \gamma_{\mathbf{x}_0}}{s_n \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} &\sim t(n-m-1) \quad \Rightarrow \quad 100(1-\alpha)\% \quad \text{IS pro } \gamma_{\mathbf{x}_0} : \\ (\hat{\mathbf{Y}}_{\mathbf{x}_0} \pm t_{1-\frac{\alpha}{2}}(n-m-1) \cdot s_n \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}) \end{aligned}$$

#### b) interval predikce pro $\mathbf{Y}_{\mathbf{x}_0}$

Bodový odhad je opět  $\hat{\mathbf{Y}}_{\mathbf{x}_0}$ , pokud  $\mathbf{Y}_{\mathbf{x}_0}$  je skutečná hodnota  $\mathbf{Y}_{\mathbf{x}}$  v bodě  $\mathbf{x} = \mathbf{x}_0$ , potom  $\mathbf{Y}_{\mathbf{x}_0}$  a  $\hat{\mathbf{Y}}_{\mathbf{x}_0}$  budou nezávislé za předpokladu, že pozorování  $\mathbf{Y}_{\mathbf{x}_0}, Y_1, \dots, Y_n$  jsou nezávislé (což předpokládáme), potom

$$D[\hat{\mathbf{Y}}_{\mathbf{x}_0} - \mathbf{Y}_{\mathbf{x}_0}] = D[\hat{\mathbf{Y}}_{\mathbf{x}_0}] - D[\mathbf{Y}_{\mathbf{x}_0}] = \sigma^2 (1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0),$$

takže

$$\frac{\hat{\mathbf{Y}}_{\mathbf{x}_0} - \mathbf{Y}_{\mathbf{x}_0}}{\sigma \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim N(0, 1) \quad \text{a} \quad \frac{\hat{\mathbf{Y}}_{\mathbf{x}_0} - \mathbf{Y}_{\mathbf{x}_0}}{s_n \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim t(n-m-1)$$

za předpokladu normality chyb.

$100(1-\alpha)\%$  IP pro  $\mathbf{Y}_{\mathbf{x}_0}$  tedy je

$$(\hat{\mathbf{Y}}_{\mathbf{x}_0} \pm t_{1-\frac{\alpha}{2}}(n-m-1) \cdot s_n \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0})$$

#### PŘÍKLAD 3.10.

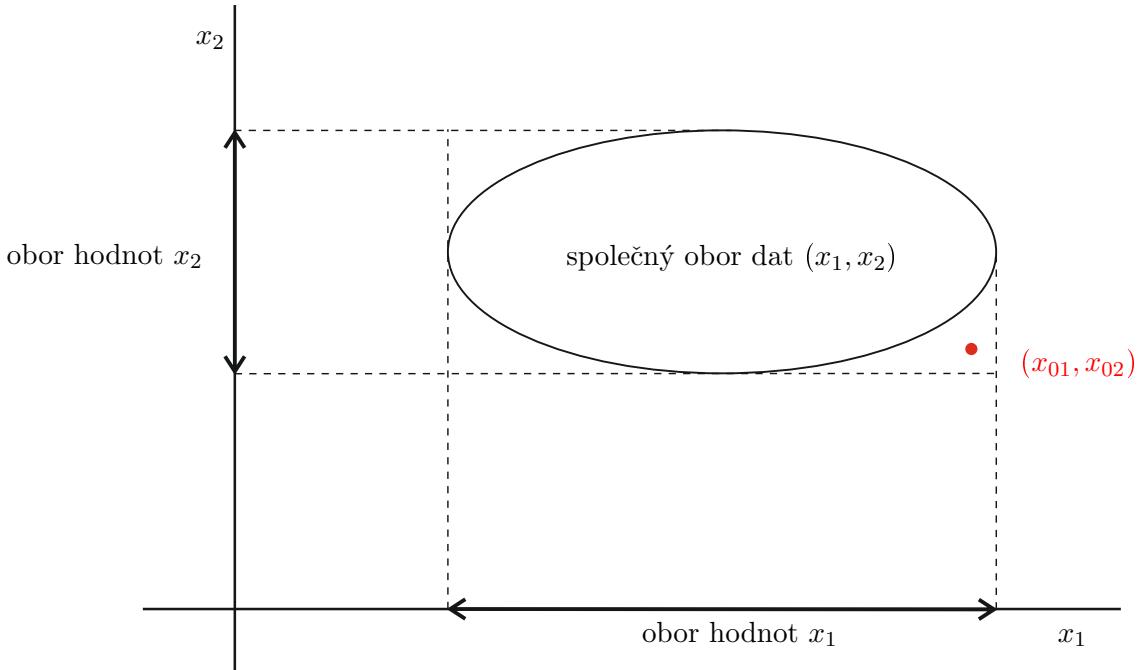
POZNÁMKA 3.11 (Extrapolace). • U jednoduché LR kvalitu predikce závisela na vzdálenosti  $x_0$  od  $\bar{x}$ .

- Je třeba si dát pozor na predikce mimo  $[x_{min}, x_{max}]$ .
- Podobné závěry platí i pro vícerozměrnou LR.
- Protože rozptyl predikce je úměrný  $\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$ , v bodech s velkými hodnotami této veličiny nebude predikce spolehlivá.
- Speciálně pokud  $\mathbf{x}_i^T$  jsou pozorovaná data, můžeme očekávat, že body s nejvyššími hodnotami  $\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0 = h_{ii}$  budou na hranici množiny, kde je predikce spolehlivá.

tzn., že vnitřek elipsoidu

$$\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0 \leq \max_{1 \leq j \leq n} h_{jj}$$

může být považován za přípustný obor predikce



Obrázek 3.1:  $(x_{01}, x_{02})$  leží uvnitř oboru hodnot pro obě  $x_1$  i  $x_2$  ale vně společného oboru původních dat.

### 3.1 Rezidua, diagnostika a transformace

- Je třeba ověřit adekvátnost modelu. Máme  $R^2, t, F$  statistiky, ty ale byly odvozeny za předpokladu linearity modelu a dalších podmínek na náhodné chyby. Pro ověření je důležitý nástroj analýza reziduí
- Je také třeba ověřit vliv jednotlivých pozorování na model - analýza odlehčích (outliers) a influenčních pozorování. (Velké reziduum pro  $i$ -té pozorování naznačuje problém s modelem, ale může to být i naopak, vlivné pozorování nemusí mít velké reziduum)
- Pokud detekujeme nějaké problémy s modelem, mohou pomoci transformace proměnných nebo metoda na korekci nekonstantního rozptylu.

#### 3.1.1 Rezidua

připomenutí:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{y}, \quad \text{kde} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{e}$$

Dále jsme ukázali:

$$\mathbb{E}[\hat{\mathbf{e}}] = 0 \quad \text{Cov}(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$$

Pokud navíc  $\mathbf{e} \sim N_n(0, \sigma^2\mathbf{I})$  potom  $\hat{\mathbf{e}} \sim N_n(0, \sigma^2(\mathbf{I}_n - \mathbf{H}))$ .

Pokud označíme  $h_{ii} = \mathbf{H}_{ii}$ ,  $\hat{\mathbf{e}}_i \sim N(0, \sigma^2(1 - h_{ii}))$ ,  $\text{Cov}(\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_j) = -\sigma^2 h_{ij}$ .

### 3 IS a t-testy pro parametry

Obecně bývá vhodnější pracovat se standardizovanými rezidui, protože  $D[\hat{e}_i] = \sigma^2(1 - h_{ii})$ , pro  $r_i = \frac{\hat{e}_i}{\sigma\sqrt{1-h_{ii}}}$  platí  $D[r_i] = 1$ .  $\sigma$  odhadneme pomocí  $s_n = \sqrt{\frac{1}{n-m-1}SSE}$ , dostaneme

$$\hat{r}_i = \frac{\hat{e}_i}{s_n\sqrt{1-h_{ii}}} \quad \text{kde } i = 1, \dots, n \quad (\text{Interně studentizované reziduum})$$

(Někdy také standardizovaná rezida)

R : rstandard(.)

Pokud  $\sigma^2$  odhadneme na základě modelu, ve kterém bylo vynecháno  $i$ -té pozorování, označíme tento odhad  $\sigma_{(-i)}^2$ , potom

$$\hat{t}_i = \frac{\hat{e}_i}{\sigma_{(-i)}^2\sqrt{1-h_{ii}}} \quad \text{kde } i = 1, \dots, n \quad (\text{Externě studentizované reziduum})$$

(Někdy také studentizované rezida)

R : rstudent(.)

Například  $\sigma_{(-i)}^2 = \frac{SSE_{(-i)}}{n-m-1}$  je nestranný odhad  $\sigma^2$  v modelu  $(-i)$ .

**POZNÁMKA 3.12.** Platí:

- Pokud  $h_{ii}$  je malé, pro velké  $n$  by se mělo  $\hat{e}_i, \hat{r}_i, \hat{t}_i$  chovat přibližně stejně a  $\hat{r}_i, \hat{t}_i \approx N(0, 1)$ .
- Pro malé  $n$  ( $n < 20$ ) a / nebo  $h_{ii} \approx 1$  je preferováno použít  $\hat{r}_i$  nebo  $\hat{t}_i$  a aktuálně bývá častěji doporučována  $\hat{t}_i$  ( $i$ -té pozorování s velkými  $h_{ii}$  může zvyšovat odhad  $\sigma^2$  a tím snižuje velikost svého rezidua).
- $h_{ii}$  hraje zásadní roli v diagnostice modelu, probereme teď jeho základní vlastnosti.

**Leverage**  $h_{ii}$  - potenciál  $i$ -tého pozorování (leverage point - píkový bod / vzdálený bod)

- $D[\hat{e}_i] = \sigma^2(1 - h_{ii}) \geq 0 \Rightarrow h_{ii} \leq 1$ .
  - $\mathbf{H}^2 = \mathbf{H} \Rightarrow h_{ii} = \sum_{j=1}^n h_{ij}h_{ji} = \sum_{j=1}^n (h_{ij})^2$  tedy  $h_{ii} > 0$  (Dá se ukázat silnější tvrzení:  $h_{ii} \geq \frac{1}{n}$ ).
  - $\mathbf{H}^2 = \mathbf{H} \Rightarrow \mathbf{A}_{i1} = \sum_{j=1}^n h_{ij}x_{j1} = \sum_{j=1}^n h_{ij} = x_{i1} = 1$  tedy
- $$\sum_{j=1}^n h_{ij} = 1 \quad \forall j \in \hat{n}$$

- Význam  $h_{ii}$  vyplýne z následujících úvahy:

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \Rightarrow \hat{y}_i = \sum_{j=1}^n h_{ij}y_j = h_{ii}y_i + \sum_{\substack{j=1 \\ i \neq j}}^n h_{ij}y_j$$

pokud  $h_{ii} \approx 1$ , potom  $\hat{y}_i \approx y_i$  a model je nucen proložit přímku bodem  $(\mathbf{x}_i, y_i)$  i když když tam neplatí body s "velkým  $h_{ii}$ " - body s velkým potenciálem (high leverage points). Tyto body by měly být detekovány pro další zkoumání.

### 3 IS a t-testy pro parametry

- Otázka je, jaká hodnota  $h_{ii}$  je "velká".

**Heuristické pravidlo:**  $\sum_{i=1}^n h_{ii} = \text{tr}(\mathbf{H}) = m + 1$ , tzn.  $\frac{m+1}{n}$  je průměrná hodnota  $h_{ii}$ .  $i$ -té pozorování má velký potenciál jestliže  $h_{ii} > \frac{3(m+1)}{n}$ . (Stejně postupuje i jazyk R)

#### 3.1.2 Grafy reziduí

1. Ověření normality - histogramy, Q-Q plots  
tyto obrázky nezávisí na počtu nezávislých proměnných  $x$ , vše stejné jako v jednoduché LR.
2. Pro ověření funkční formy pro  $\mathbb{E}[Y_x]$  a / nebo konstantního rozptylu se nejčastěji používají:
  - a) Grafy  $\hat{e}_i, \hat{r}_i$  nebo  $\hat{t}_i$  oproti  $\mathbf{x}_j^c, j = 1, \dots, m$ , kde  $\mathbf{x}_j^c$  je  $j$ -tý sloupec  $\mathbf{X}$
  - b) Grafy  $\hat{e}_i, \hat{r}_i$  nebo  $\hat{t}_i$  oproti  $\hat{y}_i$
  - c) Partial residual plots

POZNÁMKA 3.13. Zdůvodnění:

1. Normální rovnice  $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$  implikují  $\mathbf{X}^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{X}^T\hat{\mathbf{e}}$ .

$$\text{Připomenutí: } Y_i = \beta_1 x_i + e_i, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\|^2}$$

Pokud tedy naladíme LR model bez interceptu pro  $\hat{\mathbf{e}}$  v závislosti na  $\mathbf{x}_j^c$ , směrnice přímky bude

$$\hat{\beta}_j^* = \frac{(\mathbf{x}_j^c)^T \hat{\mathbf{e}}}{\|\mathbf{x}_j^c\|^2} = 0$$

Graf  $\hat{e}_i, \hat{r}_i, \hat{t}_i$  oproti  $\mathbf{x}_j^c$  by měl dávat náhodně rozptýlené body kolem osy  $x$ . (bez trendů,  $\hat{r}_i, \hat{t}_i$  uvnitř  $\pm 2$ ) Pokud tomu tak není, může to naznačovat nelinearitu v  $\mathbf{x}_j$  nebo nekonstantní rozptyl.

2. Ukázali jsme  $\sum_{i=1}^n \hat{y}_i \hat{e}_i = 0$  pro LM bez interceptu pro  $\hat{e}_i$  oproti  $\hat{y}_i$  tedy platí

$$\hat{\beta} = \frac{\hat{\mathbf{e}}^T \hat{\mathbf{y}}}{\|\hat{\mathbf{y}}\|^2} = 0$$

Body by opět měly být náhodně rozptýlené kolem osy  $x$

- Trychtýřovitý tvar indikuje nekonstantní rozptyl.
- Trendy indikují nelinearitu.

### 3.1.3 Partial residual plot

- I když grafy  $\hat{e}_i$  oproti  $\mathbf{x}_j^c$  a  $\hat{\mathbf{y}}$  mohou indukovat nedostatky modelu, nemusí být zřejmé, jaké ty nedostatky jsou.
- V SLR graf  $\hat{e}_i$  oproti  $x_i$  lze použít pro detekci nelinearity
- Ale v MLR tyto grafy, stejně jako scatterploty, mohou být zavádějící, protože  $\hat{\mathbf{e}}$  závisí na všech prediktorech, nemusí být tedy izolován efekt dané proměnné při odstranění efektů ostatních.
- Pro zkoumané efekty  $j$ -té proměnné lze použít partial rezidual plots - lze je chápat jako jeho ekvivalent scatterplotu v SLR

**Definice 3.14.**

$$\hat{e}_j^* = \hat{\mathbf{e}} + \hat{\beta}_j \mathbf{x}_j^c,$$

kde  $\hat{\mathbf{e}}$  je vektor reziduů modelu,  $\hat{\beta}_j$  je LSE parametr  $\beta_j$ ,  $\mathbf{x}_j^c$  je  $j$ -tý sloupec  $\mathbf{X}$

Partial residual plot (PRP): graf  $\hat{\mathbf{e}}$  oproti  $\mathbf{x}_j^c$ ,  $j = 1, \dots, m$  pokud je model správný, měly by být body náhodně rozmištěné kolem přímky se směrnicí  $\hat{\beta}_j$ .

**Zdůvodnění:** Vztah mezi  $\hat{e}_j^*$  a  $\mathbf{x}_j^c$  má formu SLR bez interceptu, pokud je model správný,  $\hat{e}_i, i = 1, \dots, n$ , splňují podmínu  $\mathbb{E}\hat{e}_i = 0$  a  $D\hat{e}_i = \sigma^2(1 - h_{neco})$ . Má tedy smysl uvažovat RM pro  $\hat{e}_j^*$  oproti  $\mathbf{x}_j^c$  ( $\hat{e}_j^* = \gamma_j \mathbf{x}_j^c + \mathbf{e}$ ).

Pro odhad koeficientů platí:

$$\hat{\gamma}_j = \frac{(\hat{e}_j^* \mathbf{x}_j^c)}{\|\mathbf{x}_j^c\|^2} = \frac{(\hat{\mathbf{e}} + \hat{\beta}_j \mathbf{x}_j^c)^T \mathbf{x}_j^c}{\|\mathbf{x}_j^c\|^2} = \frac{\hat{\mathbf{e}}^T \mathbf{x}_j^c + \hat{\beta}_j \|\mathbf{x}_j^c\|^2}{\|\mathbf{x}_j^c\|^2} = \hat{\beta}_j,$$

protože  $\hat{\mathbf{e}}^T \mathbf{x}_j^c = 0$ .

(2 příklady - pdf 79-93 uprostřed str 6)

POZNÁMKA 3.15. PRPs jsou někdy kritizovány za nadhodnocování efektu  $\mathbf{x}_j^c$ .

Alternativa: **partial regression plot (added variable plot)**.

Motivace: Ptáme se, zda přidat novou proměnnou do modelu a chtěli bychom dohadnout její efekt.

Budeme tedy uvažovat rozšířený model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \gamma \mathbf{w} + \mathbf{e},$$

kde  $\mathbf{w}$  je nový vektor regresorů. Model lze rozepsat jako

$$\mathbf{Y} = [\mathbf{X}\mathbf{w}] \begin{pmatrix} \boldsymbol{\beta} \\ \gamma \end{pmatrix} + \mathbf{e} = \mathbf{X}_w + \boldsymbol{\beta}_w + \mathbf{e}.$$

Použitím normálních rovnic pro  $\mathbf{X}_w$  lze odvodit formulí pro  $\hat{\gamma}$

$$\hat{\gamma} = \frac{\hat{\mathbf{e}}^T (\mathbb{I} - \mathbb{H}) \mathbf{w}}{\|(\mathbb{I} - \mathbb{H}) \mathbf{w}\|^2}. \quad (\#)$$

$\hat{\gamma}$  je směrnice RM pro  $\hat{\mathbf{e}}$  v závislosti na  $\mathbf{w}_{res} = (\mathbb{I} - \mathbb{H}) \mathbf{w}$  (rezidua modelu pro  $\mathbf{w}$  v závislosti na  $\mathbf{X}$ ).

### 3 IS a t-testy pro parametry

Ted' naopak uvažujme, že  $\mathbf{w}$  je sloupec původní  $\mathbf{X}$ , řekněme  $\mathbf{x}_j^c$  a ozn.  $\mathbf{X}_{(-j)}$  matici  $\mathbf{X}$  bez sloupce  $j$ . V předchozím modelu pomožme  $\mathbf{X} = \mathbf{X}_{(-j)}$  a  $\mathbf{w} = \mathbf{x}_j^c$ . Potom LSE  $\hat{\beta}_j$  parametru  $\beta_j$  je

$$\hat{\beta}_j = \frac{\hat{\mathbf{e}}_{(-j)}^T \mathbf{x}_{j,res}^c}{\|\mathbf{x}_{j,res}^c\|^2},$$

kde  $\hat{\mathbf{e}}_{(-j)}$  jsou rezidua modelu bez  $\mathbf{x}_j^c$ ,  $\mathbf{x}_{j,res}^c = (\mathbb{I} - \mathbb{H})\mathbf{x}_j^c$ , tedy jsou to rezidua modelu pro  $\mathbf{x}_j^c$  v závislosti na ostatních proměnných, tedy  $\mathbf{X}_{(-j)}$  (v  $\mathbf{x}_{j,res}^c$  je tedy odstraněn efekt ostatních regresorů).

$\hat{\beta}_j$  je směrnice RM pro  $\hat{\mathbf{e}}_{(-j)}$  v závislosti na  $\mathbf{x}_{j,res}$   $\implies$   
**added variable plot:** graf  $\hat{\mathbf{e}}_{(-j)}$  proti  $\mathbf{x}_{j,res}, j = 1, \dots, m$ .

Pokud je model správný, body by měly být náhodně rozptýlené kolem přímky se směrnicí  $\hat{\beta}_j$  procházející počátkem. Pokud závislost na  $\mathbf{x}_j^c$  není lineární, projeví se to odklonem bodů od přímky.

**POZNÁMKA 3.16.** Ze vztahu (#) je vidět, že MLR může být chápána jako posloupnost SLR, kde postupně vytváříme modely pro novou proměnnou s použitím reziduů modelu pro předcházející proměnné.

#### PRESS rezidua (PRESS residuals, deleted residuals)

- pokud budeme chtít model použít nejen k vysvětlení vztahu mezi proměnnými, ale také pro predikci, hodila by se míra vyjadřující jak dobře model predikuje (doposud jsme zkoumali jen jak dobře popisuje)
- šlo by použít IS nebo IP, to bychom ale předem museli znát body, ve kterých chceme predikovat
- nejjednodušší přístup, jak měřit prediktivní přesnost modelu by byl analýza reziduů pro predikce hodnot v nových bodech  $\mathbf{x}$ , obecně ale nemáme data  $y$  v těchto bodech
- jedna možnost je použít data, která máme k dispozici

**Postup:** Vynecháme jedno pozorování, naladíme model bez tohoto pozorování a porovnáme predikovanou a pozorovanou hodnotu pro vynechané pozorování.

Předpokládáme, že vynecháme  $i$ -té pozorování. Ozn.  $\hat{\beta}_{(-i)}$  odhad  $\beta$  v modelu s vynechaným  $i$ -tým pozorováním ( $M_{(-i)}$ ) a  $\hat{y}_{(-i)}$  predikovanou hodnotu modelem  $M_{(-i)}$  v bodě  $\mathbf{x}_i^T$ , tzn.

$$\hat{y}_{(-i)} = \mathbf{x}_i^T \hat{\beta}_{(-i)}.$$

Potom

$$\hat{\mathbf{e}}_{(-i)} = y_i - \hat{y}_{(-i)}, \quad i = 1, \dots, n$$

nazýváme  $i$ -té **PRESS reziduum**.

PRESS =  $\sum_{i=1}^n \hat{e}_{-i}^2$  je užitečná míra přesnosti predikce.

**POZNÁMKA 3.17.** Otázka je, jak počítat  $\hat{e}_{(-i)}, i = 1, \dots, n$ .

- pro velké  $n$  se zdá, že to bude náročný problém, protože pro každé  $i \in \hat{n}$  musíme naladit nový model

### 3 IS a t-testy pro parametry

- naštěstí to není nutné, ukážeme totiž, že

$$\hat{e}_{(-i)} = \frac{\hat{e}_i}{1 - h_{ii}},$$

tzn. všechna  $\hat{e}_{(-i)}$  lze snadno spočítat pomocí reziduí a hodnot  $h_{ii}$  z původního (plného) modelu.

Označme

$$\begin{aligned} \mathbf{x}_i^T & - i\text{-tý řádek matice } \mathbf{X} \\ \mathbf{X}_{(-i)} & - matici \mathbf{X} \text{ bez } i\text{-tého řádku} \end{aligned}$$

**Věta 3.18.** Jestliže  $h_{ii} \neq 1$ , potom

$$[\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)}] = \mathbf{X}^T \mathbf{X}^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_{ii}},$$

kde  $h_{ii}$  je  $i$ -tý diagonální prvek matice  $\mathbb{H}$ .

*Důkaz.* Nejdříve ukážeme

$$\mathbf{X}^T \mathbf{X} = \mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)} + \mathbf{x}_i \mathbf{x}_i^T \quad (+)$$

Kvůli značení předpokládáme  $i = n$  (toho se dá vždy dosáhnout permutací řádků  $\mathbf{X}$ ). Potom

$$(\mathbf{X}^T \mathbf{X})_{ij} = \sum_{k=1}^n x_{ki} x_{kj} = \sum_{k=1}^{n-1} x_{ki} x_{kj} + x_{ni} x_{nj}.$$

$i, j$ -tý prvek  $\mathbf{X}_{(-k)}^T \mathbf{X}_{(-k)}$  je  $\sum_{k=1}^{n-1} x_{ki} x_{kj}$

$i, j$ -tý prvek  $\mathbf{x}_n \mathbf{x}_n^T$  je  $x_{ni} x_{nj}$ , tzn. (+) platí.

**Věta 3.19** (Sherman-Morrison-Woodbury (z LA)). Nechť  $\mathbb{A}$  je  $n \times n$  invertibilní matice a nechť  $\mathbf{z}$  je  $n \times 1$  sloupcový vektor. Jestliže  $\mathbf{z}^T \mathbb{A}^{-1} \mathbf{z} \neq 1$ , potom matice  $\mathbb{B} = \mathbb{A} - \mathbf{z} \mathbf{z}^T$  je invertibilní a platí

$$\mathbb{B}^{-1} = \mathbb{A}^{-1} + \frac{\mathbb{A}^{-1} \mathbf{z}^T \mathbf{z} \mathbb{A}^{-1}}{1 - \mathbf{z}^T \mathbb{A}^{-1} \mathbf{z}}.$$

Položme  $\mathbb{A} = \mathbf{X}^T \mathbf{X}$ ,  $\mathbf{z} = \mathbf{x}_i$ ,  $\mathbb{B} = \mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)}$ . Pak  $\mathbb{B} = \mathbb{A} - \mathbf{z} \mathbf{z}^T$ ,  $\mathbb{A}$  je invertibilní a

$$\mathbf{z}^T \mathbb{A}^{-1} \mathbf{z} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i = \left( \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right)_{ii} = h_{ii} \neq 1.$$

Užitím věty dostaneme

$$(\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1} = (\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_{ii}}.$$

□

**Věta 3.20.** Nechť  $\hat{e}_{(-i)}$  je  $i$ -té PRESS reziduum. Potom

$$\hat{e}_{(-i)} = \frac{\hat{e}_i}{1 - h_{ii}}, \quad i = 1, \dots, n.$$

### 3 IS a t-testy pro parametry

Důkaz. Nechť  $\hat{\beta}_{(-i)}$  je odhad  $\beta$  v modelu  $M_{-i}$ , tzn.

$$\hat{\beta}_{(-i)} = \left( \mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)} \right)^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)},$$

tedy  $\mathbf{y}_{(-i)}$  je  $\mathbf{y}$  bez  $i$ -té složky  $y_i$ . Tzn.

$$\begin{aligned} \hat{y}_{(-i)} &= \mathbf{x}_i^T \hat{\beta}_{(-i)} = \mathbf{x}_i^T \left( \mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)} \right)^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} = \\ &= \left[ \left( \mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)} \right)^{-1} = (\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_{ii}}. \right] = \\ &= \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} + \frac{1}{1 - h_{ii}} \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} = \\ &= S_1 + \frac{1}{1 - h_{ii}} S_2. \end{aligned}$$

Protože  $\mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} = \mathbf{X}^T \mathbf{y} - y_i \mathbf{x}_i$ , dostaneme

$$\begin{aligned} S_1 &= \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y} - y_i \mathbf{x}_i) = \mathbf{x}_i^T \underbrace{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}}_{\hat{\beta}} - y_i \underbrace{\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i}_{h_{ii}} = \\ &= \mathbf{x}_i^T \hat{\beta} - h_{ii} y_i = \hat{y}_i - h_{ii} y_i. \end{aligned}$$

Podobně

$$S_2 = \underbrace{\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i}_{h_{ii}} \underbrace{\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{y}}_{\hat{y}_i} = y_i \underbrace{\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i}_{h_{ii}} \underbrace{\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i}_{h_{ii}} = h_{ii} \hat{y}_i - y_i h_{ii}^2$$

takže

$$\hat{y}_{(-i)} = \hat{y}_i - h_{ii} y_i + \frac{1}{1 - h_{ii}} (h_{ii} \hat{y}_i - y_i h_{ii}^2).$$

Celkem tedy

$$\begin{aligned} \hat{e}_{-i} &= y_i - \hat{y}_{(-i)} = y_i (1 + h_{ii}) - \hat{y}_i - \frac{1}{1 - h_{ii}} (h_{ii} \hat{y}_i - y_i h_{ii}^2) = \\ &= \frac{1}{1 - h_{ii}} (y_i (1 - h_{ii}^2) - \hat{y}_i (1 - h_{ii}) - h_{ii} \hat{y}_i + y_i h_{ii}^2) = \frac{1}{1 - h_{ii}} (y_i - \hat{y}_i) = \frac{\hat{e}_i}{1 - h_{ii}} \end{aligned}$$

□

Budeme potřebovat podobné formule pro  $\hat{\beta} - \hat{\beta}_{(-1)}$  a  $SSE_{(-1)}$ .

**Věta 3.21.** 1) Nechť  $\hat{\beta}_{(-1)}$  značí LSE parametru  $\beta$  v modelu bez  $i$ -tého pozorování. Potom platí

$$\hat{\beta} - \hat{\beta}_{(-1)} = \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \hat{e}_i}{1 - h_{ii}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \hat{e}_{(-1)}.$$

2) Pro součet residuálních čtverců  $SSE_{(-1)}$  v modelu bez  $i$ -tého pozorování platí

$$SSE_{(-1)} = \sum_{j=1}^n \hat{e}_j^2 - \frac{\hat{e}_i^2}{1 - h_{ii}}.$$

### 3 IS a t-testy pro parametry

*Důkaz.* 1) Stejně jako v důkazu předchozí věty platí, že

$$\hat{\beta}_{(-1)} = S_1 + \frac{1}{1 - h_{ii}} S_2,$$

kde  $S_1 = \hat{\beta} - y_i(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$  a  $S_2 = \mathbf{X}^T \mathbf{X}^{-1} \mathbf{x}_i \hat{y}_i - y_i(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i h_{ii}$ , tedy

$$\begin{aligned} \hat{\beta} - \hat{\beta}_{(-1)} &= y_i(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i - \frac{1}{1 - h_{ii}} ((\mathbf{X}^T \mathbf{X})^{-1} x_i \hat{y}_i - y_i(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i h_{ii}) = \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \left( y_i - \frac{\hat{y}_i - y_i h_{ii}}{1 - h_{ii}} \right) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \left( \frac{y_i - y_i h_{ii} - \hat{y}_i + y_i h_{ii}}{1 - h_{ii}} \right) = \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \left( \frac{y_i - \hat{y}_i}{1 - h_{ii}} \right), \end{aligned}$$

kde  $\left( \frac{y_i - \hat{y}_i}{1 - h_{ii}} \right) = \frac{\hat{e}_i}{1 - h_{ii}} = \hat{e}_{(-1)}$ .

2)

$$\begin{aligned} \text{SSE}_{(-1)} &= (\mathbf{y}_{(-1)} - \mathbf{x}_{(-1)}^T \hat{\beta}_{(-1)})^T (\mathbf{y}_{(-1)} - \mathbf{x}_{(-1)}^T \hat{\beta}_{(-1)}) = \sum_{\substack{j=1 \\ j \neq i}}^n (y_j - \mathbf{x}_j^T \hat{\beta}_{(-1)})^2 = \\ &= \sum_{j=1}^n (y_j - \mathbf{x}_j^T \hat{\beta}_{(-1)})^2 - (y_i - \mathbf{x}_i^T \hat{\beta}_{(-1)})^2. \end{aligned}$$

Z bodu 1) víme, že  $\hat{\beta}_{(-1)} = \hat{\beta} - \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \hat{e}_i}{1 - h_{ii}}$ , tzn.

$$\text{SSE}_{(-1)} = \sum_{j=1}^n \left( y_j - \mathbf{x}_j^T \hat{\beta} + \frac{\mathbf{x}_j^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \hat{e}_i}{1 - h_{ii}} \right)^2 - \left( y_i - \mathbf{x}_i^T \hat{\beta} + \frac{\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \hat{e}_i}{1 - h_{ii}} \right)^2.$$

Protože  $\mathbf{x}_j^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i = h_{ij}$ , dostaneme

$$\begin{aligned} \text{SSE}_{(-1)} &= \sum_{j=1}^n \left( \hat{e}_j + \frac{h_{ij} \hat{e}_i}{1 - h_{ii}} \right)^2 - \left( \hat{e}_i + \frac{h_{ii} \hat{e}_i}{1 - h_{ii}} \right)^2 = \underbrace{\sum_{j=1}^n \left( \hat{e}_j + \frac{h_{ij} \hat{e}_i}{1 - h_{ii}} \right)^2}_{A} - \frac{\hat{e}_i^2}{(1 - h_{ii})^2}, \\ A &= \sum_{j=1}^n \hat{e}_j^2 + \frac{2 \hat{e}_i}{1 - h_{ii}} \underbrace{\sum_{j=1}^n h_{ij} \hat{e}_j}_{0} + \frac{\hat{e}_i^2}{(1 - h_{ii})^2} \underbrace{\sum_{j=1}^n h_{ij}^2}_{h_{ii}}. \end{aligned}$$

Protože pak  $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ , tak  $\mathbf{H}\hat{\mathbf{y}} = \mathbf{H}^2\mathbf{y} = \mathbf{H}\mathbf{y}$  a tedy  $\mathbf{H}\hat{\mathbf{e}} = \mathbf{H}(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{H}\mathbf{y} - \mathbf{H}\hat{\mathbf{y}} = 0$  a tedy

$$\text{SSE}_{(-1)} = \sum_{j=1}^n \hat{e}_j^2 + \frac{\hat{e}_i^2}{(1 - h_{ii})^2} (h_{ii} - 1) = \sum_{j=1}^n \hat{e}_j^2 - \frac{\hat{e}_i^2}{1 - h_{ii}}.$$

□

### 3 IS a t-testy pro parametry

**Důsledek 3.22.** V modelu  $(**)$  s  $m + 1$  parametry  $\beta$  a bez  $i$ -tého pozorování platí, že

$$\mathbb{E} [\text{SSE}_{(-1)}] = (n - m - 2)\sigma^2,$$

takže

$$\widehat{\sigma^2}_{(-1)} = \frac{\text{SSE}_{(-1)}}{n - m - 2}$$

je nestranný odhad  $\sigma^2$ . Dále pak

$$\widehat{\sigma^2}_{(-1)} = \frac{(1 - h_{ii})(n - m - 1)s_n^2 - \widehat{e}_i^2}{(1 - h_{ii})(n - m - 2)} = \frac{1}{n - m - 2} \left( \text{SSE} - \frac{\widehat{e}_i^2}{1 - h_{ii}} \right),$$

kde  $s_n^2 = \frac{1}{n-m-1} \text{SSE}$  (pro plný model).

**Důkaz.** Protože  $\mathbb{E} [\widehat{e}_i^2] = D\widehat{e}_i = \sigma^2(1 - h_{ii})$ , dostaneme dle předchozí věty

$$\begin{aligned} \mathbb{E} [\text{SSE}_{(-1)}] &= \sum_{j=1}^n \sigma^2(1 - h_{jj}) - \sigma^2 = \sigma^2 \left[ (n - 1) - \underbrace{\sum_{j=1}^n h_{jj}}_{h\mathbf{H}=\mathbf{m+1}} \right] = \sigma^2(n - m - 2) \\ \widehat{\sigma^2}_{(-1)} &= \frac{1}{n - m - 2} \text{SSE}_{(-1)} = \frac{1}{n - m - 2} \left( \underbrace{\sum_{j=1}^n \widehat{e}_j^2}_{\text{SSE}=(n-m-1)s_n^2} - \frac{\widehat{e}_i^2}{1 - h_{ii}} \right) = \frac{1}{n - m - 2} \frac{(1 - h_{ii})\text{SSE} - \widehat{e}_i^2}{1 - h_{ii}}. \end{aligned}$$

□

**Poznámka 3.23.** Dě se ukázat, že  $\text{SSE}_{(-1)}$  a  $\widehat{e}_i$  jsou nezávislé náhodné veličiny. Protože  $\frac{\text{SSE}_{(-1)}}{\sigma^2} \sim \chi^2(n - m - 2)$  a  $\frac{\widehat{e}_i}{\sigma\sqrt{1-h_{ii}}} \sim \mathcal{N}(0, 1)$ , dostaneme  $\frac{\widehat{e}_i}{\widehat{\sigma^2}_{(-1)}\sqrt{1-h_{ii}}} \sim t(n - m - 2)$ .

**Tvrzení 3.24.** Uvažujme model  $(**)$ , kde  $h(X) = m + 1$  a  $\mathbf{e} \sim \mathcal{N}_m(0, \sigma^2 I_m)$ . Nechť pro  $i \in \hat{n}$  platí, že  $h_{ii} \neq 1$ . Potom  $i$ -té reziduum

$$\widehat{t}_i \sim t(n - m - 2).$$

**Poznámka 3.25.**  $\widehat{t}_i$  lze použít pro test hypotézy, zda je  $i$ -té pozorování odlehlé (outlier), tedy

$$H_0 : i\text{-té pozorování není odlehlé v modelu } M$$

$$H_1 : i\text{-té pozorování je odlehlé v } M,$$

kde odlehlé značí odlehlé vzhledem k  $M : \mathbf{Y} \sim \mathcal{N}_m(\mathbf{X}\beta, \sigma^2 I_m)$ :

- a) střední hodnota  $i$ -tého pozorování se nerovná té dané modelem,
- b) pozorovaná hodnota  $Y_i$  je neobvyklá za platnosti  $M$ .

$H_0$  zamítneme, pokud

$$|\widehat{t}_i| > t_{1-\frac{\alpha}{2}}(n - m - 2) \approx u_{1-\frac{\alpha}{2}} \doteq 2 \text{ pro } \alpha = 0.05 \text{ a } n \text{ velká.}$$

Pokud test použijeme na všechna pozorování, je potřeba aplikovat nějakou korekci na vícenásobné testování, např. Bonferroni.

### 3 IS a t-testy pro parametry

POZNÁMKA 3.26. Vztah  $\hat{e}_{(-1)}$  a  $\hat{t}_i$ :

$$\hat{e}_{(-1)} = \frac{\hat{e}_i}{1 - h_{ii}} \Rightarrow \mathbb{E}\hat{e}_{(-1)} = 0 \quad \wedge \quad D\hat{e}_{(-1)} = \frac{\sigma^2}{1 - h_{ii}}.$$

Standardizované PRESS reziduum

$$\frac{\hat{e}_{(-1)}}{\sqrt{D\hat{e}_{(-1)}}} = \frac{\frac{\hat{e}_i}{1 - h_{ii}}}{\frac{\sigma}{\sqrt{1 - h_{ii}}}} = \frac{\hat{e}_i}{\sigma\sqrt{1 - h_{ii}}} = r_i.$$

Pokud použijeme  $\widehat{\sigma^2}_{(-1)}$  jako odhad  $\sigma^2$ , pak **studentizovaná PRESS rezidua**

$$\frac{\hat{e}_i}{\widehat{\sigma}_{(-1)}\sqrt{1 - h_{ii}}} = \hat{t}_i.$$

POZNÁMKA 3.27.  $\hat{e}_{(-1)} = \frac{\hat{e}_i}{1 - h_{ii}}$ , a proto pokud  $i$ -té pozorování má velký potenciál  $h_{ii}$ , bude  $\hat{e}_{(-1)}$  mnohem větší, než  $\hat{e}_i$ , pozorování s velkým  $h_{ii}$  jsou dobře modelována, ale měřeno  $\hat{e}_{(-1)}$  mohou špatně predikovat. To je další ukázka fit/prediction dilema.

Stejný efekt nastává také pro

$$\hat{\beta}_i - \hat{\beta}_{(-1)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \hat{e}_{(-1)}.$$

Rozdíl může být "malý", pokud je "fit" dobrý, ale může být také "velký", pokud je  $h_{ii}$  velké.

## 3.2 Míry influence

- I pro perfektní model mohou dva různé vzorky  $(\mathbf{x}, \mathbf{y})$  a  $(\mathbf{x}', \mathbf{y}')$  vést k různým závěrům,
- většinou máme k dispozici jen originální data,
- bude nás zajímat vliv  $i$ -tého řádku  $\mathbf{x}$  na model,
- už víme, že velké  $h_{ii}$  indikuje, že  $i$ -té pozorování má velký vliv a velká rezidua naznačují možnou neadekvátnost modelu,
- míry, které zavedeme, budou kombinovat tyto dva faktory,
- použijeme přístup z PRESS residní, tzn. budeme sledovat jak velký vliv má vynechání  $i$ -tého pozorování na  $\hat{\beta}$  a  $\hat{y}$ .

### DFBETAS

$\hat{\beta} - \hat{\beta}_{(-1)}$  měří vliv vynechání  $i$ -tého pozorování na odhad  $\hat{\beta}$  (bude základem pro naši analýzu). Připomeňme nyní vztah

$$\hat{\beta} - \hat{\beta}_{(-1)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \frac{\hat{e}_i}{1 - h_{ii}}.$$

### 3 IS a t-testy pro parametry

**a) vliv i-tého pozorování na  $\beta_j$ :**

$$\beta_j - \beta_{(-1)j} = \frac{r_{ij}\hat{e}_i}{1 - h_{ii}}, \quad \text{kde } r_{ji} \text{ je } (j, i)\text{tý prvek matice } \mathbb{R} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^T.$$

i-té pozorování budeme považovat za influenční na  $\beta_j$ , pokud  $\hat{\beta}_j - \hat{\beta}_{(-1)j}$  bude velká. Protože  $\hat{\beta}_j$  je náhodná veličina, "velké" bychom měli měřit relativně vzhledem k s.f.  $(\hat{\beta}_j, \text{což je } \sigma\sqrt{v_j}, v_j = (\mathbf{X}^T \mathbf{X})_{jj}^{-1})$ . Pokud ji odhadneme pomocí  $\hat{\sigma}_{(-1)}\sqrt{v_j}$ , dostaneme definici

$$\text{DFBETAS}_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{(-1)j}}{\hat{\sigma}_{(-1)}\sqrt{v_j}} = \frac{r_{ji}\hat{e}_i}{\sqrt{v_j}\hat{\sigma}_{(-1)}(1 - h_{ii})} = \frac{r_{ji}}{\sqrt{v_j}} \frac{\hat{t}_i}{\sqrt{1 - h_{ii}}},$$

kde  $\hat{t}_i$  je ext. studentizované reziduum. Kombinuje efekt velkého rezidua  $\hat{t}_i$  a velkého  $h_{ii}$ . Jedna možnost pro limitní hodnoty: i-té pozorování je považováno za influenční na oblasti  $\beta_j$ , pokud

$$|\text{DFBETAS}_{j,i}| > \frac{2}{\sqrt{n}}.$$

Máme  $(m + 1) \times n$  hodnot pro srovnání, zjednodušíme to.

**b) Vliv i-tého pozorování na celý vektor  $\hat{\beta}$ :** spočívá v použití nejaké normy na vektor  $\hat{\beta} - \hat{\beta}_{(-1)}$ . Cook navrhnu

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(-1)})^T \mathbf{M}(\hat{\beta} - \hat{\beta}_{(-1)})}{(m + 1)c},$$

kde  $\mathbf{M}$  je PD matice a  $c$  normalizační konstanta. Nejužívanější volba je  $\mathbf{M} = \mathbf{X}^T \mathbf{X}$  a  $x = s_n^2$ . Cookova vzdálenost se potom spočítá jako

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(-1)})^T \mathbf{X}^T \mathbf{X}(\hat{\beta} - \hat{\beta}_{(-1)})}{(m + 1)s_n^2}.$$

$$D_i = \frac{1}{(m + 1)s_n^2} \left( \frac{\hat{e}_i}{1 - h_{ii}} \right)^2 \underbrace{\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}_I = \frac{1}{m + 1} \frac{h_{ii}}{1 - h_{ii}} \frac{\hat{e}_i^2}{s_n^2(1 - h_{ii})}.$$

Výpočetní formule je potom ve tvaru

$$D_i = \frac{\hat{r}_i^2}{m + 1} \left( \frac{h_{ii}}{1 - h_{ii}} \right).$$

POZNÁMKA 3.28.  $100(1 - \alpha)\%$  simultání IS pro  $\beta$  je

$$C(\alpha) = \left\{ \beta \mid \frac{(\hat{\beta} - \beta)^T \mathbf{x}^T \mathbf{x}(\hat{\beta} - \beta)}{(m + 1)s_n^2} \leq F_{1-\alpha}(m + 1, n - m - 1) \right\},$$

tzn.

$$\hat{\beta}_{(-1)} \in C(\alpha) \Leftrightarrow D_i \leq F_{1-\alpha}(m + 1, n - m - 1).$$

To je motivace pro **RULE OF THUMB**:

$$i\text{-té pozorování je influenční, jestliže } D_i > F_{\frac{1}{2}}(m + 1, n - m - 1)$$

(pro většinu  $m, n$  je  $F_{\frac{1}{2}} \approx 1$ , zjednodušení pravidla  $D_i > 1$ ).

### 3 IS a t-testy pro parametry

POZNÁMKA 3.29. Také platí, že

$$D_i = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-1)})^T (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-1)})}{(m+1)s_n^2},$$

tzn. dá se chápat jako míra influence na celkovou predikci.

## DFFITS

- vliv  $i$ -tého pozorování na  $\hat{y}_i$

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(-1)}}{\hat{\sigma}_{(-1)}\sqrt{h_{ii}}} = \dots = \hat{t}_i \sqrt{\frac{h_{ii}}{1-h_{ii}}}.$$

RULE OF THUMB:  $i$ -té pozorování je influenční, pokud  $|\text{DFFITS}| > 3\sqrt{\frac{m+1}{n-m-1}}$ .

POZNÁMKA 3.30 (Míry influence v R).

- DFBETAS – `dfbetas()`

- DFFITS – `dffits()`
- Cookova vzdálenost  $D_i$  – `cooks.distance()`
- Leverage  $h_{ii}$  – `hotvalues()`
- a vše shrnuje funkce `influence.measures()` (má navíc covariance ratio)

Používané pravidlo:  $i$ -té pozorování je influenční, pokud:

$$\begin{aligned} &= |\text{DFBETAS}| > 1, \quad |\text{DFFITS}| > 3\sqrt{\frac{m+1}{n-m-1}} \\ &D_i > F_{0.5}(m+1, n-m-1), \quad h_{ii} > 3\frac{m+1}{n} \end{aligned}$$

## 3.3 Transformace

Pokud není splněný některý z předpokladů modelu: linearita, normalita chyb, homoskedasticita, jednou z možností je pokusit se transformovat nějaké proměnné, aby transformovaný model tyto předpoklady alespoň „přibližně“ splňoval.

### 3.3.1 Transformace vysvětlované proměnné $y$

Hledáme funkci  $h(\cdot)$  tak, aby model  $Y_i^* = h(Y_i) = \beta_0 + \sum_{j=1}^m x_{ij}\beta_j + e_i$  splňoval předpoklady.

#### 3 hlavní důvody pro transformaci $Y$

1. Transformace škály měření tak, aby pokrývala celé  $\mathbb{R}$ , což může odstranit problémy s podmínkami na  $\beta$ .

Např. studie kapacity plic (FEV data,  $FEV > 0$ ):

- Chtěli bychom, aby model nepredikoval záporné hodnoty ( $\implies$  restrikční podmínky na parametr  $\beta$ ).
- Lze obejít modelování  $y^* = \log FEV$ .

### 3 IS a t-testy pro parametry

Pokud  $y$  jsou počty a 0 je možná hodnota, často se používá  $y^* = \log(y+1)$  nebo obecně  $y^* = \log(y+c)$

2. Transformace  $Y$ , aby její rozdělení bylo „více“ normální.

Typicky to znamená pokusit se udělat rozdělení hodnot  $y$  více symetrické. Často se setkáváme s rozděleními vychýlenými vpravo (obvykle se to stává, pokud naměříme nějakou fyzikální veličinu, která může nabývat pouze kladných hodnot).

Transformace  $y^* = \log y$  nebo  $y^* = y^\lambda$ ,  $\lambda < 1$  budou redukovat toto vychýlení.

Typický postup: Začít s hodnotou  $\lambda$  blízko 1, pak snižovat hodnotu  $\lambda$ , dokud není dosaženo „přibližně“ symetrie reziduí.

3. Možná nejzásadnější motivace je pokusit se dosáhnout konstantního rozptylu přes všechna pozorování.

Např. pro fyzikální veličinu s kladnými hodnotami se často stane, že rozptyl bude malý pro  $\mu \approx 0$  a větší pro  $\mu$  velké (už je z důvodu, že obor hodnot  $y$  je omezen na kladné hodnoty). Říkáme tomu **positive mean-variance relationship**.

Nepřesnost měření kladných veličin se také často vyjadřuje pomocí koeficientu variace

$$CV(Y) = \frac{s.d.Y}{\mathbb{E}[Y]}.$$

Často bývá více konstantní mezi případy než s.d. Variabilitu vyjadřuje relativně spíše než absolutně. Matematicky to znamená, že  $D[Y] = \Phi \mathbb{E}[Y]^2 - \Phi \mu^2$  pro nějaké  $\Phi$ .

4. Pro odstranění vztahu  $\mathbb{E}[Y]$  a  $D[Y]$  se často používají mocninné transformace  $y^* = y^\lambda$  (pro  $y > 0$ )

$$\begin{array}{ccccccccc} \text{Transformace:} & \leftarrow & \dots & y^3 & y^2 & y & \sqrt{y} & \log y & \frac{1}{\sqrt{y}} & \frac{1}{y} & \frac{1}{y^2} & \dots & \rightarrow \\ \text{Box-Cox}\lambda: & \leftarrow & dots & 3 & 2 & 1 & \frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} & -1 & -2 & \dots & \rightarrow \end{array}$$

- Pokud  $D[Y]$  klesá s rostoucí  $\mathbb{E}[Y]$
- Pokud  $D[Y]$  roste s rostoucí  $\mathbb{E}[Y]$

OBECNĚ:

Předpokládejme vztah  $D[Y] = \Psi V(\mu)$  a uvažujeme transformaci  $y^* = h(y)$ . Taylorův rozvoj 1. řádu funkce  $h(y)$  v bodě  $\mu$

$$y^* = h(y) \approx h(\mu) + h'(\mu)(y - \mu)$$

z čehož plyne, že  $D[Y^*] \simeq (h'(\mu))^2 \cdot D[Y]$  Transformace  $y^* = h(y)$  tedy bude přibližně stabilizovat rozptyl, pokud  $h'(y)$  je úměrné  $(D[Y])^{-\frac{1}{2}} = V^{-\frac{1}{2}}(\mu)$

- Pokud  $V(\mu) = \mu^2 \Rightarrow$  stabilizující transformace je  $\log(y) = h(y)$  protože  $h'(\mu) = \frac{1}{\mu}$

### 3 IS a t-testy pro parametry

- Pokud  $V(\mu) = \mu \Rightarrow$  stabilizující transformace je  $h(y) = \sqrt{y}$  protože  $h'(y) = \frac{1}{2\sqrt{\mu}}$

$$\left( h(\mu) = \int \frac{d\mu}{\sqrt{V(\mu)}} \right)$$

- Asi nejvíce užívanou transformací je  $y^* = \log(y)$ , jedním z důvodů je i dobrá interpretabilita parametru  $\beta$

Interpretace parametrů LM

1. Klasický LM:

$$\mathbb{E}[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

jednorozměrná změna proměnné  $x_j \Rightarrow$  změnu  $\mathbb{E}[Y]$  o  $\beta_j$  jednotek (při ostatních proměnných stálých).

$$\begin{pmatrix} \mathbf{X} = (1, x_1, \dots, x_m) & \mathbf{X}_{\text{new}} = (1, x_1, \dots, x_j + 1, \dots, x_m) \\ \downarrow & \downarrow \\ \mathbb{E}[Y] & \mathbb{E}[Y_{\text{new}}] \\ \mathbb{E}[Y_{\text{new}}] - \mathbb{E}[Y] = \beta_j \end{pmatrix}$$

2. LM pro logY:

$$\log Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + e \quad \text{kde } e \sim \mathcal{N}(0, \sigma^2)$$

Pokud je to správný model, znamená to, že  $\log Y \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow Y \sim \mathcal{LN}(\mu, \sigma^2)$  a tedy  $\mathbb{E}[Y] = e^{\mu + \frac{\sigma^2}{2}}$ .

Predikce pro  $\mathbb{E}[\log Y]$  je  $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m$ .

Predikce pro  $\mathbb{E}[Y]$  bude  $e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m + \frac{\sigma^2}{2}}$ .

Uvazujme opět jednotkovou změnu p.  $x_j$  ( $x_j \rightarrow x_j + 1$ )

$$\frac{\mathbb{E}[Y_{\text{new}}]}{\mathbb{E}[Y]} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j x_j + \hat{\beta}_j + \cdots + \hat{\beta}_m x_m + \frac{\sigma^2}{2}}}{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m + \frac{\sigma^2}{2}}} = e^{\hat{\beta}_j}$$

jednotková změna proměnné  $x_j \Rightarrow$  multiplikativní změna  $\mathbb{E}[Y] e^{\hat{\beta}_j}$ -krát.

Jinak zapsáno:  $100(e^{\hat{\beta}_j} - 1)$  je procentní změna  $\mathbb{E}[Y]$  spojená s jednotkovou změnou  $x_j$

- z dosazení vyplývá, že odhad parametru  $\beta$  a  $\sigma^2$  lze získat použitím transformovaného modelu (2)
- Protože ale transformovaný model neobsahuje intercept (první sloupec M je  $(\sqrt{w_1}, \dots, \sqrt{w_n})^T$ ), nefunguje klasický rozklad součtu čtverců a F statistika nelze definovat obvyklým způsobem, stejně jako  $R^2$  (viz. regrese skrz počátek)
- nicméně princip „extra sum of squares“ funguje, ať má model intercept nebo ne: např. celkový F-test lze provést pomocí statistiky

$$F_w = \frac{\frac{\text{SSE}_R - \text{SSE}_F}{m}}{\frac{s_w^2}{s_w^2}},$$

### 3 IS a t-testy pro parametry

kde  $SSE_F$  je reziduální součet čtverců  $s_w^2$  plného modelu a  $SSE_R$  je reziduální součet čtverců redukovaného transformovaného modelu  $\mathbb{Z} = \mathbb{M}_0\beta_0 + \mathbf{e}$ ,  $\mathbb{M}_0 = (\sqrt{w_1}, \dots, \sqrt{w_n})^T$ .

Pokud mají chyby normální rozdělení, platí za  $H_0 : \beta_1 = \dots = \beta_m = 0$ , že  $F_w \sim F(m, n - m - 1)$  a  $H_0$  zamítáme, pokud  $F_w > F_{1-\alpha}(m, n - m)$ .

- str. 110 (\*)
- pro analýzu reziduí je třeba uvažovat vhodné grafy reziduí:
  - máme dva vektory reziduí:

$$\begin{aligned}\hat{e}_i &\text{ v původním modelu (1)} \\ \hat{\varepsilon}_i &\text{ v transformovaném modelu (2)}\end{aligned}$$

a tedy dvě možnosti

- pro kontrolu konstantního rozptylu lze uvažovat i standardizovaná nebo studentizovaná rezidua (pomocí bodu 4) a 5) věty lze ukázat, že jsou v obou modelech stejná)
- je třeba být opatrny oproti jakým hodnotám budeme rezidua zobrazovat
- grafy  $\hat{\varepsilon}_i$  proti sloupcům  $\mathbb{M}$  a predikovaným hodnotám  $\hat{z}$  jsou OK, neboť např.

$$\sum_{i=1}^n \hat{z}_i \hat{\varepsilon}_i = 0$$

(jsou OG, měl by být vidět rozstřelený oblak kolem osy x).

- dosazením  $\hat{\varepsilon}_i = \sqrt{w_i} \cdot \hat{e}_i$  a  $\hat{z}_i = \sqrt{w_i} \cdot \hat{y}_i$  dostaneme  $\sum_{i=1}^n w_i \hat{e}_i \hat{y}_i = 0$ , tzn. graf  $\hat{e}_i$  proti  $\hat{y}_i$  bude zavádějící
  - graf  $\sqrt{w_i} \cdot \hat{e}_i$  proti  $\sqrt{w_i} \cdot \hat{y}_i$  je ale v pořádku
  - podobné závěry platí i pro grafy  $\hat{e}_i$  proti  $\mathbf{x}_j^c, i = 1, \dots, m$ .
- (\*): přirozené je definovat  $R^2 = \varrho^2(\hat{\mathbf{z}}, \mathbf{z})$ , kde  $\varrho(\hat{\mathbf{z}}, \mathbf{z})$  je výběrový korelační koeficient, pro  $\mathbb{W} = \mathbb{I}$  dostaneme standardní  $R^2$

**Poznámka 3.31.** • pokud jsou váhy neznámé, bylo by třeba je odhadnout společně s  $\beta$  a  $\sigma^2$  z dat

- to ale není obecně možné, protože máme více parametrů než dat
- někdy to možné je, pokud máme další informace o rozdělení chyb (tvar kovarianční matice atd.)

**Poznámka 3.32.** Celý postup WLS lze použít i na případ  $\mathbf{e} \sim \mathcal{N}_m(0, \sigma^2 \mathbb{W})$ , kde  $\mathbb{W}$  je známá, ale není diagonální. Protože  $\mathbb{W}$  je symetrická, ex. regulární  $\mathbb{K}$  tak, že  $\mathbb{W} = \mathbb{K} \mathbb{K}^T$ . Stejná transformace jako u WLS opět vede na transformovaný model, kde  $\varepsilon \sim \mathcal{N}_m(0, \sigma^2 \mathbb{I}_m)$ .

### 3.4 Korelované chyby

- Zejména v časových nebo ekonomických datech se často objevuje korelace jednotlivých hodnot.
- potom není splněn předpoklad nezávislosti chyb
- tento stav je třeba detektovat (někdy pomohou grafy reziduí)
- modely pro korelovaná data: **Analýza časových řad**

Pokud je přítomna autokorelace a chyby mají konstantní rozptyl, platí:

1. OLS odhad  $\hat{\beta}$  je nestranný, ale neplatí Gauss-Markovova věta, tzn.  $\hat{\beta}$  nemá nejmenší rozptyl.
2.  $MSE = \frac{1}{n-m-1} SSE$  (odhad  $\sigma^2$ ) může být podstatně menší než skutečná hodnota  $\sigma^2$ , což může dávat falešný pocit přesnosti.
3. V důsledku bodu 2) mohou být zvětšeny hodnoty T statistik, takže testy o parametrech a IS nefungují.
4. Protože jsou chyby nezávislé, F-testy a t-testy nejsou přesně platné ani když jsou chyby normální.

#### 3.4.1 Durbin-Watson statistika

Omezíme se na pozorování získaná v čase  $t = 1, 2, \dots, n$  a případ, že chyby  $e_t$  splňují podmínky autoregresního procesu 1. řádu (AR1), tj.

$$e_t = \varrho e_{t-1} + u_t, \quad |\varrho| < 1,$$

kde  $\varrho$  je autokorelační koeficient,  $u_t \sim \mathcal{N}(0, \sigma_n^2)$  jsou nezávislé v  $t = 1, \dots, n$  a  $u_t$  je nezávislá na  $e_t, t \geq 1$ . Častěji pro data časových řad platí  $\varrho > 0$  (pozitivní autokorelace).

Pro test  $H_0 : \varrho = 0$  vs.  $H_1 : \varrho > 0$  se užívá **Durbin-Watsonova statistika**

$$d = \frac{\sum_{t=2}^n (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^n \hat{e}_t^2},$$

kde  $\hat{e}_t$  jsou rezidua modelu LR. Pokud zamítneme  $H_0$ , odhadne se  $\varrho$  pomocí

$$\hat{\varrho} = \frac{\sum_{t=2}^n \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^n \hat{e}_t^2}.$$

**Poznámka 3.33.** Dá se ukázat, že  $d \approx 2(1 - \hat{\varrho})$ :

$$\sum_{t=2}^n (\hat{e}_t - \hat{e}_{t-1})^2 = \sum_{t=2}^n \hat{e}_t^2 + \sum_{t=2}^n \hat{e}_{t-1}^2 - 2 \sum_{t=2}^n \hat{e}_t \hat{e}_{t-1} \approx 2 \left( \sum_{t=2}^n \hat{e}_t^2 - \sum_{t=2}^n \hat{e}_t \hat{e}_{t-1} \right),$$

Z Cauchy-Schwartzovy nerovnosti  $\Rightarrow \frac{\sum_{t=2}^n \hat{e}_t \hat{e}_{t-1}}{\sum_{t=1}^n \hat{e}_t^2}$  leží přibližně v  $(-1, 1)$ , tzn.  $d$  leží přibližně v  $(0, 4)$ . Dále

$$\hat{\varrho} \approx 1 \implies d \approx 0 \quad \text{a} \quad \hat{\varrho} \approx 0 \implies d \approx 2,$$

tzn. pro malé hodnoty  $d$  budeme zamítat  $H_0$ , pro velké hodnoty nebudeme zamítat. Kritické hodnoty určené Durbinem a Watsonem jsou tabulované.

### 3 IS a t-testy pro parametry

#### Test:

1. spočítat hodnotu  $d$
2. nalézt kritické hodnoty  $(d_L, d_U)$  pro dané  $n$  a  $m + 1$
3.
  - a) zamítnout  $H_0$ , pokud  $d < d_L$
  - b) nezamítnout  $H_0$ , pokud  $d > d_U$
  - c) pro  $d_L < d < d_U$  test nerozhodne

POZNÁMKA 3.34. Pro test  $H_0 : \varrho = 0$  vs.  $H_1 : \varrho < 0$  lze použít popsaný test pro  $d' = 4 - d$ .  
Metody pro korekci autokorelace: **Cochrane-Orcutt**.