

# **01RAD**

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

11. ledna 2021

# Obsah

<b>1 Jednorozměrná lineární regrese</b>	<b>1</b>
1.1 Odhad parametrů . . . . .	2
1.1.1 Data s předpokladem normality dat . . . . .	2
1.1.2 Data bez předpokladu normality . . . . .	3
1.1.3 Vlastnosti odhadů $\hat{\beta}_0, \hat{\beta}_1, s_n^2$ . . . . .	5
1.2 Gauss - Markov theorem . . . . .	9
1.3 IS pro $\beta_0, \beta_1$ . . . . .	10
1.3.1 TH pro $\beta_0, \beta_1$ . . . . .	11
1.4 ANOVA přístup pro testování . . . . .	12
1.5 Regrese skrz počátek . . . . .	17
1.5.1 Odhady a testy v případě $\beta_0 = 0$ . . . . .	17
1.6 Predikce . . . . .	19
1.7 Základní procedury pro ověření linearity . . . . .	21
1.8 Grafy reziduí . . . . .	24
<b>2 Vícerozměrná lineární regrese</b>	<b>25</b>
2.1 Odhad parametrů . . . . .	27
2.1.1 Odhad parametru $\sigma^2$ . . . . .	27
2.1.2 Vlastnosti odhadů $\hat{\beta}, s_n^2$ . . . . .	28
2.1.3 Vlastnosti vektoru reziduů $\hat{e}$ . . . . .	31
2.2 Gauss - Markov theorem . . . . .	32
2.3 Testování modelu - tabulka ANOVA . . . . .	33
2.3.1 Celkový F-test (overall F-test) . . . . .	33
2.3.2 Koeficient (vícenásobné) determinace $R^2$ . . . . .	37
2.4 IS a t-testy pro parametry . . . . .	38
2.5 Obecná lineární hypotéza . . . . .	38
2.6 Predikce . . . . .	41
<b>3 Rezidua, diagnostika a transformace</b>	<b>44</b>
3.1 Rezidua . . . . .	44
3.1.1 Vlastnosti potenciálu $h_{ii}$ . . . . .	45
3.2 Grafy reziduí . . . . .	45
3.2.1 Partial residual plot . . . . .	46
3.2.2 Partial regression plot . . . . .	47
3.3 PRESS rezidua (PRESS residuals, deleted residuals) . . . . .	48
3.4 Míry influence . . . . .	52
3.4.1 DFBETAS . . . . .	53
3.4.2 DFFITS . . . . .	54
3.5 Transformace . . . . .	54
3.5.1 Transformace vysvětlované proměnné $y$ . . . . .	55

*Obsah*

3.5.2	Box-Cox transformace . . . . .	56
3.5.3	Transformace vysvětlujících proměnných x . . . . .	58
3.6	Korelované chyby . . . . .	63
3.6.1	Durbin-Watson statistika . . . . .	63
3.7	Metody výběru modelu . . . . .	66

# 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 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 1.1. 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

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad (*) \quad i \in \hat{n}.$$

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  $e_i \sim (0, \sigma^2)$ . Z toho důvodu však splňuje podmínky  $\mathbb{E}[e_i] = 0$ ,  $D[e_i] = \sigma^2$  pro  $\forall i \in \hat{n}$  (homoskedasticita). Měřením získáme data  $(x_1, y_1), \dots, (x_n, y_n)$  a cílem statistické analýzy je určit, zda je model  $(*)$  schopen popsat pozorovanou variabilitu u  $y$ .

V **prvním kroce** 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 pro  $\forall i \in \hat{n}$  **naměřená data  $y_i$  s predikovanou hodnotou lineární regrese  $\hat{y}(x_i)$** . To nám umožňuje posoudit adekvátnost modelu.

Pro proložení dat přímkou existuje několik způsobů. Zásadní ovšem bude znalost rozdělení  $e_i$  (v tomto případě i  $Y_i$ ), i když apriori není zřejmé, proč znát rozdělení a nikoliv  $\beta_0, \beta_1$ .

Máme dvě možnosti,

1. odhadnout  $\beta_0, \beta_1$  pomocí metody nezávisející na rozdělení chyb, nebo
2. udělat věrohodnostní předpoklad o rozdělení chyb, odhadnout  $\beta_0, \beta_1$  a následně ověřit předpoklad.

POZNÁMKA 1.2. Speciální důležitý případ je  $e_i \sim \mathcal{N}(0, \sigma^2)$ , který při MLE odhadu  $\beta_0, \beta_1$  vede na metodu nejmenších čtverců, která může být použita bez ohledu na rozdělení chyb.

## 1.1 Odhad parametrů

### 1.1.1 Data s předpokladem normality dat

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

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

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

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

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

Jednotlivé MLE odhad parametrů pak mají následující tvar

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

Nyní najdeme odhad parametru  $\sigma^2$

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

## 1 Jednorozměrná lineární regrese

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 odhadы chyb  $e_i$ , jejich analýza hráje významnou roli v ověření předpokladů rozdělení chyb.

### Odhad $\sigma^2$

Pro odhad  $\sigma^2$  se používá častěji statistika s názvem **standardní chyba regrese** (*standard error*) ve tvaru

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

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.

#### 1.1.2 Data bez předpokladu normality

V tomto případě jsou tedy  $e_1, \dots, e_n$  nekorelované,  $e_1, \dots, e_n \sim (0, \sigma^2)$ . Pro odhad  $\beta_0, \beta_1$  lze použít minimalizaci S (metodou nejmenších čtverců), což je rozumné provedení, když si uvědomíme geometrickou interpretaci.

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 stejně 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–Markovově větě.

PŘÍKLAD 1.3. Nechť  $e_1, \dots, e_n$  jsou *iid* s hustotou

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

Potom je hustota  $Y_i$  rovna

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

## 1 Jednorozměrná lineární regrese

a věrohodnostní funkce  $L$  a  $l$  mají tvar

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

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

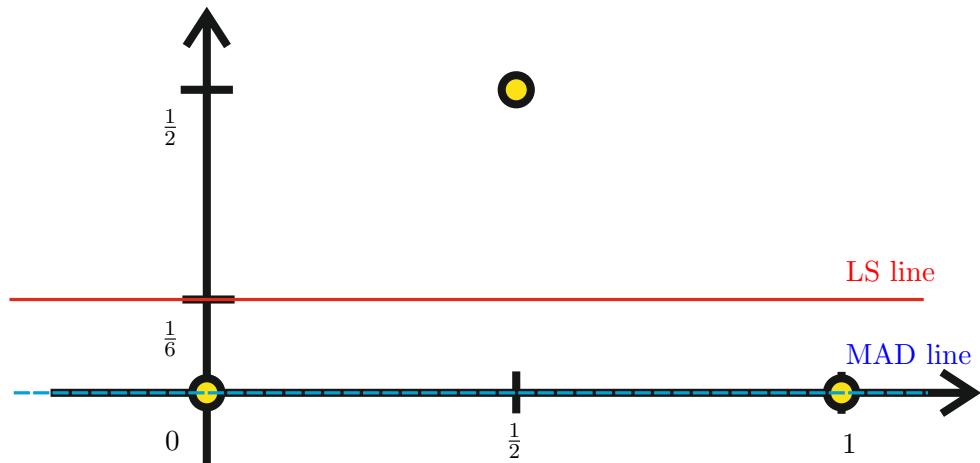
MLE odhad parametrů  $\beta_0, \beta_1$  získáme minimalizací

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

které nazýváme **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 1.4.** I když  $s_n^2$  je nestranný odhad  $\sigma^2$ ,  $s_n$  je vychýlený odhad  $\sigma$ ! Je to obecná vlastnost odhadů (nestranných) rozptylů, neboť pokud je  $s^2$  nestranný odhad  $\sigma^2$ , pak  $\mathbb{E}[s] \leq \sigma$ .

*Důkaz.* Uvažujme náhodnou veličinu  $X$ , pro kterou platí, že  $D[X] < +\infty$ . Po dosazení  $X = s$  do známé rovnice  $\mathbb{E}[X^2] = D[X] + \mathbb{E}[X]^2$  dostaneme vztah

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

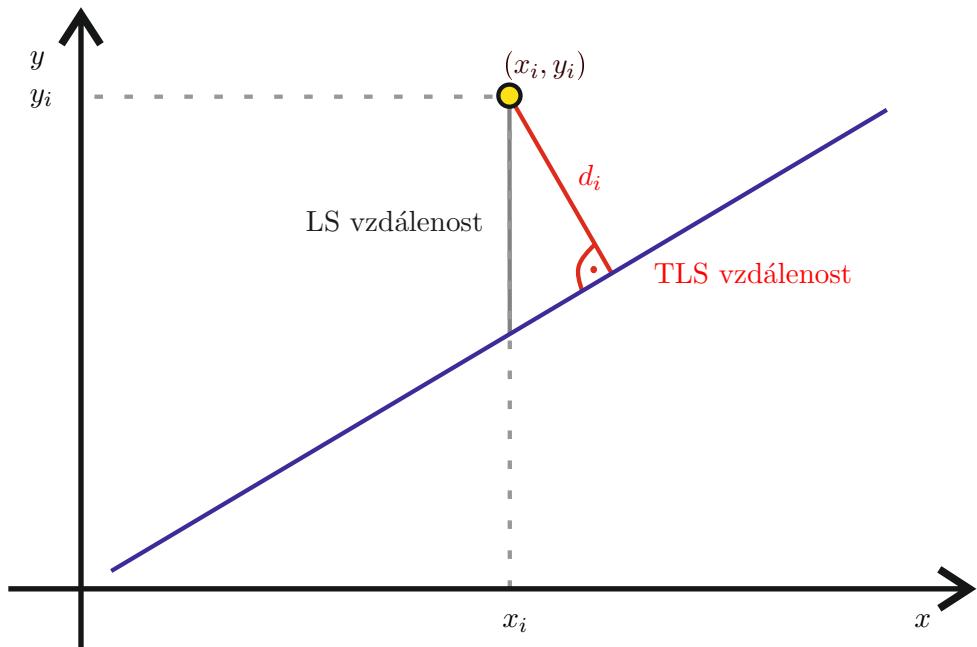
kde  $\mathbb{E}[s]^2 \leq \sigma^2$ ,  $\mathbb{E}[s] \leq \sigma$  a rovnost nastává, pokud  $D[s] = 0$ .

□

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

## 1 Jednorozměrná lineární regrese

**Poznámka 1.5.** Předpokládali jsme, že hodnoty  $x_i$  jsou dány přesně, což nemusí být vždy pravda. Často jsou obě veličiny ( $x, y$ ) měřeny nepřesně. Existují EIV models „error in variable“, v těchto modelech jsou často preferovány jiné odhadování než LSE. Populární metoda je dále **total least squares** (*orthogonal least squares*). Zde minimalizujeme  $\sum_{i=1}^n d_i^2$ , kde  $d_i$  je vzdálenost bodu a přímky (kolmice na přímku protínající bod). To znamená, že neupřednostňujeme veličinu  $x$ , ale přistupujeme k  $x$  a  $y$  rovnoměrně.



**Poznámka 1.6.** V literatuře se někdy  $x$  uvažují jako realizace náhodné veličiny (ne vždy se  $x$  nastavuje předem, nebo je jasně dané).

Model má potom tvar

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

Pro většinu výsledků prezentovaných v této přednášce ale není podstatné, zda je  $x$  chápáno jako pevné nebo náhodné. Důkazy většinou fungují s podmíněnými výrazy ( $\mathbb{E}, D, \dots$ ) při dané hodnotě  $x$  místo nepodmíněných. Větší pozornost je naproti tomu potřeba u odvození asymptotických rozdělení odhadů.

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

**Věta 1.7.** Nechť  $\hat{\beta}_0, \hat{\beta}_1$  jsou LSE odhadování parametrů  $\beta_0, \beta_1$  v lineárním modelu

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i \in \hat{n},$$

kde  $e_i$  jsou nezávislé náhodné veličiny (postačí i nekorelovanost) se stejným rozptylem  $\sigma^2$ .

Potom platí, že

1.  $\mathbb{E}[\hat{\beta}_0] = \beta_0, \quad \mathbb{E}[\hat{\beta}_1] = \beta_1$ , (nestranné odhadování),

## 1 Jednorozměrná lineární regrese

2.  $D[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$ , 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  $e_i \sim \mathcal{N}(0, \sigma^2)$ ,  $\forall i \in \hat{n}$ , potom  $\hat{\beta}_j \sim \mathcal{N}(\beta_j, D[\hat{\beta}_j])$ ,  $j \in \{0, 1\}$ .

*Důkaz.* 1. Upravíme  $\hat{\beta}_1$ :

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i x_i - n \bar{x}_n \bar{y}_n}{\sum_{i=1}^n x_i^2 - n \bar{x}_n^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = \\ &= \frac{1}{S_{xx}} \left( \sum_{i=1}^n (x_i - \bar{x}_n)y_i - \bar{y}_n \sum_{i=1}^n (x_i - \bar{x}_n) \right) = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n)y_i.\end{aligned}\quad (1.1)$$

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

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

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. Jelikož  $Y_i$  jsou nezávislé, můžeme spočítat rozptyl jako

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

3. Zde už nemáme nezávislé náhodné veličiny, proto musíme počítat i s kovariancemi:

$$\begin{aligned}D[\hat{\beta}_0] &= D[\bar{Y}_n - \hat{\beta}_1 \bar{x}_n] = D[\bar{Y}_n] + \bar{x}_n^2 D[\hat{\beta}_1] - 2 \bar{x}_n \text{Cov}(\bar{Y}_n, \hat{\beta}_1) = \\ &= \frac{\sigma^2}{n} + \frac{\bar{x}_n^2 \sigma^2}{S_{xx}} - 2 \bar{x}_n \text{Cov}(\bar{Y}_n, \hat{\beta}_1).\end{aligned}$$

Ted' už nám stačí ukázat, že  $\text{Cov}(\bar{Y}_n, \hat{\beta}_1) = 0$ .

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

Z toho už vyplývá, že

$$\text{Cov}(\bar{Y}_n, \hat{\beta}_1) = 0 = \frac{\sigma^2}{n S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n).$$

## 1 Jednorozměrná lineární regrese

4. Protože

$$\begin{aligned}\hat{\beta}_1 &= \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) Y_i, \\ \hat{\beta}_0 &= \frac{1}{n} \sum_{i=1}^n Y_i - \hat{\beta}_1 \bar{x}_n,\end{aligned}$$

pak je  $\hat{\beta}_0$  i  $\hat{\beta}_1$  LK nezávislých normálních náhodných veličin  $Y_i$ . Z toho vyplývá, že mají normální rozdělení, kde  $\mathbb{E}$  a  $D$  jsme už vypočítali.

□

**Věta 1.8.** 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  $\mathbb{E}(\hat{Y}_i) = \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i = \mathbb{E}Y_i$ , platí, že

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

Dostáváme tak

$$\begin{aligned}A &= \sum_{i=1}^n D(Y_i - \hat{Y}_i) = \sum_{i=1}^n [D(Y_i) + D(\hat{Y}_i) - 2\text{Cov}(Y_i, \hat{Y}_i)] = \\ &= n\sigma^2 + \sum_{i=1}^n D(\hat{Y}_i) - 2 \sum_{i=1}^n \text{Cov}(Y_i, \hat{Y}_i)\end{aligned}\tag{1.2}$$

Rozepíšeme

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

kde

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

a tedy

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

## 1 Jednorozměrná lineární regrese

Následně máme

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

kde za  $\hat{\beta}_1$  dosadíme podle (1.1). Tedy

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

Dosazením do (1.2) 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 1.9.** Nechť platí předpoklady věty 1 a nechť  $e_1, \dots, e_n$  iid  $\mathcal{N}(0, \sigma^2)$ . Potom platí, že

- 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 1.10. Spočetli jsme

$$\hat{\sigma}^2(\hat{\beta}_0) \xrightarrow{\text{ozn.}} D(\hat{\beta}_0) = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}} \right], \quad (1.3)$$

$$\hat{\sigma}^2(\hat{\beta}_1) \xrightarrow{\text{ozn.}} D(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}. \quad (1.4)$$

Nestranné odhady jsou

$$\sigma^2(\hat{\beta}_0) = s_n^2 \left[ \frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}} \right] = s_n^2 \delta_0 \quad \text{a} \quad \sigma^2(\hat{\beta}_1) = \frac{s_n^2}{S_{xx}} = s_n^2 \delta_1,$$

kde  $\delta_0$  a  $\delta_1$  jsou tzv. *variance multiplication factors*.

Odhady směrodatné odchylyky 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.

## 1.2 Gauss - Markov theorem

- Pokud mají chyby normální rozdělení, pak 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

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i \in \hat{n}. \quad (*)$$

**Definice 1.11.** 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  $i \in \hat{n}$ .

**Věta 1.12** (Gauss-Markov theorem). *Nechť  $e_1, \dots, e_n$  v modelu (\*) jsou nekorelované a mají stejný rozptyl  $D(e_i) = \sigma^2$ ,  $i \in \hat{n}$ . Potom LSE  $\hat{\beta}_j$ ,  $j \in \{0, 1\}$  je BLUE parametru  $\beta_j$ .*

*Důkaz.* Ukážeme pro  $\beta_1$ , pro  $\beta_0$  je důkaz podobný. Nechť tedy  $\hat{\beta}_1 = \sum_{i=1}^n c_i Y_i$ , 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

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

To musí platit pro libovolná  $\beta_0, \beta_1$ , a proto 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  $\sum_{i=1}^n c_i x_i = 1$ .

Sestavíme Lagrangeovu funkci  $L = \sum_{i=1}^n c_i^2 - 2\lambda_1 \left( \sum_{i=1}^n c_i \right) - 2\lambda_2 \left( \sum_{i=1}^n c_i x_i - 1 \right)$  (konstanta 2 před  $\lambda_i$  je zde z toho důvodu, aby výpočet vypadal lépe, ale není nutná).

## 1 Jednorozměrná lineární regrese

$$\begin{aligned}\frac{\partial L}{\partial c_i} &= 2c_i - 2\lambda_1 - 2\lambda_2 x_i = 0, \quad i \in \hat{n}, \\ \frac{\partial L}{\partial \lambda_1} &= -2 \left( \sum_{i=1}^n c_i \right) = 0, \\ \frac{\partial L}{\partial \lambda_2} &= -2 \left( \sum_{i=1}^n c_i x_i - 1 \right) = 0.\end{aligned}$$

Sečteme prvních  $n$  rovnic:

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

Sečteme dále prvních  $n$  rovnic vynásobených  $x_i$ :

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

Dosadíme za  $\lambda_1, \lambda_2$  a dostaneme

$$c_i + \frac{\bar{x}_n}{S_{xx}} - \frac{x_i}{S_{xx}} = 0 \quad \Rightarrow \quad c_i = \frac{x_i - \bar{x}_n}{S_{xx}} \quad \text{a} \quad \hat{\beta}_1 = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}_n) Y_i,$$

což je LSE. □

**Poznámka 1.13.** Ukázali jsme pouze, že to je stacionární bod, že je tam i minimum ukážeme v obecnější větě ve vícerozměrné regresi.

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

IS poskytuje jistou „míru přesnosti“ bodových odhadů. Pro jejich konstrukci ale 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 parametrů  $\beta$  jsou lineární funkce  $Y_i, i \in \hat{n}$ , což umožňuje aplikovat CLT a dostat asymptotickou normalitu odhadů  $\hat{\beta}_0, \hat{\beta}_1$ .

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

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

## 1 Jednorozměrná lineární regrese

POZNÁMKA 1.14.

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

Můžeme ukázat, že

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

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

To znamená, že  $\mathbb{P}\left[-t_{1-\frac{\alpha}{2}}(n-2) \leq \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}(\hat{\beta}_i)} \leq t_{1-\frac{\alpha}{2}}(n-2)\right] = 1 - \alpha$ . Vyjádřením  $\beta_i$  dostaneme

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

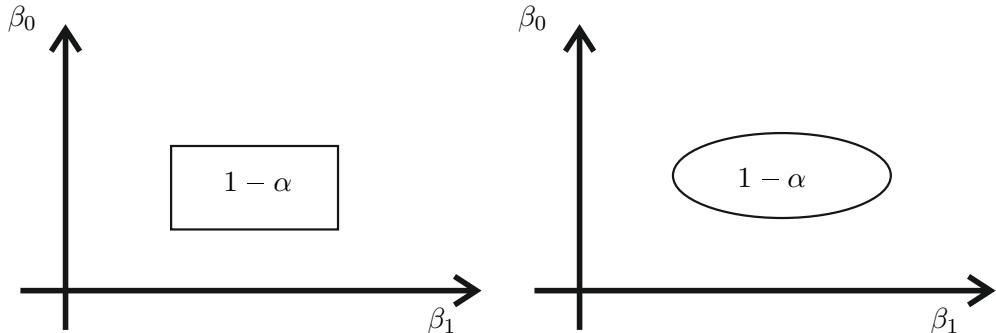
a tedy  $(\hat{\beta}_i \pm t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}(\hat{\beta}_i))$  je  $100(1 - \alpha)\%$  IS pro  $\beta_i$ ,  $i \in \{0, 1\}$ .

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

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

POZNÁMKA 1.15. 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 1.16. Někdy se konstruují simultánní IS pro oba parametry.



Zmíníme podrobněji u vícerozměrné regrese.

### 1.3.1 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. Otázkou zůstává, zda je  $x$  užitečná ve vysvětlení variability v  $y$ , chceme tedy rozhodnout mezi dvěma modely:

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

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

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

## 1 Jednorozměrná lineární regrese

POZNÁMKA 1.17. Tyto závěry jsou správné pouze za předpokladu, že model je lineární!

- Nezamítnutí  $H_0$  nemusí znamenat, že  $x$  není užitečná, může to pouze indikovat, že vztah mezi  $y$  a  $x$  není lineární.
- Zamítnutí  $H_0$  naopak říká, že existuje lineární trend mezi  $x$  a  $y$ , ale mohou tam být i jiné typy závislostí.

Pro konstrukci testů využijeme odvozené IS.

POZNÁMKA 1.18. 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 testu na hladině  $\alpha$ .

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

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

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

A zapsáno dohromady

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

POZNÁMKA 1.19. 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 1.20. 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 tedy zamítá  $H_0$ , pokud

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

### Test významnosti interceptu

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

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

## 1.4 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).

## 1 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

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

což značí *regression 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 podíl variability v  $(y_1, \dots, y_n)$  vysvětlené regresním modelem. Statistika  $R^2$  se nazývá **koeficient determinace** (*coefficient of determination*) a pro každý model by měla 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

$$F = \frac{(n-2)R^2}{1-R^2}. \quad (1.5)$$

**Lemma 1.21.** 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  $\frac{\partial S}{\partial \hat{\beta}_0} = 0$  dostaneme

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

## 1 Jednorozměrná lineární regrese

2. Z bodu 1) plynne, že  $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$ , podělením  $n$  dostaneme dokazované tvrzení.
3. Z rovnice  $\frac{\partial S}{\partial \hat{\beta}_1} = 0$  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

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

□

**Věta 1.22.** *Předpokládejme, že SST  $\neq 0$ . Potom platí*

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

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

tedy  $R^2$  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

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

Z lemmatu 1.21 vyplývá, že

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

neboť

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

Z toho potom dokazujeme jednotlivé body věty.

## 1 Jednorozměrná lineární regrese

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

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

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

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

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

5. Z definice F podle (1.5) 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 1.23.**
1. Z bodů 5 a 6 vyplývá, že použití libovolné statistiky  $T_n, R^2$  nebo F vede na ekvivalentní test významnosti regrese.
  2.  $R^2$  poskytuje hrubou představu o kvalitě modelu, čím je blíže 1, tím lépe přímka prokládá data (nicméně je třeba jisté obezřetnosti, jak uvidíme později).
  3. F lze chápat jako statistiku pro test významnosti velkých hodnot  $R^2$ .

Výsledky se většinou uvádí v tabulce ANOVA:

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

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

**POZNÁMKA 1.24.**  $H_0 : \beta_1 = 0$  je zamítnuta, pokud  $F > F_{1-\alpha}(1, n-2)$ . V tomto jednorozměrném případě je to ekvivalentní t-testu, neboť  $F = T^2$ .

## 1 Jednorozměrná lineární regrese

**Věta 1.25.** 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 1.26. Proto se v tabulce ANOVA 1.4 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  $n$  reziduí  $\hat{e}_1, \dots, \hat{e}_n$  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.  $\text{SST} = \sum_{i=1}^n (y_i - \bar{y}_n)^2$ , a proto musí  $y_i - \bar{y}_n$  splňovat  $\sum_{i=1}^n (y_i - \bar{y}_n) = 0$ , tudíž 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 1.22 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  $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$ , a tedy  $(\hat{\beta}_1 - \beta_1) \frac{\sqrt{S_{xx}}}{\sigma} \sim \mathcal{N}(0, 1)$ . 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 tvrzení 1.9) 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 1.27. R<sup>2</sup> statistika - pozor na zjednodušení kvality modelu.

1. Nízké hodnoty R<sup>2</sup> 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<sup>2</sup> oproti modelům s různými  $x$ .
2. Velké hodnoty R<sup>2</sup> 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<sup>2</sup> 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 R<sup>2</sup>, než modely s „malou“ směrnicí.

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

1. „velké“ R<sup>2</sup>,
2. „velké“ F nebo |T| hodnoty,
3. „malé“ hodnoty  $s_n^2$  vzhledem k  $\bar{y}_n$ .

Další kritéria budeme probírat později.

## 1 Jednorozměrná lineární regrese

PŘÍKLAD 1.28. Velká hodnota  $R^2$  indikuje přibližně lineární vztah mezi  $x$  a  $y$ , ale vysoký stupeň korelace nemusí znamenat příčinný vztah. Uvedeme nyní ríklad na datech z let 1924-1937. Mějme

$y_i$  - počet mentálních onemocnění na 100000 obyvatel Anglie.

$x_i$  - počet rádií v populaci.

Určíme parametry modelu  $y_i = \beta_0 + \beta_1 x_i + e_i$  jako

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

tzv. zjišťujeme 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 1.29. 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.

## 1.5 Regrese skrz počátek

Existují případy, kdy přípustný model vyžaduje  $\beta_0 = 0$ , tj.  $Y_i = \beta_1 x_i + e_i$ ,  $i \in \hat{n}$ .

1. Na základě fyzikálních úvah je předem známo, že

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

Potom tedy nemá smysl odhadovat  $\beta_0$ , protože to obecně sníží přesnost odhadu  $\sigma^2$ , a tedy i  $\beta_1$ .

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

POZNÁMKA 1.30. V praktických situacích si často nemůžeme být jisti, že model platí i blízko počátku. Část statistiků trvá na přítomnosti interceptu v modelu, i když je nevýznamný.

Položit  $\beta_0$  apriori může být 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  $x = 0$ .

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

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

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

## 1 Jednorozměrná lineární regrese

Pokud  $e_1, \dots, e_n$  iid  $\mathcal{N}(0, \sigma^2)$ , potom

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 \quad \text{a} \quad D[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^n x_i^2} \quad \Rightarrow \quad \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n x_i^2}\right)$$

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

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

kde  $100(1-\alpha)\%$  IS pro  $\beta_1$  je

$$\left( \hat{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-1) \frac{s_n}{\sqrt{\sum x_i^2}} \right).$$

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 reziduů  $\sum_{i=1}^n (y_i - \hat{y}_i)$  nemusí být 0, a tedy  $\bar{y}_n \neq \bar{y}_i$ . Odvodíme nový rozklad, který platí v obou případech, dokážeme ho ale jen pro  $\beta_0 = 0$ .

**Věta 1.31.** V modelu s  $\beta_0 = 0$  platí, že

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

Důkaz.

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

Z rovnice  $\frac{dS}{d\beta_1} = 0$  dostaneme  $\sum_{i=1}^n (y_i - \hat{\beta}_1 x_i) x_i = 0$  a po vynásobení obou stran rovnic  $\hat{\beta}_1$  již máme

$$\sum_{i=1}^n (y_i - \hat{y}_i) \hat{y}_i = 0.$$

□

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

$$R^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} \quad \Leftrightarrow \quad 1 - R^2 = \frac{\sum_{i=1}^n y_i^2 - \sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} = \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n y_i^2}$$

## 1 Jednorozměrná lineární regrese

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

$$F = \frac{\sum_{i=1}^n \hat{y}_i^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \frac{\hat{\beta}_1 \sum_{i=1}^n x_i^2}{s_n^2} = T^2.$$

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

**POZNÁMKA 1.32.** Tato definice  $R^2$  se ale v praxi moc nepoužívá, protože neumožňuje přímé srovnání modelů s interceptem a bez něj.

$$\beta_0 = 0 : R^2 = 1 - \frac{\text{SSE}}{\sum_{i=1}^n y_i^2}, \quad \beta_0 \neq 0 : R^2 = 1 - \frac{\text{SSE}}{\sum_{i=1}^n (y_i - \bar{y}_n)^2}.$$

Obecně ale  $\sum_{i=1}^n (y_i - \bar{y}_n)^2 < \sum_{i=1}^n y_i^2$ ,  $R^2$  v modelu s  $\beta_0 = 0$  tedy bude větší, než  $R^2$  modelu s  $\beta_0 \neq 0$ , i když jsou jejich SSE srovnatelné.

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

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

Tabulka 1.1: Tabulka ANOVA pro  $\beta_0 = 0$ .

## 1.6 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.

## 1 Jednorozměrná lineární regrese

### 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  $\sigma^2$  statistikou  $s_n^2$  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  $\mathcal{N}(0, \sigma^2)$ , platí, že

$$\hat{Y}_0 \sim \mathcal{N} \left( \mu_0, \underbrace{\sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}_{\hat{\sigma}^2(\hat{Y}_0)} \right),$$

a tedy

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

Celkově tedy dostaváme

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

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

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

**POZNÁMKA 1.33.** 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  $(\min x_i, \max x_i)$ .

### Ad 2)

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

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

## 1 Jednorozměrná lineární regrese

Odhad tohoto 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 1.34.** Přesnost predikce

- a) roste s rostoucím  $n$  a rostoucím rozsahem  $x$  naměřeným pomocí  $S_{xx}$ ,
- b) 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**:

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

**POZNÁMKA 1.35.** Odvozené výsledky platí za předpokladu normality chyb. Protože jsou ale za podmínek regularity odhady  $\hat{\beta}_0, \hat{\beta}_1$  asymptoticky normální, IS pro  $E[Y_0]$  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 1.36** (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.

## 1.7 Základní procedury pro ověření linearity

- 1) Prozkoumání **scatter plotu** dvojic  $(x_i, y_i)$ . Příklad lze vidět na obrázku 1.1. Takový scatter plot může indikovat, že lepší model bude

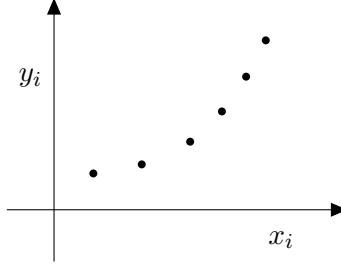
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i.$$

Scatter plot ale může být zavádějící, pokud je odklon od linearity způsoben spíše chybějící proměnnou, než polynomiální závislostí na  $x$ .

- 2) **Analyza hodnot testovacích statistik.**

- Např. malá hodnota  $R^2$  společně s významnou hodnotou 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.

## 1 Jednorozměrná lineární regrese



Obrázek 1.1: Scatter plot naměřených dat.

- 3) **Obrázky reziduů.** Je to efektivní diagnostický nástroj. Rezidua odhadují, kolik variability v datech zůstane po odstranění lineární části v  $x$ . Dá se také očekávat, že jejich hodnoty budou užitečné pro detekci odchylek od normality.

**PŘÍKLAD 1.37.** Analýza scatter plotů a obrázků reziduů 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  $x_i$ .

### Ad 3) - Analýza reziduů

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

**Věta 1.38.** Nechť  $\hat{e}_i$  jsou rezidua modelu (\*) odhadnutého metodou nejmenších čtverců. Potom platí, že

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

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

*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  $\mathbb{E}\hat{Y}_i = \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i = \mathbb{E}Y_i$ .  
2.

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

## 1 Jednorozměrná lineární regrese

3.

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

Podobně bychom dostali

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

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

4.

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

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

$$\begin{aligned}1) &\Rightarrow \mathbb{E}\hat{e}_i = 0 \\ 2) &\Rightarrow \mathbb{D}\hat{e}_i = \sigma_{\hat{e}_i}^2 \\ &\Rightarrow \frac{\hat{e}_i}{\sigma_{\hat{e}_i}} \sim \mathcal{N}(0, 1)\end{aligned}$$

□

**POZNÁMKA 1.39.** Z bodu 3) věty plyne, že  $\mathbb{C}\text{ov}(\hat{e}_i, \hat{e}_j) \approx 0$  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 je proto odhadnout  $\sigma^2$  pomocí  $s_n^2$ . Potom by se měla **standardizovaná rezidua**

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

pro velká  $n$  opět chovat jako náhodná veličina z  $\mathcal{N}(0, 1)$ .

**POZNÁMKA 1.40.**  $\hat{e}_i$  se užívají pro grafickou analýzu.

Existuje ale i jiná třída reziduí, tzv. PRESS rezidua.

Označme  $\hat{\beta}_{0(-i)}, \hat{\beta}_{1(-i)}$  odhad 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.

## 1.8 Grafy reziduí

- Histogram reziduí – umožní 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  $\Phi^{-1}\left((i - \frac{1}{2})\frac{1}{n}\right)$ ,  $i \in \hat{n}$ . Body by měly ležet přibližně na přímce ( $\mathbb{E}(e_i) \approx \Phi^{-1}\left((i - \frac{1}{2})\frac{1}{n}\right)$  pro normální chyby).
- Použití: ověření normality, detekce odlehlých 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  $\hat{e}_i(\hat{r}_i)$  a  $\hat{Y}_i$  by měly být nekorelované, pokud platí model (\*). To znamená, že graf  $\hat{r}_i \times \hat{y}_i$  by měl být náhodně rozptýlený kolem osy  $x$ , navíc  $\hat{r}_i$  by měla ležet v  $(-3, 3)$  ( $\hat{r}_i \approx \mathcal{N}(0, 1)$ ). (doplnit obrázky)
  - Standardizovaná rezidua  $\times$  poradí pozorování – možná detekce řadové korelace mezi pozorováními. (doplnit obrázky)

## 2 Vícerozměrná lineární regrese

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  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)^T$  (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

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

jako **matici modelu** (regresní matici, *design matrix*). Dostaneme tak model ve tvaru (důležitém)

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times (m+1)} \boldsymbol{\beta}_{(m+1) \times 1} + \mathbf{e}_{n \times 1}. \quad (**)$$

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

Věrohodnostní funkce je potom ve tvaru

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

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

Pro pevné  $\sigma^2$  je

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

je opět pomocí derivací, ukážeme algebraický přístup.

## 2 Vícerozměrná lineární regrese

**Věta 2.1.** Uvažujme model (\*\*), a nechť  $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ . Potom  $\hat{\boldsymbol{\beta}}$  je MLE  $\boldsymbol{\beta}$  právě tehdy, když  $\hat{\boldsymbol{\beta}}$  je řešením soustavy rovnic

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

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

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

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

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

má platit i pro  $\hat{\boldsymbol{\beta}}$ :

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

a tedy

$$g(\boldsymbol{\beta}) - g(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} - 2\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + \hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = (\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \quad (2.1)$$

$$= (\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}))^T (\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})) = \langle \mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \rangle \geq 0, \quad \forall \boldsymbol{\beta}, \quad (2.2)$$

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

$\Rightarrow$  Předpokládejme, že  $\hat{\boldsymbol{\beta}}_1$  minimalizuje  $g(\boldsymbol{\beta})$  (je tedy MLE). To potom znamená, že  $g(\hat{\boldsymbol{\beta}}_1) \leq g(\boldsymbol{\beta})$ ,  $\forall \boldsymbol{\beta}$ , speciálně  $g(\hat{\boldsymbol{\beta}}_1) \leq g(\hat{\boldsymbol{\beta}})$ , kde  $\hat{\boldsymbol{\beta}}$  je řešení soustavy  $\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$ . Z rovnice (2.2) vyplývá, že  $g(\hat{\boldsymbol{\beta}}_1) \geq g(\hat{\boldsymbol{\beta}})$ . Celkem tedy  $g(\hat{\boldsymbol{\beta}}_1) = g(\hat{\boldsymbol{\beta}})$ . Dosazením do (2.2) dostaneme, že

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

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

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

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

Aby byl důkaz korektní, je třeba ukázat, že soustava  $\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$  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 2.2.** Soustava lineárních rovnic  $\mathbf{A}\mathbf{x} = \mathbf{y}$  má řešení právě tehdy, když  $\langle \mathbf{y}, \mathbf{z} \rangle = 0$  pro všechna  $\mathbf{z}$  splňující  $\mathbf{A}\mathbf{z} = \mathbf{0}$ .

**Věta 2.3.** Soustava normálních rovnic  $\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$  má vždy alespoň jedno řešení.

## 2 Vícerozměrná lineární regrese

*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 dostáváme  $\langle \mathbf{X}^T \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{y}, \mathbf{X} \mathbf{z} \rangle = 0$ . Obecně totiž platí, že  $\langle \mathbf{X}, \mathbf{A} \mathbf{y} \rangle = \langle \mathbf{A}^T \mathbf{X}, \mathbf{y} \rangle$ .  $\square$

**Poznámka 2.4.** Z vět vyplývá, že MLE  $\hat{\beta}$  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 odhad  $\beta$ .

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

**Věta 2.5.** Matice  $\mathbf{X}^T \mathbf{X}$  je nesingulární 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.  $\square$

**Poznámka 2.6.** Pokud  $\mathbf{X}_{n \times (m+1)}$ ,  $n > m+1$ ,  $h(\mathbf{X}) = m+1$ ,  $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 I_m)$ , pak existuje jednoznačné řešení normálních rovnic  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ .

**Poznámka 2.7.** • 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 později.

## 2.1 Odhad parametrů

### 2.1.1 Odhad parametru $\sigma^2$

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

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

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

## 2 Vícerozměrná lineární regrese

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

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

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  $s_n^2, s_n$ .

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

**Věta 2.8.** Nechť  $\hat{\beta}$  je OLS odhad parametru  $\beta$  v modelu (\*\*), kde  $h(\mathbf{X}) = m + 1$  a  $e_1, \dots, e_n$  nezávislé,  $e_i \sim (0, \sigma^2)$ . Potom platí, že

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  $e_i \sim \mathcal{N}(0, \sigma^2), i \in \hat{n}$ , 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.

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

2. Označíme vektor  $\mathbf{Y}$  velikosti  $(n \times 1)$  jako náhodný vektor,  $\text{Cov}(\mathbf{Y}) = \Sigma$ . Pokud  $\mathbf{A}_{m,n}$  je matice, potom platí  $\text{Cov}(\mathbf{AY}) = \mathbf{A}\Sigma\mathbf{A}^T$ .

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

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

3. Nejdříve přepíšeme vektor reziduů  $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = \mathbf{Y} - \hat{\mathbf{Y}}$ . Pak

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{HY}, \quad (2.3)$$

kde  $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$  je tzv. **projekční matice**.

Pak  $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ . Dále platí

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

takže

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

Zřejmě pak

$$\begin{aligned} \mathbf{H}^T &= \mathbf{H}, \\ \mathbf{H}^2 &= [\mathbf{X} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] [\mathbf{X} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] = \mathbf{X} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \mathbf{H}, \\ (\mathbf{I}_n - \mathbf{H})^2 &= \mathbf{I}_n - \mathbf{H}. \end{aligned}$$

## 2 Vícerozměrná lineární regrese

$\mathbf{H}$  je tedy symetrická a idempotentní. Spočítáme dále SSE jako

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

kde  $g_{ij}$  je  $(i, j)$ -tý prvek matice  $(\mathbf{I}_n - \mathbf{H})$ . Zbývá spočítat  $\mathbb{E}(\text{SSE})$ :

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

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

4. Jelikož  $\hat{\beta}$  je LK  $Y_1, \dots, Y_n$ , které jsou nezávislé a normálně rozdělené

$$\Rightarrow \hat{\beta} \sim \mathcal{N}_{m+1}(\beta, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}).$$

□

POZNÁMKA 2.9. Vlastnosti projekční maticy:

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

**Věta 2.10.** Nechť  $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$  je LM (\*\*), kde  $h(\mathbf{X}) = m+1$  a  $\mathbf{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ . 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  $\mathbf{C} \in \mathbb{R}^{r, m+1}$  takové, že  $h(\mathbf{e}) = r$ . Potom kvadratická forma

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

Důkaz. 1. Rozepíšeme

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

a tedy  $\hat{\beta} - \beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{e}$ . Dále víme, že  $\hat{\mathbf{e}} = (\mathbf{I}_n - \mathbf{H})\mathbf{e}$  a vektor  $(\hat{\beta} - \beta, \hat{\mathbf{e}})^T$  lze zapsat jako

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

kde  $\mathbf{Z}$  je funkcií pouze  $(e_1, \dots, e_n) = \mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n) \Rightarrow \mathbf{Z}$  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)

## 2 Vícerozměrná lineární regrese

Poznámka 2.11.  $\mathbf{B}\mathbf{B}^T = \mathbf{I}_n - \mathbf{H}$  je singulární, protože  $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = 0$   
 Stačí nám tedy ukázat, že  $\text{Cov}(\hat{\beta}_i, \hat{e}_j) = 0$  pro  $i = 0, \dots, m$  a  $j \in \hat{n}$   
 spočtěme  $\text{Cov}(\mathbf{Z})$ :

$$\begin{aligned} \text{Cov}(\mathbf{Z}) &= \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} \text{Cov}(\mathbf{e}) (\mathbf{A}^T \mathbf{B}^T) = \sigma^2 \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} (\mathbf{A}^T \mathbf{B}^T) = \sigma^2 \begin{pmatrix} \mathbf{A}\mathbf{A}^T & \mathbf{A}\mathbf{B}^T \\ \mathbf{B}\mathbf{A}^T & \mathbf{B}\mathbf{B}^T \end{pmatrix} \\ &\left( \text{Cov}(\hat{\beta}_i, \hat{e}_j) \right)_{\substack{i=0, \dots, m \\ j \in \hat{n}}} = \mathbf{A}\mathbf{B}^T = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) = \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = 0 \end{aligned}$$

2. 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 jim 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 jsme ukázali, že

$$(n-m-1) \frac{s_n^2}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} = \frac{1}{\sigma^2} \mathbf{e}^T (\mathbf{I}_n - \mathbf{H}) \mathbf{e}.$$

Protože je  $(\mathbf{I}_n - \mathbf{H})$  symetrická a idempotentní, tak

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

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

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

**Věta 2.12.**  $\mathbf{V} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$  a  $\mathbf{Q}$  je ortogonální matice, potom  $\mathbf{Q}\mathbf{V} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$ .

To znamená, že  $\mathbf{q}$  je vektor nezávislých  $\mathcal{N}(0, \sigma^2)$  veličin  $(\mathbf{q} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n))$  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  $\mathcal{N}(0, 1)$  veličin.

3. Z předchozí věty víme, že

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

a z bodu 1) nezávislost

$$T_i = \frac{\hat{\beta}_i - \beta_i}{s_n \sqrt{v_i}} = \frac{\frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{v_i}}}{\frac{s_n}{\sigma}} \sim t(n-m-1) \tag{2.4}$$

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

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

Stačí tedy ukázat, že pokud  $\mathbf{Z} \sim \mathcal{N}_r(\mathbf{0}, \Sigma)$ , potom  $\mathbf{Z}^T\Sigma^{-1}\mathbf{Z} \sim \chi^2(r)$ .

Protože  $\Sigma$  je pozitivně definitní, existuje regulární matice  $\mathbf{R}$  taková, že  $\Sigma = \mathbf{R}\mathbf{R}^T$ . Definujme  $\mathbf{U} := \mathbf{R}^{-1}\mathbf{Z}$ . Potom  $\mathbb{E}\mathbf{U} = \mathbf{R}^{-1}\mathbb{E}[\mathbf{Z}] = \mathbf{0}$ .

Dále

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

tedy  $\mathbf{U} \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$ , takže složky  $\mathbf{U}$  jsou nezávislé  $\mathcal{N}(0, 1)$  rozdělené náhodné veličiny. Pak

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

a věta je dokázána pro  $\mathbf{Z} = \mathbf{C}(\hat{\beta} - \beta)$  a  $\Sigma = \sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T$ .

□

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

**Věta 2.13.** 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, že

1.  $\mathbb{E}[\hat{\mathbf{e}}] = \mathbf{0}$ ,
2.  $\text{Cov}(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$ ,
3. pokud navíc  $\mathbf{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ , potom  $\hat{\mathbf{e}} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$ ,
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.

$$\mathbb{E}[\hat{\mathbf{e}}] = (\mathbf{I}_n - \mathbf{H}) \cdot \mathbb{E}[\mathbf{e}] = (\mathbf{I}_n - \mathbf{H}) \cdot \mathbf{0} = \mathbf{0}$$

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  $\mathbf{e} \Rightarrow \hat{\mathbf{e}} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$

4. Soustava normálních rovnic  $\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{y}$  lze zapsat jako  $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}$ . Pro první rovnici musí platit

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

kde pro model s interceptem je  $x_{i1}$  vektor jedniček. Pro  $\hat{\beta}$  tedy platí

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

## 2 Vícerozměrná lineární regrese

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

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

a přenásobením zleva  $\hat{\beta}^T$  dostaneme

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

□

**Poznámka 2.14.** 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

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

## 2.2 Gauss - Markov theorem

Pro **normální chyby**, tj.  $e_i \text{ iid } \mathcal{N}(0, \sigma^2)$ , je OLS  $\hat{\beta}$  MLE, tzn. je eficientní MVVE parametr  $\beta$ .

Pro **chyby 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 2.15.** Nechť  $\beta$  je vektor regresních parametrů v lineárním modelu (LM). Řekneme, že  $\hat{\beta}$  je **lineární odhad**  $\beta$ , jestliže každý  $\beta_i$  je LK pozorování  $Y_i, i \in \hat{n}$ , tedy

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

V maticovém zápisu

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

pro  $i = 0, \dots, m$  a  $j \in \hat{n}$ .

**Poznámka 2.16.** 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 2.17** (Gauss-Markov). *Uvažujeme model  $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ , kde matice  $\mathbf{X}$  má plnou hodnost,  $e_i, i \in \hat{n}$  jsou nekorelované a  $e_i \sim \mathcal{N}(0, \sigma^2)$ . Potom OLS odhad  $\hat{\beta}$  je BLUE parametru  $\beta$  (best linear unbiased estimation).*

*Důkaz.* Nechť  $\hat{\beta} = \mathbf{A}\mathbf{Y}$  je lineární odhad  $\beta$ . Aby byl nestranný musí platit  $\mathbb{E}[\hat{\beta}] = \beta$ , tzn.

$$\mathbb{E}[\mathbf{A}\mathbf{Y}] = \mathbf{A}\mathbb{E}[\mathbf{Y}] = \mathbf{A}\mathbf{X}\beta \stackrel{!}{=} \beta,$$

## 2 Vícerozměrná lineární regrese

tedy  $(\mathbf{A}\mathbf{X} - \mathbf{I}_{m+1})\boldsymbol{\beta} = 0$ . Protože to musí platit  $\forall \boldsymbol{\beta} \in \mathbb{R}^{m+1}$ , dostáváme  $\mathbf{A}\mathbf{X} - \mathbf{I}_{m+1} = 0$ , nebo ekvivalentně  $\mathbf{A}\mathbf{X} = \mathbf{I}_{m+1}$ .

Spočteme kovarianční matici  $\hat{\boldsymbol{\beta}}$

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

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  $\hat{\boldsymbol{\beta}} = \mathbf{A}\mathbf{Y}$ , který minimalizuje rozptyl, musí platit  $\mathbf{D} = \mathbf{0}$ , bude věta dokázána.

Dosazením dostaneme:

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

a z podmínek nerovnosti

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

To znamená, že

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

a pro diagonální prvky platí, že

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

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

## 2.3 Testování modelu - tabulka ANOVA

### 2.3.1 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 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), jejímž jedním efektem jsou velké standardní chyby odhadů. To může vést k akceptování všech koeficientů jako 0, i když je model celkově významný (uvidíme na příkladu).

Bylo by dobré mít jednu statistiku pro test

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

## 2 Vícerozměrná lineární regrese

ANOVA přístup pro jednorozměrnou regresi naznačuje, že statistika

$$F = \frac{\frac{\text{SSR}}{m}}{s_n^2}$$

by mohla být užitečná (vyplýne i z obecnějších přístupů k testování později).

**Značení:**

Označíme  $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$  jako průměr j-tého sloupce matice  $\mathbf{X}$ ,

$$\bar{\mathbf{X}} = \begin{pmatrix} \bar{x}_0 & \bar{x}_1 & \cdots & \bar{x}_m \\ \vdots & \vdots & & \vdots \\ \bar{x}_0 & \bar{x}_1 & \cdots & \bar{x}_m \end{pmatrix}_{n \times m+1}$$

a  $(\mathbf{X}_c)_{ij}$  centrované matice regresorů, kde  $(\mathbf{X}_c)_{ij} = x_{ij} - \bar{x}_j$ ,  $i \in \hat{n}, j = 1, \dots, m$ .

**Věta 2.18.** *V modelu  $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ , kde  $e_i$  jsou nekorelované a  $e_i \sim \mathcal{N}(0, \sigma^2)$  pro  $i \in \hat{n}$  platí*

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

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

*Důkaz.*

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \sum_{j=1}^m x_{ij} \hat{\beta}_j \\ \frac{\partial \text{SSE}}{\partial \hat{\beta}_0} &= -2 \sum_{i=1}^n \left( y_i - (\hat{\beta}_0 + \sum_{j=1}^m x_{ij} \hat{\beta}_j) \right) = 0 \Rightarrow \hat{\beta}_0 = \bar{y} - \sum_{j=1}^m \bar{x}_j \hat{\beta}_j \end{aligned}$$

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

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

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

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

**Věta 2.19.** *Nechť  $Z = \mathbf{Y}^T \mathbf{A} \mathbf{Y}$  je kvadratická forma a nechť  $\mathbb{E} \mathbf{Y} = \mu$  a  $\text{Cov} \mathbf{Y} = \Sigma$ . Potom platí, že*

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

## 2 Vícerozměrná lineární regrese

Nejdříve zjednodušíme matici  $\mathbf{A}$  :

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

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

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

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

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

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

**Poznámka 2.20.** Pokud  $\boldsymbol{\beta}_s = 0$ , potom  $\mathbb{E} \left( \frac{\text{SSR}}{m} \right) = \sigma^2 = \mathbb{E} s_n^2$ , takže  $\boldsymbol{\beta}_s \neq 0$  implikuje, že  $\mathbb{E} \left( \frac{\text{SSR}}{m} \right) > \sigma^2$ , tedy velké hodnoty  $F = \frac{\text{SSR}/m}{s_n^2}$  budou znamenat zamítnutí  $H_0 : \boldsymbol{\beta}_s = 0$ . Budeme proto potřebovat rozdělení  $F$  za platnosti  $H_0$ .

**Věta 2.21.** Nechť v modelu  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$  (\*\*\*) jsou  $e_1, \dots, e_n$  iid  $\mathcal{N}(0, \sigma^2)$ . Pokud  $\boldsymbol{\beta}_s = 0$ , tj.  $\beta_1 = \beta_2 = \dots = \beta_m = 0$ , potom

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

*Důkaz.* V důkazu minulé věty jsme ukázali

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

a potřebujeme rozepsat  $\hat{\mathbf{Y}}$ , použijeme rovnost (2.3):

$$\hat{\mathbf{Y}} = \mathbf{HY} = \mathbf{H}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = \mathbf{H}(\mathbf{1}\beta_0 + \mathbf{X}_v \boldsymbol{\beta}_s + \mathbf{e}) = \beta_0 \underbrace{\mathbf{H}\mathbf{1}}_{=\mathbf{1}} + \mathbf{H}\mathbf{X}_v \underbrace{\boldsymbol{\beta}_s}_{=\mathbf{0}} + \mathbf{He} = \beta_0 \mathbf{1} + \mathbf{He}$$

$$\text{SSR} = (\beta_0 \mathbf{1}^T + \mathbf{e}^T \mathbf{H}) \left( \mathbf{I}_n - \frac{1}{n} \mathbf{B} \right) (\beta_0 \mathbf{1} + \mathbf{He}) = \mathbf{e}^T \mathbf{H} \left( \mathbf{I}_n - \frac{1}{n} \mathbf{B} \right) \mathbf{He},$$

## 2 Vícerozměrná lineární regrese

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

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

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

Pro matici  $\mathbf{C}$  platí

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

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

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

Z věty o spektrálním rozkladu plyne existence  $\mathbf{Q}$  OG a diagonální  $\mathbf{\Lambda}$  tak, že

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

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

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

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

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

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

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

kde  $\mathbf{w} = (\mathbf{I}_n - \frac{1}{n}\mathbf{B}) \mathbf{He} \equiv \mathbf{Ke}$ ,  $\hat{\mathbf{e}} = (\mathbf{I}_n - \mathbf{H})\mathbf{e} \equiv \mathbf{Le}$ . Stačí tedy ukázat, že  $\mathbf{w}$  a  $\hat{\mathbf{e}}$  jsou nezávislé vektory. Víme, že

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

## 2 Vícerozměrná lineární regrese

tzn. má vícerozměrné normální rozdělení. Rozepíšeme kovarianci jako

$$\mathbb{C}\text{ov} \begin{pmatrix} \mathbf{w} \\ \hat{\mathbf{e}} \end{pmatrix} = \begin{pmatrix} \mathbf{K} \\ \mathbf{L} \end{pmatrix} \mathbb{C}\text{ov}(\mathbf{e})(\mathbf{K}^T, \mathbf{L}^T) = \sigma^2 \begin{pmatrix} \mathbf{K}\mathbf{K}^T & \mathbf{K}\mathbf{L}^T \\ \mathbf{L}\mathbf{K}^T & \mathbf{L}\mathbf{L}^T \end{pmatrix}. \quad (2.5)$$

Pokud je výraz  $\mathbf{KL}^T$  z rovnice (2.5) roven nule, pak jsou  $\mathbf{w}$  a  $\hat{\mathbf{e}}$  nezávislé. Pro  $\mathbf{KL}^T$  platí, že

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

□

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

POZNÁMKA 2.22. 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		
		$\mathcal{R}^2$	$\bar{\mathcal{R}}^2$	

### 2.3.2 Koeficient (vícenásobné) determinace $\mathcal{R}^2$

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

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

protože

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

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

POZNÁMKA 2.23.  $\mathcal{R}^2$  je možno zvětšovat přidáváním nových proměnných  $x$ , i když jsou statisticky nevýznamné. (Pro  $n$  LN proměnných  $x$  a  $n$  pozorování dostaneme „perfect fit“, tedy přeúčtení.) Vysvětlení:

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

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

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

S rostoucím  $m$  klesá SSE, ale i  $n - m - 1$ , což dává určitou kompenzacii.

## 2.4 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, viz (2.4):

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

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

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

S jejich pomocí lze odvodit kritický obor pro t-test

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

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

- V případě kolinearity může být model 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.

**Poznámka 2.25.** Statistiky  $F$ ,  $\mathcal{R}^2$  a  $T$  jsou užitečné pro rozkrytí efektů jednoduchých proměnných, nemohou být ale používány úplně automaticky.

## 2.5 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

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

## 2 Vícerozměrná lineární regrese

Poznámka 2.26. Jak volit  $\mathbf{b}$  a  $\mathbf{C}$ :

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

$$H_0 : \beta_j = 0.$$

c) V modelu  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$  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  $\mathbf{C}\boldsymbol{\beta}$  a **redukovaný model** (reduced model) za předpokladu, že platí  $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b}$ .

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

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

**Příklad 2.27.** 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

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

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

**Věta 2.28.** Nechť v modelu (\*\*\*) platí, že  $e_1, \dots, e_n$  jsou nezávislé a  $e_i \sim \mathcal{N}(0, \sigma^2)$ . Označme  $SSE_F$  reziduální s.č. plného modelu a  $SSE_R$  reziduální s.č. modelu, kde platí  $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b}$ . Potom je za platnosti  $H_0$  splněno

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

*Důkaz.* Nejdříve si dokážeme následující lemma:

**Lemma 2.29.** Označme  $\hat{\boldsymbol{\beta}}_F$  a  $\hat{\boldsymbol{\beta}}_R$  LSE parametru  $\boldsymbol{\beta}$  v plném a redukovaném modelu. Potom platí

1.  $\hat{\boldsymbol{\beta}}_F = \hat{\boldsymbol{\beta}}_R - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \mathbf{A} (\mathbf{C} \hat{\boldsymbol{\beta}}_F - \mathbf{b})$ , kde  $\mathbf{A} = (\mathbf{C}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T)^{-1}$
2.  $\Delta SSE = SSE_R - SSE_F = (\mathbf{C} \hat{\boldsymbol{\beta}}_F - \mathbf{b})^T \mathbf{A} (\mathbf{C} \hat{\boldsymbol{\beta}}_F - \mathbf{b})$ .

## 2 Vícerozměrná lineární regrese

*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  $\mathbf{C}\beta = \mathbf{b}$ . Sestavíme Lagrangeovu funkci

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

a tedy

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

Z první rovnice dostáváme

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

a dosadíme do druhé

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

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

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

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

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

Dosadíme  $\beta = \hat{\beta}_R$ :

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

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

□

Vrátíme se tedy k původnímu důkazu věty. Nejdříve ukážeme, že  $\frac{\Delta SSE}{\sigma^2} \sim \chi^2(r)$  za platnosti  $H_0 : \mathbf{C}\beta = \mathbf{b}$ .

Za  $H_0 : \mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta_R, \sigma^2 \mathbf{I}_n)$  a  $\hat{\beta}_F = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ , tzn.

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

## 2 Vícerozměrná lineární regrese

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

Tedy

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

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

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

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

□

POZNÁMKA 2.30. Použitím rozkladu  $\text{SST} = \text{SSE} + \text{SSR}$  dostaneme

$$\Delta\text{SSE} = \text{SSR}_F - \text{SSR}_R.$$

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

$$\text{SSR}_F = \text{SSR}_R + \Delta\text{SSE},$$

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

Např., pokud  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{m-1}, 0)$ , tzn.  $\beta_m = 0$  a skutečný model má  $\boldsymbol{\beta} = \boldsymbol{\beta}_F$ , potom  $\Delta\text{SSE}$  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 2.31. Joint confidence region viz. Ex 5.30 str. 239

## 2.6 Predikce

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

### a) predikce $\mathbb{E}[Y_{\mathbf{x}_0}]$

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

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

tzn.  $D[\hat{Y}_{\mathbf{x}_0}] = \mathbf{x}_0^T \cdot D[\hat{\boldsymbol{\beta}}] \cdot \mathbf{x}_0 = \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$  a může být odhadnut pomocí

$$\hat{\sigma}^2(\hat{Y}_{\mathbf{x}_0}) = s_n^2 [\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0] \quad (\text{rozptyl predikce}).$$

## 2 Vícerozměrná lineární regrese

Speciálně pokud  $\mathbf{x}_0^T = \mathbf{x}_i^T$  ( $i$ -tý řádek matice  $\mathbf{X}$ ), tak

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

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

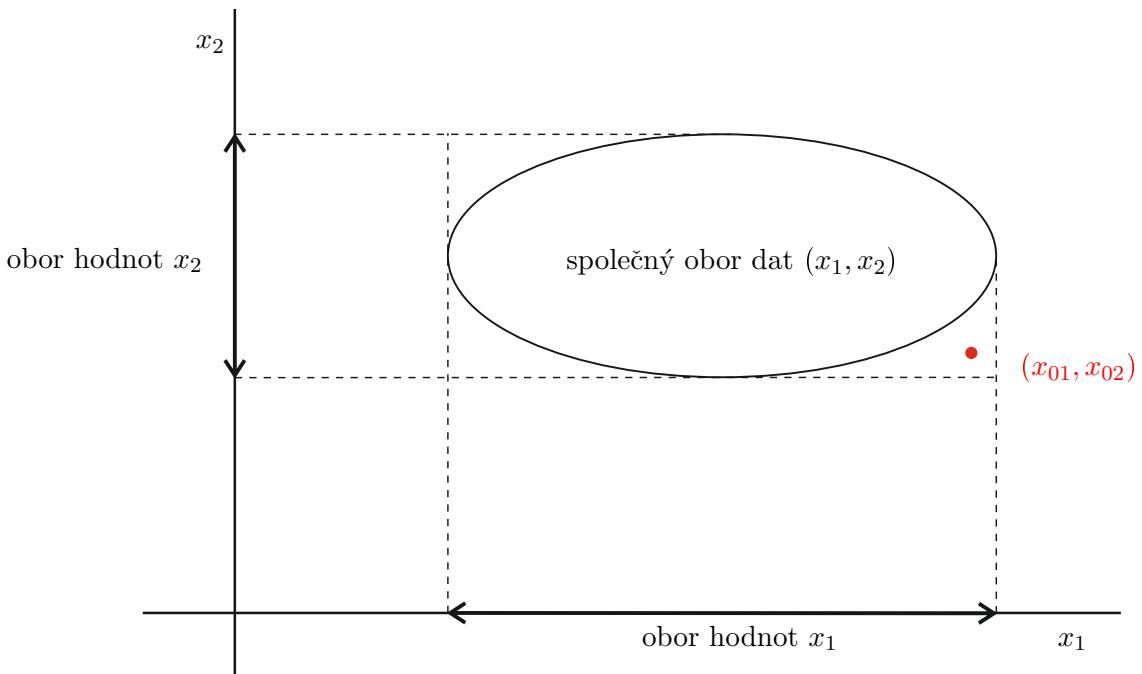
$$\frac{\hat{Y}_{\mathbf{x}_0} - \mu_{\mathbf{x}_0}}{\sigma \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim \mathcal{N}(0, 1)$$

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

$$\frac{\hat{Y}_{\mathbf{x}_0} - \mu_{\mathbf{x}_0}}{s_n \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim t(n - m - 1)$$

čímž získáme  $100(1 - \alpha)\%$  interval spolehlivosti pro  $\mu_{\mathbf{x}_0}$

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



Obrázek 2.1:  $(x_{01}, x_{02})$  leží uvnitř oboru hodnot pro obě  $x_1$  i  $x_2$  ale vně společného oboru původních dat.

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

## 2 Vícerozměrná lineární regrese

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

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

takže

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

za předpokladu normality chyb.  
100(1 -  $\alpha$ )% IP pro  $Y_{\mathbf{x}_0}$  tedy je

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

**POZNÁMKA 2.32.** Extrapolace

- U jednoduché LR kvalita 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  $\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = h_{ii}$  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.

# 3 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 Rezidua

Připomenutí: víme už, že platí následující vztahy:

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

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

Dále jsme ukázali

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

Pokud navíc  $\mathbf{e} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$  potom  $\hat{\mathbf{e}} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$ . Když označíme  $h_{ii} = \mathbf{H}_{ii}$ , pak  $\hat{e}_i \sim \mathcal{N}_n(0, \sigma^2(1 - h_{ii}))$  a  $\text{Cov}(\hat{e}_i, \hat{e}_j) = -\sigma^2 h_{ij}$ .

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

$$\hat{r}_i = \frac{\hat{e}_i}{s_n \sqrt{1 - h_{ii}}}, \quad \text{kde } i \in \hat{n},$$

kterým se říká interně studentizovaná rezidua (někdy také standardizovaná rezidua). V R se jedná o funkci `rstandard()`.

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

$$\hat{t}_i = \frac{\hat{e}_i}{\hat{\sigma}_{(-i)}^2 \sqrt{1 - h_{ii}}}, \quad \text{kde } i \in \hat{n}.$$

Říká se jim externě studentizovaná rezidua (někdy také studentizovaná rezidua). V R se jedná o funkci `rstudent()`.

Například  $\hat{\sigma}_{(-i)}^2 = \frac{1}{n-m-2} \text{SSE}_{(-i)}$  je nestranný odhad  $\sigma^2$  v modelu bez  $i$ -tého pozorování.

**POZNÁMKA 3.1.** Platí:

- Pokud je  $h_{ii}$  malé, pro velké  $n$  by se měly  $\hat{e}_i, \hat{r}_i, \hat{t}_i$  chovat přibližně stejně a  $\hat{r}_i, \hat{t}_i \approx \mathcal{N}(0, 1)$ .

### 3 Rezidua, diagnostika a transformace

- Pro malé  $n$  ( $n < 20$ ) a/nebo  $h_{ii} \approx 1$  je preferováno použít  $\hat{r}_i$  nebo  $\hat{t}_i$ . 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).
- V anglické literatuře se označuje  $h_{ii}$  jako **leverage** – potenciál  $i$ -tého pozorování (leverage point = píkový bod / vzdálený bod).  $h_{ii}$  hraje zásadní roli v diagnostice modelu, proto ted' probereme jeho základní vlastnosti.

#### 3.1.1 Vlastnosti potenciálu $h_{ii}$

- $D[\hat{e}_i] = \sigma^2(1 - h_{ii}) \geq 0 \Rightarrow h_{ii} \leq 1$ .
- $\mathbf{H}^2 = \mathbf{H} \Rightarrow h_{ii} = \sum_{j=1}^n h_{ij}h_{ji} = \sum_{j=1}^n (h_{ij})^2$ , tedy  $h_{ii} > 0$ . Dá se ukázat i silnější tvrzení:  $h_{ii} \geq \frac{1}{n}$ .
- $\mathbf{H}\mathbf{X} = \mathbf{X} \Rightarrow \sum_{j=1}^n h_{ij}x_{j1} = \sum_{j=1}^n h_{ij} = x_{i1} = 1$  tedy  $\sum_{j=1}^n h_{ij} = 1 \quad \forall j \in \hat{n}$  (v modelu s interceptem).
- 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  $h_{ii} \approx 1$ , pak  $\hat{y}_i \approx y_i$  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.
- Otázka je, jaká hodnota  $h_{ii}$  je „velká“.

#### Heuristické pravidlo

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

## 3.2 Grafy reziduí

### A) Ověření normality – histogramy, Q-Q plots

Tyto obrázky nezávisí na počtu nezávislých proměnných  $x$ , vše stejně jako v jednoduché LR. Můžeme použít testy normality jako např. Shapiro-Wilk, Anderson-Darling a další.

- B) Pro ověření funkční formy pro  $\mathbb{E}[Y_x]$  a/nebo konstantního rozptylu se nejčastěji používají:
- 1) 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  $\mathbf{X}$ ,
  - 2) grafy  $\hat{e}_i, \hat{r}_i$  nebo  $\hat{t}_i$  oproti  $\hat{y}_i$ ,
  - 3) partial residual plots.

Mezi testy konstantního rozptylu patří např. Breuch-Pagan nebo Levene test.

Poznámka 3.2. Zdůvodnění:

### 3 Rezidua, diagnostika a transformace

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

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

Pokud tedy naladíme LR model bez interceptu pro  $\hat{\mathbf{e}}$  v závislosti na  $\mathbf{x}_j^c$ , odhad 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  $\approx \pm 2$ ). 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

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

Body by opět měly být náhodně rozptýlené kolem osy  $x$ . Případný trychtířovitý tvar indikuje nekonstantní rozptyl, trendy pak indikují nelinearitu.

#### 3.2.1 Partial residual plot

- I když grafy  $\hat{\mathbf{e}}$  oproti  $\mathbf{x}_j^c$  a  $\hat{\mathbf{y}}$  mohou indikovat nedostatky modelu, nemusí být zřejmé, jaké tyto nedostatky jsou.
- V SLR lze graf  $\hat{e}_i$  oproti  $x_i$  použít pro detekci nelinearity.
- V MLR mohou být tyto grafy (stejně jako scatterploty) 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 3.3.** Definujme

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

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

**Partial residual plot** (PRP) je graf  $\hat{\mathbf{e}}_j^*$  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  $\hat{\mathbf{e}}_j^*$  a  $\mathbf{x}_j^c$  má formu SLR bez interceptu. Pokud je model správný,  $\hat{e}_i, i \in \hat{n}$ , splňuje podmítku

$$\mathbb{E}\hat{e}_i = 0 \quad \text{a} \quad D\hat{e}_i = \sigma^2(1 - h_{ii}).$$

Má tedy smysl uvažovat regresní model pro  $\hat{\mathbf{e}}_j^*$  oproti  $\mathbf{x}_j^c$  ( $\hat{e}_j^* = \gamma_j \mathbf{x}_j^c + \mathbf{e}$ ).

### 3 Rezidua, diagnostika a transformace

Pro odhad koeficientu platí, že

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

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

**Poznámka 3.4.** Partial residual ploty jsou někdy kritizovány za nadhodnocování efektu  $\mathbf{x}_j^c$ . Alternativou mohou být **partial regression plots** (added variable plots).

#### 3.2.2 Partial regression plot

**Motivace:** Ptáme se, zda přidat novou proměnnou do modelu a chtěli bychom odhadnout 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 formuli pro  $\hat{\gamma}$

$$\hat{\gamma} = \frac{\hat{\mathbf{e}}^T (\mathbf{I} - \mathbf{H}) \mathbf{w}}{\|(\mathbf{I} - \mathbf{H}) \mathbf{w}\|^2}, \quad (3.1)$$

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

Ted' naopak uvažujme, že  $\mathbf{w}$  je sloupec původní  $\mathbf{X}$ , např.  $\mathbf{x}_j^c$  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  $\mathbf{x}_j^c$ ,  $\mathbf{x}_{j,res}^c = (\mathbf{I} - \mathbf{H})\mathbf{x}_j^c$ . Jedná se tedy o rezidua modelu pro  $\mathbf{x}_j^c$  v závislosti na ostatních proměnných, tedy  $\mathbf{X}_{(-j)}$  (v  $\mathbf{x}_{j,res}^c$  je odstraněn efekt ostatních regresorů).

$\hat{\beta}_j$  je směrnice regresního modelu pro  $\hat{\mathbf{e}}_{(-j)}$  v závislosti na  $\mathbf{x}_{j,res}^c$

$\Rightarrow$  **added variable plot:** graf  $\hat{\mathbf{e}}_{(-j)}$  proti  $\mathbf{x}_{j,res}^c, j = 1, \dots, m$ . V R se jedná o funkci `avPlots()` z knihovny `car`.

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 3.5.** Ze vztahu (3.1) je vidět, že MLR může být chápána jako posloupnost SLR, kde postupně vytváříme modely pro novou proměnnou s použitím reziduí modelu pro předcházející proměnné.

### 3.3 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 reziduí pro predikce hodnot v nových bodech  $\mathbf{x}$ , obecně ale nemáme data  $y$  v těchto bodech. Jedna možnost je tak použít data, která máme k dispozici.

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

**Definice 3.6.** Předpokládáme, že vymečáme  $i$ -té pozorování. Ozn.  $\hat{\beta}_{(-i)}$  odhad  $\beta$  v modelu s vymechaným  $i$ -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

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

nazýváme  $i$ -té **PRESS reziduum**.

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

**POZNÁMKA 3.7.** Otázka je, jak počítat  $\hat{e}_{(-i)}$ ,  $i \in \hat{n}$ . Pro velké  $n$  se zdá, že to bude náročný problém, protože pro každé  $i \in \hat{n}$  musíme nalaďit nový model. 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.

Zavedeme následující značení:

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

**Věta 3.8.** Jestliže  $h_{ii} \neq 1$ , potom

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

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

*Důkaz.* Nejdříve ukážeme

$$\begin{aligned} \mathbf{X}^T \mathbf{X} &= \mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)} + \mathbf{x}_i \mathbf{x}_i^T \\ \text{s rozměry } (m+1) \times (m+1) &= (m+1) \times (m+1) + (m+1) \times (m+1). \end{aligned} \tag{3.2}$$

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

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

### 3 Rezidua, diagnostika a transformace

$i, j$ -tý prvek  $\mathbf{X}_{(-n)}^T \mathbf{X}_{(-n)}$  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. (3.2) platí.

**Věta 3.9** (Sherman-Morrison-Woodbury (z LA)). *Nechť  $\mathbf{A}$  je  $n \times n$  invertibilní matici a nechť  $\mathbf{z}$  je  $n \times 1$  sloupcový vektor. Jestliže  $\mathbf{z}^T \mathbf{A}^{-1} \mathbf{z} \neq 1$ , potom matici  $\mathbf{B} = \mathbf{A} - \mathbf{z} \mathbf{z}^T$  je invertibilní a platí*

$$\mathbf{B}^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1} \mathbf{z} \mathbf{z}^T \mathbf{A}^{-1}}{1 - \mathbf{z}^T \mathbf{A}^{-1} \mathbf{z}}. \quad (3.3)$$

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

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

Užitím věty a dosazením do (3.3) 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 3.10.** Nechť  $\hat{e}_{(-i)}$  je  $i$ -té PRESS reziduum. Potom

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

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

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

kde  $\mathbf{y}_{(-i)}$  je  $\mathbf{y}$  bez  $i$ -té složky  $y_i$ . To znamená, že

$$\begin{aligned} \hat{y}_{(-i)} &= \mathbf{x}_i^T \hat{\beta}_{(-i)} = \mathbf{x}_i^T (\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} = \\ &= \left[ (\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1} = (\mathbf{X}^T \mathbf{X})^{-1} + \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1}}{1 - h_{ii}} \text{ viz věta 3.8} \right] = \\ &= \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} + \frac{1}{1 - h_{ii}} \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)} = \\ &= S_1 + \frac{1}{1 - h_{ii}} S_2. \end{aligned}$$

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

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

Podobně

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

### 3 Rezidua, diagnostika a transformace

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  $\hat{\beta} - \hat{\beta}_{(-i)}$  a  $\text{SSE}_{(-i)}$ .

**Věta 3.11.** 1) Nechť  $\hat{\beta}_{(-i)}$  značí LSE parametru  $\beta$  v modelu bez  $i$ -tého pozorování. Potom platí

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

2) Pro součet residualních čtverců  $\text{SSE}_{(-i)}$  v modelu bez  $i$ -tého pozorování platí

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

*Důkaz.* 1) Stejně jako v důkazu předchozí věty 3.10 platí, že

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

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

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

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

2)

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

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

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

### 3 Rezidua, diagnostika a transformace

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

$$\begin{aligned} \text{SSE}_{(-i)} &= \sum_{j=1}^n \left( \hat{e}_j + \frac{h_{ij}\hat{e}_i}{1-h_{ii}} \right)^2 - \left( \hat{e}_i + \frac{h_{ii}\hat{e}_i}{1-h_{ii}} \right)^2 = \underbrace{\sum_{j=1}^n \left( \hat{e}_j + \frac{h_{ij}\hat{e}_i}{1-h_{ii}} \right)^2}_A - \frac{\hat{e}_i^2}{(1-h_{ii})^2}, \\ A &= \sum_{j=1}^n \hat{e}_j^2 + \frac{2\hat{e}_i}{1-h_{ii}} \underbrace{\sum_{j=1}^n h_{ij}\hat{e}_j}_0 + \frac{\hat{e}_i^2}{(1-h_{ii})^2} \underbrace{\sum_{j=1}^n h_{ij}^2}_{h_{ii}}. \end{aligned}$$

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

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

□

**Důsledek 3.12.** V modelu (\*\*) s  $m+1$  parametry  $\beta$  a bez  $i$ -tého pozorování platí, že

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

takže

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

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

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

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

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

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

□

**Poznámka 3.13.** Dá se ukázat, že  $\text{SSE}_{(-i)}$  a  $\hat{e}_i$  jsou nezávislé náhodné veličiny. Protože  $\frac{\text{SSE}_{(-i)}}{\sigma^2} \sim \chi^2(n-m-2)$  a  $\frac{\hat{e}_i}{\sigma\sqrt{1-h_{ii}}} \sim \mathcal{N}(0, 1)$ , dostaneme  $\frac{\hat{e}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}} \sim t(n-m-2)$ .

**Tvrzení 3.14.** 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 (externě) studentizované reziduum

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

### 3 Rezidua, diagnostika a transformace

POZNÁMKA 3.15.  $\hat{t}_i$  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

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

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

POZNÁMKA 3.16. Vztah  $\hat{e}_{(-i)}$  a  $\hat{t}_i$ :

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

Standardizované PRESS reziduum

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

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

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

POZNÁMKA 3.17.  $\hat{e}_{(-i)} = \frac{\hat{e}_i}{1-h_{ii}}$ , a proto, pokud  $i$ -té pozorování má velký potenciál  $h_{ii}$ , bude  $\hat{e}_{(-i)}$  mnohem větší, než  $\hat{e}_i$ , pozorování s velkým  $h_{ii}$  jsou dobře modelována, ale měřeno  $\hat{e}_{(-i)}$  mohou špatně predikovat. To je další ukázka fit/prediction dilema.

Stejný efekt nastává také pro

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

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

## 3.4 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, která nemusí být možné rozdělit na trénovací a validační/testovací. Bude nás proto zajímat, jaký vliv má  $i$ -té pozorování ( $i$ -tý řádek matice  $\mathbf{X}$ ) na model.

Už víme, že velké  $h_{ii}$  indikuje, že  $i$ -té pozorování má velký vliv, a velká rezidua naznačují možnou neadekvátnost modelu. Zavedeme míry, které budou oba dva faktory kombinovat. Využijeme k tomu přístup z PRESS reziduů, tzn. budeme sledovat, jak velký vliv má vynechání  $i$ -tého pozorování na  $\hat{\beta}$  a  $\hat{\mathbf{y}}$ .

### 3.4.1 DFBETAS

Vliv vynechání  $i$ -tého pozorování na odhad  $\hat{\beta}$  měří rozdíl

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

který bude základem pro naši analýzu.

**a) vliv  $i$ -tého pozorování na  $\hat{\beta}_j$  :**

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

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

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

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 ovšem velké množství hodnot pro srovnání – celkem  $(m + 1) \times n$ . Proto tuto metodu zjednodušíme.

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

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

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

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

Dosazením dostaneme

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

Výpočetní formule je potom ve tvaru

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

### 3 Rezidua, diagnostika a transformace

POZNÁMKA 3.18.  $100(1 - \alpha)\%$  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.

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

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

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

Pro většinu  $m, n$  je  $F_{\frac{1}{2}} \approx 1$ , pravidlo tak lze zjednodušit na  $D_i > 1$ .

POZNÁMKA 3.19. Také platí, že

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

tzn.  $D_i$  se dá chápout jako míra influence na celkovou predikci.

#### 3.4.2 DFFITS

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

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

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

POZNÁMKA 3.20. Míry influence v R:

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

**Shrnutí používaných pravidel:**  $i$ -té pozorování je influenční, pokud

$$|\text{DFBETAS}_{ij}| > 1, \quad |\text{DFFITS}_i| > 3\sqrt{\frac{m+1}{n-m-1}},$$

$$D_i > F_{\frac{1}{2}}(m+1, n-m-1), \quad h_{ii} > 3\frac{m+1}{n}.$$

## 3.5 Transformace

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

### 3.5.1 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 ( $\Rightarrow$  restrikční podmínky na parametr  $\beta$ ).
- Lze obejít modelování  $y^* = \log FEV$ .

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 jen 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  $D[Y] = \varphi \mathbb{E}[Y]^2 = \varphi \mu^2$  pro nějaké  $\varphi$ .

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

Transformace:	$\leftarrow$	...	$y^3$	$y^2$	$y$	$\sqrt{y}$	$\log y$	$\frac{1}{\sqrt{y}}$	$\frac{1}{y}$	$\frac{1}{y^2}$	...	$\rightarrow$
Box-Cox $\lambda$ :	$\leftarrow$	...	3	2	1	$\frac{1}{2}$	0	$-\frac{1}{\sqrt{2}}$	-1	-2	...	$\rightarrow$

- Pokud  $D[Y]$  klesá s rostoucí  $\mathbb{E}[Y]$ , budeme klesat s mocninou  $\lambda$ .
- Pokud  $D[Y]$  roste s rostoucí  $\mathbb{E}[Y]$ , budeme  $\lambda$  snižovat.

OBECNĚ: Předpokládejme vztah  $D[Y] = \varphi V(\mu)$  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  $(D[Y]^{-1/2} = V^{-1/2}(\mu))$ .

### 3 Rezidua, diagnostika a transformace

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

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

- Asi nejvíce užívanou transformací je  $y^* = \log(y)$ . Jedním z důvodů je i dobrá interpretabilitnost parametru  $\beta$ .

Interpretace parametrů LM

1. Klasický LM:

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

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  $\log Y$ :

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

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

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

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

Pak 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  $x_j$

#### 3.5.2 Box-Cox transformace

- Pokud chyby nemají normální rozdělení, hledáme transformaci  $Y$ , která by nejenom linearizovala model, ale také transformovala chyby, aby byly přibližně normální.
- Jako užitečná se ukazuje následující třída transformací (power family):

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{pokud } \lambda \neq 0 \\ \log y, & \text{pokud } \lambda = 0 \end{cases},$$

které předpokládají, že data  $y$  jsou pouze kladná. (Pokud ne, můžeme přičíst konstantu ke všem pozorováním a analyzovat takto posunutá data.)

### 3 Rezidua, diagnostika a transformace

POZNÁMKA 3.21.  $\lim_{\lambda \rightarrow 0} \frac{y^{\lambda}-1}{\lambda} = \log y$

- Pro nalezení vlastního  $\lambda$  budeme předpokládat, že transformované veličiny  $Y_i^{(\lambda)}, i \in \hat{n}$ , splňují postačující podmínky RM, tj.

$$Y_i^{(\lambda)} = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{e} \quad \text{kde } \mathbf{e} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n) \quad \left( Y_i^{(\lambda)} \sim \mathcal{N}(\mathbf{X}_i^T \boldsymbol{\beta}, \sigma^2) \right)$$

Úkol je odhadnout zároveň  $\lambda, \boldsymbol{\beta}, \sigma^2$ , použijeme MLE. Pomocí transformace získáme hustotu

$$f_{Y_i}(y_i) = f_{Y_i^{(\lambda)}}(y_i^{(\lambda)}) \cdot \frac{dy_i^{(\lambda)}}{dy_i} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i^{(\lambda)} - \mu_i)^2} \cdot y_i^{\lambda-1}, \quad \text{kde } \mu_i = \mathbb{E}[Y_i^{(\lambda)}] = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Věrohodnostní funkce pro pozorování  $y_1, \dots, y_n$  bude mít tvar

$$L = \prod_{i=1}^n f_{Y_i}(y_i) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^2 e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i^{(\lambda)} - \mu_i)^2} \cdot J(\lambda), \quad \text{kde } J(\lambda) = \prod_{i=1}^n y_i^{\lambda-1} = \left( \prod_{i=1}^n y_i \right)^{\lambda-1}$$

Dále vyjádříme log-likelihood:

$$l = \ln L = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i^{(\lambda)} - \mu_i)^2 + \ln J(\lambda).$$

Věrohodnostní rovnice nemají explicitní analytické řešení. Pro nalezení MLE si všimneme, že pro pevné  $\lambda$  je  $l$  proporcionální logaritmus věrohodnosti pro odhad  $(\boldsymbol{\beta}, \sigma^2)$  na základě  $\mathbf{y}^{(\lambda)} = (y_1^{(\lambda)}, \dots, y_n^{(\lambda)})^T$ , tedy

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\lambda) &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}^{(\lambda)} \\ \hat{\sigma}^2(\lambda) &= \frac{1}{n} \sum_{i=1}^n (y_i^{(\lambda)} - \hat{y}_i^{(\lambda)})^2 = \frac{1}{n} (\mathbf{y}^{(\lambda)})^T (\mathbf{I}_n - \mathbf{H}) \mathbf{y}^{(\lambda)}, \quad \text{kde } \hat{y}_i^{(\lambda)} = \mathbf{x}_i^T \boldsymbol{\beta}^{(\lambda)}. \end{aligned}$$

Dosazením do  $l$  dostaneme její hodnotu maximalizovanou vzhledem k  $(\boldsymbol{\beta}, \sigma^2)$ , tzv. profite log-likelihood

$$l_p^{(\lambda)} = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \hat{\sigma}^2(\lambda) - \frac{n}{2} + \ln J(\lambda) = C - \frac{n}{2} \ln \hat{\sigma}^2(\lambda) + (\lambda - 1) \sum_{i=1}^n \ln y_i.$$

POZNÁMKA 3.22. Kvůli komplikované závislosti  $l_p$  na  $\lambda$  bude třeba numerická metoda pro maximalizaci. Lze přepsat do tvaru, kde bude možné využít metody LR:

$$\begin{aligned} l_p(\lambda) &= C - \frac{n}{2} \ln \hat{\sigma}^2(\lambda) - \frac{n}{2} \ln J(\lambda)^{2/n} = C - \frac{n}{2} \ln \frac{\hat{\sigma}^2(\lambda)}{(J^{\frac{1}{n}}(\lambda))^2} \\ J^{\frac{1}{n}}(\lambda) &= \left[ \left( \prod_{i=1}^n y_i \right)^{\frac{1}{n}} \right]^{\lambda-1} = (\bar{y})^{\lambda-1}, \quad \text{kde } \bar{y} \text{ je geometrický průměr}. \end{aligned}$$

Dosazením zpátky do  $l_p(\lambda)$  dostáváme

$$\begin{aligned} l_p(\lambda) &= C - \frac{n}{2} \ln \frac{\hat{\sigma}^2(\lambda)}{(\bar{y})^{\lambda-1}} = C - \frac{n}{2} \ln s_\lambda^2, \\ \text{kde } s_\lambda^2 &= \frac{\hat{\sigma}^2(\lambda)}{(\bar{y})^{\lambda-1}} = \frac{1}{n} \sum_{i=1}^n \left( \frac{y_i^{(\lambda)}}{(\bar{y})^{\lambda-1}} - \frac{\hat{y}_i^{(\lambda)}}{(\bar{y})^{\lambda-1}} \right)^2. \end{aligned}$$

### 3 Rezidua, diagnostika a transformace

Tedy  $s_\lambda^2$  je reziduální součet čtverců (SSE) v modelu  $\frac{y_i^{(\lambda)}}{(y)^{\lambda-1}}$  v závislosti na  $\mathbf{x}_i^T$  (tzn.  $s_\lambda^2$  lze snadno získat pomocí funkce  $\ln()$ ). Celkem:  $\max l_p(\lambda) \Leftrightarrow \min s_\lambda^2$ .

#### Algoritmus pro hledání vhodného $\lambda$ :

1. Zvolit oblast hodnot  $\lambda$ ,  $I = [\lambda_{min}, \lambda_{max}]$ , a body  $\lambda \in I$  (typicky  $I = [-2, 2]$  a 10-20 rovnoramě rozdělených bodů).
2. Naladit model  $\frac{y^{(\lambda)}}{(y)^{\lambda-1}} \sim \chi$  a spočítat  $\frac{1}{n} \text{SSE} = s_\lambda^2$ .
3. Z grafu  $(\lambda, s_\lambda^2)$  vybrat  $\hat{\lambda}$ , které minimalizuje  $s_\lambda^2$ .
4. Pro zvolené  $\hat{\lambda}$  naladit model  $y^{(\hat{\lambda})} \sim \chi$  a pokračovat standardní analýzou.

#### IS pro $\lambda$

Snadno lze odvodit LRT test pro test  $H_0 : \lambda = \lambda_0$ . ( $H_0 : \lambda = 1$  zde je třeba transformace, pokud zamítneme  $H_0 \Rightarrow$  transformace pomocí  $\hat{\lambda}$ ).

LRT statistika:  $\Lambda = -2 \ln \frac{L(\lambda_0)}{L(\hat{\lambda})} = 2(l_p(\hat{\lambda}) - l_p(\lambda_0))$ . Víme, že  $\Lambda \xrightarrow{L_p} \chi^2(1)$ . Invertováním příslušné oblasti LRT testu, dostaneme as.  $100(1 - \alpha)\%$  IS pro  $\lambda$ :

$$\begin{aligned} \chi_{1-\alpha}^2 &\geq \Lambda \\ \chi_{1-\alpha}^2 &\geq 2\left(\frac{n}{n} \ln s_{\lambda_0}^2 - \frac{n}{n} \ln s_{\hat{\lambda}}^2\right) \\ \chi_{1-\alpha}^2 &\geq n \ln \frac{s_{\lambda_0}^2}{s_{\hat{\lambda}}^2} \end{aligned}$$

Pokud  $\hat{\lambda}$  je MLE  $\lambda$ , as.  $100(1 - \alpha)\%$  IS pro  $\lambda$  je:

$$\left\{ \lambda \in \mathbb{R} \mid n \cdot \ln \frac{s_{\lambda_0}^2}{s_{\hat{\lambda}}^2} \leq \chi_{1-\alpha}^2(1) \right\}.$$

**POZNÁMKA 3.23.** Kvůli jednoduchosti interpretace se často doporučuje zaokrouhlit  $\hat{\lambda}$  na nejbližší  $\frac{1}{4}$  nebo  $\frac{1}{2}$ .

Příklad data TREES

#### 3.5.3 Transformace vysvětlujících proměnných $x$

- Pokud diagnostika modelu naznačuje, že vztah mezi  $\mathbf{y}$  a  $\mathbf{X}$  není lineární pro jeden nebo více regresorů, může být vhodné přeformovat model pomocí transformací proměnných  $\mathbf{x}$ .
- Předpokládejme, že v modelu  $Y = \beta_0 + \sum_{j=1}^m \beta_j x_j + e$  máme podezření na nelinearitu v  $j$ -té proměnné  $x_j$ .
- Jednou z možností jak postupovat je nahrazení  $x_j$  proměnnou  $z_j = f(x_j)$ , model tedy bude  $Y = \beta_0 + \beta_1 x_1 + \dots + \beta_j z_j + \dots + \beta_m x_m + e$ .
- Pokud je  $f$  známé, jedná se o model LR a lze ho analyzovat standardně. Pokud je tato transformace vhodná, mělo by se to projevit ve zlepšení statistik  $R^2$ ,  $t$ ,  $F$  a zlepšení grafu rezidií pro  $z_j$  oproti těm pro  $x_j$ .

### 3 Rezidua, diagnostika a transformace

- Bohužel  $f$  většinou známá není, možný přístup je parametrizovat nějak tuto funkci a pak odhadnout tyto parametry společně s  $\beta$ .

– Typická parametrizace:

$$z_j = x_j^\lambda, \quad \text{kde } \lambda \in \mathbb{R} \quad \text{vhodné.}$$

( $x_j >$  potom  $\lambda \in \mathbb{R}$ , pokud  $x_j$  může být záporné, množina hodnot  $\lambda$  je omezená)

– Aproximace  $f$  pomocí polynomu vhodného stupně, tzn.

$$z_j = \sum_{k=1}^l r_k x_j^k, \quad \text{kde } r_k \text{ musí být odhadnuty,}$$

– další možností je použití trigonometrických funkcí nebo splines (piecewise polynomials).

Výsledný model ale v tomto případě nebude lineární v parametrech  $\beta_j, j = 0, \dots, m$  a  $r_k, k = 1, \dots, l$ .

**Zaměříme se na  $z_j = x_j^\lambda$ :**

- možnost je opět zvolit jistou množinu hodnot  $\lambda$ , naladit modely pro všechna  $\lambda$  a vybrat model s nejlepší shodou s daty, např. s nejmenší SSE nebo největší  $R^2$  nebo  $F$
- může být časově náročné, můžeme minout vhodnou hodnotu  $\lambda$ , pokud nebyla v původní množině (nevíme jak  $R^2, F, SSE$  závisí na  $\lambda$ )

#### Box-Tidwell metoda

Předpokládejme, že  $\lambda$  se příliš neliší od  $\lambda = 1$ . Taylorův rozvoj 1. řádu kolem  $\lambda = 1$  dává

$$x^\lambda \approx x^1 + (\lambda - 1) \frac{dx^\lambda}{d\lambda} \Big|_{\lambda=1}, \quad \frac{dx^\lambda}{d\lambda} = x^\lambda \ln x c \frac{dx^\lambda}{d\lambda} \Big|_{\lambda=1} = x \ln x, \quad x^\lambda \approx x + (\lambda - 1)x \ln x.$$

Dosazením do modelu

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \dots + \beta_{j-1} x_{j-1} + \beta_j (x_j + (\lambda - 1)x_j \ln x_j) + \dots + \beta_m x_m + e \\ &= \beta_0 + \sum_{k=1}^m \beta_k x_k + \underbrace{\beta_j (\lambda - 1)}_{\beta_{m+1}(\lambda)} x_j \ln x_j + e \end{aligned}$$

máme lineární model pro parametry  $\beta_k$ ,  $0 \leq k \leq m+1$ , protože  $\beta_{m+1}(\lambda - 1)\beta_j$  můžeme  $(\lambda, \beta_j)$  odhadnout následovně:

- 1) naladíme původní model a spočteme LSE  $\hat{\beta}_j$  parametru  $\beta_j$ ,
- 2) naladíme rozšířený model s  $x_{m+1} = x_j \ln x_j$  a spočteme  $\hat{\beta}_{m+1}$ ,
- 3) z rovnosti  $\hat{\beta}_{m+1} = (\hat{\lambda} - 1)\hat{\beta}_j$  dostaneme  $\hat{\lambda} = \frac{\hat{\beta}_{m+1}}{\hat{\beta}_j} + 1$ .

Tento postup umožňuje testovat potřebu transformace

$$H_0 : \lambda = 1 \text{ vs. } H_1 : \lambda \neq 1$$

pomocí t-testu pro  $H_0 : \beta_{m+1} = 0$ .

### 3 Rezidua, diagnostika a transformace

POZNÁMKA 3.24. Pokud model s  $\hat{\lambda}$  vypadá neadekvátně, lze postupovat iterativně a získat posloupnost  $\hat{\lambda}(l)$ ,  $l \geq 1$ ,  $\hat{\lambda}(0) = \hat{\lambda}$  a rozvineme  $x_j^\lambda$  kolem  $\hat{\lambda}(0)$ , tzn.

$$x_j^\lambda \approx x_j^{\hat{\lambda}(0)} + (\lambda - \hat{\lambda}(0)) x_j^{\hat{\lambda}(0)} \ln x_j \text{ dosazením do rovnice modelu}$$

$$Y = \beta_0 = \sum_{\substack{k=1 \\ k \neq j}}^m \beta_k x_k + \beta_j x_j^{\hat{\lambda}(0)} + \underbrace{\beta_j(\lambda - \hat{\lambda}(0))}_{\beta_{m+1}} x_j^{\hat{\lambda}(0)} \ln x_j + e$$

naladíme tento model s a bez přidané proměnné  $x_{m+1} = x_j^{\hat{\lambda}(0)} \ln x_j$ . Označíme  $\hat{\beta}_j(1)$  a  $\hat{\beta}_{m+1}(1)$  příslušné odhady. Potom

$$\hat{\lambda}(1) = \hat{\lambda}(0) + \frac{\hat{\beta}_{m+1}(1)}{\hat{\beta}_j(1)}.$$

Můžeme dále iterovat do konvergence nebo skončit po pevném počtu iterací.

POZNÁMKA 3.25. Další užívané transformace v  $\mathbf{x}$ ,  $\mathbf{Y}$ :

**a) centrované proměnné:**  $\mathbf{X}_C$  tak, že  $(\mathbf{X}_C)_{ij} = x_{ij} - \bar{x}_j$ ,  $i \in \hat{n}$ ,  $j \in \hat{m}$ , kde  $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$  je

průměr  $j$ -tého sloupce matice  $\mathbf{X}$ .  $\mathbf{Y}_C$ ,  $(\mathbf{Y}_C)_i = y_i - \bar{y}$ ... nevím

**b) centrované a škálované proměnné:** škálování sloupců tak, aby jejich norma byla 1, tzn.

každý prvek  $j$ -tého sloupce matice  $\mathbf{X}$  podělíme  $s_j = \left( \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \right)^{\frac{1}{2}}$ . Centrované a škálované matice  $\mathbf{X}_{SC}$  pak bude

$$\mathbf{X}_{SC} = \mathbf{X}_C \mathbf{S}, \quad \mathbf{S} = \text{diag}\left(\frac{1}{s_1}, \dots, \frac{1}{s_m}\right).$$

Model pak bude

$$\mathbf{Y}_C = \mathbf{X}_{SC} \boldsymbol{\beta}_s + e.$$

Lze použít i  $\mathbf{Y}_{SC}$ , tedy centrované a škálované  $\mathbf{Y}$ .

### Vážené nejmenší čtverce (weight least squares WLS)

Budeme nyní předpokládat, že chyby  $e_i$  jsou normální, nezávislé, ale  $D(e_i) = \sigma_i^2$  závisí na  $i$ . Konkrétně tedy  $\sigma_i^2 = \frac{\sigma^2}{w_i}$ , kde  $w_i > 0$ ,  $i \in \hat{n}$  se nazývají váhy.

Uvažujeme tedy model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \text{ kde } \mathbf{e} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{W}) \text{ a } \mathbf{W} = \text{diag}\left(\frac{1}{w_1}, \dots, \frac{1}{w_n}\right). \quad (3.4)$$

Pokud jsou váhy  $w_i$  známé, lze MLE odhadovat parametr  $\boldsymbol{\beta}$  a  $\sigma^2$  následovně:

$$\mathbf{W} = \mathbf{K}\mathbf{K}^T, \text{ kde } \mathbf{K} = \mathbf{W}^{\frac{1}{2}} = \text{diag}\left(\frac{1}{\sqrt{w_1}}, \dots, \frac{1}{\sqrt{w_n}}\right).$$

Definujeme  $\mathbf{Z} = \mathbf{K}^{-1}\mathbf{Y}$ ,  $\mathbf{M} = \mathbf{K}^{-1}\mathbf{X}$  a  $\boldsymbol{\varepsilon} = \mathbf{K}^{-1}\mathbf{e}$ . Potom dostaneme model

$$\mathbf{Z} = \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ kde } \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n), \quad (3.5)$$

protože

$$\text{Cov}(\boldsymbol{\varepsilon}) = \mathbf{K}^{-1}\sigma^2 \mathbf{W}(\mathbf{K}^{-1})^T = \sigma^2 \mathbf{K}^{-1} \mathbf{K} \mathbf{K}^T (\mathbf{K}^T)^{-1} = \sigma^2 I_n.$$

### 3 Rezidua, diagnostika a transformace

Transformační vektor je tedy ve tvaru  $\mathbf{Z} = (\sqrt{w_1}Y_1, \dots, \sqrt{w_n}Y_n)^T$ . To už je standardní model LR, na kterém platí

$$\hat{\beta}_w = (\mathbf{M}^T \mathbf{M})^{-1} \mathbf{M}^T \mathbf{z} = (\mathbf{X}^T \underbrace{(\mathbf{K}^{-1})^T \mathbf{K}^{-1}}_{=(\mathbf{K}\mathbf{K}^T)^{-1}} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{K}^{-1})^T \mathbf{K}^{-1} \mathbf{Y} = (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-1} \mathbf{Y}.$$

Dále pak

$$\widehat{\sigma}_w^2 = \frac{1}{n} \sum_{i=1}^n (z_i - \hat{z}_i)^2 = \frac{1}{n} \sum_{i=1}^n w_i (y_i - \hat{y}_i)^2 = \frac{1}{n} \text{SSE}_w,$$

kde  $\text{SSE}_w$  je vážený součet čtverců,  $z_i = \sqrt{w_i}y_i$  a  $\hat{z}_i = \sqrt{w_i}\mathbf{x}_i^T \hat{\beta} = \sqrt{w_i}\hat{y}_i$ . Dále platí, že

a)  $\mathbb{E}\hat{\beta}_w = (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-1} \underbrace{\mathbb{E}\mathbf{Y}}_{\mathbf{X}\beta} = \hat{\beta}$ , kde  $\hat{\beta}_w$  je nestranný odhad  $\hat{\beta}$ ,

b)  $\mathbb{E}\left(\frac{\text{SSE}_w}{n-m-1}\right) = \sigma^2$ , tedy  $s_w^2 = \frac{\text{SSE}_w}{n-m-1}$  je nestranný odhad  $\sigma^2$ .

**Věta 3.26.** Nechť  $\hat{\beta}_w$  je WLS odhad  $\hat{\beta}$ , jestliže  $\text{Cov}(\mathbf{e}) = \sigma^2 \mathbf{W} = \sigma^2 \text{diag}\left(\frac{1}{w_1}, \dots, \frac{1}{w_n}\right)$ . Potom platí, že

1)  $\text{Cov}(\hat{\beta}_w) = \sigma^2 (\mathbf{X} \mathbf{W}^{-1} \mathbf{X})^{-1}$ ,

2) nechť  $\delta_i$  je  $i$ -tý diagonální prvek  $(\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1}$ . Jestliže  $e_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{w_i}\right)$ ,  $i \in \hat{n}$ , potom

$$T_i = \frac{\hat{\beta}_{w,i} - \beta_i}{s_w \sqrt{\delta_i}} \sim t(n-m-1),$$

3) pro  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}_w$  platí, že  $\mathbb{E}\hat{\mathbf{Y}}_m = \mathbf{X}\mathbf{B}$  a  $\text{Cov}(\hat{\mathbf{Y}}_w) = \sigma^2 \mathbf{X} (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T$ ,

4) nechť  $\hat{e}_w = \mathbf{Y} - \hat{\mathbf{Y}}_m$  jsou rezidua v modelu 3.4 a  $\hat{\varepsilon}_w = \mathbf{Z} - \hat{\mathbf{Z}} = \mathbf{Z} - \mathbf{M}\hat{\beta}_w$  jsou rezidua v transformovaném modelu 3.5. Potom

$$\hat{\varepsilon}_w = \sqrt{\mathbf{W}^{-1}} \hat{e}_w = \mathbf{W}^{-\frac{1}{2}} \hat{e}_w \quad a \quad \mathbb{E}(h e_w) = \mathbb{E}(\hat{\varepsilon}_w) = 0,$$

5) nechť  $\mathbf{H}_w = \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X}^{-1})^{-1} \mathbf{X}^T \mathbf{W}^{-1}$  je vážená projekční matici. Potom

$$\hat{\mathbf{e}}_w = (I - \mathbf{H}_w)\mathbf{e} \quad a \quad \text{Cov}(\hat{\mathbf{e}}_w) = \sigma^2 (I - \mathbf{H}_w) \mathbf{W}.$$

To znamená, že

$$\text{Cov}(\hat{\varepsilon}_w) = \sigma^2 \mathbf{W}^{-\frac{1}{2}} (I - \mathbf{H}_w) \mathbf{W}^{\frac{1}{2}}.$$

Důkaz. 1)

$$\text{Cov}(\hat{\beta}_w) = (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-1} \underbrace{\text{Cov} \mathbf{Y} \mathbf{W}^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1}}_{=\sigma^2 \mathbf{W}} = \sigma^2 (\mathbf{X}^T \mathbf{W}^{-1} \mathbf{X})^{-1},$$

2)  $D\hat{\beta}_{w,i} = \sigma^2 \delta_i$ , tzn.  $\frac{\hat{\beta}_{w,i} - \beta_i}{\sigma \sqrt{\delta_i}} \sim \mathcal{N}(0, 1)$  a víme, že  $\hat{\beta}_{w,i}$  a  $s_w^2$  jsou nezávislé,  $\frac{s_w^2(n-m-1)}{\sigma^2} \sim \chi^2(n-m-1)$ . Z toho vyplývá, že

$$\frac{\hat{\beta}_{w,i} - \beta_i}{s_w \sqrt{\delta_i}} \sim t(n-m-1).$$

### 3 Rezidua, diagnostika a transformace

3)  $\mathbb{E}\hat{\mathbf{Y}}_w = \mathbf{X}\mathbb{E}\hat{\boldsymbol{\beta}}_w = \mathbf{X}\hat{\boldsymbol{\beta}}$ ,  $\text{Cov}(\hat{\mathbf{Y}}_w) = \mathbf{X}\text{Cov}\hat{\boldsymbol{\beta}}_w\mathbf{X}^T = \sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$ .

4) Protože  $\mathbf{Z} = \mathbf{W}^{-\frac{1}{2}}\mathbf{Y}$  a  $\mathbf{M} = \mathbf{W}^{-\frac{1}{2}}\mathbf{X}$ , dostaneme

$$\begin{aligned}\hat{\varepsilon}_w &= \mathbf{Z} - \hat{\mathbf{Z}} = \mathbf{W}^{-\frac{1}{2}}\mathbf{Y} - \mathbf{W}^{-\frac{1}{2}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w) = \mathbf{W}^{-\frac{1}{2}}\hat{\varepsilon}_w, \\ \mathbb{E}\hat{\mathbf{e}}_w &= \mathbb{E}(\mathbf{Y} - \hat{\mathbf{Y}}_w) = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0} \quad \Rightarrow \quad \mathbb{E}\hat{\varepsilon} = 0.\end{aligned}$$

5)

$$\begin{aligned}\hat{\mathbf{e}}_w &= \mathbf{Y} - \hat{\mathbf{Y}}_w = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} - \mathbf{X}(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}\underbrace{(\mathbf{X}\boldsymbol{\beta} + \mathbf{e})}_{\mathbf{Y}} = \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} + \mathbf{e} - \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}}_{\mathbf{H}_w}\mathbf{e} = (I - \mathbf{H}_w)\mathbf{e}, \\ \text{Cov}(\hat{\mathbf{e}}_w) &= (I - \mathbf{H}_w)\text{Cov } \mathbf{e}(I - \mathbf{H}_w)^T = \\ &= \sigma^2(I - \mathbf{X}(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1})\mathbf{W}(I - \mathbf{W}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T) = \\ &= \sigma^2\mathbf{W} - \sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T - \sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T + \sigma^2\mathbf{X}(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T = \\ &= \sigma^2(I - \mathbf{H}_w)\mathbf{W}.\end{aligned}$$

$$\text{Cov}(\hat{\varepsilon}) = \mathbf{W}^{-\frac{1}{2}}\text{Cov}(\hat{\varepsilon})\mathbf{W}^{-\frac{1}{2}} = \sigma^2\mathbf{W}^{-\frac{1}{2}}(I - \mathbf{H}_w)\mathbf{W}^{\frac{1}{2}}.$$

□

- z dosazení vyplývá, že odhad parametru  $\boldsymbol{\beta}$  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

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

kde  $\text{SSE}_F$  je reziduální součet čtverců  $s_w^2$  plného modelu a  $\text{SSE}_R$  je reziduální součet čtverců redukovaného transformovaného modelu  $\mathbb{Z} = \mathbf{M}_0\boldsymbol{\beta}_0 + \mathbf{e}$ ,  $\mathbf{M}_0 = (\sqrt{w_1}, \dots, \sqrt{w_n})^T$ .

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:

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

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)

### 3 Rezidua, diagnostika a transformace

- je třeba být opatrný oproti jakým hodnotám budeme rezidua zobrazovat
- grafy  $\hat{\varepsilon}_i$  proti sloupcům  $\mathbf{M}$  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  $\mathbb{W} = \mathbf{I}$  dostaneme standardní  $R^2$

POZNÁMKA 3.27. • , pokud jsou váhy neznámé, bylo by třeba je odhadnou 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 maticy atd.)

POZNÁMKA 3.28. Celý postup WLS lze použít i na případ  $\mathbf{e} \sim \mathcal{N}_m(0, \sigma^2 \mathbf{W})$ , kde  $\mathbf{W}$  je známá, ale není diagonální. Protože  $\mathbf{W}$  je symetrická, ex. regulární  $\mathbf{K}$  tak, že  $\mathbf{W} = \mathbf{KK}^T$ . Stejná transformace jako u WLS opět vede na transformovaný model, kde  $\varepsilon \sim \mathcal{N}_m(0, \sigma^2 \mathbf{I}_m)$ .

## 3.6 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, že

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 dát 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.

### 3.6.1 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.

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

### 3 Rezidua, diagnostika a transformace

kde  $\varrho$  je autokorelační koeficient,  $u_t \sim \mathcal{N}(0, \sigma_n^2)$  jsou nezávislé v  $t \in \hat{n}$  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 3.29. 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  $\Rightarrow \frac{\sum_{t=2}^n \hat{e}_t \hat{e}_{t-1}}{\sum_{t=1}^n \hat{e}_t^2} \leq 1$  leží přibližně v  $(-1, 1)$ , tzn.  $d$  leží přibližně v  $(0, 4)$ . Dále

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

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

**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 nerozhodné

POZNÁMKA 3.30. 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**.

A) Mallows  $C_p$ , AIC, BIC

Kritéria beroucí více v poraz počet použitých regresorů. Lze je použít i pro **nevnořené modely!**

- Mallows  $C_p$

$$C_p = \frac{\text{SSE}_p}{\hat{\sigma}^2} - n + 2p, \quad \hat{\sigma}^2 = \frac{\text{SSE}_T}{n - T}$$

**Vlastnosti  $C_p$ :**

- 1) Snadno se počítá,  $\text{SSE}_p$  a  $\hat{\sigma}^2$  jsou implementované.
- 2) Pokud je  $\hat{\sigma}^2$  konzistentní odhad  $\sigma^2$  (nazávisející na  $p$ ), má  $C_p$  následující interpretaci: Porovnává, co zbývá vysvětlit pomocí modelů s  $p$  a  $T$  parametry, zvýhodňuje počet dostupných dat a penalizuje počet parametrů, které je třeba odhadnout.
- 3) Při zvyšování počtu regresorů:  $q|\sigma^2$  je konst.,  $\text{SSE}_p$  klesá,  $p$  roste ( $C_p$  se snaží sladit dvě protichůdná kritéria).
- 4)  $C_T = T$ .

### 3 Rezidua, diagnostika a transformace

5) Pokud je správný model s  $p$  parametry, dá se ukázat, že  $C_p \approx p$  pro  $n >> T$ .

6) V praxi se volí model s nejmenším  $C_p$  ve skupině modelů. (Obrázek)

POZNÁMKA 3.31. Pro dobrou interpretaci je třeba spočítat  $C_p$  pro všechny nebo většinu podmnožin regresorů.

- Akaikeho informační kritérium AIC

Obecná definice je

$$AIC = -2l(\hat{\theta}) + 2p^*,$$

kde  $\hat{\theta}$  je MLE odhad v modelu,  $l$  je log-věrohodnostní funkce a  $p^*$  je počet parametrů, které je třeba odhadnout ( $p^* = p + 1$ , jelikož počítáme i  $\sigma^2$ ).

Pro náš model LR:

$$\begin{aligned} L(\beta, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)\right] \\ l(\beta, \sigma^2) &= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2} \frac{(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)}{\sigma^2} \\ AIC &= -2l(\hat{\beta}, \hat{\sigma}^2) + 2p^* = n \ln 2\pi + \ln \hat{\sigma}^2 + \frac{(\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)}{\hat{\sigma}^2} + 2p^*, \end{aligned}$$

ale  $\hat{\sigma}^2 = \frac{1}{n} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) = \frac{SSE}{n}$ , tedy

$$AIC = \underbrace{n \ln 2\pi + n}_{C} + n \ln \frac{SSE}{n} + 2p^*$$

nebo alternativně  $AIC = n \ln \frac{SSE}{n} + 2p^*$ .

POZNÁMKA 3.32. – hledáme model s minimální hodnotou AIC

– AIC není mírou kvality modelu, je užitečná pro porovnávání modelů

#### AIC v R:

- `AIC(.)` počítá  $AIC = n \ln 2\pi + n + n \ln \frac{SSE}{n} + 2p^*$ , kde  $p^*$  je počet parametrů  $\beta, \sigma^2$  (včetně interceptu)
- `extractAIC(.)` počítá  $AIC = +n \ln \frac{SSE}{n} + 2p^*$ , kde  $p^*$  je jen počet parametrů  $\beta$  (včetně interceptu)

- (Schwarzovo) bayesovské informační kritérium BIC

Z definice

$$BIC = -2l(\hat{\theta}) + p^* \ln n.$$

Více penalizuje počet parametrů  $\Rightarrow$  vybírá jednodušší modely s jednodušší interpretací, než AIC. BIC vyžaduje významnější příspěvek proměnné, aby byla zařazena do modelu. (AIC se zanořuje za predikci, BIC je kompromis mezi interpretovatelností a predikcí.)

#### BIC v R:

- `BIC(.)` nebo `AIC(.)`, `extractAIC(.)` s volbou `k = log(nobs(fit))`.

Příklad - data HALD.

## B) PRESS statistika

Pokud je pro nás důležitá kvalita predikve, lze použít pro srovnání modelů statistiku

$$PRESS = \sum_{i=1}^n \hat{e}_{(-i)}^2 = \sum_{i=1}^n \left( \frac{\hat{e}_i}{1 - h_{ii}} \right)^2.$$

Vybírá se model s minimální hodnotou této statistiky.

### 3.7 Metody výběru modelu

#### 1) Vyhodnocení všech možných modelů

Pro  $T$  dostupných regresorů tzn. nalaďit  $2^T$  modelů. Pak je porovnat pomocí nějakého kritéria. Náročné pro velká  $T$  (například  $T = 20$  znamená 1024 modelů).

#### 2) Zpětná eliminace (backward elimination)

Začneme s plným modelem a v každém kroku odstraníme jednu proměnnou, která nejméně přispívá modelu (měřeno F stat) nebo jejíž odstranění znamená největší zlepšení modelu (měřeno AIC).

##### algoritmus:

- 1) Nalaďme model se všemi proměnnými.
- 2) Pro každou proměnnou spočteme částečnou F statistiku (nebo  $t$ -statistiku) jako by právě byla přidána do modelu, tzn. za předpokladu, že ostatní proměnné tam už jsou.
- 3) Pokud je nějaká F -statistika menší, než kritická hodnota  $F_{\text{OUT}}$ , vynecháme z modelu proměnnou s nejnižší hodnotou F.  $F_{\text{OUT}} = F_{1-\alpha_{\text{OUT}}}(1, n-p)$ , kde  $p$  je aktuální počet regresorů v modelu, včetně interceptu,  $\alpha_{\text{OUT}} = 0.05, 0.1, \dots$
- 4) Opakujeme kroky 2) a 3), dokud všechny částečné F statistiky nejsou větší, než příslušná kritická hodnota  $F_{\text{OUT}}$ , tzn. nelze už vyřadit žádnou proměnnou.

Poznámka 3.33. místo F lze používat AIC.

#### 3) Dopředná regrese (forward regression)

Začneme pouze s interceptem (nebo nutným minimálním modelem) a v každém kroku přidáme jednu proměnnou, která má za následek největší zlepšení modelu (největší nárůst F nebo největší pokles AIC).

Tato metoda neumožňuje odstranit proměnnou, která už do modelu byla přidána.

##### algoritmus:

- 1) Nalaďme minimální model.
- 2) Pro každou dostupnou proměnnou spočteme F statistiku pro test významnosti jího přidání do modelu.
- 3) Pokud ně, která z této F statistik překračuje kritickou hodnotu  $F_{\text{in}}$ , přidáme do modelu proměnnou s nejvyšší hodnotou F statistiky.
- 4) Opakujeme kroky 2) a 3), dokud všechny F -statistiky pro zbývající proměnné nebudou menší, než  $F_{\text{in}}$  nebo dokud nezbyde žádná proměnná na přidání do modelu.

Poznámka 3.34. I když tento postup zjednoduší výběr modelu, často bohužel vede na zařazení proměnných, které nemají významný příspěvek, jakmile jsou zařazeny další proměnné.

#### 4) Postupná regrese (stepwise regression)

Kombinace dvou předchozích metod. V každém kroku algoritmu přidáváme jednu proměnnou a poté zkонтrolujeme, zda není možné nějakou odebrat. Budeme potřebovat dvě kritické hodnoty  $F_{\text{in}}$ ,  $F_{\text{OUT}}$  (pro použití F statistiky).

##### algoritmus:

- 1) Nalaďme minimální model.
- 2) Zjistíme, zda přidání nějaké další proměnné může zlepšit model (F nebo AIC). Pokud ano, přidáme do modelu proměnnou, která má za následek největší zlepšení modelu (nebo největší pokles AIC).
- 3) V novém modelu zjistíme, zda nelze některou proměnnou vynechat (opět pomocí AIC

### *3 Rezidua, diagnostika a transformace*

nebo F). Pokud ano, vynecháme proměnnou, jejíž vyřazení má za následek největší zlepšení modelu (nebo největší pokles AIC).

- 4) Opakujeme kroky 2) a 3) do té doby, až nebude možné přidat ani ubrat žádnou proměnnou.

**Princip marginality:**

- 1) Pokud jsou v modelu vyšší mocniny nějakého regressoru, měly by tam být obsaženy i všechny jeho nižší mocniny (i když jsou případně nevýznamné).
- 2) Pokud je v modelu obsažena interakce dvou regressorů, měly by tam být i oba individuální regresory.
- 3) S každou interakcí vyššího řádu by měl model obsahovat i všechny interakce řádu nižšího. ( $a : b : c \rightarrow a : b, a : c, b : c$ ).

Poznámka 3.35. Jakmile nalezneme optimální model, je třeba řádně ověřit adekvátnost.