01RAD

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

21. září 2023

Obsah

Předmluva

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 01MAB234, 01LAB12, 01MIP a 01MAS.

Materiál byl sestaven na základě zápisků k přednáškám, které proběhly v zimním semestru akademického roku 2020/2021 na Fakultě jaderné a fyzikálně inženýrské ČVUT v Praze. Chtěl bych tímto velmi poděkovat Martinu Kovandovi, Michaele Maškové a Filipu Bárovi za neocenitelnou pomoc při přepisování ručně psaných zápisků do TeXu.

Doporučená literatura:

1) Golberg, M. Cho, H.A.: Introduction to Regression Analysis. WITpress, Southampton 2010.
2) Víšek, J. Á.: Statistická analýza dat. Vydavatelství ČVUT v Praze, Praha 1998.
3) Zvára, K.: Regrese. Matfyzpress, Praha 2008.
4) Olive, D.: Linear Regression. Springer, 2017.

1 Jednorozměrná lineární regrese

Začneme nejjednodušším modelem lineární regrese, tedy model s jedinou vysvětlující proměnnou. 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 (to nastává například v případě, kdy se x nastavuje na předem dané úrovni a následně se k němu změří odpovídající y), zatímco y se obvykle měří s chybou. Chyba může být náhodná a y tedy budeme chápat jako náhodnou veličinu, kterou budeme značit Y. Pro dvojice $(x_1, Y_1), ..., (x_n, Y_n)$ se zavádí model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i \in \widehat{n}, \tag{1.1}$$

kde

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

Budeme předpokládat, že náhodné chyby $e_1, \ldots, e_n$ jsou nezávislé (někdy bude stačit nekorelovanost) a splňují podmínky $\mathsf{E}[e_i] = 0$, $\mathsf{Var}[e_i] = \sigma^2$ pro $\forall i \in \widehat{n}$, což budeme zkráceně značit $e_i \sim (0, \sigma^2)$. Vlastnost stejného rozptylu náhodných chyb se nazývá *homoskedasticita*.
Měřením získáme data $(x_1, y_1), ..., (x_n, y_n)$ a cílem statistické analýzy je určit, zda je model (1.1) schopen popsat pozorovanou variabilitu proměnné y. Prvním krokem je odhad neznámých parametrů $\beta_0, \beta_1, \sigma^2$. Jedná se vlastně o proložení dat přímkou ve tvaru

$$\widehat{y}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x\,.$$

Porovnání pozorovaných hodnot y_i a hodnot $\widehat{y}(x_i)$ predikovaných modelem nám pak umožňuje posoudit adekvátnost modelu.
Pro proložení dat přímkou existuje několik způsobů, zásadní ovšem bude znalost rozdělení e_i a tedy i Y_i. Máme v zásadě dvě možnosti, jak postupovat:

1. odhadnout β_0, β_1 pomocí metody nezávisející na rozdělení chyb
2. udělat věrohodný předpoklad o rozdělení chyb, odhadnout β_0, β_1 a následně náš předpoklad ověřit.

Poznámka 1.1. Speciální důležitý případ je předpoklad normality náhodných chyb, tedy $e_i \sim N(0, \sigma^2)$. Tento předpoklad vede při hledání maximálně věrohodných odhadů (MLE) parametrů β_0, β_1 na metodu nejmenších čtverců, která může být použita bez ohledu na rozdělení chyb.

1.1 Odhady parametrů

1.1.1 Odhady parametrů za předpokladu normality náhodných chyb

Nejdříve budeme předpokládat, že náhodné chyby mají normální rozdělení, tedy $e_1, ..., e_n$ *iid* $N(0, \sigma^2)$. To následně v našem modelu implikuje, že $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ a $Y_1, ..., Y_n$ jsou nezávislé.

MLE odhady

Odvodíme maximálně věrohodné odhady neznámých parametrů. Díky nezávislosti $Y_1, ..., Y_n$ a jejich normálnímu rozdělení mají věrohodnostní funkce a její logaritmus tvar

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

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

Z uvedeného vztahu je vidět, že pro pevné $\sigma^2 > 0$ je maximalizace funkce ℓ ekvivalentní minimalizaci funkce S, kde

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

Za předpokladu normality tedy MLE odhady parametrů β_0, β_1 splývají s odhady tzv. metodou nejmenších čtverců. Pro nalezení argumentu minima funkce S spočteme její první derivace a položíme je rovny 0, získáme tak soustavu rovnic

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

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

které lze přepsat do tvaru

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

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

Z první rovnice pak dostaneme

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

a dosazením do druhé dostaneme výraz

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

Vyjádřením β_1 pak získáme jednotlivé MLE odhady parametrů v následujícím tvaru

$$\widehat{\beta}_0 = \overline{y}_n - \widehat{\beta}_1 \overline{x}_n \quad \text{a} \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{x}_n \overline{y}_n}{\sum_{i=1}^{n} x_i^2 - n\overline{x}_n^2}. \tag{1.2}$$

Hledejme nyní odhad parametru σ^2. Vrátíme se k logaritmu věrohodnostní funkce a její derivací podle σ^2 získáme rovnici

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

Vyjádřením σ^2 a dosazením odhadů $\widehat{\beta}_0, \widehat{\beta}_1$ dostaneme výraz

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \frac{1}{n} SSE,$$

kde $\widehat{y}_i = \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ je predikce modelu (odhad $\mathsf{E}[Y_i]$) a zkratka SSE je odvozena z anglického *sum of the squares of errors*. Rozdíl $\widehat{e}_i = y_i - \widehat{y}_i$ nazýváme i-té reziduum. Velikost reziduí indikuje, jak dobře odhadnutá přímka odpovídá datům. Rezidua jsou vlastně odhady chyb e_i, jejich analýza bude hrát významnou roli v ověření předpokladů o rozdělení chyb.
Pro odhad parametru σ^2 se častěji používá statistika

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

která je nestranným odhadem parametru σ^2 (pro libovolné rozdělení e_i), zatímco MLE odhad $\widehat{\sigma}^2$ je vychýlený odhad i pro normální rozdělení chyb. Statistika

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

se pak nazývá standardní chyba regrese (*standard error*).

Poznámka 1.2. Poznamenejme, že i když s_n^2 je nestranný odhad parametru σ^2, s_n už je vychýlený odhad směrodatné odchylky σ! Jedná se o obecnou vlastnost nestranných odhadů rozptylů, neboť pokud je s^2 nestranný odhad σ^2, pak $\mathsf{E}\, s \leqslant \sigma$, jak lze nahlédnout z následující úvahy.
Uvažujme náhodnou veličinu X, pro kterou platí, že $\mathsf{Var}\, X < +\infty$. Po dosazení $X = s$ do známé rovnice $\mathsf{E}[X^2] = \mathsf{Var}\, X + (\mathsf{E}\, X)^2$ dostaneme vztah

$$\mathsf{E}[s^2] = \mathsf{Var}\, s + (\mathsf{E}\, s)^2,$$

kde $\mathsf{E}[s^2] = \sigma^2$ z nestrannosti a $\mathsf{Var}\, s \geqslant 0$. Platí tedy $(\mathsf{E}\, s)^2 \leqslant \sigma^2$, zapsáno ekvivalentně $\mathsf{E}\, s \leqslant \sigma$ a rovnost nastává právě tehdy, když $\mathsf{Var}\, s = 0$. Například pro normální chyby má statistika s_n^2 χ^2-rozdělení s nenulovým rozptylem a tedy $\mathsf{E}\, s_n < \sigma$.

1.1.2 Odhady parametrů bez předpokladu rozdělení náhodných chyb

Uvažujme teď případ, kdy jsou tedy $e_1, ..., e_n$ nekorelované a $e_1, ..., e_n \sim (0, \sigma^2)$. Pro náhodné chyby tedy neuvažujeme žádné konkrétní rozdělení, pouze nulovost střední hodnoty a konstantní rozptyl. Pro odhad parametrů β_0, β_1 lze v tomto případě také použít minimalizaci součtu čtverců chyb S. Že je to rozumná procedura, vyplyne z geometrické interpretace funkce S.
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 této přímky (viz obrázek 1.2) a

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

tedy součet čtverců těchto vzdáleností, je míra udávající, jak dobře přímka prokládá data. Dává smysl vybrat takovou přímku, která minimalizuje S. Minimalizací S získáme stejné formule pro odhady $\widehat{\beta}_0, \widehat{\beta}_1$ jako u MLE odhadů pro normální data. Protože se ale nyní nejedná o maximálně věrohodné odhady, nazývají se odhady metodou nejmenších čtverců (LSE - *least squares estimators*). Jak už bylo zmíněno dříve, existuje více měr vhodnosti přímky. Použití LSE pro libovolné rozdělení chyb má dvě následující zdůvodnění:

1. pro normálně rozdělené chyby splývájí odhady LSE s MLE,
2. LSE odhady jsou navíc BLUE (best linear unbiased estimators), tedy nestranné lineární odhady s minimálním rozptylem (jak ukážeme v Gauss–Markovově větě).

Příklad 1.1. Uvažujme případ, kdy náhodné chyby mají tzv. Laplaceovo rozdělení, tedy $e_1, ..., e_n$ jsou *iid* s hustotou

$$f(\varepsilon) = \frac{1}{2} \mathrm{e}^{-|\varepsilon|}, \quad \varepsilon \in \mathbb{R}.$$

V modelu (1.1) potom pro hustotu veličiny Y_i platí

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

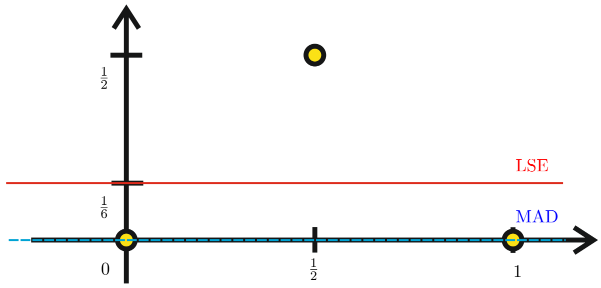
a věrohodnostní funkce L a její logaritmus ℓ mají tvar

$$L = \frac{1}{2^n} \mathrm{e}^{-\sum\limits_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|}, \qquad \ell = -n \ln 2 - \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|.$$

MLE odhady parametrů β_0, β_1 pak získáme minimalizací funkce

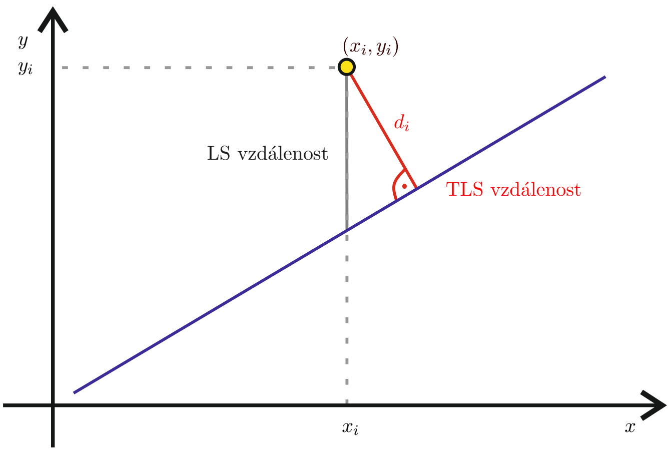
$$A = A(\beta_0, \beta_1) = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|.$$

Díky tvaru funkce A se tyto odhady někdy nazývají MAD odhady (*minimum absolute deviation*). Tyto odhady budou obecně jiné než LSE odhady. Pro ilustraci uvažujme 3 body $(0, 0)$, $(1, 0)$, $(\frac{1}{2}, \frac{1}{2})$ a spočtěme MLE a LSE odhady. Minimalizací funkce A lze ukázat, že MLE (MAD) odhady mají hodnotu $\widehat{\beta}_0 = \widehat{\beta}_1 = 0$ ($A(\widehat{\beta}_0, \widehat{\beta}_1) = 0.5$) a rovnice přímky modelu je v tomto případě $\widehat{y} = 0$. Pro výpočet LSE odhadů dosadíme hodnoty $\overline{x}_n = \frac{1}{2}, \overline{y}_n = \frac{1}{6}, \sum\limits_{i=1}^{n} x_i^2 = \frac{5}{4}, \sum\limits_{i=1}^{n} x_i y_i = \frac{1}{4}$ do vzorců (1.2) a dostaneme odhady $\widehat{\beta}_1 = 0, \widehat{\beta}_0 = \frac{1}{6}$. Přímka modelu je v tomto případě $\widehat{y} = \frac{1}{6}$. Situace je znázorněna na obrázku 1.1.



Obrázek 1.1: LSE a MLE (MAD) přímky proložené body $(0, 0)$, $(1, 0)$, $(\frac{1}{2}, \frac{1}{2})$.

Poznámka 1.3. V našem modelu jsme předpokládali, že hodnoty x_i jsou dány přesně, což nemusí být vždy pravda. Často jsou obě veličiny (x, y) měřeny nepřesně. V těchto tzv. EIV modelech (*error in variable*) jsou často preferovány jiné odhady než LSE. Jedna z populárních metod je např. metoda totálních nejmenších čtverců (*total least squares* někdy *ortogonal least squares*). Tato metoda minimalizuje $\sum_{i=1}^{n} d_i^2$, kde d_i je kolmá vzdálenost bodu (x_i, y_i) od regresní přímky (viz obrázek 1.2). Tato metoda tedy neupřednostňuje veličinu x, jako LSE, ale přistupuje k oběma veličinám x a y rovnocenně.



Obrázek 1.2: Vzdálenosti minimalizované metodami nejmenších a totálních nejmenších čtverců.

Poznámka 1.4. V literatuře se někdy hodnoty proměnné x uvažují jako realizace náhodné veličiny (ne vždy se x nastavuje předem nebo je jasně dané). Model má potom tvar

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

Pro většinu výsledků prezentovaných v této přednášce ale není podstatné, zda je x chápáno jako pevné nebo náhodné. Důkazy většinou fungují s podmíněnými výrazy (E, Var, ...) při dané hodnotě x místo nepodmíněných. Nicméně větší pozornost je třeba u odvozování asymptotických rozdělení odhadů.

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

Věta 1.1. Nechť $\widehat{\beta}_0, \widehat{\beta}_1$ jsou LSE odhady parametrů β_0, β_1 v lineárním modelu

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

kde $e_i \sim (0, \sigma^2)$ jsou nezávislé náhodné veličiny (postačí i nekorelovanost). Potom platí:

1) $\mathsf{E}[\widehat{\beta}_0] = \beta_0, \quad \mathsf{E}[\widehat{\beta}_1] = \beta_1$, (nestranné odhady),
2) $\mathsf{Var}[\widehat{\beta}_1] = \frac{\sigma^2}{S_{xx}}, \quad$ kde $S_{xx} = \sum\limits_{i=1}^{n} (x_i - \overline{x}_n)^2$,
3) $\mathsf{Var}[\widehat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}_n^2}{S_{xx}}\right)$.
4) Pokud navíc platí, že $e_i \sim N(0, \sigma^2),\ \forall i \in \hat{n}, \quad$ potom $\quad \widehat{\beta}_j \sim N(\beta_j, \mathsf{Var}[\widehat{\beta}_j]),\ j \in \{0, 1\}$.

Důkaz. 1) Upravíme $\widehat{\beta}_1$ do tvaru, se kterým se bude lépe pracovat:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum\limits_{i=1}^{n} y_i x_i - n\overline{x}_n\overline{y}_n}{\sum\limits_{i=1}^{n} x_i^2 - n\overline{x}_n^2} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum\limits_{i=1}^{n} (x_i - \overline{x}_n)^2} = \\ &= \frac{1}{S_{xx}} \left(\sum_{i=1}^{n} (x_i - \overline{x}_n) y_i - \overline{y}_n \sum_{i=1}^{n} (x_i - \overline{x}_n) \right) = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}_n) y_i, \end{aligned} \tag{1.3}$$

neboť $\sum\limits_{i=1}^{n} (x_i - \overline{x}_n) = \sum\limits_{i=1}^{n} x_i - n\overline{x}_n = 0$. Střední hodnota odhadu $\widehat{\beta}_1$ (pokud se na $\widehat{\beta}_1$ budeme dívat jako na náhodnou veličinu, nahradíme realizace y_i veličinami Y_i) má potom tvar

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

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

$$\begin{aligned}\mathsf{E}[\widehat{\beta}_0] &= \mathsf{E}[\overline{Y}_n - \widehat{\beta}_1 \overline{x}_n] = \mathsf{E}[\overline{Y}_n] - \overline{x}_n\, \mathsf{E}[\widehat{\beta}_1] = \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}[Y_i] - \overline{x}_n \beta_1 \\ &= \beta_0 + \frac{\beta_1}{n} \sum_{i=1}^{n} x_i - \overline{x}_n \beta_1 = \beta_0.\end{aligned}$$

2) Jelikož Y_i jsou nezávislé a platí $\mathsf{Var}[Y_i] = \sigma^2$, můžeme spočítat rozptyl odhadu $\widehat{\beta}_1$ jako

$$\mathsf{Var}[\widehat{\beta}_1] = \mathsf{Var}\left[\frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}_n) Y_i\right] = \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2\, \mathsf{Var}[Y_i] = \frac{\sigma^2 S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}.$$

3) Při výpočtu rozptylu odhadu $\widehat{\beta}_0$ už je třeba užít obecnější vztah pro výpočet rozptylu součtu náh. veličin obsahující i kovarianci

$$\begin{aligned}\mathsf{Var}\,[\widehat{\beta}_0] &= \mathsf{Var}\,[\overline{Y}_n - \widehat{\beta}_1\overline{x}_n] = \mathsf{Var}\,[\overline{Y}_n] + \overline{x}_n^2\,\mathsf{Var}\,[\widehat{\beta}_1] - 2\overline{x}_n\,\mathsf{Cov}\,(\overline{Y}_n, \widehat{\beta}_1) = \\ &= \frac{\sigma^2}{n} + \frac{\overline{x}_n^2\sigma^2}{S_{xx}} - 2\overline{x}_n\,\mathsf{Cov}\,(\overline{Y}_n, \widehat{\beta}_1).\end{aligned}$$

Teď už nám stačí ukázat, že $\mathsf{Cov}\,(\overline{Y}_n, \widehat{\beta}_1) = 0$. S využitím vlastností kovariance dostaneme

$$\begin{aligned}\mathsf{Cov}\,(\overline{Y}_n, \widehat{\beta}_1) &= \mathsf{Cov}\left(\overline{Y}_n, \frac{1}{S_{xx}}\sum_{i=1}^n (x_i - \overline{x}_n)Y_i\right) = \frac{1}{S_{xx}}\sum_{i=1}^n (x_i - \overline{x}_n)\,\mathsf{Cov}\,(\overline{Y}_n, Y_i), \\ \mathsf{Cov}\,(\overline{Y}_n, Y_i) &= \mathsf{Cov}\left(\frac{1}{n}\sum_{j=1}^n Y_j, Y_i\right) = \frac{1}{n}\sum_{j=1}^n \mathsf{Cov}\,(Y_j, Y_i) = \frac{1}{n}\,\mathsf{Cov}\,(Y_i, Y_i) = \frac{1}{n}\,\mathsf{Var}\,Y_i = \frac{\sigma^2}{n},\end{aligned}$$

neboť díky nekorelovanosti veličin Y_i víme, že $\mathsf{Cov}\,(Y_j, Y_i) = 0$ pro $j \neq i$. Dosazením zpět následně získáme

$$\mathsf{Cov}\,(\overline{Y}_n, \widehat{\beta}_1) = \frac{\sigma^2}{nS_{xx}}\sum_{i=1}^n (x_i - \overline{x}_n) = 0.$$

4) Protože

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

jsou odhady $\widehat{\beta}_0$ i $\widehat{\beta}_1$ lineární kombinací nezávislých normálních náhodných veličin Y_i. Z toho vyplývá, že mají normální rozdělení a příslušné střední hodnoty a rozptyly jsme spočetli výše. $\square$

Věta 1.2. Za předpokladu předchozí věty platí

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

tedy s_n^2 je nestranný odhad σ^2.

Důkaz. Dosadíme za statistiku s_n^2

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

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

$$\mathsf{E}\,(Y_i - \widehat{Y}_i)^2 = \mathsf{Var}\,(Y_i - \widehat{Y}_i),$$

neboť
$$\mathsf{Var}\,(Y_i-\widehat{Y}_i)=\mathsf{E}\,(Y_i-\widehat{Y}_i)^2-\underbrace{\Big(\mathsf{E}\,(Y_i-\widehat{Y}_i)\Big)^2}_{=0}.$$
Dostáváme tak
$$A=\sum_{i=1}^n \mathsf{Var}\,(Y_i-\widehat{Y}_i)=\sum_{i=1}^n\big[\mathsf{Var}\,(Y_i)+\mathsf{Var}\,(\widehat{Y}_i)-2\,\mathsf{Cov}\,(Y_i,\widehat{Y}_i)\big]=$$
$$=n\sigma^2+\sum_{i=1}^n \mathsf{Var}\,(\widehat{Y}_i)-2\sum_{i=1}^n \mathsf{Cov}\,(Y_i,\widehat{Y}_i). \tag{1.4}$$
Pro rozptyl $\widehat{Y}_i$ platí
$$\mathsf{Var}\,\widehat{Y}_i=\mathsf{Var}\,(\widehat{\beta}_0+\widehat{\beta}_1x_i)=\mathsf{Var}\,\widehat{\beta}_0+x_i^2\,\mathsf{Var}\,\widehat{\beta}_1+2x_i\,\mathsf{Cov}\,(\widehat{\beta}_0,\widehat{\beta}_1),$$
kde
$$\mathsf{Cov}\,(\widehat{\beta}_0,\widehat{\beta}_1)=\mathsf{Cov}\,(\overline{Y}_n-\widehat{\beta}_1\overline{x}_n,\widehat{\beta}_1)=\underbrace{\mathsf{Cov}\,(\overline{Y}_n,\widehat{\beta}_1)}_{=0\text{ (viz. dříve)}}-\overline{x}_n\underbrace{\mathsf{Var}\,(\widehat{\beta}_1)}_{\frac{\sigma^2}{S_{xx}}}=-\frac{\sigma^2\overline{x}_n}{S_{xx}}, \tag{1.5}$$
a tedy
$$\mathsf{Var}\,\widehat{Y}_i=\sigma^2\left[\frac{1}{n}+\frac{\overline{x}_n^2}{S_{xx}}+x_i^2\frac{1}{S_{xx}}-\frac{2x_i\overline{x}_n}{S_{xx}}\right]=\sigma^2\left[\frac{1}{n}+\frac{(x_i-\overline{x}_n)^2}{S_{xx}}\right].$$
Pro první sumu ve vztahu (1.4) tedy platí
$$\sum_{i=1}^n \mathsf{Var}\,\widehat{Y}_i=\sigma^2+\frac{\sigma^2}{S_{xx}}\underbrace{\sum_{i=1}^n(x_i-\overline{x}_n)^2}_{=S_{xx}}=2\sigma^2.$$
Pro kovarianci veličin $Y_i,\widehat{Y}_i$ platí
$$\mathsf{Cov}\,(Y_i,\widehat{Y}_i)=\mathsf{Cov}\,(Y_i,\widehat{\beta}_0+\widehat{\beta}_1x_i)=\mathsf{Cov}\,(Y_i,\widehat{\beta}_0)+x_i\,\mathsf{Cov}\,(Y_i,\widehat{\beta}_1).$$
S využitím vztahů (1.2) dostaneme
$$\mathsf{Cov}\,(Y_i,\widehat{\beta}_1)=\frac{1}{S_{xx}}\sum_{j=1}^n(x_j-\overline{x}_n)\underbrace{\mathsf{Cov}\,(Y_i,Y_j)}_{=0\text{ pro }i\neq j}=\frac{\sigma^2(x_i-\overline{x}_n)}{S_{xx}},$$
$$\mathsf{Cov}\,(Y_i,\widehat{\beta}_0)=\mathsf{Cov}\,(Y_i,\overline{Y}_n-\overline{x}_n\widehat{\beta}_1)=\mathsf{Cov}\,(Y_i,\overline{Y}_n)-\overline{x}_n\,\mathsf{Cov}\,(Y_i,\widehat{\beta}_1)=\frac{\sigma^2}{n}-\frac{\overline{x}_n\sigma^2(x_i-\overline{x}_n)}{S_{xx}}.$$
Dosazením získáme
$$\mathsf{Cov}\,(Y_i,\widehat{Y}_i)=\frac{\sigma^2}{n}-\frac{\overline{x}_n\sigma^2(x_i-\overline{x}_n)}{S_{xx}}+\frac{x_i\sigma^2(x_i-\overline{x}_n)}{S_{xx}}=\frac{\sigma^2}{n}+\frac{\sigma^2}{S_{xx}}(x_i-\overline{x}_n)^2$$
a pro druhou sumu ve vztahu (1.4) tak platí
$$\sum_{i=1}^n \mathsf{Cov}\,(Y_i,\widehat{Y}_i)=\sigma^2+\frac{\sigma^2}{S_{xx}}\sum_{i=1}^n(x_i-\overline{x}_n)^2=2\sigma^2.$$

Dosazením do (1.4) dostaneme

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

a celkově tedy

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

□

Tvrzení 1.1. Nechť platí předpoklady věty 1.1 a nechť $e_1, ..., e_n$ jsou *iid* $N(0, \sigma^2)$. Potom platí:

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

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

Důkaz. Vyplyne z obecnějšího tvrzení pro vícerozměrnou regresi. □

Poznámka 1.5. Spočetli jsme

$$\sigma^2(\widehat{\beta}_0) \overset{ozn.}{=\!=} \mathsf{Var}\,(\widehat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}_n^2}{S_{xx}}\right], \tag{1.6}$$

$$\sigma^2(\widehat{\beta}_1) \overset{ozn.}{=\!=} \mathsf{Var}\,(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}. \tag{1.7}$$

Nestranné odhady těchto rozptylů získáme dosazením odhadu s_n^2 za σ^2,

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

kde δ_0 a δ_1 jsou tzv. *variance multiplication factors*. Odhady směrodatné odchylky veličin $\widehat{\beta}_0$ a $\widehat{\beta}_1$ pak jsou

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

říká se jim standardní chyby odhadů $\widehat{\beta}_0$ a $\widehat{\beta}_1$, a budou hrát zásadní roli při konstrukci intervalů spolehlivosti a testů hypotéz.

1.2 Gauss - Markov theorem

Pokud mají chyby normální rozdělení, pak odhady metodou nejmenších čtverců (LSE) $\widehat{\beta}_0, \widehat{\beta}_1$ parametrů β_0, β_1 jsou zároveň maximálně věrohodné odhady těchto parametrů, o kterých víme, že jsou eficientní. Otázkou je, jaké je opodstatnění použití LSE, pokud chyby nejsou normální. Ukážeme, že LSE jsou BLUE (best linear unbiased estimators), tedy lineární nestranné odhady s minimálním rozptylem. Je ale třeba poznamenat, že můžou existovat nelineární nebo vychýlené odhady parametrů β_0, β_1, které jsou eficientnější než LSE, pokud se rozdělení chyb liší výrazně od normálního (tím se zabývá robustní regresní analýza). Začneme definicí lineárního odhadu parametru.

Definice 1.1. Lineární odhad parametru β, založený na pozorováních $Y_1, \ldots, Y_n$, je statistika tvaru

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

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

Uvažujme opět model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i \in \widehat{n}. \tag{1.8}$$

Věta 1.3 (Gauss-Markov theorem)**.** Nechť $e_1, ..., e_n$ v modelu (1.8) jsou nekorelované, mají stejný rozptyl $\mathsf{Var}\,(e_i) = \sigma^2$ a $\mathsf{E}\, e_i = 0$, $i = 1, \ldots, n$. Potom LSE $\widehat{\beta}_j$, $j \in \{0, 1\}$, je BLUE parametru β_j.

Důkaz. Ukážeme pro β_1, pro β_0 je důkaz podobný. Nechť tedy $\widehat{\beta}_1 = \sum\limits_{i=1}^{n} c_i Y_i$, pak

$$\mathsf{Var}\, \widehat{\beta}_1 = \sum_{i=1}^{n} c_i^2\, \mathsf{Var}\, Y_i = \sigma^2 \sum_{i=1}^{n} c_i^2.$$

Aby byl odhad $\widehat{\beta}_1$ nestranný, musí platit $\mathsf{E}\, \widehat{\beta}_1 = \beta_1$, tedy

$$\mathsf{E}\, \widehat{\beta}_1 = \sum_{i=1}^{n} c_i\, \mathsf{E}\, Y_i = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \overset{!}{=} \beta_1.$$

To musí platit pro libovolná β_0, β_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 β_1 s minimálním rozptylem je tedy redukováno na minimalizaci $\sum\limits_{i=1}^{n} c_i^2$ za vazebných podmínek $\sum\limits_{i=1}^{n} c_i = 0$ a $\sum\limits_{i=1}^{n} c_i x_i = 1$. 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 λ_j je zde kvůli snazšímu výpočtu, ale není nutná) a její derivace položíme rovny nule,

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

Sečtením prvních n rovnic získáme

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

Sečtením prvních n rovnic vynásobených x_i, dosazením $\sum_{i=1}^{n} c_i x_i = 1$ a $\lambda_1 = -\lambda_2 \overline{x}_n$ dostaneme postupně

$$\begin{aligned}
\sum_{i=1}^{n} c_i x_i - \lambda_1 \sum_{i=1}^{n} x_i - \lambda_2 \sum_{i=1}^{n} x_i^2 &= 0, \\
\lambda_1 \sum_{i=1}^{n} x_i + \lambda_2 \sum_{i=1}^{n} x_i^2 &= 1, \\
-\lambda_2 \overline{x}_n \cdot n\overline{x}_n + \lambda_2 \sum_{i=1}^{n} x_i^2 &= 1, \\
\lambda_2 \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}_n^2 \right) &= 1,
\end{aligned}$$

a tedy

$$\lambda_2 = \frac{1}{S_{xx}} \quad \text{a} \quad \lambda_1 = -\frac{\overline{x}_n}{S_{xx}}.$$

Dosadíme za λ_1, λ_2 do rovnice pro derivaci Lagrangeovy funkce podle c_i a dostaneme

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

Z výsledného tvaru odhadu $\widehat{\beta}_1$ vidíme, že lineární nestranný odhad s minimálním rozptylem je LSE (viz vztah (1.3)). □

Poznámka 1.6. Ukázali jsme pouze, že to je stacionární bod Lagrangeovy funkce, že je tam i minimum, ukážeme v obecnější větě ve vícerozměrné regresi.

1.3 Intervaly spolehlivosti pro β_0, β_1

Intervaly spolehlivosti (IS) poskytují 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 paramatrů β jsou lineární funkcí $Y_i, i \in \widehat{n}$, což umožňuje aplikovat centrální limitní větu a dostat asymptotickou normalitu odhadů $\widehat{\beta}_0, \widehat{\beta}_1$.

Uvažujme model $Y_i = \beta_0 + \beta_1 x_i + e_i$, kde e_i jsou *iid* $N(0, \sigma^2)$. Už víme, že

$$\widehat{\beta}_i \sim N\left(\beta_i, \sigma^2(\widehat{\beta}_i)\right), \quad \frac{(n-2)s_n^2}{\sigma^2} \sim \chi^2(n-2) \text{ a nezávisí na } \widehat{\beta}_0, \widehat{\beta}_1. \tag{1.9}$$

Poznámka 1.7. Připomeňme definici Studentova rozdělení. Nechť náhodné veličiny X, Y jsou nezávislé a platí $X \sim N(0,1)$ a $Y \sim \chi^2(n)$. Potom říkáme, že náhodná veličina T,

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n),$$

má Studentovo rozdělení s n stupni volnosti.

Z vlastností (1.9) a definice Studentova rozdělení vyplývá, že

$$T_i := \frac{\frac{\widehat{\beta}_i - \beta_i}{\sigma(\widehat{\beta}_i)}}{\frac{s_n}{\sigma}} = \frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}(\widehat{\beta}_i)} \sim t(n-2), \quad i \in \{0,1\}, \tag{1.10}$$

neboť $\sigma(\widehat{\beta}_i) = \sigma\sqrt{\delta_i}$ a $\widehat{\sigma}(\widehat{\beta}_i) = s_n\sqrt{\delta_i}$. Díky vlastnostem α-kvantilů to znamená, že

$$\mathsf{P}\left[-t_{1-\frac{\alpha}{2}}(n-2) \leqslant \frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}(\widehat{\beta}_i)} \leqslant t_{1-\frac{\alpha}{2}}(n-2)\right] = 1-\alpha.$$

Vyjádřením β_i dostaneme

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

a tedy $\left(\widehat{\beta}_i \pm t_{1-\frac{\alpha}{2}}(n-2)\widehat{\sigma}(\widehat{\beta}_i)\right)$ je $100(1-\alpha)\%$ interval spolehlivosti pro β_i, $i \in \{0,1\}$.

Dosazením konkrétních tvarů standardních chyb $\widehat{\sigma}(\widehat{\beta}_i)$ pak dostaneme následující intervaly spolehlivosti pro jednotlivé parametry

- $100(1-\alpha)\%$ IS pro β_0: $\widehat{\beta}_0 \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot s_n\sqrt{\frac{1}{n} + \frac{\overline{x}_n^2}{S_{xx}}}$,
- $100(1-\alpha)\%$ IS pro β_1: $\widehat{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot s_n\frac{1}{\sqrt{S_{xx}}}$.

Poznámka 1.8. Z tvarů intervalů spolehlivosti lze pozorovat, že interval spolehlivosti pro β_0 bude ve většině praktických případů širší než interval spolehlivosti pro β_1, tzn. směrnice je obecně odhadnuta s větší přesností než absolutní člen (intercept).

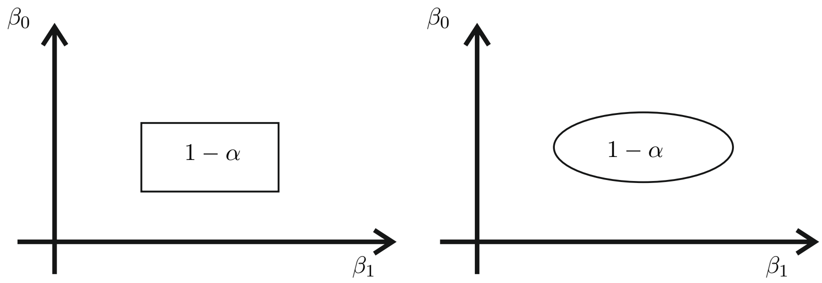
Poznámka 1.9. Někdy se konstruují simultánní intervaly spolehlivosti pro oba parametry. Ty mohou mít pro náš model např. tvar obdélníku nebo elipsy, viz obrázek 1.3. Podrobněji se o tom zmíníme u vícerozměrné regrese.

1.4 Testování hypotéz pro β_0, β_1

Chtěli bychom ověřit platnost předpokladu lineárního vztahu mezi x a y. Předpokládejme nyní, že model je lineární, a že x je jediná dostupná vysvětlující proměnná. Otázkou 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 proměnná x nevysvětluje nic z variability y a není v modelu významná. Pokud zamítneme H_0, znamená to, že proměnná x v modelu významná je.



Obrázek 1.3: Možné tvary simultánních intervalů spolehlivosti

Poznámka 1.10. Tyto závěry jsou správné pouze za předpokladu, že je model lineární!

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

Pro konstrukci testů využijeme odvozené tvary intervalů spolehlivosti. Připomeňme z matematické statistiky, že pokud máme testovat hypotézu $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ na základě náhodného výběru $X_1, \dots, X_n$ a máme k dispozici $100(1-\alpha)\%$ interval spolehlivosti pro θ, pak množina $W = \{\boldsymbol{x} | \theta_0 \notin (\underline{\theta}, \overline{\theta})\}$ je kritický obor testu dané hypotézy na hladině α.
V našem případě hypotézu $H_0 : \beta_1 = 0$ tedy zamítneme, pokud $0 \notin \left(\widehat{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot \frac{s_n}{\sqrt{S_{xx}}}\right)$, tzn.

$$\begin{aligned} \text{buď } \widehat{\beta}_1 + t_{1-\frac{\alpha}{2}}(n-2) \cdot \frac{s_n}{\sqrt{S_{xx}}} < 0 \quad &\Leftrightarrow \quad \widehat{\beta}_1 \frac{\sqrt{S_{xx}}}{s_n} < -t_{1-\frac{\alpha}{2}}(n-2) \\ \text{nebo } \widehat{\beta}_1 - t_{1-\frac{\alpha}{2}}(n-2) \cdot \frac{s_n}{\sqrt{S_{xx}}} > 0 \quad &\Leftrightarrow \quad \widehat{\beta}_1 \frac{\sqrt{S_{xx}}}{s_n} > t_{1-\frac{\alpha}{2}}(n-2). \end{aligned}$$

Zapsáno dohromady, hypotézu $H_0 : \beta_1 = 0$ zamítneme pokud

$$|T_1| = |\widehat{\beta}_1| \frac{\sqrt{S_{xx}}}{s_n} > t_{1-\frac{\alpha}{2}}(n-2).$$

Poznámka 1.11. Intuitivní interpretace: $|T_1| = |\widehat{\beta}_1| \frac{\sqrt{S_{xx}}}{s_n} = \frac{|\widehat{\beta}_1|}{\widehat{\sigma}(\widehat{\beta}_1)}$ je převrácená hodnota relativní chyby $\frac{\widehat{\sigma}(\widehat{\beta}_1)}{|\widehat{\beta}_1|}$. Pokud je β_1 dobře odhadnuto, očekáváme malý rozptyl $\widehat{\sigma}(\widehat{\beta}_1)$ vzhledem k absolutní hodnotě β_1, tedy hodnota statistiky T_1 bude velká. Odvozený t-test tedy říká, že zamítneme H_0, pokud je relativní chyba odhadu malá.

Poznámka 1.12. Někdy dopředu známe kandidáta b_1 jako hodnotu parametru β_1 a chtěli bychom testovat $H_0 : \beta_1 = b_1$ vs. $H_1 : \beta_1 \neq b_1$. Test bude zamítnout H_0, pokud

$$\left|\widehat{\beta}_1 - b_1\right| \cdot \frac{\sqrt{S_{xx}}}{s_n} > t_{1-\frac{\alpha}{2}}(n-2).$$

Věnujme se ještě krátce testu 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 $y = \beta_0 + \beta_1 x + e$. S využitím příslušného intervalu spolehlivosti dostaneme stejným postupem jako výše následující pravidlo: zamítnout H_0 pokud

$$|T_0| = \frac{|\widehat{\beta}_0|}{\widehat{\sigma}(\widehat{\beta}_0)} = |\widehat{\beta}_0| \frac{1}{s_n \sqrt{\frac{1}{n} + \frac{\overline{x}_n^2}{S_{xx}}}} > t_{1-\frac{\alpha}{2}}(n-2).$$

1.5 ANOVA přístup pro testování

Odvodili jsme t-test významnosti regresních koeficientů, nyní odvodíme ekvivalentní F-test, který může být zobecněn na test celkové významnosti vícerozměrného regresního modelu (testy významnosti jednotlivých koeficientů mohou být totiž zavádějící).
Myšlenkou metody (analýza rozptylu ANOVA) je určit, kolik variability v pozorováních $(y_1, y_2, ..., y_n)$ je „vysvětleno“ regresním modelem (přímkou). Míru variability v datech budeme vyjadřovat pomocí tzv. celkového součtu čtverců (*total sum of squares*) ve tvaru

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

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

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

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

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

kde poslední suma se nazývá regresní součet čtverců (*regression sum of squares*). Podíl

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

tak vyjadřuje podíl variability v $(y_1, ..., y_n)$ vysvětlené regresním modelem. Statistika R^2 se nazývá **koeficient determinace** *(coefficient of determination)* a z její definice plyne, že pro dobrý model by měla být hodnota $R^2 \approx 1$.
Ukážeme, že R^2 je druhá mocnina výběrového korelačního koeficientu mezi vektory $\boldsymbol{x} = (x_1, \ldots, x_n)$ a $\boldsymbol{y} = (y_1, \ldots, y_n)$, což dává statistice R^2 význam míry „dobré shody“.
Pokud bychom znali rozdělení pravděpodobnosti statistiky R^2, nabízí se její použití pro test $H_0 : \ \beta_1 = 0$, kterou bychom zamítli, pokud bude $R^2 \approx 1$. Protože každá monotonní funkce R^2 vede na ekvivalentní test, budeme uvažovat statistiku (proč uvidíme z následující věty)

$$F = \frac{(n-2)R^2}{1-R^2}. \tag{1.11}$$

Věta 1.4. Předpokládejme, že v modelu (1.8) je splněno $SST \neq 0$. Potom platí

1) $0 \leqslant R^2 \leqslant 1$,
2) $R^2 = 1 - \frac{SSE}{SST}$, kde $SSE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ je reziduální součet čtverců (*sum of squared errors*),
3) $R^2 = 1 \iff \widehat{y}_i = y_i, \forall i \in \widehat{n}$, (tj. všechna pozorování leží na přímce),
4) pokud označíme $\boldsymbol{x} = (x_1, ..., x_n)$ a $\boldsymbol{y} = (y_1, ..., y_n)$, potom

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

tedy R^2 je druhá mocnina výběrového korelačního koeficientu vektorů $\boldsymbol{x}, \boldsymbol{y}$,
5) $F = \frac{(n-2)R^2}{1-R^2} = \frac{SSR}{s_n^2} = T_1^2$, kde T_1 je statistika definovaná ve vztahu (1.10) za předpokladu $H_0: \ \beta_1 = 0$,
6) pokud jsou chyby $e_1, ..., e_n$ *iid* $N(0, \sigma^2)$ a platí $H_0: \ \beta_1 = 0$, potom $F \sim F(1, n-2)$.

Poznámka 1.13. 1. Z bodů 5) a 6) vyplývá, že použití libovolné statistiky T_1, 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.

Důkaz věty bude založen na rozkladu

$$\sum_{i=1}^{n}(y_i - \overline{y}_n)^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y}_n)^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2, \quad \text{neboli} \quad SST = SSR + SSE,$$

pro jehož ověření budeme potřebovat následující pomocné tvrzení.

Lemma 1.1. Nechť $\widehat{e}_i = y_i - \widehat{y}_i$ značí rezidua, kde $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ a $\widehat{\beta}_0, \widehat{\beta}_1$ jsou LSE v modelu (1.8). Potom platí

$$\sum_{i=1}^{n}\widehat{e}_i = 0, \qquad \overline{\widehat{y}}_n = \overline{y}_n, \qquad \sum_{i=1}^{n}\widehat{e}_i\widehat{y}_i = 0.$$

Důkaz. Připomeňme, že LSE odhady $(\widehat{\beta}_0, \widehat{\beta}_1)$ jsou získány minimalizací funkce $S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$, jsou tedy stacionárním bodem této funkce.
1) Z rovnice $\frac{\partial S}{\partial \beta_0}(\widehat{\beta}_0, \widehat{\beta}_1) = 0$ dostaneme

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

2) Z předchozí rovnosti plyne, že $\sum_{i=1}^{n}\widehat{y}_i = \sum_{i=1}^{n} y_i$, podělením n tedy dostaneme druhé dokazované tvrzení.
3) Z rovnice $\frac{\partial S}{\partial \beta_1}(\widehat{\beta}_0, \widehat{\beta}_1) = 0$ dostaneme

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

a tedy

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

□

Důkaz Věty 1.4. Ukažme nejdříve platnost rozkladu $SST = SSR + SSE$. Z lemmatu 1.1 vyplývá, že

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

neboť

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

Teď přejdeme k důkazům jednotlivých tvrzení věty.
1) Díky rozkladu $SST = SSE + SSR$ a nezápornosti součtu čtverců platí

$$0 \leqslant R^2 = \frac{SSR}{SST} \leqslant \frac{SST}{SST} = 1.$$

2) Protože $SSR = SST - SSE$, dostaneme

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

3) Z bodu 2) plyne, že $R^2 = 1$ právě tehdy, když $SSE = 0$. Dále pak

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = 0 \ \Leftrightarrow \ y_i = \widehat{y}_i, \ \forall i \in \widehat{n}.$$

4) Díky vztahu $\widehat{\beta}_0 = \overline{y}_n - \widehat{\beta}_1 \overline{x}_n$ platí $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i = \overline{y}_n - \widehat{\beta}_1 \overline{x}_n + \widehat{\beta}_1 x_i = \overline{y}_n + \widehat{\beta}_1 (-x_i - \overline{x}_n)$. Proto pak

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

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

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

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

5) Z definice statistiky F (viz (1.11)) a statistiky R^2 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 $T_1 = \widehat{\beta}_1 \frac{\sqrt{S_{xx}}}{s_n}$, dostaneme

$$T_1^2 = \frac{\widehat{\beta}_1^2 S_{xx}}{s_n^2} = \frac{SSR}{s_n^2} = F.$$

6) Ze vztahu (1.10) víme, že za platnosti $H_0: \ \beta_1 = 0$ je splněno $T_1 \sim t(n-2)$. Tvrzení $F = T_1^2 \sim F(1, n-2)$ pak plyne přímo z vlastností Studentova a Fischerova rozdělení známých z teorie pravděpodobnosti. □

Výsledky se většinou uvádí ve formě tabulky analýzy rozptylu, jejíž tvar je prezentován v Tabulce 1.1. V této formě tabulky *Source* označuje zdroj součtu čtverců (regresní, reziduální

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

Tabulka 1.1: Tabulka ANOVA

a celkový), *df* počet stupňů volnosti příslušný danému součtu čtverců, *SS* součet čtverců a *MS* ($\mathrm{MS} = \frac{\mathrm{SS}}{\mathrm{df}}$) průměr čtverců (mean squares).

Poznámka 1.14. Hypotéza $H_0: \beta_1 = 0$ je zamítnuta, pokud $F > F_{1-\alpha}(1, n-2)$. V tomto jednorozměrném případě je to ekvivalentní t-testu, neboť $F = \mathrm{T}^2$.

Věta 1.5. Nechť $e_1, ..., e_n$ jsou *iid* $N(0, \sigma^2)$. Za platnosti $H_0: \ \beta_1 = 0$ je splněno, že

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

Poznámka 1.15. Proto se v tabulce ANOVA 1.1 uvádí stupně volnosti po řadě $1, n-2, n-1$. Používají se však i v případě jiného rozdělení chyb. Vysvětlit si to lze takto:

1. $SSE = \sum_{i=1}^{n} \widehat{e}_i^2$, na n reziduí $\widehat{e}_1, ..., \widehat{e}_n$ máme 2 podmínky, $\sum_{i=1}^{n} \widehat{e}_i = 0$ a $\sum_{i=1}^{n} x_i \widehat{e}_i = 0$. To znamená, že zbývá $n-2$ stupňů volnosti.
2. $SST = \sum_{i=1}^{n} (y_i - \overline{y}_n)^2$, protože $y_1, \ldots, y_n$ musí splňovat $\sum_{i=1}^{n} (y_i - \overline{y}_n) = 0$, máme jednu podmínku a zbývá $n-1$ stupňů volnosti.
3. $SSR = SST - SSE$ a počet stupňů volnosti je roven $(n-1) - (n-2) = 1$.

Důkaz. V důkazu věty 1.4 jsme ukázali, že $SSR = \widehat{\beta}_1^2 S_{xx}$, takže $\frac{SSR}{\sigma^2} = \left(\frac{\widehat{\beta}_1 \sqrt{S_{xx}}}{\sigma}\right)^2$. Dále víme, že $\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$, a tedy $(\widehat{\beta}_1 - \beta_1)\frac{\sqrt{S_{xx}}}{\sigma} \sim N(0,1)$. Za platnosti $H_0: \beta_1 = 0$ tedy s pomocí definice χ^2-rozdělení dostaneme

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

Z tvrzení 1.1 plyne, že $\frac{SSE}{\sigma^2} = \frac{(n-2)s_n^2}{\sigma^2} \sim \chi^2(n-2)$ a nezávisí na $\widehat{\beta}_1$. Z toho vyplývá, že $\frac{SSR}{\sigma^2}$ a $\frac{SSE}{\sigma^2}$ jsou nezávislé. Protože ale

$$\frac{SST}{\sigma^2} = \frac{SSR}{\sigma^2} + \frac{SSE}{\sigma^2},$$

tedy $\frac{SST}{\sigma^2}$ je součtem dvou nezávislých χ^2-rozdělení s jedním a $n-2$ stupni volnosti, dostaneme

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

□

Poznámka 1.16. Při použití R^2 statistiky je třeba dát pozor na příliš zjednodušené hodnocení kvality modelu.
Nízké hodnoty R^2 nemusí znamenat, že regresní model není významný. V datech jen může být velké množství nevysvětlitelné náhodné variability. Například opakované hodnoty regresoru x snižují hodnotu R^2 oproti modelům s různými x.
Naopak velké hodnoty R^2 mohou být způsobeny velkým měřítkem dat (S_{xx} je velké). Platí totiž, že

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

což je rostoucí funkce S_{xx}. Velký rozptyl hodnot $x_1, ..., x_n$ tak může mít za následek velkou hodnotu R^2 a přitom nic neříká o kvalitě modelu.
Střední hodnota $\mathsf{E}\,(R^2)$ je také rostoucí funkcí β_1^2. Modely s „velkou“ směrnicí tedy budou mít obecně větší hodnotu statistiky R^2, než modely s „malou“ směrnicí.
Při hodnocení kvality modelů budeme tedy potřebovat více kritérií. Mezi ně patří například

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

Další použitelná kritéria uvedeme později.

Příklad 1.2. Velká hodnota R^2 indikuje přibližně lineární vztah mezi x a y, ale vysoký stupeň korelace nemusí znamenat příčinný vztah. Uvažujme například data z let 1924-1937, kde veličina y_i vyjadřuje počet mentálních onemocnění na 100000 obyvatel Anglie a veličina x_i počet rádií v populaci. Pokud naladíme lineární model $y_i = \beta_0 + \beta_1 x_i + e_i$, dostaneme následující odhady parametrů β_0, β_1 a hodnotu R^2 statistiky

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

Zjišťujeme tedy velmi významný lineární vztah mezi x a y. Závěr by mohl být, že rostoucí počet rádií způsobuje nárůst mentálních 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 a lepší diagnostické procedury umožňovaly identifikovat více lidí s mentálními problémy.

Poznámka 1.17. (korelace × příčinnost)
Uveďme schematicky různé typy příčinnosti a její vztah ke korelovanosti.

a)

$$x \longrightarrow y$$
$$x \longleftarrow y$$

Causal link (příčinná spojitost)

I když je příčinná spojitost mezi x a y, korelace samotná nám neřekne, zda x ovlivňuje y nebo naopak.

b)

$$\begin{tikzcd} & z \arrow[dl] \arrow[dr] & \\ x & & y \end{tikzcd}$$

Hidden cause (skrytá příčinnost)

Skrytá veličina z ovlivňuje x i y, což způsobuje jejich korelovanost.

c)

$$\begin{tikzcd} & z \arrow[dr] & \\ x \arrow[rr] & & y \end{tikzcd}$$

Confounding factor (zavádějící faktor)

Skrytá proměnná z i x ovlivňují y, výsledky tedy závisí i na z.

d)

$$x \qquad y$$

Coincidence (shoda okolností)

Korelace je náhodná.

1.6 Regrese skrz počátek

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

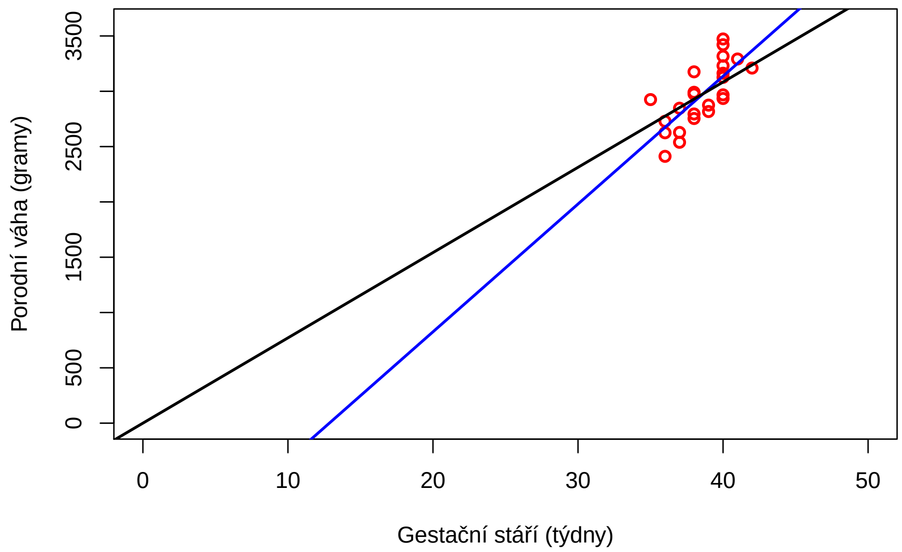
$$Y_i = \beta_1 x_i + e_i, \;\; i \in \hat{n}.$$

Například na základě fyzikálních úvah je předem známo, že

$$\mathsf{E}\left[Y_0\right] = \beta_0 = 0.$$

Potom tedy nemá smysl odhadovat β_0, protože to obecně sníží přesnost odhadu σ^2, a tedy i β_1. Nebo na začátku předpokládáme, že $\beta_0 \neq 0$ a t-test nezamítne hypotézu $\mathrm{H}_0 : \beta_0 = 0$, potom může být β_0 z modelu odstraněno. V této sekci se tedy budeme krátce zabývat modely bez interceptu.

Poznámka 1.18. 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 β_0 apriorně může být chybné, i když $\mathsf{E}\left[Y_0\right] = 0$. Pokud totiž nevíme jistě, že model je lineární na okolí 0, volba $\beta_0 = 0$ může vést k vychýleným odhadům β_1, pokud jsou nezávislé proměnné daleko od $x = 0$. Na obrázku 1.4 je zobrazena závislost porodní váhy dítěte na gestačním stáří a příslušné modely lineární regrese s interceptem a bez interceptu. I když je zřejmé, že v čase nula bude hmotnost dítěte nulová, model na okolí nuly být lineární nemusí. Naopak lze předpokládat, že hmotnost nejdříve roste výrazně pomaleji než těsně před porodem. Model procházející počátkem tedy povede na vychýlený odhad parametru β_1. Poznamenejme ještě, že funkce `lm()` v prostředí R vrátí pro model procházející počátkem hodnotu $R^2 = 0{,}99$ a pro model s interceptem $R^2 = 0{,}55$, i když jsou oba modely srovnatelné z hlediska součtu čtverců chyb. Proč tomu tak je, nastíníme dále v této sekci.



Obrázek 1.4: Modely lineární regrese bez interceptu a s interceptem

1.6.1 Odhady a testy v případě $\beta_0 = 0$

LSE odhad parametru β_1 dostaneme minimalizací $S = \sum\limits_{i=1}^{n}(y_i - \beta_1 x_i)^2$ ve tvaru

$$\widehat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} y_i x_i}{\sum\limits_{i=1}^{n} x_i^2}.$$

Pokud jsou $e_1, ..., e_n$ *iid* $N(0, \sigma^2)$, potom z vlastností

$$\mathsf{E}\,[\widehat{\beta}_1] = \beta_1 \quad \text{a} \quad \mathsf{Var}\,[\widehat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

vyplyne, že

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right) \quad \text{a} \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{SSE}{n-1}$$

je nestranný odhad σ^2. Dále $\frac{SSE}{\sigma^2} \sim \chi^2(n-1)$ a nezávisí na $\widehat{\beta}_1$. Hypotézu $\mathrm{H}_0 : \beta_1 = 0$ lze otestovat za pomoci statistiky

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

a $100(1-\alpha)\%$ interval spolehlivosti pro β_1 je

$$\left(\widehat{\beta}_1 \pm \mathrm{t}_{1-\frac{\alpha}{2}}(n-1)\frac{s_n}{\sqrt{\sum x_i^2}}\right).$$

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

Věta 1.6. V modelu s $\beta_0 = 0$ platí, že

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

Důkaz. Jednoduchou úpravou dostaneme

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

zbývá tedy ukázat, že poslední člen je roven 0. Z rovnice $\frac{\mathrm{d}S}{\mathrm{d}\beta_1}(\widehat{\beta}_1) = 0$ dostaneme $\sum\limits_{i=1}^{n}(y_i - \widehat{\beta}_1 x_i)x_i = 0$ a díky rovnosti $\widehat{y}_i = \widehat{\beta}_1 x_i$ získáme, po vynásobení obou stran rovnic $\widehat{\beta}_1$,

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

□

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

$$R^2 = \frac{\sum\limits_{i=1}^{n} \widehat{y}_i^2}{\sum\limits_{i=1}^{n} y_i^2} \quad \text{a tedy} \quad 1 - R^2 = \frac{\sum\limits_{i=1}^{n} y_i^2 - \sum\limits_{i=1}^{n} \widehat{y}_i^2}{\sum\limits_{i=1}^{n} y_i^2} = \frac{\sum\limits_{i=1}^{n} \widehat{e}_i^2}{\sum\limits_{i=1}^{n} y_i^2}.$$

Definujme $F := \frac{(n-1)R^2}{1-R^2}$, potom

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

Vztah mezi statistikami R^2, F a T^2 je tedy stejný jako v případě modelu s $\beta_0 \neq 0$.

Poznámka 1.19. Tato definice R^2 statistiky se ale v praxi moc nepoužívá, protože neumožňuje přímé srovnání modelů s interceptem a bez něj. Připomeňme tvar R^2 pro oba modely

$$\beta_0 = 0: \quad R^2 = 1 - \frac{SSE}{\sum\limits_{i=1}^{n} y_i^2}, \qquad \beta_0 \neq 0: \quad R^2 = 1 - \frac{SSE}{\sum\limits_{i=1}^{n}(y_i - \overline{y}_n)^2}.$$

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

Definice vhodné R^2 statistiky pro případ modelu s $\beta_0 = 0$ vyvolává jistou kontroverzi a existuje několik verzí. Možná volba je $R^2 = (\varrho(\boldsymbol{y}, \widehat{\boldsymbol{y}}))^2$, kde $\widehat{\boldsymbol{y}} = (\widehat{y}_1, ..., \widehat{y}_n)$, protože tento vztah platí i pro případ $\beta_0 \neq 0$. Další možnost je srovnat modely pomocí hodnot s_n^2 (preferujeme model s nejnižší hodnotou s_n^2).
Tvar tabulky ANOVA pro případ modelu s $\beta_0 = 0$ je prezentován v tabulce 1.2.

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

Tabulka 1.2: Tabulka ANOVA pro model s $\beta_0 = 0$.

1.7 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 $\mu_0 = \mathsf{E}\, Y_0$ v bodě x_0,
2) predikce hodnoty nového pozorování Y_0 v bodě x_0.

Pro oba typy použijeme bodový odhad

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0,$$

intervalové odhady se ale budou lišit.

Ad 1)

Protože je $\mu_0 = \beta_0 + \beta_1 x_0$ vlastně parametr, lze pro něj odvodit intervaly spolehlivosti (za předpokladu normality chyb). Za tímto účelem spočteme $\mathsf{E}\, \widehat{Y}_0$ a $\mathsf{Var}\,(\widehat{Y}_0)$. Pro střední hodnotu platí, díky nestrannosti odhadů $\widehat{\beta}_0, \widehat{\beta}_1$,

$$\mathsf{E}\, \widehat{Y}_0 = \mathsf{E}\,(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0 = \mu_0\,.$$

Dosazením formule pro odhad $\widehat{\beta}_0$ dostaneme $\widehat{Y}_0 = \overline{Y}_n + \widehat{\beta}_1(x_0 - \overline{x}_n)$ a

$$\begin{aligned} \mathsf{Var}\, \widehat{Y}_0 &= \mathsf{Var}\,(\overline{Y}_n) + (x_0 - \overline{x}_n)^2\, \mathsf{Var}\,(\widehat{\beta}_1) + 2(x_0 - \overline{x}_n)\, \mathsf{Cov}\,(\overline{Y}_n, \widehat{\beta}_1) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \overline{x}_n)^2}{S_{xx}} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{S_{xx}} \right] \triangleq \sigma^2(\widehat{Y}_0), \end{aligned}$$

neboť $\mathsf{Cov}\,(\overline{Y}_n, \widehat{\beta}_1) = 0$, jak bylo ukázáno v důkazu věty 1.1. Nahrazením σ^2 statistikou s_n^2, dostaneme odhad rozptylu $\mathsf{Var}\,(\widehat{Y}_0)$ ve tvaru

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

Hodnota $\widehat{\sigma}(\widehat{Y}_0)$ se obvykle nazývá *standardní chyba predikce v bodě* x_0 . Jsou-li $e_1, ..., e_n$ navíc *iid* $N(0, \sigma^2)$, dostáváme, že platí

$$\widehat{Y}_0 \sim N\left(\mu_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{S_{xx}}\right]\right),$$

a tedy

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

Celkově tak dostáváme

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

Vyjádřením získáme $100(1-\alpha)$ % IS pro μ_0 ve tvaru

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

Poznámka 1.20. Ze tvaru intervalu spolehlivosti je vidět, že bude nejkratší pro $x_0 = \overline{x}_n$ a s rostoucí vzdáleností $|x_0 - \overline{x}_n|$ se prodlužuje. Speciálně, čím dále jsme od oblasti, kde jsou naše data x, tím méně spolehlivé jsou naše predikce. Je tedy třeba opatrnosti při predikci hodnot Y mimo interval $(\min_i\{x_i\}, \max_i\{x_i\})$.

Ad 2)

Intervalové odhady pro Y_0 nebudou intervaly spolehlivosti, protože Y_0 není parametr. Říká se jim *intervaly predikce*. Pro jejich konstrukci potřebujeme znát hodnotu rozptylu rozdílu $Y_0 - \widehat{Y}_0$. Pokud je nové pozorování Y_0 nezávislé na Y_i, $i \in \widehat{n}$, potom

$$\mathsf{Var}\,(Y_0 - \widehat{Y}_0) = \underbrace{\mathsf{Var}\,Y_0}_{\sigma^2} + \mathsf{Var}\,\widehat{Y}_0 + 0 = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{S_{xx}}\right],$$

neboť predikce $\widehat{Y}_0$ je funkcí pouze $Y_1, \ldots, Y_n$ a je tedy nezávislá na Y_0. Odhad spočteného rozptylu bude s_p^2, kde

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

Za předpokladu normality chyb pak platí

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

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

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

Poznámka 1.21. Ze tvaru spočtených intervalových odhadů je vidět, že přesnost predikce

a) roste s rostoucím n a rostoucím rozsahem pozorování x, měřeným pomocí S_{xx},

b) klesá s rostoucím vzdáleností x_0 od $\overline{x}_n$, tedy $|x_0 - \overline{x}_n|$.

Pokud můžeme předem zvolit $x_1, ..., x_n$, lze přesnost predikce zvýšit volbou dostatečně rozptýlených hodnot x. To ale může zvyšovat hodnotu R^2 statistiky a někdy vést k horšímu modelu. To indikuje *základní rozpor v regresní analýze*:

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

Poznámka 1.22. Odvozené výsledky platí za předpokladu normality chyb. Protože jsou ale za podmínek regularity odhady $\widehat{\beta}_0, \widehat{\beta}_1$ asymptoticky normální, intervaly spolehlivosti pro $\mathsf{E}\, Y_0$ budou použitelné pro velká n. Intervaly predikce pro hodnotu Y_0 ale závisí na normalitě chyb i pro velká n, mohou tedy být nepřesné pro nenormální chyby.

1.8 Ověření adekvátnosti modelu

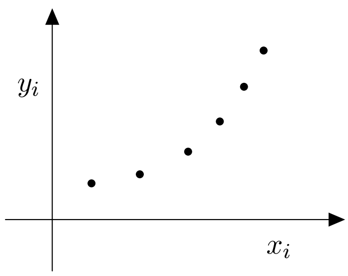
Ověření adekvátnosti modelu je důležitá součást analýzy dat. Měla by být provedena dříve, než budeme interpretovat parametry modelu nebo přijímat nějaké závěry založené na modelu. Všechny výsledky týkající se β_0, β_1 byly odvozeny za předpokladu *linearity modelu* a některé za předpokladu *normality chyb*. Bylo by tedy dobré mít nějaké metody pro ověření linearity případně normality.

Základní procedury:

1) *Prozkoumání scatter plotu dvojic* (x_i, y_i).
   Příklad lze vidět na obrázku 1.5, který může indikovat, že lepší model bude

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

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



Obrázek 1.5: Scatter plot naměřených dat.

2) *Analýza hodnot testovacích statistik.*
   Např.
   - malá hodnota R^2 společně s významnou hodnotou t-statistiky pro parametr β_1 obecně naznačuje, že skutečný model obsahuje i jiné proměnné x,
   - velká hodnota R^2 a významná t-statistika ale samo o sobě neznamená, že je model lineární.

3) *Obrázky reziduí.*
Jde o 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 linearity nebo normality.

Poznámka 1.23. Analýza scatter plotů a obrázků reziduí je dost subjektivní. Bylo by dobré mít nějaký objektivní analytický nástroj pro ověření linearity modelu. Bohužel nejsou téměř žádné k dispozici a pro většinu dat jsou v praxi nejvíce využívány metody zmíněné v bodech 1) - 3). Jinak tomu může být u navržených experimentů typu industriálních nebo klinických studií, kde existuje doporučený analytický test, tzv. *lack of fit* test (LOFT). Ten např. předpokládá, že máme více pozorování pro jednu hodnotu x_i.

Ad 3) - Analýza reziduí

Věnujme se teď podrobněji analýze reziduí. Intuitivně, pokud je náš model správný, měla by se rezidua chovat jako náhodný výběr z rozdělení $N(0, \sigma^2)$. Pokud se bude zdát, že se tak nechovají, bude to znamenat neadekvátnost modelu. Později ukážeme užitečné grafické nástroje, začněme ale vlastnostmi reziduí.

Věta 1.7. Nechť $\widehat{e}_i$ jsou rezidua modelu (1.8) odhadnutého metodou nejmenších čtverců. Potom platí, že

1) $\mathsf{E}\,\widehat{e}_i = 0, \quad i \in \widehat{n},$
2) $\mathsf{Var}\,\widehat{e}_i = \sigma^2_{\widehat{e}_i} = \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(x_i - \overline{x}_n)^2}{S_{xx}}\right)\right], \qquad (\approx \sigma^2 \text{ pro velká } n)$
3) $\mathsf{Cov}\,(\widehat{e}_i, \widehat{e}_j) = -\sigma^2 \left[\frac{1}{n} + \frac{(\overline{x}_n - x_i)(\overline{x}_n - x_j)}{S_{xx}}\right], \quad i \neq j,$
4) $\mathsf{Cov}\,(\widehat{e}_i, \widehat{Y}_i) = 0, \quad i \in \widehat{n}.$
5) jsou-li navíc $e_1, ..., e_n$ *iid* $N(0, \sigma^2)$, platí $\widehat{Z}_i = \dfrac{\widehat{e}_i}{\sigma_{\widehat{e}_i}} \sim N(0, 1)$.

Důkaz. 1) Pro střední hodnotu reziduí $\widehat{e}_i = Y_i - \widehat{Y}_i$ platí $\mathsf{E}\,(\widehat{e}_i) = \mathsf{E}\,Y_i - \mathsf{E}\,\widehat{Y}_i$. Z nestrannosti odhadů parametrů β_0, β_1 ale dostaneme

$$\mathsf{E}\,\widehat{Y}_i = \mathsf{E}\,(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i = \mathsf{E}\,Y_i.$$

2) Pro rozptyl reziduí platí

$$\mathsf{Var}\,\widehat{e}_i = \mathsf{Var}\,(Y_i - \widehat{Y}_i) = \mathsf{Var}\,Y_i + \mathsf{Var}\,\widehat{Y}_i - 2\,\mathsf{Cov}\,(Y_i, \widehat{Y}_i) = \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(x_i - \overline{x}_n)^2}{S_{xx}}\right)\right],$$

neboť

$$\mathsf{Var}\,\widehat{Y}_i = \mathsf{Cov}\,(Y_i, \widehat{Y}_i) = \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \overline{x}_n)^2}{S_{xx}}\right], \tag{1.12}$$

jak bylo ukázáno v důkazu věty 1.2.
3) Pro kovarianci reziduí $\widehat{e}_i, \widehat{e}_j$, $i \neq j$, dostaneme díky nezávislosti Y_i, Y_j

$$\mathsf{Cov}\,(\widehat{e}_i, \widehat{e}_j) = \mathsf{Cov}\,(Y_i - \widehat{Y}_i, Y_j - \widehat{Y}_j) = \underbrace{\mathsf{Cov}\,(Y_i, Y_j)}_{=0} - \mathsf{Cov}\,(Y_i, \widehat{Y}_j) - \mathsf{Cov}\,(\widehat{Y}_i, Y_j) + \mathsf{Cov}\,(\widehat{Y}_i, \widehat{Y}_j).$$

Pro poslední člen platí (s využitím tvrzení věty 1.1 a vztahu (1.5))

$$\begin{aligned}\operatorname{Cov}(\widehat{Y}_i,\widehat{Y}_j) &= \operatorname{Cov}(\widehat{\beta}_0+\widehat{\beta}_1 x_i, \widehat{\beta}_0+\widehat{\beta}_1 x_j)\\ &= \underbrace{\operatorname{Var}(\widehat{\beta}_0)}_{\sigma^2\left[\frac{1}{n}+\frac{\overline{x}_n^2}{S_{xx}}\right]} +(x_i+x_j)\underbrace{\operatorname{Cov}(\widehat{\beta}_0,\widehat{\beta}_1)}_{-\frac{\sigma^2\overline{x}_n}{S_{xx}}} + x_i x_j \underbrace{\operatorname{Var}(\widehat{\beta}_1)}_{\frac{\sigma^2}{S_{xx}}} =\\ &= \sigma^2\left[\frac{1}{n}+\frac{\overline{x}_n^2}{S_{xx}}-\frac{(x_i+x_j)\overline{x}_n}{S_{xx}}+\frac{x_i x_j}{S_{xx}}\right] = \sigma^2\left[\frac{1}{n}+\frac{(x_i-\overline{x}_n)(x_j-\overline{x}_n)}{S_{xx}}\right].\end{aligned}$$

Podobně bychom dostali

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

celkem tedy

$$\operatorname{Cov}(\widehat{e}_i,\widehat{e}_j) = -\sigma^2\left[\frac{1}{n}+\frac{(x_i-\overline{x}_n)(x_j-\overline{x}_n)}{S_{xx}}\right].$$

4) Pro kovarianci $\widehat{e}_i, \widehat{Y}_i$ pak s pomocí vztahu (1.12) dostaneme

$$\operatorname{Cov}(\widehat{e}_i,\widehat{Y}_i) = \operatorname{Cov}(Y_i-\widehat{Y}_i,\widehat{Y}_i) = \operatorname{Cov}(Y_i,\widehat{Y}_i) - \operatorname{Var}\widehat{Y}_i = 0.$$

5) Jsou-li náhodné chyby $e_1,...,e_n$ *iid* $N(0,\sigma^2)$, znamená to, že i-té reziduum $\widehat{e}_i$ má normální rozdělení, neboť se dá zapsat jako lineární kombinace nezávislých veličin $Y_1,...,Y_n$ s normálním rozdělením. Z bodu 1) navíc víme, že $\operatorname{E}\widehat{e}_i = 0$ a z bodu 2), že $\operatorname{Var}\widehat{e}_i = \sigma^2_{\widehat{e}_i}$. Celkem tedy

$$\frac{\widehat{e}_i}{\sigma_{\widehat{e}_i}} \sim N(0,1).$$

□

Poznámka 1.24. Z bodu 3) věty plyne, že $\operatorname{Cov}(\widehat{e}_i,\widehat{e}_j)\approx 0$ pro velká n. Pokud jsou tedy e_i *iid* $N(0,\sigma^2)$, měla by se standardizovaná rezidua $\widehat{Z}_i = \frac{\widehat{e}_i}{\sigma_{\widehat{e}_i}}$ chovat pro velká n jako náhodný výběr z $N(0,1)$ rozdělení. V praxi ale budeme potřebovat odhad σ^2 pro výpočet $\widehat{Z}_i$. Nejznámější procedura je nahradit σ^2 pomocí odhadu s_n^2. Potom se

$$\widehat{r}_i := \frac{\widehat{e}_i}{s_n\sqrt{1-\left(\frac{1}{n}+\frac{(x_i-\overline{x}_n)^2}{S_{xx}}\right)}}$$

nazývají *studentizovaná rezidua* (někdy také *standardizovaná*). Opět, pro velká n by se $\widehat{r}_i$ měla chovat jako náhodný výběr z $N(0,1)$.

Poznámka 1.25. Rezidua $\widehat{e}_i$ nebo $\widehat{r}_i$ se užívají pro grafickou analýzu. Existují ale i jiné třídy reziduí, např. tzv. PRESS rezidua, která jsou používána i v negrafických metodách zkoumání reziduí. Např. pro vyjádření toho, jak přesný je uvažovaný model pro predikci. Základní myšlenka je následující. Vynecháme z dat i-té pozorování a spočteme odhady $\widehat{\beta}_{0(-i)}, \widehat{\beta}_{1(-i)}$ parametrů β_0,β_1. Pak i-té PRESS reziduum je definováno jako

$$\widehat{e}_{(-i)} = y_i - \widehat{y}_{(-i)}, \quad \text{kde } \widehat{y}_{(-i)} = \widehat{\beta}_{0(-i)} + x_i\widehat{\beta}_{1(-i)}.$$

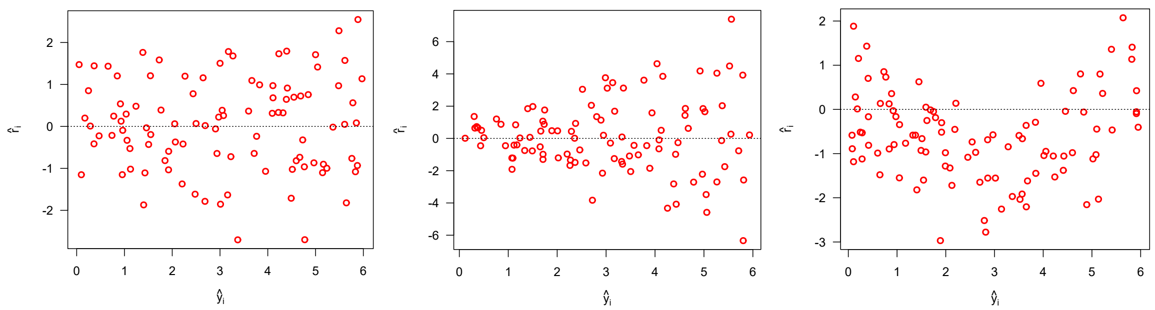
Porovnáváme tedy napozorovanou hodnotu y_i s predikovanou hodnotou $\widehat{y}_{(-i)}$ v bodě x_i. Podrobněji se jim budeme věnovat později, poznamenejme jen, že PRESS znamená *predicted residual error sum of squares* a platí PRESS$=\sum\limits_{i=1}^{n}(y_i-\widehat{y}_{(-i)})^2$.

1.9 Grafy reziduí

Shrňme nyní nejužívanější grafické nástroje pro analýzu reziduí.

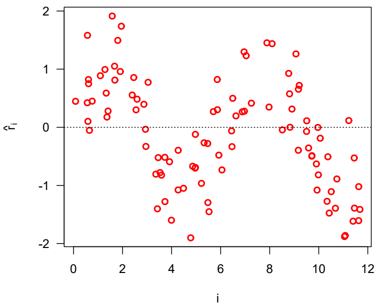
- **Histogram reziduí:** umožní náhled normality reziduí.
- **Kvantilový graf (Q-Q plot) studentizovaných reziduí:** rezidua seřadíme dle velikosti: $\widehat{r}_{(1)} \leqslant \widehat{r}_{(2)} \leqslant \ldots \leqslant \widehat{r}_{(n)}$ a vyneseme oproti teoretickým kvantilům $\Phi^{-1}\left((i-\frac{1}{2})\frac{1}{n}\right)$, $i \in \widehat{n}$. Body by měly ležet přibližně na přímce neboť se dá ukázat, že $\mathsf{E}\left(e_{(i)}\right) \approx \Phi^{-1}\left((i-\frac{1}{2})\frac{1}{n}\right)$ pro normálně rozdělené chyby.

  Použití: ověření normality, detekce odlehlých pozorování.
- **Studentizovaná rezidua proti jednotlivým vysvětlujícím proměnným x:** rezidua $\widehat{r}_i$ nezávisí na σ, graf $\widehat{r}_i \times x_i$ lze použít pro detekci nelinearity nebo nekonstantního rozptylu. V ideálním případě by měla být rezidua rovnoměrně rozptýlena a neměl by být patrný žádný očividný trend nebo vzor.
- **Studentizovaná rezidua $\widehat{r}_i$ proti predikovaným hodnotám $\widehat{y}_i$:** platí $\mathsf{Cov}\,(\widehat{e}_i, \widehat{Y}_i) = 0$, tedy $\widehat{e}_i(\widehat{r}_i)$ a $\widehat{Y}_i$ by měly být nekorelované, pokud platí model (1.8). To znamená, že graf $\widehat{r}_i \times \widehat{y}_i$ by měl být náhodně rozptýlený kolem osy x, navíc $\widehat{r}_i$ by měla ležet v $(-3, 3)$ ($\widehat{r}_i \approx N(0,1)$). Typické situace jsou znázorněny na obrázku 1.6.
- **Studentizovaná rezidua proti pořadí pozorování:** možná detekce řadové korelace mezi pozorováními (viz např. obrázek 1.7).



(a) (b) (c)

Obrázek 1.6: Studentizovaná rezidua $\widehat{r}_i$ proti predikovaným hodnotám $\widehat{y}_i$: (a) ideální obrázek, (b) nestejný rozptyl, (c) nelinearita.



Obrázek 1.7: Studentizovaná rezidua $\widehat{r}_i$ proti pořadí pozorování i.

2 Vícerozměrná lineární regrese

Uvažujme nyní situaci, kdy máme k dispozici více vysvětlujících proměnných. Předpokládejme tedy, že kromě y_i máme pro každé $i \in \widehat{n}$ k dispozici také m nezávislých proměnných $x_{i1}, x_{i2}, \ldots, x_{im}$. Model lineární regrese pak bude mít tvar

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

kde $e_1, \ldots, e_n$ jsou *nezávislé (nekorelované)* chyby a $e_i \sim (0, \sigma^2)$. Na základě pozorování $(x_{i1}, \ldots, x_{im}, y_i)$, $i \in \widehat{n}$, chceme odhadnout neznámý vektor parametrů $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^T$ (jedná se vlastně o proložení dat $m+1$ dimenzionální nadrovinou). Budeme předpokládat, že $n > m + 1$, tj., že máme více dat než parametrů. Model nejdříve přepíšeme do maticového zápisu. Označme

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

a

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

Matice $\boldsymbol{X}$ se nazývá *matice modelu* (regresní matice, *design matrix*). Model vícerozměrné lineární regrese pak lze zapsat ve tvaru

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times(m+1)} \boldsymbol{\beta}_{(m+1)\times 1} + \boldsymbol{e}_{n\times 1}. \tag{2.1}$$

2.1 Odhady parametrů

Pro odvození MLE odhadu parametru $\boldsymbol{\beta}$ budeme předpokládat, že $e_1, \ldots, e_n$ jsou nezávislé a $e_i \sim N(0, \sigma^2)$, tzn. $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ a $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$. S využitím označení

$$\mu_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}, \qquad \boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T = \boldsymbol{X\beta},$$

pak dostaneme věrohodnostní funkce ve tvaru

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

Z odvozeného tvaru věrohodnostní funkce je vidět, že pro pevné σ^2 je

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

Minimalizace funkce $g(\boldsymbol{\beta})$ by šla opět provést pomocí derivací, v této sekci ale ukážeme algebraický přístup.

Věta 2.1. Uvažujme model (2.1) a nechť $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. Potom $\widehat{\boldsymbol{\beta}}$ je MLE parametru $\boldsymbol{\beta}$ právě tehdy, když $\widehat{\boldsymbol{\beta}}$ je řešením soustavy rovnic (*soustava normálních rovnic*)

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}. \tag{2.2}$$

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

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

Důkaz. ($\Leftarrow$:) Ukážeme, že každé řešení $\widehat{\boldsymbol{\beta}}$ soustavy (2.2) minimalizuje funkci $g(\boldsymbol{\beta})$ a je tedy MLE odhadem. Pro každé $\boldsymbol{\beta}$ platí, že

$$\begin{aligned} g(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}, \end{aligned}$$

kde jsme využili toho, že $\widehat{\boldsymbol{\beta}}$ řeší soustavu (2.2) a tedy $\boldsymbol{y}^T\boldsymbol{X} = \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}$. Odvozený vztah musí platit i pro $\widehat{\boldsymbol{\beta}}$,

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

a s využitím vlastností skalárního součinu dostaneme

$$\begin{aligned} g(\boldsymbol{\beta}) - g(\widehat{\boldsymbol{\beta}}) &= \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= \left(\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right)^T\left(\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right) = \langle \boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\,;\, \boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\rangle \geqslant 0, \quad \forall \boldsymbol{\beta}, \end{aligned} \tag{2.3}$$

to znamená, že $\widehat{\boldsymbol{\beta}}$ minimalizuje $g(\boldsymbol{\beta})$ a je tedy MLE odhadem parametru $\boldsymbol{\beta}$.

($\Rightarrow$:) Předpokládejme, že $\widehat{\boldsymbol{\beta}}_1$ minimalizuje $g(\boldsymbol{\beta})$ (je tedy MLE). To potom znamená, že $g(\widehat{\boldsymbol{\beta}}_1) \leqslant g(\boldsymbol{\beta})$, $\forall \boldsymbol{\beta}$, speciálně $g(\widehat{\boldsymbol{\beta}}_1) \leqslant g(\widehat{\boldsymbol{\beta}})$, kde $\widehat{\boldsymbol{\beta}}$ je řešení soustavy (2.2). Z rovnice (2.3) vyplývá, že $g(\widehat{\boldsymbol{\beta}}_1) \geqslant g(\widehat{\boldsymbol{\beta}})$, celkem tedy dostáváme $g(\widehat{\boldsymbol{\beta}}_1) = g(\widehat{\boldsymbol{\beta}})$. Dosazením do (2.3) dostaneme, že

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

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

$$\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{0},$$

což, díky rovnosti $\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$, implikuje

$$\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{X}^T\boldsymbol{y}$$

a $\widehat{\boldsymbol{\beta}}_1$ tedy je řešením soustavy (2.2). $\square$

Poznámka 2.1. Aby byl důkaz korektní, je třeba ukázat, že soustava $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$ má vždy alespoň 1 řešení. Pokud existuje $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, není co dokazovat, řešení máme přímo. Co když je ale $\boldsymbol{X}^T\boldsymbol{X}$ singulární?

Lemma 2.1. Soustava lineárních rovnic $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ má řešení právě tehdy, když $\langle \boldsymbol{y}; \boldsymbol{z}\rangle = 0$ pro všechna $\boldsymbol{z}$ splňující $\boldsymbol{A}^T\boldsymbol{z} = \boldsymbol{0}$.

Věta 2.2. Soustava normálních rovnic $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$ má vždy alespoň jedno řešení.

Důkaz. Musíme ukázat, že $\langle \boldsymbol{X}^T\boldsymbol{y}, \boldsymbol{z}\rangle = 0$ pro všechny vektory $\boldsymbol{z}$ splňující $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{z} = \boldsymbol{0}$. Podmínka $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{z} = \boldsymbol{0}$ ale implikuje, že

$$\langle \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{z}; \boldsymbol{z}\rangle = \langle \boldsymbol{X}\boldsymbol{z}; \boldsymbol{X}\boldsymbol{z}\rangle = 0,$$

a tedy $\boldsymbol{X}\boldsymbol{z} = \boldsymbol{0}$. Celkem tedy díky vlastnostem skalárního součinu dostáváme $\langle \boldsymbol{X}^T\boldsymbol{y}; \boldsymbol{z}\rangle = \langle \boldsymbol{y}; \boldsymbol{X}\boldsymbol{z}\rangle = 0$. Obecně totiž platí, že $\langle \boldsymbol{X}; \boldsymbol{A}\boldsymbol{y}\rangle = \langle \boldsymbol{A}^T\boldsymbol{X}; \boldsymbol{y}\rangle$. □

Poznámka 2.2. Z vět vyplývá, že MLE odhad parametru $\boldsymbol{\beta}$ může být nalezen řešením $m+1$ lineárních rovnic o $m+1$ neznámých. Málokdy ale existuje analytické řešení, často je třeba použít numerické metody. Matice $\boldsymbol{X}^T\boldsymbol{X}$ může být v praktických aplikacích špatně podmíněná, což ovlivňuje numerickou přesnost odhadu $\widehat{\boldsymbol{\beta}}$. Proto se často užívají metody jako Choleského rozklad, QR rozklad nebo singulární rozklad (SVD).
Odvodili jsme to pro normální chyby, minimalizace $g(\boldsymbol{\beta})$ lze ale použít i pro jiné druhy chyb, potom se odhad $\widehat{\boldsymbol{\beta}}$ nazývá **ordinary least squares estimate (OLS)** (obyčejné nejmenší čtverce). Jedná se o nejužívanější metodu pro odhad parametru $\boldsymbol{\beta}$.

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

Věta 2.3. Matice $\boldsymbol{X}^T\boldsymbol{X}$ je nesingulární právě tehdy, když jsou sloupce matice $\boldsymbol{X}$ lineárně nezávislé.

Důkaz. ($\Leftarrow$: (sporem)) Nechť jsou sloupce $\boldsymbol{X}$ lineárně nezávislé a matice $\boldsymbol{X}^T\boldsymbol{X}$ singulární, tzn. $\exists \boldsymbol{c} \neq \boldsymbol{0}$ takové, že $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{c} = \boldsymbol{0}$. Potom

$$0 = \langle \boldsymbol{c}, \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{c}\rangle = \langle \boldsymbol{X}\boldsymbol{c}, \boldsymbol{X}\boldsymbol{c}\rangle \quad \Rightarrow \quad \boldsymbol{X}\boldsymbol{c} = \boldsymbol{0},$$

tedy

$$\sum_{j=0}^{m} c_j \boldsymbol{x}_j^c = \boldsymbol{0},$$

kde $\boldsymbol{c} = (c_0, \dots, c_m)^T$ a $\boldsymbol{x}_j^c$ je j-tý sloupec matice $\boldsymbol{X}$. To ale znamená, že sloupce matice $\boldsymbol{X}$ jsou lineárně závislé, což je spor.

($\Rightarrow$: (sporem)) Předpokládejme, že $\boldsymbol{X}^T\boldsymbol{X}$ je regulární a sloupce $\boldsymbol{X}$ jsou lineárně závislé. Potom existuje $\boldsymbol{c} \neq \boldsymbol{0}$ takové, že $\boldsymbol{X}\boldsymbol{c} = \boldsymbol{0}$, a tedy i $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{c} = \boldsymbol{0}$. Z toho vyplývá, že $\boldsymbol{X}^T\boldsymbol{X}$ je singulární, což je spor. □

Poznámka 2.3. Pokud tedy uvažujeme model s maticí regresorů $\boldsymbol{X}_{n\times(m+1)}$ a platí $n > m+1$, $h(\boldsymbol{X}) = m+1$ a $\boldsymbol{e} \sim (\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$, z předcházejících vět vyplývá, že existuje jednoznačné řešení normálních rovnic a má tvar $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

Poznámka 2.4. Pokud jsou sloupce matice $\boldsymbol{X}$ lineárně závislé, je $\boldsymbol{X}^T\boldsymbol{X}$ singulární, což je většinou detekováno numerickou metodou výpočtu $\widehat{\boldsymbol{\beta}}$. Horší situace nastává, pokud jsou sloupce $\boldsymbol{X}$ „téměř" lineárně závislé (tzv. *multikolinearita*) , což způsobuje problémy při výpočtu $\widehat{\boldsymbol{\beta}}$, protože je $\boldsymbol{X}^T\boldsymbol{X}$ „téměř" singulární. Jak multikolinearitu detekovat probereme později.

2.1.1 Odhad parametru σ^2

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

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

kde $\widehat{y}_i = (\boldsymbol{X}\widehat{\boldsymbol{\beta}})_i = \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}$, $i \in \widehat{n}$ a $\boldsymbol{x}_i^T$ značí i-tý řádek matice $\boldsymbol{X}$. Protože se jedná o vychýlený odhad, používá se obecně nestranný odhad

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

pro parametr σ^2 a $s_n = \sqrt{s_n^2}$ jako odhad σ (už není nestranný). Pro náhodné chyby bez předpokladu normality, $e_i \sim (0, \sigma^2)$, se také používají statistiky s_n^2 a s_n.

2.1.2 Vlastnosti odhadů $\widehat{\boldsymbol{\beta}}, s_n^2$

Věta 2.4. Nechť $\widehat{\boldsymbol{\beta}}$ je OLS odhad parametru $\boldsymbol{\beta}$ v modelu (2.1), kde $h(\boldsymbol{X}) = m+1$ a náhodné chyby $e_1, \ldots, e_n$ jsou nekorelované, $e_i \sim (0, \sigma^2)$, $i \in \widehat{n}$. Potom platí, že

1) $\mathsf{E}\,(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ (tj. $\widehat{\boldsymbol{\beta}}$ je nestranný odhad $\boldsymbol{\beta}$)
2) $\mathsf{Cov}\,(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
3) $\mathsf{E}\,(s_n^2) = \sigma^2$ (tj. $s_n^2 b$ je nestranný odhad σ^2)
4) Pokud navíc $e_i \sim N(0, \sigma^2), i \in \widehat{n}$, potom

$$\widehat{\boldsymbol{\beta}} \sim N_{m+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

Speciálně $\widehat{\beta}_i \sim N(\beta_i, \sigma^2 v_i)$, kde v_i je i-tý diagonání prvek matice $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Důkaz. 1) Z předpokladu $h(\boldsymbol{X}) = m+1$ vyplývá (viz Poznámka 2.3), že $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Aplikací střední hodnoty a faktu $\mathsf{E}\,\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$ získáme

$$\mathsf{E}\,\widehat{\boldsymbol{\beta}} = \mathsf{E}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\right] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathsf{E}\,\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

2) Z teorie pravděpodobnosti víme, že pokud je $\boldsymbol{Y}_{(n\times 1)}$ náhodný vektor s kovarianční maticí $\mathsf{Cov}\,(\boldsymbol{Y}) = \Sigma$ a $\boldsymbol{A}_{(m\times n)}$ je reálná matice, potom platí $\mathsf{Cov}\,(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\Sigma\boldsymbol{A}^T$.
Z předpokladů věty plyne, že $\widehat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}$, kde $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, tedy $\widehat{\boldsymbol{\beta}}$ je lineární kombinací $Y_1, \ldots, Y_n$, a protože $\mathsf{Cov}\,(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}_n$, dostaneme

$$\mathsf{Cov}\,\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

3) Pro výpočet střední hodnoty odhadu $s_n^2 = \frac{1}{n-(m+1)} SSE$ nejdříve přepíšeme vektor reziduí $\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$. Pro vektor $\widehat{\boldsymbol{Y}}$ platí

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}, \tag{2.4}$$

kde $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ je tzv. *projekční matice*, a tedy $\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$. Dále platí

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

takže

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

Projekční matice $\boldsymbol{H}$ je symetrická a idempotentní,

$$\begin{aligned} \boldsymbol{H}^T &= \boldsymbol{H}, \\ \boldsymbol{H}^2 &= \left[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right]\left[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right] = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H} \end{aligned}$$

a platí pro ni $(\boldsymbol{I}_n - \boldsymbol{H})^2 = \boldsymbol{I}_n - \boldsymbol{H}$. Pro součet čtverců reziduálních chyb SSE tak dostaneme

$$SSE = (\boldsymbol{Y} - \widehat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}) = \widehat{\boldsymbol{e}}^T\widehat{\boldsymbol{e}} = \boldsymbol{e}^T(\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{e} = \boldsymbol{e}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{e} = \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}e_ie_j,$$

kde g_{ij} je (i,j)-tý prvek matice $\boldsymbol{I}_n - \boldsymbol{H}$. Využitím nekorelovanosti náhodných chyb ($\mathsf{Cov}\,(e_i, e_j) = \mathsf{E}\,(e_ie_j) = 0$ pro $i \neq j$) pak pro střední hodnotu $\mathsf{E}\,(SSE)$ dostaneme

$$\mathsf{E}\,(SSE) = \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}\,\mathsf{E}\,(e_ie_j) = \sum_{i=1}^{n} g_{ii}\,\mathsf{Var}\,e_i = \sigma^2\sum_{i=1}^{n} g_{ii}\,.$$

Pro výpočet sumy využijeme známé vlastnosti stopy matice ($\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n,n}$, $\mathrm{tr}(\boldsymbol{A} + \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B} + \boldsymbol{A})$, $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A})$)

$$\begin{aligned} \sum_{i=1}^{n} g_{ii} &= \mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = \mathrm{tr}(\boldsymbol{I}_n) - \mathrm{tr}(\boldsymbol{H}) = n - \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\ &= n - \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) = n - \mathrm{tr}(\boldsymbol{I}_{m+1}) = n - (m+1)\,. \end{aligned}$$

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

4) Jelikož $\widehat{\boldsymbol{\beta}}$ je lineární kombinací $Y_1, \ldots, Y_n$, které jsou nezávislé a normálně rozdělené, dostaneme

$$\widehat{\boldsymbol{\beta}} \sim N_{m+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}),$$

neboť lineární kombinace složek vektoru s vícerozměným normálním rozdělením má opět vícerozměrné normální rozdělení a příslušnou střední hodnotu a kovariační matici jsme již spočítali. □

Poznámka 2.5. Shrňme ještě dokázané vlastnosti projekční matice:

- $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, $\quad\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, $\quad\widehat{\boldsymbol{e}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{e}$,

- $\boldsymbol{H}^T = \boldsymbol{H}, \quad (\boldsymbol{I}_n - \boldsymbol{H})^T = (\boldsymbol{I}_n - \boldsymbol{H}), \quad$ (symetrie)
- $\boldsymbol{H}^2 = \boldsymbol{H}, \quad (\boldsymbol{I}_n - \boldsymbol{H})^2 = \boldsymbol{I}_n - \boldsymbol{H}, \quad$ (idempotentnost)
- $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}, \quad \operatorname{tr}(\boldsymbol{H}) = \sum\limits_{i=1}^{n} h_{ii} = m+1,$
- $\boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{H}) = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{H} = \boldsymbol{0}.$

Věta 2.5. Nechť $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ je lineární model (2.1), kde $h(\boldsymbol{X}) = m+1$ a $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Potom platí:

1) $\widehat{\boldsymbol{\beta}}$ a s_n^2 jsou nezávislé náhodné veličiny.
2) $(n-m-1)\dfrac{s_n^2}{\sigma^2} = \dfrac{SSE}{\sigma^2} \sim \chi^2(n-m-1)$.
3) Jestliže $v_i = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}$, potom $T_i = \dfrac{\widehat{\beta}_i - \beta_i}{s_n\sqrt{v_i}} \sim t(n-m-1)$.
4) Nechť $\boldsymbol{C} \in \mathbb{R}^{r,m+1}$ takové, že $h(\boldsymbol{C}) = r$. Potom kvadratická forma

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

Důkaz. 1) Odhad parametru $\boldsymbol{\beta}$ lze vyjádřit ve tvaru

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

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

$$\boldsymbol{Z} \overset{\text{ozn.}}{=\!=} \begin{pmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\boldsymbol{e}} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \\ \boldsymbol{I}_n - \boldsymbol{H} \end{pmatrix} \boldsymbol{e} = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{pmatrix} \boldsymbol{e},$$

kde $\boldsymbol{Z}$ je lineární transformací vektoru $(e_1, \ldots, e_n) = \boldsymbol{e} \sim N(0, \sigma^2\boldsymbol{I}_n)$. Z toho plyne, že $\boldsymbol{Z}$ má vícerozměrné normální rozdělení (i když degenerované, protože $\operatorname{Cov}(\boldsymbol{Z})$ je singulární, $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{I}_n - \boldsymbol{H}$ je singulární, protože $(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$).
Abychom ukázali, že $\widehat{\boldsymbol{\beta}}$ a $\widehat{\boldsymbol{e}}$ jsou nezávislé ($s_n^2 = \frac{1}{n-m-1}\widehat{\boldsymbol{e}}^T\widehat{\boldsymbol{e}}$, tedy i $\widehat{\boldsymbol{\beta}}$ a s_n^2 pak budou nezávislé), stačí ukázat, že $\operatorname{Cov}(\widehat{\beta}_i, \widehat{e}_j) = 0$ pro $i = 0, \ldots, m$ a $j \in \widehat{n}$. Spočteme $\operatorname{Cov}(\boldsymbol{Z})$,

$$\operatorname{Cov}(\boldsymbol{Z}) = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{pmatrix} \operatorname{Cov}(\boldsymbol{e}) \left(\boldsymbol{A}^T \boldsymbol{B}^T\right) = \sigma^2 \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{pmatrix} \left(\boldsymbol{A}^T \boldsymbol{B}^T\right) = \sigma^2 \begin{pmatrix} \boldsymbol{A}\boldsymbol{A}^T & \boldsymbol{A}\boldsymbol{B}^T \\ \boldsymbol{B}\boldsymbol{A}^T & \boldsymbol{B}\boldsymbol{B}^T \end{pmatrix}.$$

Ze tvaru spočtené kovariační matice je vidět, že $\operatorname{Cov}(\widehat{\beta}_i, \widehat{e}_j) = \left(\boldsymbol{A}\boldsymbol{B}^T\right)_{ij}$, $i = 0, \ldots, m$, $j \in \widehat{n}$, zbývá tedy ukázat, že $\boldsymbol{A}\boldsymbol{B}^T = \boldsymbol{0}$. Dosazením dostaneme

$$\begin{aligned} \boldsymbol{A}\boldsymbol{B}^T &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) = \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{0}. \end{aligned}$$

2) Pro důkaz druhého tvrzení budeme potřebovat dva výsledky z lineární algebry:

- *Spektrální rozklad matice:* nechť $\boldsymbol{A}_{n\times n}$ je symetrická matice. Potom existuje ortogonální matice $\boldsymbol{Q}$ a diagonální matice $\boldsymbol{\Lambda}$ tak, že $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$, sloupce $\boldsymbol{Q}$ jsou ON vlastní vektory matice $\boldsymbol{A}$ a diagonální prvky matice $\boldsymbol{\Lambda}$ jsou jim odpovídající vlastní čísla.

- Nechť $\boldsymbol{A}_{n\times n}$ je idempotentní matice, potom vlastní čísla matice $\boldsymbol{A}$ jsou pouze 0 nebo 1 a $h(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A})$.

V důkazu předchozí věty jsme ukázali, že

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

Protože je matice $(\boldsymbol{I}_n - \boldsymbol{H})$ symetrická a idempotentní, lze ji zapsat ve tvaru

$$\boldsymbol{I}_n - \boldsymbol{H} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T,$$

kde $\boldsymbol{Q}$ je ortogonální matice a $\boldsymbol{\Lambda}$ je diagonální matice s vlastními čísly matice $\boldsymbol{I}_n - \boldsymbol{H}$. Protože vlastní čísla $\boldsymbol{I}_n - \boldsymbol{H}$ jsou pouze 0 nebo 1 a $\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = h(\boldsymbol{I}_n - \boldsymbol{H}) = n - m - 1$, diagonální matice $\boldsymbol{\Lambda}$ může být zapsána ve tvaru

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

takže

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

Nyní využijeme tvrzení z pravděpodobnosti: pokud $\boldsymbol{U} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$ a $\boldsymbol{Q}$ je ortogonální matice, potom $\boldsymbol{Q}\boldsymbol{U} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$. Protože $\boldsymbol{q} = \boldsymbol{Q}^T\boldsymbol{e}$ a $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ dostaneme, že $\boldsymbol{q}$ je vektor nezávislých $N(0, \sigma^2)$ veličin $\big(\boldsymbol{q} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)\big)$ a

$$(n-m-1)\frac{s_n^2}{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{e}^T(\mathrm{I}_n - \boldsymbol{H})\boldsymbol{e} = \frac{1}{\sigma^2}\boldsymbol{q}^T\boldsymbol{\Lambda}\boldsymbol{q} = \sum_{i=1}^{n-m-1}\frac{q_i^2}{\sigma^2}$$

je suma druhých mocnin $n - m - 1$ nezávislých náhodných veličin s rozdělením $N(0, 1)$, tzn. že má rozdělení $\chi^2(n - m - 1)$.

3) Z předchozí věty víme, že

$$\frac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{v_i}} \sim N(0, 1),$$

teď jsme ukázali, že $\frac{(n-m-1)s_n^2}{\sigma^2} \sim \chi^2(n-m-1)$ a s_n^2 a $\widehat{\beta}_i$ jsou nezávislé. Z definice Studentova rozdělení tak dostaneme

$$T_i = \frac{\frac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{v_i}}}{\sqrt{\frac{\frac{(n-m-1)s_n^2}{\sigma^2}}{n-m-1}}} = \frac{\frac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{v_i}}}{\frac{s_n}{\sigma}} = \frac{\widehat{\beta}_i - \beta_i}{s_n\sqrt{v_i}} \sim t(n-m-1)\,. \tag{2.5}$$

4) Z vlastností transformace vícerozměrného normálního rozdělení plyne, že

$$\boldsymbol{C}\widehat{\boldsymbol{\beta}} \sim N_r\left(\boldsymbol{C}\boldsymbol{\beta}, \sigma^2\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T\right),$$

a tedy

$$\boldsymbol{C}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{C}\boldsymbol{\beta} \sim N_r\left(\boldsymbol{0}, \sigma^2\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T\right).$$

Stačí tedy ukázat, že pokud $\boldsymbol{Z} \sim N_r(\mathbf{0}, \Sigma)$, potom $\boldsymbol{Z}^T\Sigma^{-1}\boldsymbol{Z} \sim \chi^2(r)$. Protože Σ je pozitivně definitní, existuje regulární matice $\boldsymbol{R}$ taková, že $\Sigma = \boldsymbol{R}\boldsymbol{R}^T$. Definujme $\boldsymbol{U} := \boldsymbol{R}^{-1}\boldsymbol{Z}$. Potom $\mathsf{E}\,\boldsymbol{U} = \boldsymbol{R}^{-1}\,\mathsf{E}\,[\boldsymbol{Z}] = \mathbf{0}$. Dále

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

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

$$\boldsymbol{Z}^T\Sigma^{-1}\boldsymbol{Z} = \boldsymbol{U}^T\boldsymbol{R}^T(\boldsymbol{R}^T)^{-1}\boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{U} = \boldsymbol{U}^T\boldsymbol{U} = \sum_{i=1}^{r} \mathrm{U}_i^2 \sim \chi^2(r)\,.$$

Tvrzení věty dostaneme dosazením za $\boldsymbol{Z} = \boldsymbol{C}(\widehat{\beta} - \beta)$ a $\Sigma = \sigma^2\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T$.

□

2.1.3 Vlastnosti vektoru reziduí $\widehat{e}$

Věta 2.6. Uvažujeme model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, kde $e_1, \ldots, e_n$ jsou nekorelované a $e_i \sim (0, \sigma^2)$. Nechť $\widehat{\boldsymbol{\beta}}$ je OLS odhad parametru $\boldsymbol{\beta}$ a $\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ je vektor reziduí. Potom platí, že

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

Důkaz. Využijeme dokázanou vlastnost $\widehat{\boldsymbol{e}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{e}$ (viz Poznámka 2.5).
1) Pro střední hodnotu platí

$$\mathsf{E}\,[\widehat{\boldsymbol{e}}] = (\boldsymbol{I}_n - \boldsymbol{H}) \cdot \mathsf{E}\,[\boldsymbol{e}] = (\boldsymbol{I}_n - \boldsymbol{H}) \cdot \mathbf{0} = \mathbf{0}\,.$$

2) Pro kovarianční matici dostaneme díky symetrii a idempotentnosti matice $\boldsymbol{I}_n - \boldsymbol{H}$

$$\mathsf{Cov}\,(\widehat{\boldsymbol{e}}) = (\boldsymbol{I}_n - \boldsymbol{H})\,\mathsf{Cov}\,(\boldsymbol{e})(\boldsymbol{I}_n - \boldsymbol{H})^T = \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})\,.$$

3) Protože $\widehat{\boldsymbol{e}}$ je lineární kombinací složek vektoru $\boldsymbol{e}$, který má vícerozměrné normální rozdělení, má i vektor $\widehat{\boldsymbol{e}}$ vícerozměrné normální rozdělení a s využitím dokázaných vlastností 1) a 2) dostáváme $\widehat{\boldsymbol{e}} \sim N_n\big(\mathbf{0}, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})\big)$.
4) Soustavu normálních rovnic $\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}$ lze zapsat jako $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0}$. První rovnice má tvar

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

kde v modelu s interceptem je první sloupec matice $\boldsymbol{X}$ vektor jedniček, to znamená $x_{i1} = 1$ pro všechna i. Pro $\widehat{\boldsymbol{\beta}}$ tedy platí

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

5) Z předchozího bodu platí pro OLS odhad $\widehat{\boldsymbol{\beta}}$

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$$

a přenásobením zleva řádkovým vektorem $\widehat{\boldsymbol{\beta}}^T$ dostaneme

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

□

Poznámka 2.6. Použitím bodů 4) a 5) dostaneme pro model s interceptem následující rozklad čtverců (stejně jako u jednorozměrné regrese)

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

tedy

$$SST = SSR + SSE.$$

2.2 Gauss - Markov theorem

Pro normální náhodné chyby, tj. e_i *i.i.d.* $N(0, \sigma^2)$, je OLS odhad $\widehat{\boldsymbol{\beta}}$ maximálně věrohodným odhadem, tzn. je to eficientní odhad parametru $\boldsymbol{\beta}$ (MVUE - minimum variance unbiased estimator).
Pro náhodné chyby s jiným rozdělením ukážeme (za jistých podmínek), že OLS odhad $\widehat{\boldsymbol{\beta}}$ je BLUE (best linear unbiased estimator) parametru $\boldsymbol{\beta}$, tj. lineární nestranný odhad s minimálním rozptylem. Je ale třeba si uvědomit, že mohou existovat lepší lineární vychýlené odhady nebo nelineární odhady.

Definice 2.1. Nechť $\boldsymbol{\beta}$ je vektor regresních parametrů v lineárním modelu. Řekneme, že $\widehat{\boldsymbol{\beta}}$ je *lineární odhad* parametru $\boldsymbol{\beta}$, jestliže každé $\widehat{\beta}_j$ je lineární kombinací pozorování $Y_i, i = 1, \ldots, n$, tedy

$$\widehat{\beta}_j = \sum_{i=1}^{n} a_{ij} Y_i \quad j = 0, \ldots, m\,.$$

V maticovém zápisu

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}, \quad \text{kde} \quad \boldsymbol{A}^T = (a_{ij}),\ i = 1, \ldots, n,\ j = 0, \ldots, m\,.$$

Poznámka 2.7. Pokud v modelu (2.1) platí $h(\boldsymbol{X}) = m + 1$, potom OLS odhad $\widehat{\boldsymbol{\beta}}$ je lineární, neboť $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, tzn. matice $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

Věta 2.7 (Gauss-Markov)**.** Uvažujme model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, kde matice $\boldsymbol{X}$ má plnou hodnost, $e_i, i \in \widehat{n}$, jsou nekorelované a $e_i \sim (0, \sigma^2)$. Potom OLS odhad $\widehat{\boldsymbol{\beta}}$ je BLUE parametru $\boldsymbol{\beta}$ (best linear unbiased estimator).

Důkaz. Nechť $\widehat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}$ je lineární odhad $\boldsymbol{\beta}$. Aby byl nestranný, musí platit $\mathsf{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, tzn.

$$\mathsf{E}[\boldsymbol{A}\boldsymbol{Y}] = \boldsymbol{A}\,\mathsf{E}[\boldsymbol{Y}] = \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

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

$$\mathsf{Cov}(\widehat{\boldsymbol{\beta}}) = \mathsf{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\,\mathsf{Cov}(\boldsymbol{Y})\boldsymbol{A}^T = \boldsymbol{A}\sigma^2\boldsymbol{I}_n\boldsymbol{A}^T = \sigma^2\boldsymbol{A}\boldsymbol{A}^T.$$

Zapišme $\boldsymbol{A}$ ve tvaru $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D}$, kde $\boldsymbol{D}$ vlastně vyjadřuje rozdíl mezi maticí $\boldsymbol{A}$ a maticí pro OLS odhad. Pokud ukážeme, že pro nestranný lineární odhad $\widehat{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{Y}$, který minimalizuje rozptyl, musí platit $\boldsymbol{D} = \boldsymbol{0}$, bude věta dokázána. Dosazením dostaneme:

$$\begin{aligned}\mathsf{Cov}(\widehat{\boldsymbol{\beta}}) &= \sigma^2\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D}\right)\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D}\right)^T = \\ &= \sigma^2\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{D}^T\right].\end{aligned}$$

Z podmínky nestrannosti

$$\boldsymbol{A}\boldsymbol{X} = \left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D}\right]\boldsymbol{X} = \mathbf{I}_{m+1} + \boldsymbol{D}\boldsymbol{X} = \mathbf{I}_{m+1}$$

vyplývá $\boldsymbol{D}\boldsymbol{X} = \boldsymbol{0}$, a tedy i $\boldsymbol{X}^T\boldsymbol{D}^T = \boldsymbol{0}$. To znamená, že

$$\mathsf{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{D}^T\right]$$

a pro diagonální prvky platí, že

$$\mathsf{Var}[\widehat{\beta}_i] = \sigma^2\Big[v_i + \sum_{j=1}^{n} d_{ij}^2\Big], \quad i = 0, \ldots, m,$$

kde $v_i = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{ii}$. Protože $v_i \geqslant 0$ je pevná konstanta a $\sum_{j=1}^{n} d_{ij}^2 \geqslant 0$, znamená to, že rozptyl $\mathsf{Var}[\widehat{\beta}_i]$ je minimalizován volbou $\sum_{j=1}^{n} d_{ij}^2 = 0$, tj. $d_{ij} = 0$ pro všechna $j \in \hat{n}$. Pokud to má platit pro všechny složky vektoru $\widehat{\boldsymbol{\beta}}$, tedy pro všechna $i \in \{0, \ldots, m\}$, znamená to $\boldsymbol{D} = \boldsymbol{0}$. Celkově tedy dostáváme, že lineární nestranný odhad $\widehat{\boldsymbol{\beta}}$, který minimalizuje $\mathsf{Var}[\widehat{\beta}_i]$, $i = 0, \ldots, m$, je $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ (je to tedy OLS odhad). □

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 $\beta_1, \ldots, \beta_m$ je nenulový. Mohli bychom testovat významnost jednotlivých koeficientů, tedy hypotézy $H_0 : \beta_j = 0$, pomocí alternativy t-testu. Celková chyba I. druhu by takto ale mohla být velká, pokud máme hodně proměnných. Museli bychom hodně snížit hladinu významnosti α pro jednotlivé testy, což by mělo za následek zvýšení pravděpodobnosti chyby II. druhu (tzn. rizika 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 nulových, i když je model celkově významný (uvidíme na příkladu).

Bylo by tedy dobré mít jednu statistiku pro test

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

ANOVA přístup pro jednorozměrnou regresi naznačuje, že statistika

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

by mohla být užitečná (vyplyne to i z obecnějších přístupů k testování později). Budeme potřebovat její pravděpodobnostní rozdělení. Zavedeme následující značení:

$\overline{x}_j = \frac{1}{n}\sum\limits_{i=1}^{n} x_{ij}$ - průměr j-tého sloupce matice $\boldsymbol{X}$

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

$\boldsymbol{X}_c$ - centrovaná matice regresorů, kde $(\boldsymbol{X}_c)_{ij} = x_{ij} - \overline{x}_j$, $i = 1,\dots,n$, $j = 1,\dots,m$.

Věta 2.8. V modelu $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, kde $h(\boldsymbol{X}) = m + 1$, e_i jsou nekorelované a $e_i \sim (0, \sigma^2)$, $i \in \hat{n}$, platí

$$\mathsf{E}\left(\frac{SSR}{m}\right) = \sigma^2 + \frac{\boldsymbol{\beta}^T(\boldsymbol{X} - \overline{\boldsymbol{X}})^T(\boldsymbol{X} - \overline{\boldsymbol{X}})\boldsymbol{\beta}}{m} = \sigma^2 + \frac{\boldsymbol{\beta}_s^T \boldsymbol{X}_c^T \boldsymbol{X}_c \boldsymbol{\beta}_s}{m},$$

kde $\boldsymbol{\beta}_s = (\beta_1, \dots, \beta_m)^T$.

Důkaz. Pro vyjádření $SSR = \sum\limits_{i=1}^{n} (\hat{y}_i - \overline{y}_n)^2$ nejprve upravme predikci $\hat{y}_i$, která má předpis

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{m} x_{ij}\hat{\beta}_j\,.$$

Díky tvaru derivace součtu čtverců SSEpodle β_0,

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum_{i=1}^{n}\Big(y_i - (\beta_0 + \sum_{j=1}^{m} x_{ij}\beta_j)\Big) = 0$$

lze β_0 vyjádřit ve tvaru

$$\hat{\beta}_0 = \overline{y}_n - \sum_{j=1}^{m} \overline{x}_j\hat{\beta}_j\,.$$

Dosazením získáme vztah $\hat{y}_i - \overline{y}_n = \sum_{j=1}^{m}(x_{ij} - \overline{x}_j)\hat{\beta}_j$, $i \in \hat{n}$, zapsáno maticově:

$$\hat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}} = (\boldsymbol{X} - \overline{\boldsymbol{X}})\hat{\boldsymbol{\beta}}, \quad \text{kde } \overline{\boldsymbol{Y}} = (\overline{y}_n, \overline{y}_n, \dots, \overline{y}_n)^T_{1\times n},$$

protože první sloupec matice $\boldsymbol{X}-\overline{\boldsymbol{X}}$ je nulový. Potom

$$SSR=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y}_n)^2=(\widehat{\boldsymbol{Y}}-\overline{\boldsymbol{Y}})^T(\widehat{\boldsymbol{Y}}-\overline{\boldsymbol{Y}})=\widehat{\boldsymbol{\beta}}^T\underbrace{(\boldsymbol{X}-\overline{\boldsymbol{X}})^T(\boldsymbol{X}-\overline{\boldsymbol{X}})}_{\boldsymbol{A}}\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}^T\boldsymbol{A}\widehat{\boldsymbol{\beta}}\,.$$

Pro výpočet střední hodnoty SSR použijeme následující tvrzení z teorie pravděpodobnosti: *Nechť $Z=\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ je kvadratická forma a nechť $\mathsf{E}\,\boldsymbol{Y}=\boldsymbol{\mu}$ a $\mathsf{Cov}\,\boldsymbol{Y}=\Sigma$. Potom platí, že*

$$\mathsf{E}\,Z\;=\operatorname{tr}(\boldsymbol{A}\Sigma)+\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}.$$

Nejdříve zjednodušíme matici $\boldsymbol{A}$. Matici $\overline{\boldsymbol{X}}$ lze zapsat ve tvaru

$$\overline{\boldsymbol{X}}=\frac{1}{n}\boldsymbol{B}\boldsymbol{X},\qquad\text{kde}\qquad \boldsymbol{B}=\begin{pmatrix}1&1&\cdots&1\\ \vdots&\vdots&&\vdots\\ 1&1&\cdots&1\end{pmatrix}$$

a tedy

$$\boldsymbol{X}-\overline{\boldsymbol{X}}=\left(\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{B}\right)\boldsymbol{X}\quad\text{a}\quad(\boldsymbol{X}-\overline{\boldsymbol{X}})^T=\boldsymbol{X}^T\left(\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{B}\right).$$

Dosazením získáme

$$\boldsymbol{A}=(\boldsymbol{X}-\overline{\boldsymbol{X}})^T(\boldsymbol{X}-\overline{\boldsymbol{X}})=\boldsymbol{X}^T\left(\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{B}\right)^2\boldsymbol{X}=\boldsymbol{X}^T\left(\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{B}\right)\boldsymbol{X}\,,$$

neboť $\left(\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{B}\right)^2=\boldsymbol{I}_n-\frac{2}{n}\boldsymbol{B}+\frac{\boldsymbol{B}^2}{n^2}$ a $\frac{\boldsymbol{B}^2}{n}=\boldsymbol{B}$. Pro kovariační matici vektoru $\widehat{\boldsymbol{\beta}}$ platí $\mathsf{Cov}\,\widehat{\boldsymbol{\beta}}=\sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ a tedy

$$\boldsymbol{A}\Sigma=\sigma^2\boldsymbol{X}^T\left(\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{B}\right)\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}.$$

S využitím vztahu $\operatorname{tr}(\boldsymbol{C}\boldsymbol{D})=\operatorname{tr}(\boldsymbol{D}\boldsymbol{C})$ spočítáme stopu matice $\boldsymbol{A}\Sigma$,

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

jelikož víme, že $\boldsymbol{H}\boldsymbol{X}=\boldsymbol{X}$ a $\boldsymbol{1}=(1,\dots,1)^T$ je první sloupec $\boldsymbol{X}$, takže $\boldsymbol{H}\boldsymbol{1}=\boldsymbol{1}$, a tedy $\boldsymbol{H}\boldsymbol{B}=\boldsymbol{B}$ a stopu matice $\boldsymbol{H}$ jsme spočetli dříve. Celkem tak dostáváme

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

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

Poznámka 2.8. Pokud $\boldsymbol{\beta}_s=0$, potom $\mathsf{E}\left(\frac{SSR}{m}\right)=\sigma^2=\mathsf{E}\,s_n^2$, takže $\boldsymbol{\beta}_s\neq0$ implikuje, že $\mathsf{E}\left(\frac{SSR}{m}\right)>\sigma^2$, tedy velké hodnoty $F=\frac{SSR/m}{s_n^2}$ budou znamenat zamítnutí $H_0:\boldsymbol{\beta}_s=0$. Budeme proto potřebovat rozdělení statistiky F za platnosti H_0.

Věta 2.9. Nechť v modelu $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ tvaru (2.1) platí $h(\boldsymbol{X}) = m + 1$ a $e_1, \ldots, e_n$ jsou i.i.d. $N(0, \sigma^2)$. Pokud $\boldsymbol{\beta}_s = \boldsymbol{0}$, tj. $\beta_1 = \beta_2 = \cdots = \beta_m = 0$, potom

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

Důkaz. V důkazu minulé věty jsme ukázali

$$SSR = \widehat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{B} \right) \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{Y}}^T \left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{B} \right) \widehat{\boldsymbol{Y}}\,.$$

Tento výraz potřebujeme zjednodušit. Rozepíšeme $\widehat{\boldsymbol{Y}}$ (s využitím (2.4)),

$$\widehat{\boldsymbol{Y}} = \boldsymbol{HY} = \boldsymbol{H}(\boldsymbol{X\beta} + \boldsymbol{e}) = \boldsymbol{H}\left(\boldsymbol{1}\beta_0 + \boldsymbol{X}_v\boldsymbol{\beta}_s + \boldsymbol{e}\right) = \beta_0\boldsymbol{H1} + \boldsymbol{HX}_v\boldsymbol{\beta}_s + \boldsymbol{He} = \beta_0\boldsymbol{1} + \boldsymbol{He}\,,$$

neboť $\boldsymbol{H1} = \boldsymbol{1}$ a $\boldsymbol{\beta}_s = \boldsymbol{0}$. Pro SSR tedy platí

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

protože $\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{B}\right)\boldsymbol{1} = \boldsymbol{0}$ a $\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{B}\right)$ je symetrická. Dále platí $\boldsymbol{H} = \boldsymbol{H}^T, \boldsymbol{H}^2 = \boldsymbol{H}$ a $\boldsymbol{HB} = \boldsymbol{BH} = \boldsymbol{B}$ (protože $\boldsymbol{H1} = \boldsymbol{1}$) a celkem tedy dostáváme

$$SSR = \boldsymbol{e}^T \left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{B} \right) \boldsymbol{e} = \boldsymbol{e}^T\boldsymbol{Ce}, \quad \text{kde} \quad \boldsymbol{C} = \boldsymbol{H} - \frac{1}{n}\boldsymbol{B}\,.$$

Pro matici $\boldsymbol{C}$ platí

$$\boldsymbol{C}^T = \left(\boldsymbol{H}^T - \frac{1}{n}\boldsymbol{B}^T \right) = \left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{B} \right) = \boldsymbol{C},$$

a

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

tzn. $\boldsymbol{C}$ je symetrická a idempotentní, a proto

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

Z věty o spektrálním rozkladu navíc plyne existence ortogonání matice $\boldsymbol{Q}$ a diagonální matice $\boldsymbol{\Lambda}$ takových, že

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

neboť idempotentní matice má vlastní čísla pouze 0 a 1. Díky ortogonalitě matice $\boldsymbol{Q}$ platí $\boldsymbol{Z} = \boldsymbol{Qe} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ a

$$SSR = \boldsymbol{e}^T\boldsymbol{Q}^T \begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{Qe} = \boldsymbol{Z}^T \begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{Z} = \sum_{i=1}^{m} Z_i^2,$$

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

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

To znamená, že

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

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

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

kde $\boldsymbol{w} = \left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{B}\right)\boldsymbol{H}\boldsymbol{e} \equiv \boldsymbol{K}\boldsymbol{e}$ a vektor reziduí lze zapsat jako $\widehat{\boldsymbol{e}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{e} \equiv \boldsymbol{L}\boldsymbol{e}$. Stačí tedy ukázat, že $\boldsymbol{w}$ a $\widehat{\boldsymbol{e}}$ jsou nezávislé vektory. Sdružený vektor

$$\begin{pmatrix}\boldsymbol{w}\\ \widehat{\boldsymbol{e}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{K}\\ \boldsymbol{L}\end{pmatrix}\boldsymbol{e},$$

je lineární kombinací složek vektoru s vícerozměrným normálním rozdělením, tzn. má vícerozměrné normální rozdělení. Rozepíšeme jeho kovariční matici jako

$$\mathsf{Cov}\begin{pmatrix}\boldsymbol{w}\\ \widehat{\boldsymbol{e}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{K}\\ \boldsymbol{L}\end{pmatrix}\mathsf{Cov}(\boldsymbol{e})(\boldsymbol{K}^T, \boldsymbol{L}^T) = \sigma^2\begin{pmatrix}\boldsymbol{K}\boldsymbol{K}^T & \boldsymbol{K}\boldsymbol{L}^T\\ \boldsymbol{L}\boldsymbol{K}^T & \boldsymbol{L}\boldsymbol{L}^T\end{pmatrix}. \tag{2.6}$$

Pokud tedy ukážeme, že $\boldsymbol{K}\boldsymbol{L}^T$ je nulová matice, bude z toho plynout, že vektory $\boldsymbol{w}$ a $\widehat{\boldsymbol{e}}$ jsou nezávislé. Pro $\boldsymbol{K}\boldsymbol{L}^T$ dostaneme

$$\boldsymbol{K}\boldsymbol{L}^T = \left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{B}\right)\boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{H}) = \boldsymbol{0},$$

neboť díky idempotentnosti matice $\boldsymbol{H}$ je splněno $\boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{H}) = \boldsymbol{H} - \boldsymbol{H}^2 = \boldsymbol{0}$. □

Na základě odvozených výsledků tedy získáváme následující test:

$$\boxed{\text{zamítnout } H_0 : \beta_1 = \cdots = \beta_m = 0 \quad \text{pokud} \quad F > F_{1-\alpha}(m, n-m-1).}$$

Poznámka 2.9. Odvodili jsme pro $e_i \sim N(0, \sigma^2)$, obecně se používá, i když to nevíme. Pro velké n to může být často zdůvodněno pomocí centrální limitní věty.

Výsledky se většinou uvádějí ve formě tabulky ANOVA:

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

Význam jednotlivých zkratek je vysvětlen v popisu u Tabulky 1.1.

2.3.2 Koeficient (vícenásobné) determinace R^2

Podobně jako u jednorozměrné regrese, lze F-test chápat jako test významnosti koeficientu determinace R^2, definovaného jako

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

protože

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

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

Poznámka 2.10. Koeficient determinace R^2 je možné zvětšovat přidáváním nových proměnných x, i když jsou statisticky nevýznamné (pro n lineárně nezávislých proměnných x a n pozorování dostaneme „perfect fit“, tedy přeučení). Vysvětlení může být následující,

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

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

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

Tento upravený koeficient obsahuje určitou kompenzaci efektu rostoucího počtu proměnných, neboť s rostoucím m sice klesá SSE, ale i $n-m-1$.

2.4 Intervaly spolehlivosti a t-testy pro parametry

Pokud se model ukáže jako významný, bude nás zajímat, které koeficienty přispívají. Lze použít intervaly spolehlivosti a testy hypotéz stejně jako u jednorozměrné regrese. Výsledky jsou odvozeny pro normální chyby, v praxi se ale používají i pro jiné typy chyb (za jistých předpokladů budou platit asymptoticky, lze je tedy použít pro velká n).
Pro konstrukci použijeme dokázanou vlastnost, viz (2.5):

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

Standardním postupem získáme $100(1-\alpha)\%$ interval spolehlivosti pro β_j ve tvaru

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

S jejich pomocí lze odvodit kritický obor pro test hypotézy

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

ve tvaru (H_0 zamítneme, pokud b_j nenáleží do intervalu spolehlivosti)

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

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

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

Poznámka 2.11.

a) Pokud nejsou porušeny předpoklady modelu nebo není přítomna kolinearita, lze zvážit odstranění všech nevýznamných proměnných (dle t-testu).
b) 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).
c) Naopak, pokud má model velký počet možných proměnných, některé proměnné se mohou jevit významné, i když jsou náhodným šumem.
d) Statistiky F, R^2 a T jsou užitečné pro rozkrytí efektů jednoduchých proměnných, nemohou být ale používány úplně automaticky a je třeba postupovat obezřetně. Více detailů viz příklady z přednášky a cvičení.

2.5 Obecná lineární hypotéza

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

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

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

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

Poznámka 2.12. Uveďme pro ilustraci několik možných voleb vektoru $\boldsymbol{b}$ a matice $\boldsymbol{C}$.

a) Volba

$$\boldsymbol{b} = \boldsymbol{0} = (0, \ldots, 0)^T, \qquad \boldsymbol{C} = \left(\begin{array}{c|cccc} 0 & 1 & 0 & \ldots & 0 \\ \hline 0 & 0 & 1 & & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{array}\right)_{m\times(m+1)}$$

vede na test hypotézy

$$H_0 : \ \boldsymbol{C\beta} = \boldsymbol{0} \qquad \Leftrightarrow \qquad H_0 : \ \beta_1 = \beta_2 = \ldots = \beta_m = 0.$$

b) Volba $\boldsymbol{b} = \boldsymbol{0}$ a $\boldsymbol{C} = (0, \ldots, 0, 1, 0, \ldots, 0)$ vede na test hypotézy

$$H_0 : \ \beta_j = 0.$$

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 $\boldsymbol{C} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$ a $\boldsymbol{b} = (0,0)^T$.

Pro test obecné hypotézy H_0 naladíme 2 modely:

1) *plný model* (full model) - bez podmínek na $\boldsymbol{C\beta}$
2) *redukovaný model* (reduced model) - za předpokladu, že platí $H_0: \ \boldsymbol{C\beta} = \boldsymbol{b}$.

Označme příslušné reziduální součty čtverců SSE_F a SSE_R (bude platit $SSE_F \leqslant SSE_R$). Pokud neplatí H_0, dá se očekávat, že $\Delta SSE = SSE_R - SSE_F$ bude významně větší, než náhodná chyba σ^2. Hypotézu H_0 tedy budeme zamítat, pokud $\frac{\Delta SSE}{s_n^2}$ bude velké. To vede na zobecnění F-testu, kde testovací statistika bude mít tvar

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

a uvedené pravděpodobnostní rozdělení odvodíme pro normální chyby za platnosti hypotézy H_0.

Příklad 2.1. Odvoďme tvar navržené testovací statistiky pro hypotézu $H_0: \beta_1 = \beta_2 = \ldots = \beta_m = 0$ v modelu (2.1) (plný model). Redukovaný model bude mít tvar $Y_i = \beta_0 + e_i$, $i = 1, \ldots, n$. Pro odhad parametru β_0 v redukovaném modelu platí $\widehat{\beta}_0 = \overline{\boldsymbol{y}}$ a $SSE_R = \sum\limits_{i=1}^{n} (y_i - \overline{y}_n)^2 = SST$, tedy

$$\Delta SSE = SST - SSE_F = \sum_{i=1}^{n} (y_i - \overline{y}_n)^2 - \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = SSR.$$

Navržená statistika pak má tvar $F = \frac{\frac{SSR}{m}}{s_n^2}$, který odpovídá celkovému F-testu odvozenému dříve, pro který jsme již ukázali $F \sim F(m, n-m-1)$.

Věta 2.10. Nechť v modelu (2.1) platí, že $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ a $h(\boldsymbol{X}) = m+1$. Označme SSE_F reziduální součet čtverců plného modelu a SSE_R reziduální součet čtverců modelu, kde platí $H_0: \ \boldsymbol{C\beta} = \boldsymbol{b}$. Potom je za platnosti H_0 splněno

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

kde $\Delta SSE = SSE_R - SSE_F$.

Pro důkaz budeme potřebovat následující lemma.

Lemma 2.2. Označme $\widehat{\boldsymbol{\beta}}_F$ a $\widehat{\boldsymbol{\beta}}_R$ LSE odhady parametru $\boldsymbol{\beta}$ v plném a redukovaném modelu. Potom platí

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

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

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

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

$$L = L(\boldsymbol{\beta}, \lambda) = g(\boldsymbol{\beta}) - 2\boldsymbol{\lambda}^T(\boldsymbol{C\beta} - \boldsymbol{b}), \quad \text{kde} \quad \boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_r)^T.$$

Rozepsáním funkce $g(\boldsymbol{\beta})$ dostaneme

$$L = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta} - 2\boldsymbol{\lambda}^T\boldsymbol{C\beta} + 2\boldsymbol{\lambda}^T\boldsymbol{b}$$

a tedy

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

Z první rovnice vyjádříme odhad parametru $\boldsymbol{\beta}$ v redukovaném modelu,

$$\widehat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T\boldsymbol{\lambda} = \widehat{\boldsymbol{\beta}}_F + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T\boldsymbol{\lambda}, \tag{2.7}$$

a dosadíme do druhé

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

Pro vektor Lagrangeových multiplikátorů tak platí $\boldsymbol{\lambda} = -\left(\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T\right)^{-1}\left(\boldsymbol{C}\widehat{\boldsymbol{\beta}}_F - \boldsymbol{b}\right)$ a dosazením do rovnice (2.7) získáme

$$\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}}_F - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T\left(\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T\right)^{-1}\left(\boldsymbol{C}\widehat{\boldsymbol{\beta}}_F - \boldsymbol{b}\right) = \widehat{\boldsymbol{\beta}}_F - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T\boldsymbol{A}(\boldsymbol{C}\widehat{\boldsymbol{\beta}}_F - \boldsymbol{b}).$$

2) Z důkazu Věty 2.1 víme, že

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

Dosadíme $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_R$ a dostaneme

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

neboť díky symetrii matice $\boldsymbol{A}$ a její definici platí $\boldsymbol{A}^T\underbrace{\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T}_{=\boldsymbol{A}^{-1}}\boldsymbol{A} = \boldsymbol{A}$.

□

Důkaz Věty 2.10. Nejdříve ukážeme, že $\frac{\Delta SSE}{\sigma^2} \sim \chi^2(r)$ za platnosti $H_0 : \boldsymbol{C\beta} = \boldsymbol{b}$. Díky předpokládanému vztahu $\boldsymbol{b} = \boldsymbol{C\beta}$ platí

$$\begin{aligned}
\frac{\Delta SSE}{\sigma^2} &= \frac{(\boldsymbol{C}\widehat{\boldsymbol{\beta}}_F - \boldsymbol{b})^T\boldsymbol{A}(\boldsymbol{C}\widehat{\boldsymbol{\beta}}_F - \boldsymbol{b})}{\sigma^2} \\
&= \frac{(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})^T\boldsymbol{C}^T(\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T)^{-1}\boldsymbol{C}(\widehat{\boldsymbol{\beta}}_F - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2(r),
\end{aligned}$$

jak vyplývá z bodu 4) Věty 2.5. Bod 2) téže věty navíc uvádí rozdělení statistiky $(n-m-1)\frac{s_n^2}{\sigma^2}$ a bod 1) říká, že $\widehat{\boldsymbol{\beta}}_F$ a s_n^2 jsou nezávislé. Protože je ΔSSE funkcí pouze $\widehat{\boldsymbol{\beta}}_F$, tzn. nezávisí na s_n^2, dostaneme díky definici Fisherova rozdělení

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

□

Poznámka 2.13. Použitím rozkladu $SST = SSE + SSR$ dostaneme vztah

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

Rozdíl ΔSSE tedy může být interpretován jako nárůst regresního součtu čtverců díky neplatnosti hypotézy H_0. Dále

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

proto se ΔSSE někdy nazývá *extra sum of squares* a je to hodnota přidaná k součtu regresních čtverců SSR díky neplatnosti hypotézy H_0.
Například, pokud $\boldsymbol{\beta}_R = (\beta_0, \beta_1, \ldots, \beta_{m-1}, 0)$, tzn. $\beta_m = 0$ a skutečný model má $\boldsymbol{\beta} = \boldsymbol{\beta}_F = (\beta_0, \beta_1, \ldots, \beta_m)$, potom ΔSSE je extra regresní součet čtverců získaný díky přidání β_m do modelu. To umožňuje rozklad regresního součtu čtverců SSR plného modelu na jednotlivé části odpovídající po řadě proměnné x_1, proměnné x_2, když x_1 je už v modelu, proměnné x_3 za podmínky, že x_1 a x_2 jsou v modelu, atd. (zapsáno symbolicky $x_1, x_2|x_1, x_3|x_2x_1, \ldots$). Přesně tento rozklad provádí funkce `anova()` v prostředí R, více podrobností viz příklad z přednášky nebo cvičení.

2.6 Predikce

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

a) Intervaly spolehlivosti pro $\mathsf{E}\,(Y_{\boldsymbol{x}_0})$

Nechť $\boldsymbol{x}_0 = (1, x_{0,1}, \ldots, x_{0,m})^T$ je nový bod proměnné $\boldsymbol{x}$. Bodový odhad $\mathsf{E}\,(Y_{\boldsymbol{x}_0})$ je roven

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

tzn. $\mathsf{Var}\,(\widehat{Y}_{\boldsymbol{x}_0}) = \boldsymbol{x}_0^T \cdot \mathsf{Var}\,(\widehat{\boldsymbol{\beta}}) \cdot \boldsymbol{x}_0 = \sigma^2 \boldsymbol{x}_0^T (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0$ a může být odhadnut pomocí

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

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

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

Pro normální náhodné chyby modelu lze odvodit interval spolehlivosti pro $\mathsf{E}\,(Y_{\boldsymbol{x}_0}) = \mu_{\boldsymbol{x}_0}$. Protože $\widehat{Y}_{\boldsymbol{x}_0}$ je lineární kombinací složek náhodného vektoru s vícerozměrným normálním rozdělením, má normální rozdělení se střední hodnotou $\mathsf{E}\,(\widehat{Y}_{\boldsymbol{x}_0}) = \mu_{\boldsymbol{x}_0} = \boldsymbol{x}_0^T\boldsymbol{\beta}$ a rozptylem $\mathsf{Var}\,(\widehat{Y}_{\boldsymbol{x}_0}) = \sigma^2\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0$, to znamená

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

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

$$\frac{\widehat{Y}_{\boldsymbol{x}_0} - \mu_{\boldsymbol{x}_0}}{s_n\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim \mathrm{t}(n-m-1)\,.$$

Standardním postupem tak získáme $100(1-\alpha)\%$ interval spolehlivosti pro $\mu_{\boldsymbol{x}_0}$ ve tvaru

$$\left(\widehat{Y}_{\boldsymbol{x}_0} \pm \mathrm{t}_{1-\frac{\alpha}{2}}(n-m-1)\cdot s_n\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}\right).$$

b) Intervaly predikce pro $Y_{\boldsymbol{x}_0}$

Bodový odhad je opět $\widehat{Y}_{\boldsymbol{x}_0}$. Pokud $Y_{\boldsymbol{x}_0}$ je skutečná hodnota $Y_{\boldsymbol{x}}$ v bodě $\boldsymbol{x} = \boldsymbol{x}_0$, potom $Y_{\boldsymbol{x}_0}$ a $\widehat{Y}_{\boldsymbol{x}_0}$ budou nezávislé za předpokladu, že pozorování $Y_{\boldsymbol{x}_0}, \mathrm{Y}_1, \ldots, \mathrm{Y}_n$ jsou nezávislé (což předpokládáme). Potom

$$\mathsf{Var}\,(\widehat{Y}_{\boldsymbol{x}_0} - Y_{\boldsymbol{x}_0}) = \mathsf{Var}\,(\widehat{Y}_{\boldsymbol{x}_0}) + \mathsf{Var}\,(Y_{\boldsymbol{x}_0}) = \sigma^2(1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0),$$

takže

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

za předpokladu normality náhodných chyb modelu. Hledaný $100(1-\alpha)\%$ interval predikce pro $Y_{\boldsymbol{x}_0}$ tedy je

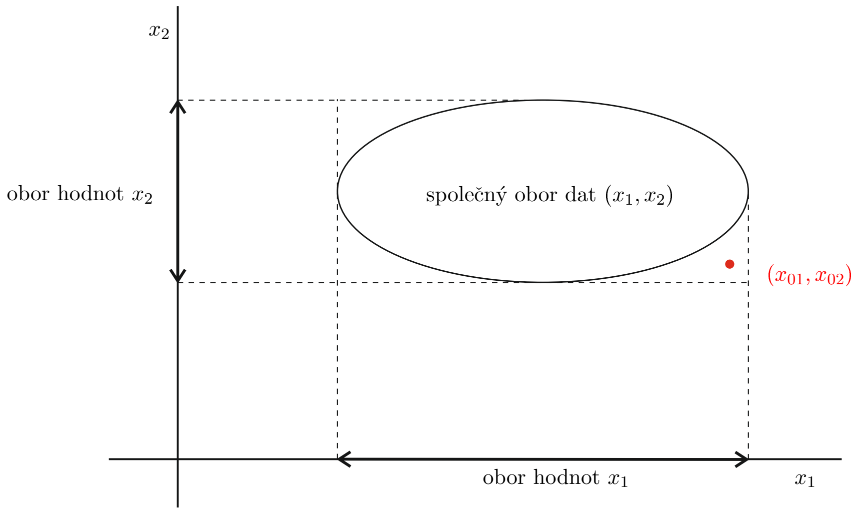
$$\left(\widehat{Y}_{\boldsymbol{x}_0} \pm \mathrm{t}_{1-\frac{\alpha}{2}}(n-m-1)\cdot s_n\sqrt{1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}\right).$$

Poznámka 2.14 (Extrapolace). U modelu jednoduché lineární regrese závisela kvalita predikce na vzdálenosti x_0 od průměru $\overline{x}_n$ a je třeba dát si pozor na predikce mimo interval $\langle x_{min}, x_{max}\rangle$.

Podobné závěry platí i pro vícerozměrný model lineární regrese. Protože rozptyl predikce je úměrný $\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0$, v bodech s velkými hodnotami této veličiny nebude predikce spolehlivá. Speciálně, pokud $\boldsymbol{x}_i^T$ jsou pozorovaná data, můžeme očekávat, že body s nejvyššími hodnotami $\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i = h_{ii}$ budou na hranici množiny, kde je predikce spolehlivá, tzn. že vnitřek elipsoidu

$$\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x} \leqslant \max_{1\leqslant j\leqslant n} h_{ii}$$

může být považován za přípustný obor pro predikci.



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

3 Rezidua, diagnostika a transformace

V této kapitole se budeme věnovat následujícím tématům:

a) Před přijímáním nějakých závěrů založených na modelu je třeba ověřit adekvátnost modelu. Máme k dispozici statistiky R^2, t, F, ty ale byly odvozeny za předpokladu linearity modelu a dalších podmínek na náhodné chyby. Pro jejich ověření je důležitým nástrojem analýza reziduí.
b) Je také třeba ověřit vliv jednotlivých pozorování na model – analýza odlehlých (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.
c) Pokud detekujeme nějaké problémy s modelem, mohou pomoci transformace proměnných nebo metoda na korekci nekonstantního rozptylu.

3.1 Rezidua

Připomeňme nejdříve již odvozené vztahy:

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

$$\widehat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{e}.$$

Dále jsme ukázali

$$\mathsf{E}\,(\widehat{\boldsymbol{e}}) = \boldsymbol{0} \quad \text{a} \quad \mathsf{Cov}\,(\widehat{\boldsymbol{e}}) = \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}).$$

Pokud navíc pro náhodné chyby platí $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ potom $\widehat{\boldsymbol{e}} \sim N_n\big(\boldsymbol{0}, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})\big)$. Když označíme $h_{ii} = \boldsymbol{H}_{ii}$, pak

$$\widehat{e}_i \sim N(0, \sigma^2(1 - h_{ii})) \quad \text{a} \quad \mathsf{Cov}\,(\widehat{e}_i, \widehat{e}_j) = -\sigma^2 h_{ij} \text{ pro } i \neq j.$$

Obecně bývá vhodnější pracovat se standardizovanými rezidui. Protože $\mathsf{Var}\,(\widehat{e}_i) = \sigma^2(1 - h_{ii})$, pro $r_i = \frac{\widehat{e}_i}{\sigma\sqrt{1-h_{ii}}}$ platí $\mathsf{Var}\,(r_i) = 1$. Pokud parametr σ odhadneme pomocí $s_n = \sqrt{\frac{1}{n-m-1}SSE}$, dostaneme

$$\widehat{r}_i = \frac{\widehat{e}_i}{s_n\sqrt{1 - h_{ii}}}, \quad i = 1, \dots, n,$$

kterým se říká *interně studentizovaná rezidua* (někdy také standardizovaná rezidua). Tato rezidua vrací v R funkce `rstandard()`.

Ukazuje se, že odlehlé pozorování může výrazně zvýšit hodnotu odhadu s_n, čímž vlastně maskuje svoje velké reziduum. Proto se někdy volí jiný postup pro standardizaci. Pokud σ^2 odhadneme na základě modelu, ve kterém bylo vynecháno i-té pozorování, označíme tento odhad $\widehat{\sigma}^2_{(-i)}$, potom

$$\widehat{t}_i = \frac{\widehat{e}_i}{\widehat{\sigma}^2_{(-i)}\sqrt{1 - h_{ii}}}, \quad i = 1, \dots, n,$$

jsou tzv. *externě studentizovaná rezidua* (někdy také studentizovaná rezidua). V R je vrací funkce `rstudent()`. Například $\widehat{\sigma}^2_{(-i)} = \frac{1}{n-m-2} SSE_{(-i)}$ je nestranný odhad σ^2 v modelu bez i-tého pozorování, jak ukážeme.

Poznámka 3.1. Platí:

a) Pokud je h_{ii} malé, pro velká n by se měla rezidua $\widehat{e}_i, \widehat{r}_i, \widehat{t}_i$ chovat přibližně stejně a $\widehat{r}_i, \widehat{t}_i$ by měla mít přibližně $N(0,1)$ rozdělení.

b) Pro malé n ($n < 20$) a/nebo $h_{ii} \approx 1$ je preferováno použití reziduí $\widehat{r}_i$ nebo $\widehat{t}_i$ a aktuálně bývá častěji doporučováno $\widehat{t}_i$ (i-té pozorování s velkými h_{ii} může zvyšovat odhad σ^2 a tím snižuje velikost svého rezidua).

c) Hodnota h_{ii} se označuje jako *potenciál i-tého pozorování* (leverage, leverage point = pákový bod / vzdálený bod). Potenciál h_{ii} hraje zásadní roli v diagnostice modelu, proto teď probereme jeho základní vlastnosti.

3.1.1 Vlastnosti potenciálu h_{ii}

a) $\mathsf{Var}\,(\widehat{e}_i) = \sigma^2(1 - h_{ii}) \geqslant 0 \quad \Rightarrow \quad h_{ii} \leqslant 1.$

b) $\boldsymbol{H}^2 = \boldsymbol{H} \Rightarrow h_{ii} = \sum_{j=1}^{n} h_{ij}h_{ji} = \sum_{j=1}^{n} (h_{ij})^2$, tedy $h_{ii} > 0$ (dá se ukázat i silnější tvrzení: $h_{ii} \geqslant \frac{1}{n}$).

c) $\boldsymbol{HX} = \boldsymbol{X} \Rightarrow \sum_{j=1}^{n} h_{ij}x_{j1} = \sum_{j=1}^{n} h_{ij} = x_{i1} = 1$ tedy $\sum_{j=1}^{n} h_{ij} = 1, \quad \forall i \in \widehat{n}$ (v modelu s interceptem).

d) Význam h_{ii} vyplyne z následujících úvahy: protože $\widehat{\boldsymbol{y}} = \boldsymbol{Hy}$,dostaneme pro i-tou složku

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{\substack{j=1 \\ i \neq j}}^{n} h_{ij}y_j.$$

Pokud $h_{ii} \approx 1$, pak díky vlastnostem zmíněným výše $\widehat{y}_i \approx y_i$ a model je nucen proložit přímku bodem $(\boldsymbol{x}_i, y_i)$, i když tam případně neplatí. Body s „velkým h_{ii}“ se nazývají body s velkým potenciálem (high leverage points) a měly by být detekovány pro další zkoumání.

e) Otázka je, jaká hodnota h_{ii} je „velká“. Pro rozhodnutí lze použít vlastnost $\sum_{i=1}^{n} h_{ii} = \mathrm{tr}(\boldsymbol{H}) = m + 1$, ze které plyne, že $\frac{m+1}{n}$ je průměrná hodnota h_{ii}. Heuristické pravidlo pak je:

$$i\text{-té pozorování má velký potenciál, jestliže } h_{ii} > \frac{3(m+1)}{n}.$$

Stejnou mez používá i funkce `influence.measures()` v R.

3.2 Grafy reziduí

A) **Ověření normality** – histogramy, kvantilové grafy (Q-Q plots)

Tyto obrázky nezávisí na počtu nezávislých proměnných x, vše stejné jako v modelu jednorozměrné lineární regrese. Používané testy normality jsou např. Shapiro-Wilk, Anderson-Darling nebo Lilliefors.

B) Pro **ověření funkční formy** pro $\mathsf{E}(Y_x)$ a/nebo konstantního rozptylu se nejčastěji používají:

1) grafy $\widehat{e}_i, \widehat{r}_i$ nebo $\widehat{t}_i$ oproti $\boldsymbol{x}_j^c$, $j = 1, ..., m$, kde $\boldsymbol{x}_j^c$ je j-tý sloupec matice $\boldsymbol{X}$,
2) grafy $\widehat{e}_i, \widehat{r}_i$ nebo $\widehat{t}_i$ oproti predikovaným hodnotám $\widehat{y}_i$,
3) partial residual plots.

Existují i testy konstantního rozptylu jako např. Breuch-Pagan nebo Levene test.

Zdůvodnění 1). Připomeňme, že v modelu bez interceptu $Y_i = \beta_1 x_i + e_i$ má odhad parametru β_1 tvar

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\|\boldsymbol{x}\|^2}.$$

Normální rovnice $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$ implikují $\boldsymbol{X}^T(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{X}^T \widehat{\boldsymbol{e}} = \boldsymbol{0}$. Pokud tedy naladíme regresní model bez interceptu pro $\widehat{\boldsymbol{e}}$ v závislosti na $\boldsymbol{x}_j^c$, odhad směrnice přímky bude

$$\widehat{\beta}_j^* = \frac{(\boldsymbol{x}_j^c)^T \widehat{\boldsymbol{e}}}{\|\boldsymbol{x}_j^c\|^2} = 0.$$

Graf reziduí $\widehat{e}_i, \widehat{r}_i, \widehat{t}_i$ oproti hodnotám j-tého regresoru, tedy $\boldsymbol{x}_j^c$, by měl dávat náhodně rozptýlené body kolem osy x (bez trendů, $\widehat{r}_i, \widehat{t}_i$ přibližně uvnitř intervalu ± 2). Pokud tomu tak není, může to naznačovat nelinearitu v $\boldsymbol{x}_j^c$ nebo nekonstantní rozptyl.

Zdůvodnění 2). Ukázali jsme $\sum_{i=1}^n \widehat{y}_i \widehat{e}_i = 0$. Pro lineární model bez interceptu pro $\widehat{e}_i$ oproti $\widehat{y}_i$ tedy platí

$$\widehat{\beta} = \frac{\widehat{\boldsymbol{e}}^T \widehat{\boldsymbol{y}}}{|\widehat{\boldsymbol{y}}|^2} = 0.$$

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

3.2.1 Partial residual plot

- I když grafy $\widehat{\boldsymbol{e}}$ oproti $\boldsymbol{x}_j^c$ a $\widehat{\boldsymbol{y}}$ mohou indikovat nedostatky modelu, nemusí být zřejmé, jaké tyto nedostatky jsou.
- V modelu jednorozměrné lineární regrese lze graf $\widehat{e}_i$ oproti x_i použít pro detekci nelinearity.
- Ve vícerozměrném modelu mohou být tyto grafy (stejně jako scatterploty) zavádějící, protože $\widehat{\boldsymbol{e}}$ závisí na všech prediktorech, nemusí být tedy izolován efekt dané proměnné při odstranění efektů ostatních.
- Pro zkoumání efektu j-té proměnné lze použít tzv. *partial rezidual plots*, které lze chápat jako ekvivalent scatterplotu v jednorozměrné regresi.

Definujme

$$\widehat{\boldsymbol{e}}_j^* = \widehat{\boldsymbol{e}} + \widehat{\beta}_j \boldsymbol{x}_j^c,$$

kde $\widehat{\boldsymbol{e}}$ je vektor reziduí modelu, $\widehat{\beta}_j$ je OLS odhad parametru β_j, $\boldsymbol{x}_j^c$ je j-tý sloupec matice $\boldsymbol{X}$.

Partial residual plot je graf hodnot $\widehat{\boldsymbol{e}}_j^*$ oproti $\boldsymbol{x}_j^c$, $j = 1, ..., m$. Pokud je model správný, měly by být body náhodně rozmístěné kolem přímky se směrnicí $\widehat{\beta}_j$. Zdůvodnění je následující. Vztah mezi $\widehat{\boldsymbol{e}}_j^*$ a $\boldsymbol{x}_j^c$ má formu jednoduchého regresního modelu bez interceptu. Pokud je model správný, $\widehat{e}_i, i \in \widehat{n}$, splňují podmínku

$$\mathsf{E}\,\widehat{e}_i = 0 \quad \text{a} \quad \mathsf{Var}\,\widehat{e}_i = \sigma^2(1 - h_{ii}).$$

Má tedy smysl uvažovat regresní model pro $\widehat{\boldsymbol{e}}_j^*$ oproti $\boldsymbol{x}_j^c$ ($\widehat{e}_j^* = \gamma_j \boldsymbol{x}_j^c + \boldsymbol{e}$). Pro odhad koeficientu γ_j platí

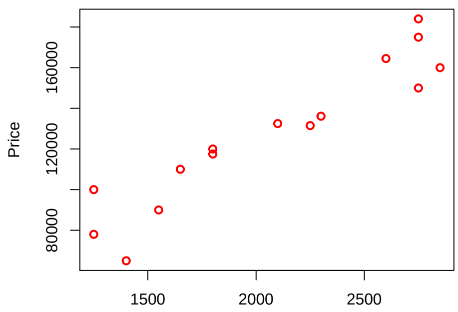
$$\widehat{\gamma}_j = \frac{(\widehat{\boldsymbol{e}}_j^*)^T \boldsymbol{x}_j^c}{\|\boldsymbol{x}_j^c\|^2} = \frac{(\widehat{\boldsymbol{e}} + \widehat{\beta}_j \boldsymbol{x}_j^c)^T \boldsymbol{x}_j^c}{\|\boldsymbol{x}_j^c\|^2} = \frac{\widehat{\boldsymbol{e}}^T \boldsymbol{x}_j^c + \widehat{\beta}_j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

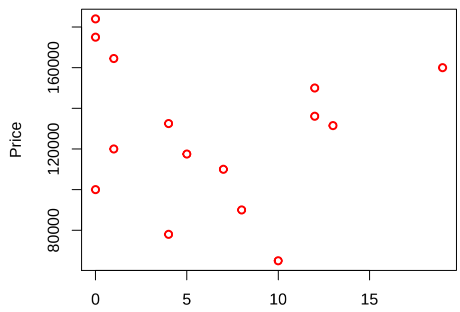
protože $\widehat{\boldsymbol{e}}^T \boldsymbol{x}_j^c = 0$.

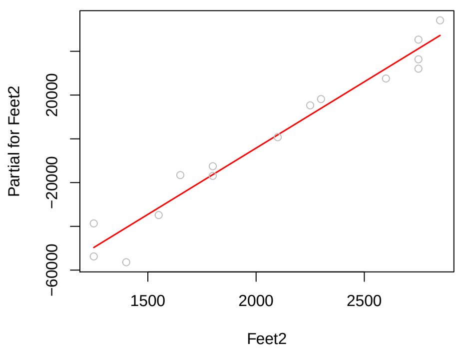
Příklad 3.1 (Housing Price Data). Pro ilustraci uvažujme data o cenách nemovitostí, které budeme chtít modelovat pomocí proměnných plocha a věk.

	Feet2	Age	Price
1	1800	1	120000
2	1650	7	110000
3	2750	12	150000
4	1550	8	90000
⋮	⋮	⋮	⋮

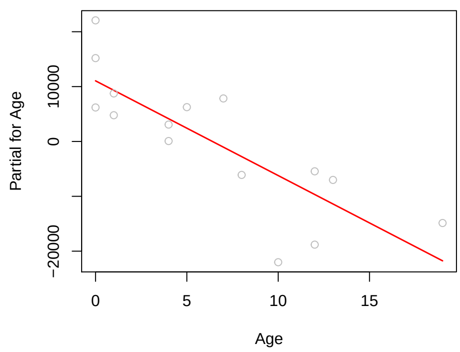
```
mod <- lm(Price ~ Feet2 + Age)
summary(mod)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13238.600   7677.183   1.724 0.110273
## Feet2          60.589      3.644  16.625 1.19e-09 ***
## Age         -1726.762    364.172  -4.742 0.000479 ***
## ---
## Residual standard error: 7763 on 12 degrees of freedom
## Multiple R-squared:  0.9586,Adjusted R-squared:  0.9517
## F-statistic:   139 on 2 and 12 DF,  p-value: 5.021e-09
```







`termplot(mod, partial.resid=TRUE, terms="Feet2")`



`termplot(mod, partial.resid=TRUE, terms="Age")`

Pokud naladíme model pomocí funkce `lm()`, vidíme, že obě proměnné jsou v modelu významné a také poměrně vysokou hodnotu koeficientu determinace R^2, tedy model dobře prokládá data. Nicméně ze scatterplotů proměnné cena oproti ploše a věku, zobrazených v první řadě obrázků, není vidět žádný zjevný lineární trend pro proměnnou věk. Pokud ale vykreslíme partial residual plots pomocí funkce `termplot()`, objeví se u proměnné věk zřetelný klesající efekt.

3.2.2 Partial regression plot

Partial residual ploty jsou někdy kritizovány za nadhodnocování efektu $\boldsymbol{x}_j^c$. Alternativou mohou být **partial regression plots** (added variable plots).
Pro odvození jejich významu začněme motivací. Ptáme se, zda přidat novou proměnnou do modelu a chtěli bychom odhadnout její efekt. Budeme tedy uvažovat rozšířený model

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

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

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

Použitím normálních rovnic pro $\boldsymbol{X}_w$ lze odvodit formuli pro $\widehat{\gamma}$

$$\widehat{\gamma} = \frac{\widehat{\mathrm{e}}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{w}}{\|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{w}\|^2}, \tag{3.1}$$

kde $\widehat{\gamma}$ je směrnice regresního modelu pro $\widehat{\boldsymbol{e}}$ v závislosti na $\boldsymbol{w}_{res} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{w}$ (tj. rezidua modelu pro $\boldsymbol{w}$ v závislosti na $\boldsymbol{X}$).
Teď uvažujme naopak, že $\boldsymbol{w}$ je sloupec původní $\boldsymbol{X}$, např. $\boldsymbol{x}_j^c$ a ozn. $\boldsymbol{X}_{(-j)}$ matici $\boldsymbol{X}$ bez sloupce j. V předchozím modelu položme $\boldsymbol{X} = \boldsymbol{X}_{(-j)}$ a $\boldsymbol{w} = \boldsymbol{x}_j^c$. Potom LSE odhad $\widehat{\beta}_j$ parametru β_j je

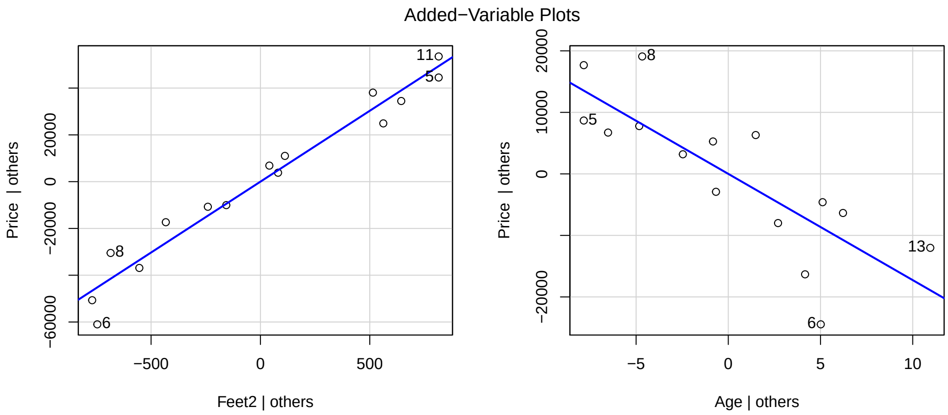
$$\widehat{\beta}_j = \frac{\widehat{\boldsymbol{e}}_{(-j)}^T\boldsymbol{x}_{j,res}^c}{\|\boldsymbol{x}_{j,res}^c\|^2},$$

kde $\widehat{\boldsymbol{e}}_{(-j)}$ jsou rezidua modelu bez $\boldsymbol{x}_j^c$ a $\boldsymbol{x}_{j,res}^c = (\boldsymbol{I}-\boldsymbol{H}_{(-j)})\boldsymbol{x}_j^c$. Jedná se tedy o rezidua modelu pro $\boldsymbol{x}_j^c$ v závislosti na ostatních proměnných, tedy $\boldsymbol{X}_{(-j)}$ (v $\boldsymbol{x}_{j,res}^c$ je odstraněn efekt ostatních regresorů). Odhad $\widehat{\beta}_j$ je vlastně směrnice regresního modelu pro $\widehat{\boldsymbol{e}}_{(-j)}$ v závislosti na rezisuích $\boldsymbol{x}_{j,res}^c$. Z toho vyplývá následující definice.
Added variable plot je graf hodnot $\widehat{\boldsymbol{e}}_{(-j)}$ proti $\boldsymbol{x}_{j,res}$, $j = 1, ..., m$. Pokud je model správný, body by měly být náhodně rozptýlené kolem přímky se směrnicí $\widehat{\beta}_j$ procházející počátkem. Pokud závislost na $\boldsymbol{x}_j^c$ není lineární, projeví se to odklonem bodů od přímky.

Poznámka 3.2. Ze vztahu (3.1) je vidět, že vícerozměrný model lineární regrese může být chápán jako posloupnost jednoduchých modelů, kde postupně vytváříme modely pro novou proměnnou s použitím reziduí modelu pro předcházející proměnné.

Příklad 3.2 (Housing Price Data).
Pro vykreslení added variable plot lze v R použít např. funkci `avPlots()` z knihovny `car`. Výstup je v tomto případě srovnatelný s příkladem 3.1.

```
mod <- lm(Price ~ Feet2 + Age)
avPlots(mod)
```



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 intervaly spolehlivosti nebo intervaly predikce, to bychom ale předem museli znát body, ve kterých chceme predikovat.
Nejjednodušší přístup, jak měřit prediktivní přesnost modelu, by byla analýza reziduí pro predikce hodnot v nových bodech $\boldsymbol{x}$, obecně ale nemáme data y v těchto bodech. Jedna možnost je použít data, která máme k dispozici.

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

Předpokládejme, že vynecháme i-té pozorování a označme $\widehat{\boldsymbol{\beta}}_{(-i)}$ odhad parametru $\boldsymbol{\beta}$ v modelu s vynechaným i-tým pozorováním ($\mathrm{M}_{(-i)}$) a $\widehat{y}_{(-i)}$ hodnotu predikovanou modelem $\mathrm{M}_{(-i)}$ v bodě $\boldsymbol{x}_i^T$, tzn. $\widehat{y}_{(-i)} = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{(-i)}$. Potom

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

nazýváme i-té *PRESS reziduum.* Součet čtverců těchto reziduí (predicted residual error sum of squares),

$$\mathrm{PRESS} = \sum_{i=1}^{n} \widehat{e}^2_{(-i)},$$

je pak užitečná míra přesnosti predikce.

Poznámka 3.3. Otázka je, jak počítat $\widehat{e}_{(-i)}, i \in \widehat{n}$. Pro velké n se zdá, že to bude náročný problém, protože pro každé $i \in \widehat{n}$ musíme naladit nový model. Naštěstí to není nutné, ukážeme totiž, že

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

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

Zavedeme následující značení:

$$\boldsymbol{x}_i^T \text{ – } i\text{-tý řádek matice } \boldsymbol{X}, \qquad \boldsymbol{X}_{(-i)} \text{ – matice } \boldsymbol{X} \text{ bez } i\text{-tého řádku.}$$

Věta 3.1. Jestliže $h_{ii} \neq 1$, potom

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

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

Důkaz. Nejdříve ukážeme, že matici $\boldsymbol{X}^T\boldsymbol{X}$ lze zapsat ve tvaru

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}_{(-i)}^T\boldsymbol{X}_{(-i)} + \boldsymbol{x}_i\boldsymbol{x}_i^T, \tag{3.2}$$

kde všechny členy mají rozměry $(m+1)\times(m+1)$. Kvůli značení předpokládáme $i = n$ (toho se dá vždy dosáhnout permutací řádků $\boldsymbol{X}$). Potom i, j-tý prvek matice $\boldsymbol{X}^T\boldsymbol{X}$ má tvar

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

Dále i, j-tý prvek matice $\boldsymbol{X}_{(-n)}^T\boldsymbol{X}_{(-n)}$ je $\sum_{k=1}^{n-1} x_{ki}x_{kj}$ a i, j-tý prvek matice $\boldsymbol{x}_n\boldsymbol{x}_n^T$ je $x_{ni}x_{nj}$, tzn. (3.2) platí.

Pro výpočet inverzní matice použijeme následující tvrzení z lineární algebry.
Věta (Sherman-Morrison-Woodburry): Nechť $\boldsymbol{A}$ je $n \times n$ invertibilní matice a nechť $\boldsymbol{z}$ je $n \times 1$ sloupcový vektor. Jestliže $\boldsymbol{z}^T\boldsymbol{A}^{-1}\boldsymbol{z} \neq 1$, potom matice $\boldsymbol{B} = \boldsymbol{A} - \boldsymbol{z}\boldsymbol{z}^T$ je invertibilní a platí

$$\boldsymbol{B}^{-1} = \boldsymbol{A}^{-1} + \frac{\boldsymbol{A}^{-1}\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{A}^{-1}}{1-\boldsymbol{z}^T\boldsymbol{A}^{-1}\boldsymbol{z}}. \tag{3.3}$$

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

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

Užitím vztahu (3.3) dostaneme

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

□

Věta 3.2. Nechť $\widehat{e}_{(-i)}$ je i-té PRESS reziduum a $h_{ii} \neq 1$. Potom

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

Důkaz. Nechť $\widehat{\boldsymbol{\beta}}_{(-i)}$ je odhad parametru $\boldsymbol{\beta}$ v modelu bez i-tého pozorování $\mathrm{M}_{(-i)}$, tedy

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

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

$$\widehat{y}_{(-i)} = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_{(-i)} = \boldsymbol{x}_i^T \left(\boldsymbol{X}_{(-i)}^T \boldsymbol{X}_{(-i)}\right)^{-1} \boldsymbol{X}_{(-i)}^T \boldsymbol{y}_{(-i)} \tag{3.4}$$

a užitím věty 3.1 dostaneme

$$\begin{aligned}\widehat{y}_{(-i)} &= \boldsymbol{x}_i^T (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_{(-i)}^T \boldsymbol{y}_{(-i)} + \frac{1}{1-h_{ii}} \boldsymbol{x}_i^T (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_{(-i)}^T \boldsymbol{y}_{(-i)} = \\ &= S_1 + \frac{1}{1-h_{ii}} S_2.\end{aligned}$$

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

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

Podobně

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

takže

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

Celkem tedy

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

□

Budeme potřebovat podobné formule pro $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}$ a $SSE_{(-i)}$.

Věta 3.3. 1) Nechť $\widehat{\boldsymbol{\beta}}_{(-i)}$ značí LSE odhad parametru $\boldsymbol{\beta}$ v modelu bez i-tého pozorování. Potom platí

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

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

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

Důkaz. 1) Stejným postupem jako v důkazu předchozí věty 3.2 (viz vztah (3.4)) lze ukázat, že

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

kde $S_1 = \widehat{\boldsymbol{\beta}} - y_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$ a $S_2 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\widehat{y}_i - y_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i h_{ii}$. Tedy

$$\begin{aligned}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)} &= y_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i - \frac{1}{1-h_{ii}}\Big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\widehat{y}_i - y_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i h_{ii}\Big) = \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\Big(y_i - \frac{\widehat{y}_i - y_i h_{ii}}{1-h_{ii}}\Big) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\Big(\frac{y_i - y_i h_{ii} - \widehat{y}_i + y_i h_{ii}}{1-h_{ii}}\Big) = \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\Big(\frac{y_i - \widehat{y}_i}{1-h_{ii}}\Big),\end{aligned}$$

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

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

$$\begin{aligned}SSE_{(-i)} &= \big(\boldsymbol{y}_{(-i)} - \boldsymbol{X}_{(-i)}\widehat{\boldsymbol{\beta}}_{(-i)}\big)^T\big(\boldsymbol{y}_{(-i)} - \boldsymbol{X}_{(-i)}\widehat{\boldsymbol{\beta}}_{(-i)}\big) = \sum_{\substack{j=1\\ j\neq i}}^{n}(y_j - \boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}}_{(-i)})^2 = \\ &= \sum_{j=1}^{n}(y_j - \boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}}_{(-i)})^2 - (y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_{(-i)})^2.\end{aligned}$$

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

$$SSE_{(-i)} = \sum_{j=1}^{n}\Big(y_j - \boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}} + \frac{\boldsymbol{x}_j^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\widehat{e}_i}{1-h_{ii}}\Big)^2 - \Big(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}} + \frac{\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\widehat{e}_i}{1-h_{ii}}\Big)^2.$$

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

$$SSE_{(-i)} = \sum_{j=1}^{n}\Big(\widehat{e}_j + \frac{h_{ij}\widehat{e}_i}{1-h_{ii}}\Big)^2 - \Big(\widehat{e}_i + \frac{h_{ii}\widehat{e}_i}{1-h_{ii}}\Big)^2 = \underbrace{\sum_{j=1}^{n}\Big(\widehat{e}_j + \frac{h_{ij}\widehat{e}_i}{1-h_{ii}}\Big)^2}_{A} - \frac{\widehat{e}_i^2}{(1-h_{ii})^2}.$$

Vyčísleme výraz A,

$$A = \sum_{j=1}^{n}\widehat{e}_j^2 + \frac{2\widehat{e}_i}{1-h_{ii}}\underbrace{\sum_{j=1}^{n}h_{ij}\widehat{e}_j}_{0} + \frac{\widehat{e}_i^2}{(1-h_{ii})^2}\underbrace{\sum_{j=1}^{n}h_{ij}^2}_{h_{ii}},$$

neboť $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ implikuje $\boldsymbol{H}\widehat{\boldsymbol{y}} = \boldsymbol{H}^2\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$, a tedy $\boldsymbol{H}\widehat{\boldsymbol{e}} = \boldsymbol{H}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{H}\boldsymbol{y} - \boldsymbol{H}\widehat{\boldsymbol{y}} = \boldsymbol{0}$ a druhá vlastnost plyne z idempotentnosti matice $\boldsymbol{H}$. Dosazením zpět dostaneme

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

□

Důsledek 3.1. V modelu (2.1) s $m+1$ parametry $\boldsymbol{\beta}$ a bez i-tého pozorování platí

$$\mathsf{E}\left(SSE_{(-i)}\right) = (n-m-2)\sigma^2,$$

takže

$$\widehat{\sigma}^2_{(-i)} = \frac{SSE_{(-i)}}{n-m-2}$$

je nestranný odhad parametru σ^2. Dále pak

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

kde $s_n^2 = \frac{1}{n-m-1} SSE$ je odhad σ^2 v plném modelu.

Důkaz. Protože $\mathsf{E}\left(\widehat{e}_i^2\right) = \mathsf{Var}\,\widehat{e}_i = \sigma^2(1-h_{ii})$, dostaneme dle předchozí věty

$$\mathsf{E}\left(SSE_{(-i)}\right) = \sum_{j=1}^{n} \sigma^2(1-h_{jj}) - \sigma^2 = \sigma^2\Big[(n-1) - \sum_{j=1}^{n} h_{jj}\Big] = \sigma^2(n-m-2),$$

neboť $\sum\limits_{j=1}^{n} h_{jj} = \operatorname{tr}(\boldsymbol{H}) = m+1$ jak bylo dokázáno dříve. Protože v modelu se všemi pozorováními platí $SSE = (n-m-1)s_n^2 = \sum\limits_{j=1}^{n} \widehat{e}_j^2$, dostaneme s využitím předchozí věty

$$\widehat{\sigma}^2_{(-i)} = \frac{SSE_{(-i)}}{n-m-2} = \frac{1}{n-m-2}\Big(\sum_{j=1}^{n} \widehat{e}_j^2 - \frac{\widehat{e}_i^2}{1-h_{ii}}\Big) = \frac{1}{n-m-2}\frac{(1-h_{ii})SSE - \widehat{e}_i^2}{1-h_{ii}}.$$

□

Poznámka 3.4. Dá se ukázat, že $SSE_{(-i)}$ a $\widehat{e}_i$ jsou nezávislé náhodné veličiny. Protože $\frac{SSE_{(-i)}}{\sigma^2} \sim \chi^2(n-m-2)$ a $\frac{\widehat{e}_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1)$, dostaneme

$$\widehat{t}_i = \frac{\widehat{e}_i}{\widehat{\sigma}_{(-i)}\sqrt{1-h_{ii}}} \sim t(n-m-2).$$

Tvrzení 3.1. Uvažujme model (2.1), kde $h(X) = m+1$ a $\boldsymbol{e} \sim N_n(0, \sigma^2 I_n)$. Nechť pro $i \in \widehat{n}$ platí, že $h_{ii} \neq 1$. Potom pro i-té (externě) studentizované reziduum platí

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

Poznámka 3.5. Reziduum $\widehat{t}_i$ lze použít pro test hypotézy, zda je i-té pozorování odlehlé (outlier), tedy

$$\begin{aligned} H_0 &: \ i\text{-té pozorování není odlehlé v modelu } M \\ H_1 &: \ i\text{-té pozorování je odlehlé v } M, \end{aligned}$$

kde odlehlé značí odlehlé vzhledem k modelu M : $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$, což může být způsobeno tím, že

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

Hypotézu H_0 zamítneme, pokud

$$|\widehat{t}_i| > t_{1-\frac{\alpha}{2}}(n-m-2).$$

Pro velká n a $\alpha = 0.05$ platí $t_{1-\frac{\alpha}{2}}(n-m-2) \approx u_{1-\frac{\alpha}{2}} \doteq 2$ a tedy rezidua $\widehat{t}_i$ by měla ležet v intervalu (-2,2). Pokud test použijeme na všechna pozorování, je potřeba aplikovat nějakou korekci na vícenásobné testování, např. Bonferroni.

Poznámka 3.6. Podívejme se na vztah vztah $\widehat{e}_{(-i)}$ a $\widehat{t}_i$. Pro PRESS rezidua $\widehat{e}_{(-i)} = \frac{\widehat{e}_i}{1-h_{ii}}$ platí

$$\mathsf{E}\,\widehat{e}_{(-i)} = 0 \quad \text{a} \quad \mathsf{Var}\,\widehat{e}_{(-i)} = \frac{\sigma^2}{1-h_{ii}}.$$

Standardizované PRESS reziduum má tedy tvar

$$\frac{\widehat{e}_{(-i)}}{\sqrt{\mathsf{Var}\,\widehat{e}_{(-i)}}} = \frac{\frac{\widehat{e}_i}{1-h_{ii}}}{\frac{\sigma}{\sqrt{1-h_{ii}}}} = \frac{\widehat{e}_i}{\sigma\sqrt{1-h_{ii}}} = r_i.$$

Pokud použijeme $\widehat{\sigma}^2_{(-i)}$ jako odhad σ^2, pak *studentizovaná PRESS rezidua* mají tvar externě studentizovaných reziduí, neboť

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

Poznámka 3.7. Ze vztahu $\widehat{e}_{(-i)} = \frac{\widehat{e}_i}{1-h_{ii}}$ plyne, že pokud i-té pozorování má velký potenciál h_{ii}, bude $\widehat{e}_{(-i)}$ mnohem větší než $\widehat{e}_i$. Pozorování s velkým h_{ii} jsou tak dobře modelována, ale měřeno $\widehat{e}_{(-i)}$ mohou špatně predikovat. To je další ukázka fit/prediction dilema.
Stejný efekt nastává také pro

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

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 $(\boldsymbol{x}, \boldsymbol{y})$ a $(\boldsymbol{x}', \boldsymbol{y}')$ 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 $\boldsymbol{X}$) na model.
Už víme, že velký potenciál h_{ii} indikuje, že i-té pozorování má velký vliv, a velká rezidua naznačují možnou neadekvátnost modelu. Zavedeme míry, které budou oba tyto faktory kombinovat. Využijeme k tomu přístup z PRESS reziduí, tzn. budeme sledovat, jak velký vliv má vynechání i-tého pozorování na odhad $\widehat{\boldsymbol{\beta}}$ a predikci $\widehat{\boldsymbol{y}}$.

3.4.1 DFBETAS a Cookova vzdálenost

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

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

který bude základem pro naši analýzu.

a) vliv i-tého pozorování na $\widehat{\beta}_j$

Pro j-tou složku rozdílu $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}$ platí

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

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

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

kde $\widehat{t}_i$ je externě studentizované reziduum. Tato míra tak kombinuje efekt velkého rezidua $\widehat{t}_i$ a velkého h_{ii}. Jedna možnost pro limitní hodnoty: i-té pozorování je považováno za influenční na odhad β_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 a budeme zkoumat vliv i-tého pozorování na celý vektor $\widehat{\boldsymbol{\beta}}$.

b) Vliv i-tého pozorování na celý vektor $\widehat{\boldsymbol{\beta}}$

Princip spočívá v použití nějaké normy na vektor $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}$. Cook navrhl

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

kde $\boldsymbol{M}$ je pozitivně definitní matice a c je normalizační konstanta. Nejužívanější volbou je $\boldsymbol{M} = \boldsymbol{X}^T\boldsymbol{X}$ a $c = s_n^2$. *Cookova vzdálenost* se potom definuje jako

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

Dosazením dostaneme

$$D_i = \frac{1}{(m+1)s_n^2}\Big(\frac{\widehat{e}_i}{1-h_{ii}}\Big)^2 \underbrace{\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\overbrace{\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}}^{I}\boldsymbol{x}_i}_{h_{ii}} = \frac{1}{m+1}\frac{h_{ii}}{1-h_{ii}}\underbrace{\frac{\widehat{e}_i^2}{s_n^2(1-h_{ii})}}_{=\widehat{r}_i^2}.$$

Výpočetní formule má potom tvar

$$D_i = \frac{\widehat{r}_i^2}{m+1}\Big(\frac{h_{ii}}{1-h_{ii}}\Big),$$

kde $\widehat{r}_i$ je interně studentizované reziduum.

Poznámka 3.8. $100(1-\alpha)\%$ simultánní interval spolehlivosti pro $\boldsymbol{\beta}$ má tvar

$$C(\alpha) = \left\{ \boldsymbol{\beta} \;\middle|\; \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{(m+1)s_n^2} \leqslant F_{1-\alpha}(m+1, n-m-1) \right\},$$

tzn.

$$\widehat{\boldsymbol{\beta}}_{(-i)} \in C(\alpha) \quad \Leftrightarrow \quad D_i \leqslant F_{1-\alpha}(m+1, n-m-1).$$

To je motivace pro následující pravidlo (*RULE OF THUMB*):

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

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

Poznámka 3.9. Také platí, že

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

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

3.4.2 DFFITS

Vliv i-tého pozorování na predikci $\widehat{y}_i$ můžeme měřit pomocí statistiky DFFITS, pro kterou platí

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

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

Poznámka 3.10. Shrňme funkce, které implementují míry influence v R:

- DFBETAS – `dfbetas()`
- DFFITS – `dffits()`
- Cookova vzdálenost D_i – `cooks.distance()`
- Leverage h_{ii} – `hatvalues()`
- vše shrnuje funkce `influence.measures()` (má navíc covariance ratio), tato funkce používá následující pravidla: i-té pozorování je influenční, pokud

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

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

3.5 Transformace

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

3.5.1 Transformace vysvětlované proměnné y

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

Tři hlavní důvody pro transformaci proměnné Y mohoubýt následující:

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

   Např. studie kapacity plic (FEV data, FEV > 0). Chtěli bychom, aby model nepredikoval záporné hodnoty, to by znamenalo klást restriktivní podmínky na parametr β. To lze obejít modelováním $y^* = \ln(\text{FEV})$.
   Pokud y jsou počty a 0 je možná hodnota, často se používá $y^* = \ln(y+1)$ nebo obecně $y^* = \ln(y+c)$

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 měříme nějakou fyzikální veličinu, která může nabývat pouze kladných hodnot). Transformace $y^* = \ln y$ nebo $y^* = y^\lambda, \lambda < 1$, budou redukovat toto vychýlení.
   Typický postup: Začít s hodnotou λ blízko 1, pak snižovat hodnotu λ, dokud není dosaženo „přibližné“ 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 μ velké (už jen z důvodu, že obor hodnot y je omezen na kladné hodnoty). Říká se tomu *positive mean-variance relationship.*

Nepřesnost měření kladných veličin se také často vyjadřuje pomocí koeficientu variace

$$CV(Y) = \frac{\text{s.d.}Y}{\mathsf{E}\,(Y)}.$$

Ten bývá často více konstantní mezi jednotlivými pozorováními, než směrodatná odchylka. Variabilitu vyjadřuje relativně, spíše než absolutně. Matematicky to znamená, že $\mathsf{Var}\,(Y) = \varphi\,\mathsf{E}\,(Y)^2 = \varphi\mu^2$ pro nějaké φ.
Pro odstranění vztahu $\mathsf{E}\,(Y)$ a $\mathsf{Var}\,(Y)$ se často používají mocninné transformace $y^* = y^\lambda$ (pro $y > 0$).

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

$\leftarrow$: Pokud $\mathsf{Var}\,(Y)$ klesá s rostoucí $\mathsf{E}\,(Y)$, budeme zvyšovat mocninu λ.
$\rightarrow$: Pokud $\mathsf{Var}\,(Y)$ roste s rostoucí $\mathsf{E}\,(Y)$, budeme λ snižovat.

Obecně předpokládejme vztah $\mathsf{Var}\,[Y] = \varphi V(\mu)$, kde V je daná funkce, a uvažujeme transformaci $y^* = h(y)$. Taylorův rozvoj 1. řádu funkce $h(y)$ v bodě μ je

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

z čehož plyne, že $\mathsf{Var}(Y^*) \simeq \big(h'(\mu)\big)^2 \cdot \mathsf{Var}(Y)$. Transformace $y^* = h(y)$ tedy bude přibližně stabilizovat rozptyl, pokud $h'(\mu)$ je úměrné $\mathsf{Var}(Y)^{-1/2} = \mathrm{V}^{-1/2}(\mu)$. Vhodnou funkci $h(\mu)$ tedy lze najít pomocí vztahu

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

Například pokud $V(\mu) = \mu^2$, stabilizující transformace je $h(y) = \ln(y)$, protože $h'(\mu) = \frac{1}{\mu}$. Nebo pokud $V(\mu) = \mu$, stabilizující transformace je $h(y) = \sqrt{y}$, protože $h'(\mu) = \frac{1}{2\sqrt{\mu}}$.
Asi nejvíce užívanou transformací je $y^* = \ln(y)$. Jedním z důvodů je i dobrá interpretovatelnost parametrů $\boldsymbol{\beta}$.

Poznámka 3.11 (Interpretace parametrů lineárního modelu).

1. Klasický lineární model:

$$\mathsf{E}(Y) = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m.$$

Jednotková změna proměnné x_j implikuje změnu $\mathsf{E}(Y)$ o β_j jednotek (při ostatních proměnných stejných). Pokud totiž označíme

$$\mathsf{E}(Y) = \boldsymbol{X\beta}, \quad \text{kde} \quad \boldsymbol{X} = (1, x_1, ..., x_m)$$

a

$$\mathsf{E}(Y_{new}) = \boldsymbol{X}_{new}\boldsymbol{\beta}, \quad \text{kde} \quad \boldsymbol{X}_{new} = (1, x_1, ..., x_j + 1, ..., x_m),$$

platí

$$\mathsf{E}(Y_{new}) - \mathsf{E}(Y) = \beta_j.$$

2. Lineární model pro transformaci $\ln Y$:

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

Pokud je to správný model, znamená to, že $\ln Y \sim N(\mu, \sigma^2)$ a veličina Y má log-normální rozdělení se střední hodnotou $\mathsf{E}(Y) = \mathrm{e}^{\mu + \frac{\sigma^2}{2}}$. To znamená, že

$$\begin{aligned} &\text{predikce pro} \quad \mathsf{E}(\ln Y) \quad \text{je} \quad \widehat{\mu} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_m x_m, \\ &\text{predikce pro} \quad \mathsf{E}(Y) \quad \text{je} \quad \mathrm{e}^{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_m x_m + \frac{\widehat{\sigma}^2}{2}}. \end{aligned}$$

Uvažujme opět jednotkovou změnu proměnné x_j ($x_j \to x_j + 1$), potom

$$\frac{\mathsf{E}(Y_{new})}{\mathsf{E}(Y)} = \frac{\mathrm{e}^{\beta_0 + \beta_1 x_1 + ... + \beta_j x_j + \beta_j + ... + \beta_m x_m + \frac{\sigma^2}{2}}}{\mathrm{e}^{\beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \frac{\sigma^2}{2}}} = \mathrm{e}^{\beta_j}.$$

To znamená, že jednotková změna proměnné x_j má za následek multiplikativní změnu střední hodnoty $\mathsf{E}(Y)$ e^{β_j}-krát. Jinak zapsáno, $100(\mathrm{e}^{\beta_j} - 1)$ je procentní změna $\mathsf{E}(Y)$ 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):

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

která předpokládá, že data y jsou pouze kladná (pokud ne, můžeme přičíst konstantu ke všem pozorováním a analyzovat takto posunutá data). Poznamenejme ještě, že funkce $y^{(\lambda)}$ je pro $\lambda = 0$ dodefinována ze spojitosti, neboť $\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \ln y$.

Pro nalezení vhodného λ budeme předpokládat, že transformované veličiny $Y_i^{(\lambda)}, i \in \hat{n}$, splňují podmínky regresního modelu, tj.

$$Y_i^{(\lambda)} = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{e}, \quad \text{kde} \quad \boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n),$$

jinak zapsáno

$$Y_i^{(\lambda)} \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2), \qquad Y_1^{(\lambda)}, \ldots, Y_n^{(\lambda)} \text{ nezávislé.}$$

Úkolem je odhadnout zároveň $\lambda, \boldsymbol{\beta}, \sigma^2$, na což použijeme metodu maximální věrohodnosti. Pomocí transformace získáme hustotu

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

Věrohodnostní funkce pro pozorování $y_1, ..., y_n$ bude mít tvar

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

Dále vyjádříme log-likelihood

$$l = \ln L = -\frac{n}{2} \ln 2\pi \underbrace{-\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i^{(\lambda)} - \mu_i)^2}_{\approx l \text{ pro LM s } \boldsymbol{y}^{(\lambda)} = (y_1^{(\lambda)}, \ldots, y_n^{(\lambda)})} + \ln J(\lambda).$$

Věrohodnostní rovnice nemají explicitní analytické řešení. Pro nalezení MLE si všimneme, že pro pevné λ je l proporcionální logaritmu věrohodnosti pro odhad $(\boldsymbol{\beta}, \sigma^2)$ na základě $\boldsymbol{y}^{(\lambda)} = (y_1^{(\lambda)}, ..., y_n^{(\lambda)})^T$ v klasickém lineárním modelu $\boldsymbol{y}^{(\lambda)} = \boldsymbol{X}^T \boldsymbol{\beta}(\lambda) + \boldsymbol{e}$. Pro pevné λ tak získáme odhady

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}^{(\lambda)},$$
$$\widehat{\sigma}^2(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(y_i^{(\lambda)} - \widehat{y}_i^{(\lambda)}\right)^2, \quad \text{kde } \widehat{y}_i^{(\lambda)} = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}(\lambda).$$

Dosazením do logaritmu věrohodnostní funkce l dostaneme po úpravě hodnotu maximalizovanou vzhledem k $(\boldsymbol{\beta}, \sigma^2)$, tzv. *profile log–likelihood*

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

Kvůli komplikované závislosti l_p na λ bude třeba numerická metoda pro maximalizaci. Celý problém lze ale přepsat do tvaru, kdy bude možné využít metody lineární regrese. Pro profile log-likelihood totiž platí

$$l_p(\lambda) = C - \frac{n}{2}\ln\widehat{\sigma}^2(\lambda) - \frac{n}{2}\ln\left(J(\lambda)\right)^{2/n} = C - \frac{n}{2}\ln\frac{\widehat{\sigma}^2(\lambda)}{(J^{\frac{1}{n}}(\lambda))^2}$$

a

$$(J(\lambda))^{\frac{1}{n}} = \left[\left(\prod_{i=1}^{n} y_i\right)^{\frac{1}{n}}\right]^{\lambda-1} = (\dot{y})^{\lambda-1}, \quad \text{kde } \dot{y} \text{ značí geometrický průměr.}$$

Dosazením zpátky do $l_p(\lambda)$ dostáváme

$$l_p(\lambda) = C - \frac{n}{2}\ln\frac{\widehat{\sigma}^2(\lambda)}{\left[(\dot{y})^{\lambda-1}\right]^2} = C - \frac{n}{2}\ln s_\lambda^2,$$

kde

$$s_\lambda^2 = \frac{\widehat{\sigma}^2(\lambda)}{\left[(\dot{y})^{\lambda-1}\right]^2} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i^{(\lambda)}}{(\dot{y})^{\lambda-1}} - \frac{\widehat{y}_i^{(\lambda)}}{(\dot{y})^{\lambda-1}}\right)^2.$$

To znamená, že s_λ^2 odpovídá reziduálnímu součtu čtverců ($\frac{1}{n}SSE$) v modelu $\frac{y_i^{(\lambda)}}{(\dot{y})^{\lambda-1}}$ v závislosti na $\boldsymbol{x}_i^T$ (tzn. s_λ^2 lze snadno získat pomocí funkce `lm()`).

Celkem tedy dostáváme, že maximalizace $l_p(\lambda)$ je ekvivalentní minimalizaci s_λ^2, zapsáno symbolicky

$$\max_\lambda l_p(\lambda) \quad \Leftrightarrow \quad \min_\lambda s_\lambda^2.$$

Algoritmus pro hledání vhodného λ

1) Zvolit oblast hodnot λ, $I = \langle\lambda_{min}, \lambda_{max}\rangle$, a body $\lambda \in I$ (typickou volbou je $I = [-2, 2]$ a 10–20 rovnoměrně rozdělených bodů).
2) Naladit model $\frac{y^{(\lambda)}}{(\dot{y})^{\lambda-1}} \sim \boldsymbol{x}$ a spočítat $\frac{1}{n}SSE = s_\lambda^2$.
3) Z grafu bodů (λ, s_λ^2) vybrat $\widehat{\lambda}$, které minimalizuje s_λ^2.
4) Pro zvolené $\widehat{\lambda}$ naladit model $y^{(\widehat{\lambda})} \sim \boldsymbol{x}$ a pokračovat standardní analýzou.

Intervaly spolehlivosti pro λ

Snadno lze odvodit test poměrem věrohodností (LRT test) pro test $\mathrm{H}_0 : \lambda = \lambda_0$. Pro volbu $\mathrm{H}_0 : \lambda = 1$, vlastně testujeme, zda je třeba transformace. Pokud zamítneme H_0, provedeme transformaci pomocí $\widehat{\lambda}$. LRT statistika má tvar

$$\Lambda = -2\ln\frac{L(\lambda_0)}{L(\widehat{\lambda})} = 2\left(l_p(\widehat{\lambda}) - l_p(\lambda_0)\right)$$

a víme, že $\Lambda \xrightarrow{L} \chi^2(1)$. Invertováním přípustné oblasti LRT testu dostaneme asymptotický interval spolehlivosti pro λ :

$$\begin{aligned} \Lambda &\leqslant \chi^2_{1-\alpha}(1) \\ 2\left(\frac{n}{2}\ln s^2_{\lambda_0} - \frac{n}{2}\ln s^2_{\widehat{\lambda}}\right) &\leqslant \chi^2_{1-\alpha}(1) \\ n \ln \frac{s^2_{\lambda_0}}{s^2_{\widehat{\lambda}}} &\leqslant \chi^2_{1-\alpha}(1). \end{aligned}$$

Pokud je tedy $\widehat{\lambda}$ MLE odhad parametru λ, asymptotický $100(1-\alpha)\%$ interval spolehlivosti pro λ má tvar

$$\left\{\lambda \in \mathbb{R} | n \cdot \ln \frac{s^2_{\lambda}}{s^2_{\widehat{\lambda}}} \leqslant \chi^2_{1-\alpha}(1)\right\}.$$

Poznámka 3.12. Kvůli jednoduchosti interpretace se často doporučuje zaokrouhlit $\widehat{\lambda}$ na nejbližší čtvrtinu nebo třetinu.

3.5.3 Transformace vysvětlujících proměnných x

Pokud diagnostika modelu naznačuje, že vztah mezi $\boldsymbol{y}$ a $\boldsymbol{X}$ není lineární pro jeden nebo více regresorů, může být vhodné přeformovat model pomocí transformací proměnných $\boldsymbol{x}$.
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 tak dostane podobu

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_j z_j + ... + \beta_m x_m + e.$$

Pokud je f známé, jedná se o model lineární regrese a lze ho analyzovat standardně. Je-li tato transformace vhodná, mělo by se to projevit ve zlepšení statistik jako je R^2, t, F a zlepšením grafu reziduí pro z_j oproti těm pro x_j. Bohužel f většinou známá není. Možný přístup je parametrizovat nějak tuto funkci a pak odhadnout tyto parametry společně s parametrem $\boldsymbol{\beta}$.
Typická parametrizace je

$$z_j = x_j^{\lambda}, \quad \lambda \in \mathbb{R},$$

kde λ je vhodná konstanta (pokud $x_j > 0$, potom $\lambda \in \mathbb{R}$, nicméně pokud může být x_j záporné, je množina hodnot λ omezená).
Lze také použít aproximaci f pomocí polynomu vhodného stupně, tzn.

$$z_j = \sum_{k=1}^{l} r_k x_j^k,$$

kde konstanty r_k musí být odhadnuty. Výsledný model ale v tomto případě nebude lineární v parametrech $\beta_j, j = 0, ..., m$ a $r_k, k = 1, ..., l$.
Další možností je použití trigonometrických funkcí nebo splines (piecewice polynomials).

Zaměříme se na transformaci $z_j = x_j^\lambda$

Jedna možnost by opět byla zvolit jistou množinu hodnot λ, naladit modely pro všechna λ a vybrat model s nejlepší shodou s daty, např. s nejmenší SSE nebo největší R^2 nebo F. To může být ale časově náročné nebo můžeme minout vhodnou hodnotu λ, pokud nebyla v původní množině (nevíme jak R^2, F, SSE závisí na λ). Popíšeme jednu metou, která se snaží hledat vhodné λ jiným způsobem.

Box-Tidwell metoda

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

$$x^\lambda \approx x^1 + (\lambda - 1)\frac{\mathrm{d}x^\lambda}{\mathrm{d}\lambda}\bigg|_{\lambda=1}, \quad \text{kde} \quad \frac{\mathrm{d}x^\lambda}{\mathrm{d}\lambda}\bigg|_{\lambda=1} = x^\lambda \ln x\Big|_{\lambda=1} = x \ln x,$$

tedy

$$x^\lambda \approx x + (\lambda - 1)x \ln x.$$

Dosazením do modelu

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

získáme lineární model pro parametry β_k, $0 \leqslant k \leqslant m+1$, a protože $\beta_{m+1} = (\lambda-1)\beta_j$, můžeme parametry λ a β_j odhadnout následovně:

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

$$\widehat{\lambda} = \frac{\widehat{\beta}_{m+1}}{\widehat{\beta}_j} + 1.$$

Tento postup umožňuje testovat potřebu transformace, tedy hypotézu

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

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

Poznámka 3.13. Pokud model s $\widehat{\lambda}$ vypadá neadekvátně, lze postupovat iterativně a získat posloupnost $\widehat{\lambda}(l)$, $l \geqslant 1$. Položíme $\widehat{\lambda}(0) = \widehat{\lambda}$ a rozvineme x_j^λ kolem $\widehat{\lambda}(0)$, tzn.

$$x_j^\lambda \approx x_j^{\widehat{\lambda}(0)} + \big(\lambda - \widehat{\lambda}(0)\big)x_j^{\widehat{\lambda}(0)} \ln x_j$$

a dosazením do rovnice modelu získáme

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

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

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

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

Poznámka 3.14. Další užívané transformace v $\boldsymbol{x}, \boldsymbol{y}$:

a) centrované proměnné: transformujeme matici $\boldsymbol{X}$ na centrovanou verzi $\boldsymbol{X}_C$ tak, že $(\boldsymbol{X}_C)_{ij} = x_{ij} - \overline{x}_j$, $i \in \widehat{n}$, $j \in \widehat{m}$, kde $\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ je průměr j-tého sloupce matice $\boldsymbol{X}$, centrovaný vektor $\boldsymbol{y}$ má tvar $(\boldsymbol{y}_C)_i = y_i - \overline{y}$. Parametry pak odhadneme následovně

1) $\widehat{\beta}_1, ..., \widehat{\beta}_m$ jsou řešením soustavy normálních rovnic $\boldsymbol{X}_C^T \boldsymbol{X}_C \boldsymbol{\beta} = \boldsymbol{X}_C^T \boldsymbol{y}_C$,

2) $\widehat{\beta}_0 = \overline{y}_n - \sum_{j=1}^{m} \widehat{\beta}_j \overline{x}_j$.

b) centrované a škálované proměnné: škálování sloupců centrované matice tak, aby jejich norma byla jedna, tzn. každý prvek j-tého sloupce matice $\boldsymbol{X}_C$ podělíme $s_j = \left(\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2\right)^{\frac{1}{2}}$. Centrovaná a škálovaná matice $\boldsymbol{X}_{SC}$ pak bude

$$\boldsymbol{X}_{SC} = \boldsymbol{X}_C \boldsymbol{S}, \qquad \boldsymbol{S} = \operatorname{diag}\left(\frac{1}{s_1}, ..., \frac{1}{s_m}\right)$$

a model má tvar

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

Lze použít i $\boldsymbol{Y}_{SC}$, tedy centrovanou a škálovanou verzi vektoru $\boldsymbol{Y}$.

3.6 Vážené nejmenší čtverce (weighted least squares - WLS)

Budeme nyní předpokládat, že chyby e_i jsou normální, nezávislé, ale jejich rozptyl $\mathsf{Var}(e_i) = \sigma_i^2$ závisí na i. Konkrétně tak, že $\sigma_i^2 = \frac{\sigma^2}{w_i}$, kde $w_i > 0$, $i = 1, \ldots, n$, se nazývají váhy. Uvažujeme tedy model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \quad \text{kde} \quad \boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{W}) \text{ a } \boldsymbol{W} = \operatorname{diag}\left(\frac{1}{w_1}, ..., \frac{1}{w_n}\right). \tag{3.5}$$

Pokud jsou váhy w_i známé, lze MLE odhady parametrů $\boldsymbol{\beta}$ a σ^2 nalézt následujícím způsobem. Označíme

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

a definujeme $\boldsymbol{Z} = \boldsymbol{K}^{-1}\boldsymbol{Y}$, $\boldsymbol{M} = \boldsymbol{K}^{-1}\boldsymbol{X}$ a $\boldsymbol{\varepsilon} = \boldsymbol{K}^{-1}\boldsymbol{e}$. Přenásobením rovnice modelu (3.5) maticí $\boldsymbol{K}^{-1}$ zleva, dostaneme model

$$\boldsymbol{Z} = \boldsymbol{M}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ kde } \boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \boldsymbol{I}_n), \tag{3.6}$$

protože

$$\mathsf{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{K}^{-1}\sigma^2 \boldsymbol{W} \left(\boldsymbol{K}^{-1}\right)^T = \sigma^2 \boldsymbol{K}^{-1}\boldsymbol{K}\boldsymbol{K}^T \left(\boldsymbol{K}^T\right)^{-1} = \sigma^2 \boldsymbol{I}_n.$$

Transformovaný vektor $\boldsymbol{Y}$ tak má tvar $\boldsymbol{Z} = (\sqrt{w_1}Y_1, \dots, \sqrt{w_n}Y_n)^T$. Transformovaný model (3.6) už je standardní model lineární regrese, ve kterém díky vztahu $(\boldsymbol{K}^{-1})^T\boldsymbol{K}^{-1} = (\boldsymbol{K}\boldsymbol{K}^T)^{-1} = \boldsymbol{W}^{-1}$ platí

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_w &= (\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T\boldsymbol{z} = \left(\boldsymbol{X}^T(\boldsymbol{K}^{-1})^T\boldsymbol{K}^{-1}X\right)^{-1}\boldsymbol{X}^T(\boldsymbol{K}^{-1})^T\boldsymbol{K}^{-1}\boldsymbol{y}\\ &= (\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{y},\\ \widehat{\sigma^2}_w &= \frac{1}{n}\sum_{i=1}^n(z_i-\widehat{z}_i)^2 = \frac{1}{n}\sum_{i=1}^n w_i(y_i-\widehat{y}_i)^2 = \frac{1}{n}SSE_w,\end{aligned}$$

kde SSE_w označuje vážený součet čtverců, $z_i = \sqrt{w_i}y_i$ a $\widehat{z}_i = \sqrt{w_i}\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}} = \sqrt{w_i}\widehat{y}_i$.
Dále platí

a) $\mathsf{E}\,\widehat{\boldsymbol{\beta}}_w = (\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\underbrace{\mathsf{E}\,\boldsymbol{Y}}_{\boldsymbol{X}\boldsymbol{\beta}} = \boldsymbol{\beta}$, tedy $\widehat{\boldsymbol{\beta}}_w$ je nestranný odhad $\boldsymbol{\beta}$,

b) $\mathsf{E}\left(\frac{SSE_w}{n-m-1}\right) = \sigma^2$, tedy $s_w^2 = \dfrac{SSE_w}{n-m-1}$ je nestranný odhad σ^2.

Věta 3.4. Nechť $\widehat{\boldsymbol{\beta}}_w$ je WLS odhad parametru $\boldsymbol{\beta}$ v modelu (3.5), tedy jestliže $\mathsf{Cov}\,(\boldsymbol{e}) = \sigma^2\boldsymbol{W} = \sigma^2\mathrm{diag}\left(\frac{1}{w_1}, \dots, \frac{1}{w_n}\right)$. Potom platí:

1) $\mathsf{Cov}\,(\widehat{\boldsymbol{\beta}}_w) = \sigma^2(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}$.

2) Nechť δ_i je i-tý diagonální prvek matice $(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}$. Jestliže $e_i \sim N\left(0, \frac{\sigma^2}{w_i}\right)$, $i \in \widehat{n}$, potom
$$T_i = \frac{\widehat{\beta}_{w,i}-\beta_i}{s_w\sqrt{\delta_i}} \sim t(n-m-1).$$

3) Pro $\widehat{\boldsymbol{Y}}_w = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_w$ platí, že $\mathsf{E}\,\widehat{\boldsymbol{Y}}_w = \boldsymbol{X}\boldsymbol{\beta}$ a $\mathsf{Cov}\,(\widehat{\boldsymbol{Y}}_w) = \sigma^2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

4) Nechť $\widehat{\boldsymbol{e}}_w = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_w$ jsou rezidua v modelu (3.5) a $\widehat{\boldsymbol{\varepsilon}}_w = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{M}\widehat{\boldsymbol{\beta}}_w$ jsou rezidua v transformovaném modelu (3.6). Potom
$$\widehat{\boldsymbol{\varepsilon}}_w = \sqrt{\boldsymbol{W}^{-1}}\widehat{\boldsymbol{e}}_w = \boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{e}}_w \quad \text{a} \quad \mathsf{E}\,(\widehat{\boldsymbol{e}}_w) = \mathsf{E}\,(\widehat{\boldsymbol{\varepsilon}}_w) = \boldsymbol{0}.$$

5) Nechť $\boldsymbol{H}_w = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}$ je vážená projekční matice. Potom
$$\widehat{\boldsymbol{e}}_w = (\boldsymbol{I}_n - \boldsymbol{H}_w)\boldsymbol{e} \quad \text{a} \quad \mathsf{Cov}\,(\widehat{\boldsymbol{e}}_w) = \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}_w)\boldsymbol{W}.$$
To znamená, že
$$\mathsf{Cov}\,(\widehat{\boldsymbol{\varepsilon}}_w) = \sigma^2\boldsymbol{W}^{-\frac{1}{2}}(\boldsymbol{I}_n - \boldsymbol{H}_w)\boldsymbol{W}^{\frac{1}{2}}.$$

Důkaz. 1) Pro kovariační matici vektoru $\widehat{\boldsymbol{\beta}}_w$ platí
$$\mathsf{Cov}\,(\widehat{\boldsymbol{\beta}}_w) = (\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\underbrace{\mathsf{Cov}\,\boldsymbol{Y}}_{=\sigma^2\boldsymbol{W}}\boldsymbol{W}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}.$$

2) Z předchozího bodu plyne $\mathsf{Var}\,\widehat{\beta}_{w,i} = \sigma^2\delta_i$, tzn. $\frac{\widehat{\beta}_{w,i}-\beta_i}{\sigma\sqrt{\delta_i}} \sim N(0,1)$ a víme, že $\widehat{\beta}_{w,i}$ 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
$$\frac{\widehat{\beta}_{w,i}-\beta_i}{s_w\sqrt{\delta_i}} \sim t(n-m-1).$$

3) Pro střední hodnotu a kovariační matici vektoru $\widehat{\boldsymbol{Y}}_w$ dostaneme

$$\mathsf{E}\,\widehat{\boldsymbol{Y}}_w = \boldsymbol{X}\,\mathsf{E}\,\widehat{\boldsymbol{\beta}}_w = \boldsymbol{X}\boldsymbol{\beta}, \quad \mathsf{Cov}\,(\widehat{\boldsymbol{Y}}_w) = \boldsymbol{X}\,\mathsf{Cov}\,\widehat{\boldsymbol{\beta}}_w\boldsymbol{X}^T = \sigma^2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

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

$$\widehat{\varepsilon}_w = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{Y} - \boldsymbol{W}^{-\frac{1}{2}}\boldsymbol{X}\widehat{\boldsymbol{\beta}}_w = \boldsymbol{W}^{-\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_w) = \boldsymbol{W}^{-\frac{1}{2}}\widehat{\boldsymbol{e}}_w.$$

Pro střední hodnoty pak platí

$$\mathsf{E}\,\widehat{\boldsymbol{e}}_w = \mathsf{E}\,(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_w) = \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \quad \Rightarrow \quad \mathsf{E}\,\widehat{\varepsilon} = \boldsymbol{0}.$$

5) Pro vektor reziduí platí

$$\widehat{\boldsymbol{e}}_w = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}_w = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_w = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}\underbrace{(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e})}_{\boldsymbol{Y}} =$$
$$= \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} - \underbrace{\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1}}_{\boldsymbol{H}_w}\boldsymbol{e} = (\boldsymbol{I}_n - \boldsymbol{H}_w)\boldsymbol{e}.$$

Pro jeho kovariační matici dostaneme

$$\begin{aligned}
\mathsf{Cov}\,(\widehat{\boldsymbol{e}}_w) &= (\boldsymbol{I}_n - \boldsymbol{H}_w)\,\mathsf{Cov}\,(\boldsymbol{e})(\boldsymbol{I}_n - \boldsymbol{H}_w)^T = \\
&= \sigma^2(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{-1})\boldsymbol{W}\big(\boldsymbol{I}_n - \boldsymbol{W}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\big) \\
&= \sigma^2\boldsymbol{W} - 2\sigma^2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T + \sigma^2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T \\
&= \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}_w)\boldsymbol{W}.
\end{aligned}$$

Následně tak

$$\mathsf{Cov}\,(\widehat{\varepsilon}_w) = \boldsymbol{W}^{-\frac{1}{2}}\,\mathsf{Cov}\,(\widehat{\boldsymbol{e}}_w)\boldsymbol{W}^{-\frac{1}{2}} = \sigma^2\boldsymbol{W}^{-\frac{1}{2}}(\boldsymbol{I}_n - \boldsymbol{H}_w)\boldsymbol{W}^{\frac{1}{2}}.$$

□

Z odvozeného vyplývá, že odhady parametrů $\boldsymbol{\beta}$ a σ^2 lze získat použitím transformovaného modelu (3.6). Protože ale transformovaný model neobsahuje intercept (první sloupec matice $\boldsymbol{M}$ je $(\sqrt{w_1}, ..., \sqrt{w_n})^T$), nefunguje klasický rozklad součtu čtverců a F statistiku nelze definovat obvyklým způsobem, stejně jako koeficient determinace R^2 (viz. regrese skrz počátek).
Nicméně princip „extra sum of squares“ funguje, ať má model intercept nebo ne. Například celkový F-test lze provést pomocí statistiky

$$F_w = \frac{\frac{SSE_R - SSE_F}{m}}{s_w^2},$$

kde SSE_F je reziduální součet čtverců s_w^2 plného modelu a SSE_R je reziduální součet čtverců redukovaného transformovaného modelu $\boldsymbol{Z} = \boldsymbol{M}_0\boldsymbol{\beta}_0 + \boldsymbol{e}$, kde $\boldsymbol{M}_0 = (\sqrt{w_1}, ..., \sqrt{w_n})^T$.
Pokud mají chyby normální rozdělení, pak za platnosti $H_0 : \beta_1 = ... = \beta_m = 0$ je splněno

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

a H_0 zamítáme, pokud $F_w > F_{1-\alpha}(m, n - m - 1)$.
Přirozené je definovat koeficient determinace předpisem $R^2 = \varrho^2(\widehat{\boldsymbol{z}}, \boldsymbol{z})$, kde $\varrho(\widehat{\boldsymbol{z}}, \boldsymbol{z})$ je výběrový korelační koeficient. Pro $\boldsymbol{W} = \boldsymbol{I}_n$ dostaneme standardní tvar statistiky R^2.

3.6.1 Analýza reziduí pro WLS

Pro analýzu reziduí je třeba uvažovat vhodné grafy reziduí, máme totiž k dispozici dva vektory reziduí,

$$\widehat{e}_i \text{ v původním modelu (3.5)}, \quad \widehat{\varepsilon}_i \text{ v transformovaném modelu (3.6)},$$

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). Při grafické analýze je třeba být opatrný oproti jakým hodnotám budeme rezidua zobrazovat. Grafy $\widehat{\varepsilon}_i$ proti sloupcům matice $\boldsymbol{M}$ a predikovaným hodnotám $\widehat{\boldsymbol{z}}$ mají své opodstatnění, neboť např. $\sum_{i=1}^{n} \widehat{z}_i \widehat{\varepsilon}_i = 0$ (jsou OG, měl by být vidět rozptýlený oblak kolem osy x). Dosazením $\widehat{\varepsilon}_i = \sqrt{w_i} \cdot \widehat{e}_i$ a $\widehat{z}_i = \sqrt{w_i} \cdot \widehat{y}_i$ dostaneme $\sum_{i=1}^{n} w_i \widehat{e}_i \widehat{y}_i = 0$, tzn. graf $\widehat{e}_i$ proti $\widehat{y}_i$ bude zavádějící. Graf $\sqrt{w_i} \cdot \widehat{e}_i$ proti $\sqrt{w_i} \cdot \widehat{y}_i$ je ale v pořádku. Podobné závěry platí i pro grafy $\widehat{e}_i$ proti $\boldsymbol{x}_j^c, i = 1, ..., m$.

Poznámka 3.15. Pokud jsou váhy neznámé, bylo by třeba je odhadnou společně s $\boldsymbol{\beta}$ a σ^2 z dat. To ale není obecně možné, protože máme více parametrů, než pozorování. Někdy to možné je, pokud máme další informace o rozdělení chyb (např. tvar kovariační matice atd.).

Poznámka 3.16. Celý postup vážených nejmenších čtverců lze použít i na případ $\boldsymbol{e} \sim N_m(0, \sigma^2 \boldsymbol{W})$, kde $\boldsymbol{W}$ je známá, ale není diagonální. Protože $\boldsymbol{W}$ je symetrická a pozitivně definitní, existuje regulární matice $\boldsymbol{K}$ tak, že $\boldsymbol{W} = \boldsymbol{K}\boldsymbol{K}^T$. Stejná transformace jako u WLS opět vede na transformovaný model, kde $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \boldsymbol{I}_n)$.

3.7 Korelované chyby

Zejména v časových nebo ekonomických datech se často objevuje korelace jednotlivých hodnot, potom ale není splněn předpoklad nezávislosti chyb a tento stav je třeba detekovat (někdy pomohou grafy reziduí). Modely pro korelovaná data se podrobněji zabývá **Analýza časových řad**, zde se omezíme na jednu metodu detekce korelovanosti.

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

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

3.7.1 Durbin-Watson statistika

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

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

kde ϱ je autokorelační koeficient, $u_t \sim N(0, \sigma_n^2)$ jsou nezávislé, $t = 1, \ldots, n$, a u_t je nezávislé na $e_t, t \geqslant 1$. Častěji pro data časových řad platí $\varrho > 0$ (pozitivní autokorelace) a pro test $H_0 : \varrho = 0$ vs. $H_1 : \varrho > 0$ se užívá *Durbin-Watsonova statistika*

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

kde $\widehat{e}_t$ jsou rezidua modelu lineární regrese. Pokud zamítneme H_0, odhadne se ϱ pomocí formule

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

Poznámka 3.17. Dá se ukázat, že $d \approx 2(1 - \widehat{\varrho})$, platí totiž

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

a $\sum_{t=1}^n \widehat{e}_t^2 \approx \sum_{t=2}^n \widehat{e}_t^2$ pro velké n díky vlastnosti $\mathsf{E}\, \widehat{e}_t = 0$. Protože $\widehat{\varrho}$ je pro velká n blízké výběrovému korelačnímu koeficientu vektorů $(\widehat{e}_2, \ldots, \widehat{e}_n)$ a $(\widehat{e}_1, \ldots, \widehat{e}_{n-1})$, plyne z Cauchy-Schwartzovy nerovnosti, že $\widehat{\varrho}$ leží přibližně v intervalu $(-1, 1)$, tzn. d leží přibližně v intervalu $(0, 4)$. Pokud bude $\widehat{\varrho} \approx 1$, bude $d \approx 0$ a pro $\widehat{\varrho} \approx 0$ bude $d \approx 2$. Tzn. pro malé hodnoty d budeme zamítat H_0, pro velké hodnoty nebudeme zamítat. Kritické hodnoty určené Durbinem a Watsonem jsou tabelované a test se provádí následovně:

Test:

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

Poznámka 3.18. Pro test $H_0 : \varrho = 0$ vs. $H_1 : \varrho < 0$ lze použít popsaný test pro $d' = 4 - d$. Jenom zmíníme, že pro korekci autokorelace se používá např. metoda *Cochrane-Orcutt*.

4 Výběr regresního modelu

Budeme se zabývat výběrem nejvhodnější množiny regresorů. Špatná specifikace modelu (použití jiného než skutečného modelu) má dva hlavní důsledky:

1) Při vynechání některých proměnných modelu, jsou odhady parametrů ostatních proměnných vychýlené.
2) Pokud je v modelu příliš mnoho proměnných, jsou obecně rozptyly odhadů pro ostatní proměnné velké.

Volba "nejlepšího"modelu je hledání kompromisu mezi dvěma kritérii:

a) přesností modelu   b) jednoduchostí modelu (parsimony)

Ideální model by měl mít nejmenší možný počet regresorů, který umožňuje adekvátní interpretaci (nebo predikci). Je nutné poznamenat, že obvykle neexistuje jednoznačně nejlepší model ani jednoznačná statistická procedura, jak ho najít.

Poznámka 4.1. Podívejme se blíže na důsledky vynechání proměnných ze skutečného, i když neznámého modelu. Předpokládejme, že skutečný model má tvar

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \quad \text{a} \quad \boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T,$$

kde

$$\boldsymbol{\beta}_1 = (\beta_0, \ldots, \beta_p)^T, \quad \boldsymbol{\beta}_2 = (\beta_{p+1}, \ldots, \beta_m)^T.$$

Podobně lze rozepsat i matici $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ a model lze tedy zapsat následovně

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$$

a pro odhad parametru $\boldsymbol{\beta}$ platí

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

Uvažujme teď redukovaný model (model bez regresorů obsažených v matici $\boldsymbol{X}_2$)

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{e}$$

ve kterém je odhad parametru $\boldsymbol{\beta}_1$ roven

$$\widehat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y}\,.$$

Víme, že $\widehat{\boldsymbol{\beta}}$ je nestranný odhad parametru $\boldsymbol{\beta}$, ale odhad $\widehat{\boldsymbol{\beta}}_1$, získaný v modelu s vynechanými proměnnými odpovídajícími parametru $\boldsymbol{\beta}_2$, už není nestranný odhad parametru $\boldsymbol{\beta}_1$, neboť

$$\mathsf{E}\,(\widehat{\boldsymbol{\beta}}_1^T) = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\,\mathsf{E}\,(\boldsymbol{Y}) = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 + \boldsymbol{A}\boldsymbol{\beta}_2,$$

protože ve skutečném modelu platí $\mathsf{E}\,(\boldsymbol{Y}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2$. Obecně platí $\boldsymbol{A} \neq \boldsymbol{0}$, takže $\mathsf{E}\,(\widehat{\boldsymbol{\beta}}_1) \neq \boldsymbol{\beta}_1$.

Pro kovariační matice obou odhadů platí

$$\mathsf{Cov}\,(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}, \qquad \mathsf{Cov}\,(\widehat{\boldsymbol{\beta}}_1) = \sigma^2 \left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}.$$

Pokud označíme $\widehat{\boldsymbol{\beta}}_1^*$ LSE odhad parametru $\boldsymbol{\beta}_1$ v modelu s parametrem $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, dá se ukázat, že matice $\mathsf{Cov}\,(\widehat{\boldsymbol{\beta}}_1^*) - \mathsf{Cov}\,(\widehat{\boldsymbol{\beta}}_1)$ je pozitivně semidefinitní. To znamená, že rozptyly složek odhadu $\widehat{\boldsymbol{\beta}}_1^*$ budou obecně větší než rozptyly složek odhadu $\widehat{\boldsymbol{\beta}}_1$, tedy vynechání proměnných zvyšuje přesnost odhadů $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$.

4.1 Kritéria pro porovnávání modelů

Předpokládejme, že máme k dispozici T proměnných (regresorů) včetně interceptu a uvažujme podmnožinu p proměnných (včetně interceptu). Uvedeme přehled kritérií, která lze použít pro porovnávání modelů.

A) Koeficient vícenásobné determinace R^2

$$R_p^2 = \frac{SSR_p}{SST} = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2}{\sum_{i=1}^{n} (y_i - \overline{y}_n)^2} = 1 - \frac{SSE_p}{SST}$$

- Při použití je třeba si uvědomit, že R_p^2 je rostoucí funkcí počtu regresorů p (SST je konstantní), tedy maxima nabyde pro $p = T$.
- Hledáme tedy model ve kterém přidání dalšího regresoru už nezpůsobí podstatný nárůst statistiky R_p^2.
- Často se používá upravený koeficient determinace

$$\overline{R}_p^2 = 1 - \frac{\frac{SSE_p}{n-p}}{\frac{SST}{n-1}}.$$

B) Střední kvadratická chyba MSE nebo její odmocnina RMSE

$$MSE_p = \frac{SSE_p}{n-p} = s_n^2, \qquad RMSE_p = s_n$$

- Hledáme model s minimální hodnotou statistiky s_n^2.

C) F-test pro vnořené modely

Pro $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ a $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ umíme významnost regresorů v $\boldsymbol{X}_2$, tedy hypotézu $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$, otestovat pomocí F-testu (extra sum of squares princip). V R lze použít funkci `anova()` pro porovnání dvou modelů,

`anova(mod1,mod2)` - pozor je třeba dát na vnořenost modelů, nebo na test významnosti jednotlivých regresorů v modelu,

`anova(mod)` - v tomto případě ale záleží na pořadí v jakém přidáváme regresory do modelu.

D) Mallows C_p, AIC, BIC

Mallowsova statistika C_p, Akaikeho informační kritérium AIC a bayesovské informační kritérium BIC jsou kritéria beroucí více v potaz počet použitých regresorů. Lze je použít i pro **nevnořené modely**!

Mallowsova statistika C_p

Tato statistika je definována vztahem

$$C_p = \frac{SSE_p}{\widehat{\sigma}^2} - n + 2p, \quad \text{kde} \quad \widehat{\sigma}^2 = \frac{SSE_T}{n-T}.$$

Vlastnosti statistiky C_p:

1. Snadno se počítá, SSE_p a $\widehat{\sigma}^2$ jsou implementované.
2. Pokud je $\widehat{\sigma}^2$ konzistentní odhad σ^2 (nezávisející 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 je $\widehat{\sigma}^2$ konstantní, SSE_p klesá, p roste (C_p se snaží sladit dvě protichůdná kritéria).
4. $C_T = T$.
5. Pokud je správný model s p parametry, dá se ukázat, že $C_p \approx p$ pro $n \gg T$.
6. V praxi se volí model s nejmenším C_p ve skupině modelů splňujících $C_p \approx p$.

Poznámka 4.2. Nevýhodou použití Mallowsovy statistiky C_p je, že pro dobrou interpretaci je třeba spočítat hodnoty C_p pro všechny nebo většinu podmnožin regresorů.

Akaikeho informační kritérium AIC

Definice AIC má pro obecný statistický model tvar

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

kde $\widehat{\theta}$ je MLE odhad parametru θ v modelu, l je logaritmus věrohodnostní funkce modelu a p^* je počet parametrů, které je třeba odhadnout (v našem případě $p^* = p + 1$, počítáme i σ^2).
Pro námi uvažovaný model lineární regrese má věrohodnostní funkce tvar

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X\beta})^T (\boldsymbol{y} - \boldsymbol{X\beta})\right]$$

a její logaritmus je

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2} \frac{(\boldsymbol{y} - \boldsymbol{X\beta})^T (\boldsymbol{y} - \boldsymbol{X\beta})}{\sigma^2}.$$

Statistika AIC má tedy tvar

$$\mathrm{AIC} = -2l(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) + 2p^* = n \ln 2\pi + n \ln \widehat{\sigma}^2 + \frac{\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)}{\widehat{\sigma}^2} + 2p^*$$

a dosazením MLE odhadu $\widehat{\sigma}^2 = \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = \frac{SSE}{n}$ dostaneme

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

Protože $n \ln 2\pi + n$ je konstanta, která pro daná data neovlivní porovnávání dvou modelů pomocí AIC, definuje se někdy alternativně

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

Poznámka 4.3. *Použití AIC:* hledáme model s minimální hodnotou AIC statistiky. Je třeba si uvědomit, že AIC není mírou kvality modelu, je ale užitečná pro porovnávání modelů.
V R lze hodnotu AIC statistiky získat pomocí následujících funkcí:

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

(Schwarzovo) bayesovské informační kritérium BIC

Statistika BIC je definována předpisem

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

Toto kritérium více penalizuje počet parametrů, jeho použití má tedy za následek výběr modelů s jednodušší interpretací, než AIC. BIC vyžaduje významnější příspěvek proměnné, aby byla zařazena do modelu.
V R lze hodnotu BIC statistiky získat pomocí funkce `BIC()` nebo pomocí funkcí `AIC(.), extractAIC(.)` s volbou `k = log(nobs(fit))`.

E) PRESS statistika

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

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

Vybírá se model s minimální hodnotou této statistiky.

4.2 Metody výběru modelu

Nyní popíšeme čtyři standardně používané metody výběru vhodných regresorů do modelu.

1) Vyhodnocení všech možných modelů

Pro T dostupných regresorů to znamená naladit 2^T modelů a pak je porovnat pomocí nějakého kritéria. To může být náročné pro velká T (například $T = 10$ dostupných proměnných 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 statistikou) nebo jejíž odstranění znamená největší zlepšení modelu (měřeno AIC).

Algoritmus:

1) Naladíme 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_{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, $\alpha_{\text{out}} = 0.05, 0.1, ...$).
4) Opakujeme kroky 2) a 3), dokud všechny částečné F statistiky nejsou větší než příslušná kritická hodnota F_{out}, tzn. nelze už vyřadit žádnou proměnnou.

Poznámka 4.4. Místo F statistiky lze používat kritérium AIC. Z modelu pak vyřazujeme tu proměnnou, jejíž vyřazení má za následek největší pokles AIC. Proceduru ukončíme, pokud už vyřazením žádné další proměnné hodonota AIC neklesne.

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) Naladíme minimální model.
2) Pro každou dostupnou proměnnou spočteme F statistiku pro test významnosti jejího přidání do modelu.
3) Pokud některá z těchto F statistik překračuje kritickou hodnotu F_{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_{in} nebo dokud nezbyde žádná proměnná na přidání do modelu.

Poznámka 4.5. I když tento postup zjednodušuje výběr modelu, často bohužel vede na zařazení proměnných, které nemají významný příspěvek, jakmile jsou zařazeny další proměnné.

4) Postupná regrese (stepwise regression)

Jedná se o kombinaci dvou předchozích metod. V každém kroku algoritmu přidáme jednu proměnnou a poté zkontrolujeme, zda není možné nějakou odebrat. Budeme potřebovat dvě kritické hodnoty F_{in}, F_{out} (pro použití F statistiky).

Algoritmus:

1) Naladíme 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 (např. největší pokles AIC).
3) V novém modelu zjistíme, zda nelze některou proměnnou vynechat (opět pomocí AIC nebo F). Pokud ano, vynecháme proměnnou, jejíž vyřazení má za následek největší zlepšení modelu (např. největší pokles AIC).
4) Opakujeme kroky 2) a 3) do té doby, až nebude možné přidat ani ubrat žádnou proměnnou.

Příklad 4.1. Pro příklady použití popsaných postupů v prostředí R odkazujeme na materiály z přednášek a cvičení.

Poznámka 4.6 (Princip marginality). Při výběru proměnných do modely by měly být dodrženy následující principy:

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

Poznámka 4.7. Jakmile nalezneme optimální model, je třeba řádně ověřit adekvátnost.

5 Kolinearita (multikolinearita)

Budeme předpokládat, že matice $\boldsymbol{X}$ má *plnou hodnost* a studovat situaci, kdy je $\boldsymbol{X}^T\boldsymbol{X}$ na pokraji singularity. V tomto případě mluvíme o *špatně podmíněné matici* $\boldsymbol{X}$. Např. výpočet $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ může být problematický (díky moderním numerickým metodám, jako je QR rozklad, už to není téměř problém), nejsou to však jediné potíže, které může špatná podmíněnost $\boldsymbol{X}$ způsobit.
Kolinearitou tedy míníme situaci, kdy alespoň jeden ze sloupců matice $\boldsymbol{X}$ je „skoro“ lineární kombinací ostatních sloupců. Otázka je, jak to vyjádřit. Jeden možný přístup:

$$\exists \boldsymbol{c} = (c_1, \dots, c_{m+1})^T, \quad \|\boldsymbol{c}\| = 1, \quad \text{tak, že} \quad \sum_{i=1}^{m+1} c_i \boldsymbol{x}_i \approx \boldsymbol{\varepsilon}, \quad \text{kde } \|\boldsymbol{\varepsilon}\| \text{ je malé.}$$

Dá se ukázat, že

$$\text{kolinearita je přítomna} \quad \Longleftrightarrow \quad \boldsymbol{X}^T\boldsymbol{X} \text{ má alespoň jedno vlastní číslo malé.}$$

Možné zdroje kolinearity jsou:

- způsob sběru dat,
- omezení v populaci, ze které byla data vybírána,
- špatná specifikace modelu (přeurčení).

Zabývejme se teď otázkou, jak kolinearitu rozpoznat. První nápad by mohl být, podívat se, jak je hodnota determinantu $\boldsymbol{X}^T\boldsymbol{X}$ blízko 0. To se ale ukazuje jako nevhodné, neboť vynásobením matice číslem lze udělat determinant libovolně velký bez ovlivnění závislosti mezi sloupci, totéž platí i pro vlastní čísla. Násobení matice číslem ale nic nezmění na poměru vlastních čísel.
Například tedy poměr největšího ku nejmenšímu vlastnímu číslu matice $\boldsymbol{X}^T\boldsymbol{X}$ lze použít jako charakteristiku. Pokud je matice singulární, alespoň jedno vlastní číslo je nulové, matice na pokraji singularity bude mít jedno vlastní číslo výrazně menší než to největší. Z historických důvodů se používá odmocnina z vlastních čísel.
Díky předpokladu $h(\boldsymbol{X}) = m + 1$ je matice $\boldsymbol{X}^T\boldsymbol{X}$ je pozitivně definitní a tedy vlastní čísla $\lambda_i > 0$, $\forall i \in \widehat{m+1}$. Předpokládejme, že $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{m+1} > 0$ a položme $s_i = \sqrt{\lambda_i}$.

Definice 5.1. Pro $j = 1, \dots, m+1$ rozumíme j-tým indexem podmíněnosti matice $\boldsymbol{X}$ veličinu $\eta_j = \dfrac{s_1}{s_j}$. *Index podmíněnosti matice* $\boldsymbol{X}$ definujeme jako

$$\kappa(\boldsymbol{X}) = \eta_{m+1} = \frac{s_1}{s_{m+1}}.$$

Věta 5.1. Nechť $\boldsymbol{PSQ}^T$ je singulární rozklad matice $\boldsymbol{X}$. Potom pro $j \in \widehat{m+1}$ platí

$$\mathsf{Var}\,(\widehat{\beta}_j) = \sigma^2 \sum_{i=1}^{m-1} \frac{q_{ji}^2}{s_i^2}\,,$$

kde q_{ij} je (i,j)-tý prvek matice $\boldsymbol{Q}$ a $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{m+1})^T$ je OLS odhad parametru $\boldsymbol{\beta}$ v modelu (2.1).

Důkaz. V důkazu využijeme následující lemma o singulárním rozkladu matice.

Lemma (singulární rozklad matice). Nechť matice $\boldsymbol{A} \in \mathbb{R}^{n,m}$, $n \geqslant m$, má hodnost $r \leqslant m$. Potom existují matice $\boldsymbol{P} \in \mathbb{R}^{n,m}$, $\boldsymbol{S} \in \mathbb{R}^{m,m}$ a $\boldsymbol{Q} \in \mathbb{R}^{m,m}$ takové, že platí

$$\boldsymbol{A} = \boldsymbol{PSQ}^T, \quad \boldsymbol{S} \text{ je diagonální}, \quad \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}_m, \quad \text{a} \quad \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{QQ}^T = \boldsymbol{I}_m.$$

Kovariační matice odhadu $\widehat{\boldsymbol{\beta}}$ má tvar $\mathsf{Cov}\,(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$. Dosazením singulárního rozkladu $\boldsymbol{X} = \boldsymbol{PSQ}^T$ dostaneme

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{QSP}^T\boldsymbol{PSQ}^T = \boldsymbol{QS}^2\boldsymbol{Q}^T$$

a pro inverzní matici platí

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} = \left(\boldsymbol{Q}^T\right)^{-1}\boldsymbol{S}^{-2}\boldsymbol{Q}^{-1} = \boldsymbol{QS}^{-2}\boldsymbol{Q}^T,$$

neboť $\left(\boldsymbol{Q}^T\right)^{-1} = \boldsymbol{Q}$ a $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T$. Pro rozptyl j-té složky odhadu tak platí

$$\mathsf{Var}\,\widehat{\beta}_j = \sigma^2 \sum_{i=1}^{m+1} q_{ji}^2 \frac{1}{s_i^2}\,.$$

□

Dokázáná věta má následující důsledky:

- pokud je alespoň jedno s_i malé, rozptyl $\widehat{\boldsymbol{\beta}}$ může být velký,
- pokud je jedno s_i malé ve srovnání s ostatními s_j, bude mít i-tý člen sumy velkou váhu a může „destabilizovat“ odhad.

Věta 5.2. V modelu $\boldsymbol{Y} \sim (\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$ platí

$$\mathsf{E}\,\|\widehat{\boldsymbol{Y}}\|^2 = \|\boldsymbol{X\beta}\|^2 + \sigma^2 h(\boldsymbol{X}).$$

Má-li $\boldsymbol{X}$ lineárně nezávislé sloupce, platí

$$\mathsf{E}\,\|\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\beta}\|^2 + \sigma^2\mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\,.$$

Důkaz. Rozepíšeme střední hodnotu $\mathsf{E}\,\|\widehat{\boldsymbol{Y}} - \boldsymbol{X\beta}\|^2$ dvěma způsoby. Jednak platí

$$\begin{aligned}\mathsf{E}\,\|\widehat{\boldsymbol{Y}} - \boldsymbol{X\beta}\|^2 &= \mathsf{E}\left(\widehat{\boldsymbol{Y}} - \boldsymbol{X\beta}\right)^T\left(\widehat{\boldsymbol{Y}} - \boldsymbol{X\beta}\right) = \mathrm{tr}\left[\left(\widehat{\boldsymbol{Y}} - \boldsymbol{X\beta}\right)\left(\widehat{\boldsymbol{Y}} - \boldsymbol{X\beta}\right)^T\right]\\ &= \mathrm{tr}[\,\mathsf{Cov}\,\widehat{\boldsymbol{Y}}] = \mathrm{tr}[\,\mathsf{Cov}\,\boldsymbol{HY}] = \mathrm{tr}[\sigma^2\boldsymbol{H}] = \sigma^2 h(\boldsymbol{X})\,.\end{aligned}$$

Roznásobením druhého členu v předchozí rovnosti pak dostaneme

$$\mathsf{E}\,\|\widehat{\boldsymbol{Y}} - \boldsymbol{X\beta}\|^2 = \mathsf{E}\,\|\widehat{\boldsymbol{Y}}\|^2 - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\,\mathsf{E}\,\widehat{\boldsymbol{Y}} + \|\boldsymbol{X\beta}\|^2 = \mathsf{E}\,\|\widehat{\boldsymbol{Y}}\|^2 - \|\boldsymbol{X\beta}\|^2.$$

Porovnáním obou rovností získáme vztah

$$\mathsf{E}\,\|\widehat{\boldsymbol{Y}}\|^2 = \|\boldsymbol{X}\boldsymbol{\beta}\|^2 + \sigma^2 h(\boldsymbol{X})\,.$$

Podobně rozepíšeme $\mathsf{E}\,\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$,

$$\begin{aligned}\mathsf{E}\,\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 &= \mathsf{E}\,\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \operatorname{tr}\left[\,\mathsf{E}\,\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\right] \\ &= \operatorname{tr}[\,\mathsf{Cov}\,\widehat{\boldsymbol{\beta}}] = \sigma^2 \operatorname{tr}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\end{aligned}$$

a

$$\mathsf{E}\,\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \mathsf{E}\,\|\widehat{\boldsymbol{\beta}}\|^2 - 2\boldsymbol{\beta}^T\,\mathsf{E}\,\widehat{\boldsymbol{\beta}} + \|\boldsymbol{\beta}\|^2 = \mathsf{E}\,\|\widehat{\boldsymbol{\beta}}\|^2 - \|\boldsymbol{\beta}\|^2.$$

Porovnáním dostaneme

$$\mathsf{E}\,\|\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\beta}\|^2 + \sigma^2 \operatorname{tr}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\,.$$

□

Poznámka 5.1. Střední hodnota $\mathsf{E}\,\|\widehat{\boldsymbol{Y}}\|^2$ závisí pouze na $h(\boldsymbol{X})$, nikoli na tom, jak moc jsou sloupce matice $\boldsymbol{X}$ lineárně nezávislé, kolinearita tu tedy nehraje významnou roli. To ale neplatí pro $\mathsf{E}\,\|\widehat{\boldsymbol{\beta}}\|^2$, protože

$$\mathsf{E}\,\|\widehat{\boldsymbol{\beta}}\|^2 = \|\boldsymbol{\beta}\|^2 + \sum_{j=0}^{m} \mathsf{Var}\,\widehat{\beta}_j$$

a $\mathsf{Var}\,\widehat{\beta}_j$ může být velké. Kolinearita tedy ovlivňuje více interpretaci než predikci.

Poznámka 5.2. Pozor si musíme dát na na změnu indexu podmíněnosti $\kappa(\boldsymbol{X})$ při různém škálování různých sloupců matice $\boldsymbol{X}$, např. změna jednotek jedné proměnné. Proto se mnohdy jednotlivé regresory centrují a škálují, aby měly stejnou normu (pokud bude norma sloupců 1, bude matice $\boldsymbol{X}_{sc}^T\boldsymbol{X}_{sc}$ odpovídat výběrové korelační matici původních regresorů). Centrování není vhodné, pokud má intercept vlastní věcnou interpretaci, neboť centrováním o intercept přijdeme.

Příklad 5.1 (Data Trees).

```
mod <- lm(Volume ~ Girth + Height)
summary(mod)
## Coeff:      Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## ---
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948, Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16

vif(mod)
##   Girth  Height
## 1.36921 1.36921

# jednotky XX  palce, stopy
C <- model.matrix(mod)
kappa(C,exact = TRUE)

## 959.377

# jednotky XX cm, metry
Cm <- cbind(rep(1,31), 2.54*Girth, 0.305*Height)
```

```
kappa(Cm,exact = TRUE)

## 514.5681

mod.scale <- lm(scale(Volume) ~ scale(Girth)
                            + scale(Height))
summary(mod.scale)

##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.813e-16  4.241e-02   0.000   1.0000
## scale(Girth)   8.988e-01  5.045e-02  17.816   <2e-16 ***
## scale(Height)  1.315e-01  5.045e-02   2.607   0.0145 *
## ---
## Residual standard error: 0.2362 on 28 degrees of freedom
## Multiple R-squared: 0.948, Adjusted R-squared: 0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16

D <- model.matrix(mod.scale)
kappa(D,exact = TRUE)

## 1.777759
```

5.1 Regrese standardizovaných veličin (regression in correlation form)

Uvažujme model s interceptem, $h(\boldsymbol{X}) = m + 1$ a $\boldsymbol{e} \sim (\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. To znamená, pro $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \sum_{j=1}^{m} x_{ij}\beta_j + e_i = \left(\beta_0 + \sum_{j=1}^{m} \overline{x}_j \beta_j\right) + \sum_{j=1}^{m} (x_{ij} - \overline{x}_j)\beta_j + e_i.$$

Podělením zatím neurčeným T_0 dostaneme

$$\frac{Y_i}{T_0} = \frac{\beta_0 + \sum_{j=1}^{m} \overline{x}_j \beta_j}{T_0} + \sum_{j=1}^{m} \frac{x_{ij} - \overline{x}_j}{T_j} \frac{T_j}{T_0}\beta_j + \frac{e_i}{T_0} = \frac{\beta_0^0}{T_0} + \sum_{j=1}^{m} x_{ij}^* \beta_j^* + e_i^*,$$

kde

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad T_j = \sqrt{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}, \quad x_{ij}^* = \frac{x_{ij} - \overline{x}_j}{T_j},$$

$$\beta_j^* = \frac{T_j}{T_0}\beta_j, \quad \beta_0^0 = \beta_0 + \sum_{j=1}^{m} \overline{x}_j \beta_j \quad \text{a} \quad e_i^* = \frac{e_i}{T_0}.$$

Označme $\boldsymbol{X}^* = \boldsymbol{X}_{sc} = (x_{ij}^*)_{n \times m}$ (sloupce matice $\boldsymbol{X}^*$ jsou ortogonální vůči vektoru $\boldsymbol{1}_n$ a mají normu 1). Normální rovnice uvažovaného modelu mají tvar ($\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}$ v klasickém modelu)

$$\begin{pmatrix} n & \boldsymbol{0}^T \\ \boldsymbol{0} & (\boldsymbol{X}^*)^T\boldsymbol{X}^* \end{pmatrix} \begin{pmatrix} \frac{\beta_0^0}{T_0} \\ \boldsymbol{\beta}^* \end{pmatrix} = \begin{pmatrix} \frac{n\overline{y}_n}{T_0} \\ \frac{(\boldsymbol{X}^*)^T\boldsymbol{y}}{T_0} \end{pmatrix},$$

jejich řešení je

$$\widehat{\beta}_0^0 = \overline{y}_n, \qquad \widehat{\boldsymbol{\beta}}^* = \left((\boldsymbol{X}^*)^T\boldsymbol{X}^*\right)^{-1} (\boldsymbol{X}^*)^T \frac{\boldsymbol{y}}{T_0}$$

a díky ortogonalitě je odhad $\widehat{\beta}_0^0$ nekorelovaný s $\widehat{\boldsymbol{\beta}}^*$.

Snadno se ukáže, že matice $(\boldsymbol{X}^*)^T\boldsymbol{X}^* = R_{xx}$ je výběrová korelační matice původních regresorů. Pokud zvolíme $T_0 = \sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}$, je vektor $(\boldsymbol{X}^*)^T \frac{\boldsymbol{y}}{T_0} = r_{xy}$ vektorem výběrových korelačních koeficientů mezi původními (i upavenými) sloupci $\boldsymbol{x}_j$ a $\boldsymbol{y}$, to znamená

$$\widehat{\boldsymbol{\beta}}^* = R_{xx}^{-1} r_{xy}.$$

Formálně lze postup chápat tak, že jsme od x_{ij}, y_i přešli k $x_{ij}^*, y_i^* = \frac{y_i - \overline{y}_n}{T_0}$, tedy k jejich standardizovaným verzím. Pro volbu $T_0 = 1$ dostaneme pouze centrované proměnné $\boldsymbol{y}$.
Model tak bude mít tvar

$$Y_i^* = \sum_{j=1}^m x_{ij}^* \beta_j^* + e_i^*, \qquad \text{kde} \quad \beta_j^* = \frac{T_j}{T_0}\beta_j, \quad e_i^* = \frac{e_i}{T_0}$$

a odhady původních parametrů získáme pomocí vztahů

$$\widehat{\beta}_j = \frac{T_0}{T_j}\widehat{\beta}_j^*, \quad \widehat{\beta}_0 = \overline{y} - \sum_{j=1}^m \widehat{\beta}_j \overline{x}_j.$$

Poznámka 5.3. Odhady $\widehat{\beta}_j^*$ se někdy nazývají *beta weights* a mají důležitou interpretaci: vztahují se k regresorům vyjádřeným bezrozměrně a ve stejném měřítku, tzn. *ukazují relativní vliv jednotlivých regresorů na y.*

Podívejme se na vztah mezi statistikou SSE^* a koeficientem determinace R^2 původního modelu (R^2 je stejná i ve standardizovaném modelu),

$$SSE^* = \sum_{i=1}^n (\widehat{e}_i^*)^2 = \sum_{i=1}^n \left(\frac{\widehat{e}_i}{T_0}\right)^2 = \frac{SSE}{T_0^2} = 1 - \left(1 - \frac{SSE}{T_0^2}\right) = 1 - R^2.$$

Dále víme, že $\mathsf{Cov}\,(\widehat{\boldsymbol{\beta}}^*) = \sigma^{*2}((\boldsymbol{X}^*)^T\boldsymbol{X}^*)^{-1} = \sigma^{2*}R_{xx}^{-1}$ a při použití odhadu parametru σ^{2*} ve tvaru $s_n^{*2} = \frac{SSE^*}{n-m-1}$ dostaneme odhad

$$\widehat{\mathsf{Cov}}\,(\widehat{\boldsymbol{\beta}}^*) = \frac{SSE^*}{n-m-1}R_{xx}^{-1} = \frac{1-R^2}{n-m-1}R_{xx}^{-1} \quad \text{a} \quad \widehat{\mathsf{Var}}\,(\widehat{\beta}_j^*) = \frac{1-R^2}{n-m-1}r_{jj},$$

kde r_{jj} je j-tý diagonální prvek matice R_{xx}^{-1}.
Dá se ukázat, že $r_{jj} = \frac{1}{1-R_j^2}$, kde R_j^2 je koeficient determinace v modelu pro j-tý sloupec matice $\boldsymbol{X}$ v závislosti na ostatních. Celkem tedy

$$\widehat{\mathsf{Var}}\,(\widehat{\beta}_j^*) = \frac{1-R^2}{n-m-1}\frac{1}{1-R_j^2},$$

nejmenší rozptyl tak dostaneme pro $R_j^2 = 0$, s rostoucí hodnotou R_j^2 rozptyl odhadu $\widehat{\beta}_j^*$ roste.
Charakteristika

$$VIF_j = \frac{1}{1-R_j^2}$$

se nazývá *inflační faktor (variance inflation factor)* a udává, kolikrát se zhorší rozptyl odhadu $\widehat{\beta}_j^*$ v důsledku korelovanosti j-tého regresoru s ostatními regresory.

Stejný význam má VIF_j i pro odhad $\widehat{\beta}_j$, neboť

$$\widehat{\beta}_j = \frac{T_0}{T_j}\widehat{\beta}_j^* \quad \Longrightarrow \quad \widehat{\mathsf{Var}}\,(\widehat{\beta}_j) = \frac{1-R^2}{n-m-1}\left(\frac{T_0}{T_j}\right)^2 \frac{1}{1-R_j^2}.$$

Poznámka 5.4. Testovací statistika pro test hypotézy $H_0 : \beta_j = 0$ je

$$T_j = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\mathsf{Var}}\,\widehat{\beta}_j}} = \frac{\frac{T_0}{T_j}\widehat{\beta}_j^*}{\sqrt{\widehat{\mathsf{Var}}\left(\frac{T_0}{T_j}\widehat{\beta}_j^*\right)}} = \frac{\widehat{\beta}_j^*}{\sqrt{\widehat{\mathsf{Var}}\,\widehat{\beta}_j^*}} = T_j^*,$$

což znamená, že t-testy jsou stejné v původním i standardizovaném modelu. Velký inflační faktor VIF_j vyžaduje pro zamítnutí H_0 větší hodnotu $|\widehat{\beta}_j^*|$, proto často t-testy vycházejí nevýznamné v přítomnosti kolinearity.

Poznámka 5.5. Při výpočtu odhadu $\widehat{\beta}_j^*$ si v R můžeme pomoci funkcí `scale()`, která provádí centrování a škálování vektoru $\boldsymbol{x}$ pomocí $\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2}$. Konstanta $\frac{1}{n-1}$ je tam navíc, což ale odhady $\widehat{\beta}_j^*$ ani index podmíněnosti matice $\boldsymbol{X}^*$ neovlivní.
Při použití funkce `scale()` se v modelu ponechává intercept, aby byly zachovány potřebné stupně volnosti.

5.2 Zjišťování (detekce) kolinearity

Základní možnosti, jak detekovat kolinearitu jsou:

A) **pomocí indexu podmíněnosti $\kappa(\boldsymbol{X}^*) = \kappa(\boldsymbol{X}_{sc})$:**

$\kappa(\boldsymbol{X}^*) > 100$ $\rightarrow$ silná kolinearita (jedinou možností je vypuštění nějakého sloupce),
$\kappa(\boldsymbol{X}^*) > \kappa$, kde $\kappa \in (10, 30)$ $\rightarrow$ použití metody na potlačení kolinearity.

B) **pomocí hodnoty VIF:**

$$VIF_j > 10 \quad \text{může indikovat problémy.}$$

Hodnota $VIF_j \geqslant 10$ odpovídá tomu, že $R_j^2 \geqslant 0.9$. Už jsme zmínili, že R_j^2 nemusí být jednoznačně svázán s mírou lineární závislosti, inflační faktor VIF tak nemusí být vždy spolehlivý. Na druhou stranu má ale jednoduchou statistickou interpretaci a snadno se počítá. Většinou představuje spolehlivou náhradu indexu podmíněnosti $\kappa(\boldsymbol{X}^*)$ (když např, balíky nepodporují analýzu vlastních čísel).

C) **další možné příznaky:**

- velké hodnoty odhadnutých parametrů $\widehat{\boldsymbol{\beta}}$,
- velké směrodatné odchylky odhadů $\widehat{\boldsymbol{\beta}}$,
- jeden nebo více odhadnutých regresních koeficientů se špatným znaménkem.

5.3 Potlačení kolinearity

Kolinearitu je možné potlačit některým z následujících způsobů:

- získáním dalších dat,
- změnou formulace modelu, někdy pomůže centrování proměnných nebo vynechání proměnných (to ale může vést k vychýleným odhadům),

- kolinearita má za následek velké rozptyly odhadů. Gauss-Markovova věta říká, že OLS odhad je BLUE parametru $\boldsymbol{\beta}$. To znamená zmenšení rozptylu odhadu $\widehat{\boldsymbol{\beta}}$ je možné jen použitím *nelineárních* nebo *vychýlených* odhadů. Jednu z takových metod probereme podrobněji.

5.3.1 Hřebenová regrese (Ridge regression)

Zůstaneme u standardizovaného modelu, tzn. uvažujeme standardizované verze $\boldsymbol{X}^*, \boldsymbol{Y}^*$. Pro jednoduchost značení budeme používat symboly $\boldsymbol{X}, \boldsymbol{Y}$ a $\boldsymbol{\beta}, \sigma^2$ místo $\boldsymbol{\beta}^*, \sigma^{2*}$ a předpokládat, že $h(\boldsymbol{X}) = m$.
Myšlenka metody hřebenové regrese spočívá v umělém zvednutí diagonály matice $\boldsymbol{X}^T\boldsymbol{X}$, čímž se potlačí kolinearita. Uvažujme matici

$$\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m,$$

kde $\delta > 0$ je malé. Ukážeme, že je tato matice regulární a tedy pozitivně definitní. Matice $\boldsymbol{X}^T\boldsymbol{X}$ je symetrická, tedy díky lemmatu o spektrálním rozkladu matice existuje ortogonální matice $\boldsymbol{Q}$ taková, že

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{S}^2\boldsymbol{Q}^T,$$

kde $\boldsymbol{S}^2$ je diagonální matice, mající na diagonále vlastní čísla, λ_i, matice $\boldsymbol{X}^T\boldsymbol{X}$. Protože $h(\boldsymbol{X}^T\boldsymbol{X}) = m$, je matice $\boldsymbol{X}^T\boldsymbol{X}$ pozitivně definitní a $\lambda_i > 0$, $\forall i \in \widehat{m}$. Platí tedy

$$\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m = \boldsymbol{Q}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)\boldsymbol{Q}^T,$$

což je ale regulární matice (všechny tři matice jsou regulární). Odtud plyne

$$(\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)$$

tzn. čísla na diagonále matice $\boldsymbol{S}^2 + \delta\boldsymbol{I}_m$ jsou vlastní čísla matice $\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m$. Neboli $\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m$ je regulární a má vlastní čísla $\lambda_i + \delta$. Jejich velikost i poměry (tedy index podmíněnosti) lze měnit volbou parametru δ - to je hlavní myšlenka hřebenové regrese.
Místo odhadu $\widehat{\boldsymbol{\beta}}^{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ tak budeme studovat odhad

$$\widehat{\boldsymbol{\beta}}^{R,\delta} = (\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,.$$

Věta 5.3. Vychýlení odhadu $\widehat{\boldsymbol{\beta}}^{R,\delta}$ je

$$bias(\widehat{\boldsymbol{\beta}}^{R,\delta}) = -\delta(\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)^{-1}\boldsymbol{\beta}.$$

Pro střední kvadratickou chybu odhadu $\widehat{\boldsymbol{\beta}}^{R,\delta}$ platí

$$MSE(\widehat{\boldsymbol{\beta}}^{R,\delta}) = (\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)^{-1}\left[\sigma^2\boldsymbol{X}^T\boldsymbol{X} + \delta^2\boldsymbol{\beta}\boldsymbol{\beta}^T\right](\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)^{-1}\,.$$

Důkaz. Pro vychýlení odhadu $\widehat{\boldsymbol{\beta}}^{R,\delta}$ dostaneme

$$\begin{aligned}
\mathsf{E}\left(\widehat{\boldsymbol{\beta}}^{R,\delta} - \boldsymbol{\beta}\right) &= \left(\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta} \\
&= \left(\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{X} - \delta\boldsymbol{I}_m\right)\boldsymbol{\beta} = -\delta\left(\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m\right)^{-1}\boldsymbol{\beta}\,.
\end{aligned}$$

Pro libovolný náhodný vektor $\boldsymbol{Z}$ je jeho střední kvadratická chyba od vektoru $\boldsymbol{h}$ rovna

$$\begin{aligned}\mathsf{E}\left((\boldsymbol{Z}-\boldsymbol{h})(\boldsymbol{Z}-\boldsymbol{h})^T\right) &= \mathsf{E}\left((\boldsymbol{Z}-\mathsf{E}\boldsymbol{Z}+\mathsf{E}\boldsymbol{Z}-\boldsymbol{h})(\boldsymbol{Z}-\mathsf{E}\boldsymbol{Z}+\mathsf{E}\boldsymbol{Z}-\boldsymbol{h})^T\right)\\ &= \mathsf{Cov}\,\boldsymbol{Z}+(\mathsf{E}\boldsymbol{Z}-\boldsymbol{h})(\mathsf{E}\boldsymbol{Z}-\boldsymbol{h})^T.\end{aligned}$$

Protože

$$\mathsf{E}\widehat{\boldsymbol{\beta}}^{R,\delta}=(\boldsymbol{X}^T\boldsymbol{X}+\delta\boldsymbol{I}_m)^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta},$$

platí

$$\widehat{\boldsymbol{\beta}}^{R,\delta}-\mathsf{E}\widehat{\boldsymbol{\beta}}^{R,\delta}=\left(\boldsymbol{X}^T\boldsymbol{X}+\delta\boldsymbol{I}_m\right)^{-1}\boldsymbol{X}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})=\left(\boldsymbol{X}^T\boldsymbol{X}+\delta\boldsymbol{I}_m\right)^{-1}\boldsymbol{X}^T\boldsymbol{e}$$

a

$$\begin{aligned}\mathsf{Cov}\left(\widehat{\boldsymbol{\beta}}^{R,\delta}\right) &= \mathsf{E}\left[\left(\boldsymbol{X}^T\boldsymbol{X}+\delta\boldsymbol{I}_m\right)^{-1}\boldsymbol{X}^T\boldsymbol{e}\boldsymbol{e}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}+\delta\boldsymbol{I}_m\right)^{-1}\right]\\ &= \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}+\delta\boldsymbol{I}_m\right)^{-1}\boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}+\delta\boldsymbol{I}_m\right)^{-1}.\end{aligned}$$

Celkem pak

$$\begin{aligned}MSE(\widehat{\boldsymbol{\beta}}^{R,\delta}) &= \mathsf{E}\left((\widehat{\boldsymbol{\beta}}^{R,\delta}-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}^{R,\delta}-\boldsymbol{\beta})^T\right)\\ &= \left(\boldsymbol{X}^T\boldsymbol{X}+\delta\boldsymbol{I}_m\right)^{-1}\left[\sigma^2\boldsymbol{X}^T\boldsymbol{X}+\delta^2\boldsymbol{\beta}\boldsymbol{\beta}^T\right]\left(\boldsymbol{X}^T\boldsymbol{X}+\delta\boldsymbol{I}_m\right)^{-1},\end{aligned}$$

neboť $\mathsf{E}\boldsymbol{Z}-\boldsymbol{h}=\mathsf{E}\left(\widehat{\boldsymbol{\beta}}^{R,\delta}-\boldsymbol{\beta}\right)$ v našem označení. □

Věta 5.4. Ve standardizovaném modelu s plnou hodností platí pro $0<\delta<\frac{2\sigma^2}{\|\boldsymbol{\beta}\|^2}$, že

$$\mathsf{Cov}\,\widehat{\boldsymbol{\beta}}^{LS}-MSE(\widehat{\boldsymbol{\beta}}^{R,\delta})\quad\text{je pozitivně definitní matice.}$$

(Existuje $\delta>0$ tak, aby složky odhadu $\widehat{\boldsymbol{\beta}}^{R,\delta}$ měly menší rozptyly než složky odhadu $wh\boldsymbol{\beta}^{LS}$.)

Důkaz. Lze najít v publikacích Zvára (2008) nebo Víšek (1995). □

Poznámka 5.6. Ukázali jsme, že $\mathsf{E}\|\widehat{\boldsymbol{\beta}}\|^2>\|\boldsymbol{\beta}\|^2\ (+\sigma^2\mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. Můžeme se snažit $\widehat{\boldsymbol{\beta}}$ nějak zkrátit, za cenu ztráty nestrannosti. *Jiná interpretace hřebenového odhadu* je následující: hledejme $\boldsymbol{\beta}$, které pro dané δ minimalizuje

$$\varphi(\boldsymbol{\beta})=\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2+\delta\|\boldsymbol{\beta}\|^2$$

(minimalizujeme tedy současně SSE a normu vektoru $\boldsymbol{\beta}$, poměr je dán hodnotou δ). Dá se ukázat, že řešením je odhad $\widehat{\boldsymbol{\beta}}^{R,\delta}$.

Věta 5.5. Nechť $\boldsymbol{PSQ}^T$ je singulární rozklad matice $\boldsymbol{X}$. Potom platí

$$\mathsf{Cov}\,(\widehat{\boldsymbol{\beta}}^{R,\delta})=\sigma^2\sum_{i=1}^m\left(\frac{s_i}{s_i^2+\delta}\right)^2\boldsymbol{q}_i\boldsymbol{q}_i^T=\sigma^2\sum_{i=1}^m\frac{\lambda_i}{(\lambda_i+\delta)^2}\boldsymbol{q}_i\boldsymbol{q}_i^T.$$

Důkaz. Použitím singulárního rozkladu, podobně jako výše, dostaneme $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{S}^2\boldsymbol{Q}^T$ a tedy

$$\left(\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m\right)^{-1} = \left(\boldsymbol{Q}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)\mathbf{Q}^T\right)^{-1} = \boldsymbol{Q}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)^{-1}\mathbf{Q}^T.$$

Potom

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{R,\delta} &= (\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}_m)^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{Q}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)^{-1}\mathbf{Q}^T\boldsymbol{Q}\boldsymbol{S}\boldsymbol{P}^T\boldsymbol{Y} \\ &= \boldsymbol{Q}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)^{-1}\boldsymbol{S}\boldsymbol{P}^T\boldsymbol{Y}\end{aligned}$$

a

$$\begin{aligned}\mathsf{Cov}\left(\widehat{\boldsymbol{\beta}}^{R,\delta}\right) &= \boldsymbol{Q}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)^{-1}\boldsymbol{S}\boldsymbol{P}^T\,\mathsf{Cov}\,(\boldsymbol{Y})\boldsymbol{P}\boldsymbol{S}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)^{-1}\boldsymbol{Q}^T \\ &= \sigma^2\boldsymbol{Q}(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)^{-1}\boldsymbol{S}^2(\boldsymbol{S}^2 + \delta\boldsymbol{I}_m)^{-1}\boldsymbol{Q}^T = \sigma^2\sum_{i=1}^{p}\frac{s_i^2}{(s_i^2+\delta)^2}\boldsymbol{q}_i\boldsymbol{q}_i^T\,.\end{aligned}$$

□

Poznámka 5.7. Pokud je matice $\boldsymbol{X}$ špatně podmíněná, bude minimálně jedno z čísel s_i^2 malé, tedy $\frac{1}{s_i^2}$ vystupující ve $\mathsf{Var}\,(\widehat{\beta}_j^{LS})$ bude velké. Ale $\frac{s_i^2}{(s_i^2+\delta)^2}$ může být zase malé! S rostoucím δ se stále více potlačuje vliv malých vlastních čísel. Parametry σ^2 ani $\boldsymbol{\beta}$ neznáme, není tedy jasné, jak velké δ lze volit.

Volba parametru δ

1) **Hřebenová stopa (ridge trace):**
   - vypočítají se odhady $\widehat{\boldsymbol{\beta}}^{R,\delta}$ pro různé hodnoty δ,
   - vykreslí se graf jednotlivých složek $\widehat{\boldsymbol{\beta}}^{R,\delta}$ v závislosti na δ,
   - doporučuje se volit takové δ, pro které se grafy „stabilizují“ (což je subjektivní).

2) **Harmonic mean estimator:**

$$\widehat{\delta} = \frac{m s_n^2}{\|\widehat{\boldsymbol{\beta}}\|^2},$$

   kde $\widehat{\boldsymbol{\beta}}$ a s_n^2 jsou klasické (OLS) odhady parametrů $\boldsymbol{\beta}, \sigma^2$.

Poznámka 5.8. Existuje spousta dalších metod pro odhad δ, jejich vlastnosti jsou většinou zkoumány jen pomocí simulačních studií. Pokud nemáme nějakou apriorní informaci o parametru $\boldsymbol{\beta}$, použití hřebenové regrese nemusí zaručit zlepšení OLS odhadů.

Příklad 5.2 (Data Cement).

```
XX <- cbind(scale(x1),scale(x2),scale(x3),scale(x4))
kappa(XX,exact = TRUE)

## 37.10634

ridge <- lm.ridge(scale(y) ~ XX, lambda = seq(0,0.3,0.001))

plot(ridge)

select(ridge)

## modified HKB estimator is 0.08499604
## modified L-W estimator is 0.05830686
## mallest value of GCV  at 0.3
```