# Lineární modely

## Ekonometrie

Jiří Neubauer

Katedra ekonometrie FVL UO Brno
kancelář 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

# Lineární regresní model

$Y$ – **vysvětlovaná veličina** nebo **odezva**
$X_1, X_2, \ldots, X_k$ – **vysvětlující proměnné** nebo **regresory**.
$i$-té pozorování vysvětlované proměnné $Y$ popíšeme rovnicí

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \tag{1}$$

kde

1. $Y_i$ je $i$-té pozorování $Y$, $i = 1, 2, \ldots, n$,
2. $x_{ij}$ je $i$-té pozorování regresoru $X_j$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$,
3. $\beta_j$, $j = 1, \ldots, k$, jsou neznámé parametry,
4. $\epsilon_i$, $i = 1, 2, \ldots, n$, jsou neznámé náhodné chyby, které vznikají při pozorování vysvětlované proměnné $Y$ a které nemůžeme přímo pozorovat ani měřit.

# Lineární regresní model

Předpokládáme, že $x_{ij}$ jsou pevně dané známé reálné hodnoty a veličiny $Y_i$ a $\epsilon_i$ jsou náhodného charakteru (náhodné veličiny). Na jejich pravděpodobnostní rozdělení klademe následující předpoklady:

(P1) Střední hodnota $E(\epsilon_i) = 0$, $i = 1, 2, \ldots, n$, tj. náhodné chyby jsou **nesystematické**.

(P2) Rozptyl $D(\epsilon_i) = \sigma^2$, $i = 1, 2, \ldots,$, tj. náhodné chyby jsou **homogenní** se stejným neznámým rozptylem $\sigma^2$.

(P3) Náhodné chyby $\epsilon_i$ jsou nezávislé.

V případě, kdy je třeba provádět testy hypotéz o neznámých parametrech a konstruovat intervaly spolehlivosti pro neznámé parametry modelu, zavádí se v LRM další předpoklad:

(P4) Náhodné chyby $\epsilon_i$ mají normální rozdělení.

Často se v lineárním regresním modelu předpokládá, že první regresor je konstanta, potom pozorované hodnoty $x_{i1} = 1, i = 1, 2, \ldots, n$ a model má tvar

$$Y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

.

# Lineární regresní model

Odhady parametrů v lineárním regresním modelu (1) provedeme, podobně jako v případě přímkové regrese, metodou nejmenších čtverců. Model nejdříve zapíšeme v maticovém tvaru. Označme

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \\ \vdots \\ \epsilon_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Lineární regresní model (1) lze potom vyjádřit jednoduchým maticovým zápisem

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Požaduje se, aby matice $\boldsymbol{X}$ měla lineárně nezávislé sloupce. Protože předpokládáme, počet regresorů je menší než počet pozorování $k < n$, je hodnost matice $\boldsymbol{X}$ rovna $k$. Odtud plyne, že matice $\boldsymbol{X}'\boldsymbol{X}$ je regulární. Odhad neznámých parametrů modelu určíme metodou nejmenších čtverců, tedy minimalizací výrazu

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

# Lineární regresní model

Odhady metodou nejmenších čtverců získáme řešením soustavy lineárních rovnic, tzv. **normálních rovnic**

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}.$$

Toto nalezneme snadno, protože matice $\boldsymbol{X}'\boldsymbol{X}$ je regulární a tedy existuje inverzní matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Řešení normálních je tvaru

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{2}$$

Odhadnutý regresní model lze psát ve tvaru

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y},$$

kde $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ je čtvercová matice typu $n \times n$. Tato matice popisuje zobrazení (lineární transformaci) pozorovaných hodnot na hodnoty vyrovnané pomocí regresního modelu.
Rezidua $e_i = Y_i - \hat{Y}_i$, tedy rozdíly mezi naměřenou hodnotou $Y_i$ a hodnotou vyrovnanou $\hat{Y}_i$, lze vyjádřit v maticové podobě jako

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}. \tag{3}$$

# Lineární regresní model

Střední hodnota odhadu $\hat{\boldsymbol{\beta}}$ je rovna

$$E(\hat{\boldsymbol{\beta}}) = E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}\right] = E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})\right] =$$
$$= E\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}\right] = \boldsymbol{\beta},$$

neboť $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ a $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$. Odhad je tedy nestranný (nevychýlený). Dále určíme

$$\text{var}(\hat{\boldsymbol{\beta}}) = \text{var}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y})\right] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{var}(\boldsymbol{Y})\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]' =$$
$$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

protože $\text{var}(\boldsymbol{Y}) = \text{var}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) = \text{var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$.

# Lineární regresní model

Zavedeme **reziduální součet čtverců**

$$S_e^2 = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Reziduální součet čtverců normovaný konstantou $n - k$ je nevychýleným odhadem rozptylu $\sigma^2$. Tedy nestranný odhad $\sigma^2$ je roven

$$\hat{\sigma}^2 = s_e^2 = \frac{1}{n-k} S_e^2 = \frac{1}{n-k} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

# Intervaly spolehlivosti a testy významnosti regresních parametrů

Předpokládejme, že náhodné chyby $\epsilon_i, i = 1, \ldots, n$ v lineárním regresním modelu mají normální rozdělení s nulovou střední hodnotou a rozptylem $\sigma^2$. Potom mají odhady $\hat{\beta}_j,\ j = 1, 2, \ldots, k$ regresních parametrů $\beta_j$ normální rozdělení a platí

$$\hat{\beta}_j \sim N(\beta_j, D(\hat{\beta}_j)),$$

kde rozptyly $D(\hat{\beta}_j)$ jsou dány vztahy: $D(\hat{\beta}_1) = \sigma^2 v_{11}$, $D(\hat{\beta}_2) = \sigma^2 v_{22}, \ldots, D(\hat{\beta}_k) = \sigma^2 v_{kk}$, přičemž $v_{11}, v_{22}, \ldots, v_{kk}$ jsou prvky na hlavní diagonále matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Pokud nahradíme $\sigma^2$ reziduálním rozptylem $s^2$, získáme odhady rozptylů regresních parametrů $\hat{D}(\hat{\beta}_j) = s_e^2 v_{jj}$. Druhé odmocniny těchto odhadů

$$s(\hat{\beta}_j) = \sqrt{s_e^2 v_{jj}}$$

se nazývají **směrodatné chyby** odhadů regresních parametrů.

# Intervaly spolehlivosti a testy významnosti regresních parametrů

Při konstrukci intervalů spolehlivosti pro parametry $\beta_j$ regresního modelu (1) ze statistik $t = (\hat{\beta}_j - \beta_j)/s(\hat{\beta}_j),\ j = 1, 2, \ldots, k$, které mají Studentovo rozdělení s $n - k$ stupni volnosti. Oboustranné intervaly spolehlivosti při riziku odhadu $\alpha$ jsou dány vztahem

$$\hat{\beta}_j - t_{1-\alpha/2}(n-k) \cdot s(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{1-\alpha/2}(n-k) \cdot s(\hat{\beta}_j),$$

kde $t_{1-\alpha/2}(n-k)$ označuje kvantil Studentova rozdělení.

Testování statistické významnosti regresních koeficientů
$H$: $\beta_j = 0$, alternativní hypotéza $A$: $\beta_j \neq 0$. Při platnosti nulové hypotézy má statistika

$$t = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} = \frac{\hat{\beta}_j - 0}{s(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)} \tag{4}$$

Studentovo rozdělení s $n - k$ stupni volnosti. Parametr se považuje za statisticky významný na hladině významnosti $\alpha$, pokud hodnota testové statistiky $t$ je v absolutní hodnotě větší než kritická hodnota vyjádřená kvantilem Studentova rozdělení $t_{1-\frac{\alpha}{2}}(n-k)$.

# Test významnosti regresního modelu

Označme $S_Y = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2$ celkovou variabilitu vysvětlované proměnné, $S_T = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - n\bar{Y}^2$ variabilitu vysvětlované proměnné popsanou daným regresním modelem, $S_e = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$ reziduální součet čtverců. Platí

$$S_Y = S_T + S_e.$$

Testuje se nulová hypotéza $H_0$: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$ proti alternativní hypotéze $H_1$: $\beta_j \neq 0$ pro alespoň jedno $j = 2, 3, \ldots, k$. Testové kritérium je statistika

$$F = \frac{S_T^2}{k-1} : \frac{S_e^2}{n-k}, \tag{5}$$

která má při platnosti nulové hypotézy Fisherovo-Snedecorovo rozdělení $F$ s $k-1$ a $n-k$ stupni volnosti. Regresní model se považuje za statisticky významný na hladině významnosti $\alpha$, pokud hodnota testové statistiky $F$ je větší než kritická hodnota vyjádřená kvantilem $F_{1-\alpha}(k-1, n-k)$ daného $F$ rozdělení.

# Index determinace

Vhodnost zvoleného modelu lze vyjádřit pomocí tzv. **indexu (koeficientu) determinace**

$$R^2 = \frac{S_T^2}{S_Y^2} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

Tento odhad je asymptoticky nestranný, nicméně pro malé výběry nadhodnocuje skutečnou těsnost závislosti a je závislý na počtu parametrů regresního modelu. Lze provést jeho korekci

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-k},$$

čím získáme odhad nestranný.

# Predikce v LRM

Nalezené odhady $\hat{\beta}_1, \ldots, \hat{\beta}_k$ parametrů $\beta_1, \ldots, \beta_k$ regresního modelu (1) lze použít k **odhadu regresní funkce** $y$ v daném bodě $\boldsymbol{x}_0 = (x_{01}, x_{02}, \ldots, x_{0k})'$, tedy při hodnotách regresorů $X_1 = x_{01}, X_2 = x_{02}, \ldots, X_k = x_{0k}$

$$\hat{y} = \hat{y}(\boldsymbol{x}_0) = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} = \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}.$$

Platí $D[\hat{y}(\boldsymbol{x}_0)] = D(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$. Nahrazením neznámé hodnoty $\sigma^2$ jejím odhadem $s_e^2$ obdržíme odhad rozptylu. Směrodatná chyba hledaného odhadu je potom $s(\hat{y}(\boldsymbol{x}_0)) = s_e\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}$. Oboustranný interval spolehlivosti má tvar

$$\hat{y}(\boldsymbol{x}_0) - t_{1-\alpha/2}(n-k)\cdot s(\hat{y}(\boldsymbol{x}_0)) < y(\boldsymbol{x}_0) < \hat{y}(\boldsymbol{x}_0) + t_{1-\alpha/2}(n-k)\cdot s(\hat{y}(\boldsymbol{x}_0)).$$

Zajímá-li nás interval spolehlivosti pro **predikci (předpověď)** vysvětlované veličiny $Y$ v bodě $\boldsymbol{x}_0 = (x_{01}, x_{02}, \ldots, x_{0k})'$, tedy interval spolehlivosti pro $Y_0 = y(\boldsymbol{x}_0) + \epsilon_0$, kde $\epsilon_0$ je náhodná chyba tohoto pozorování v bodě $\mathbf{x}_0$, dostaneme s využitím uvedeného modelu

$$\hat{y}(\boldsymbol{x}_0) - t_{1-\alpha/2}(n-k)\cdot s_0 < Y_0 < \hat{y}(\boldsymbol{x}_0) + t_{1-\alpha/2}(n-k)\cdot s_0,$$

kde
$s_0$ je směrodatná chyba odhadu $Y_0$, tedy směrodatná chyba veličiny $\hat{y}(\boldsymbol{x}_0) + e_0$, která je rovna $s_0 = s_e\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}$. Využilo se zde faktu, že
$D(\hat{Y}_0) = D(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} + \epsilon_0) = D(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}) + D(\epsilon_0) = \sigma^2\left(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0\right)$.

# Tvorba regresního modelu

Při sestavování regresního modelu obvykle požadujeme, aby všechny odhadnuté regresní parametry byly na zvolené hladině významnosti statisticky významné, viz (4), a aby celkový model byl statisticky významným, viz (5). Předpokládejme, že máme nějaký regresní model. Mohou nastat následující situace:

1. Regresní model i všechny regresní parametry jsou statisticky významné.
2. Regresní model i všechny regresní parametry jsou statisticky nevýznamné.
3. Regresní model je statisticky významný, ale některé regresní parametry vychází nevýznamné.
4. Regresní model je statisticky významný, všechny regresní parametry jsou statisticky nevýznamné.

# Ověření vhodnosti regresního modelu

Mezi základní předpoklady lineárního regresního modelu patří:

1. Vztah mezi vysvětlovanou proměnnou a regresory je lineární.
2. Chybová složka $\epsilon$ má nulovou střední hodnotu a konstantní rozptyl $\sigma^2$.
3. Náhodné chyby jsou nekorelované.
4. Náhodné chyby mají normální rozdělení

# Analýza reziduí

Rezidua jsou rozdíly mezi pozorovanými hodnotami vysvětlované proměnné a hodnotami, které vychází z odhadnutého regresního modelu

$$e_i = Y_i - \hat{Y}_i, \ \ i = 1, 2, \ldots, n.$$

- **Normovaná reziduа** jsou definována jako

$$d_i = \frac{e_i}{s_e}, \ \ i = 1, 2, \ldots . n,$$

mají nulovou střední hodnotu a přibližně jednotkový rozptyl. Velké hodnoty reziduí (řekněme $d_i > 3$) indikují potenciální odlehlou hodnotu, tzv. **outlier**.

- **Standardizovaná** rezidua

$$r_i = \frac{e_i}{s_e \sqrt{(1 - h_{ii})}}, \ \ i = 1, 2, \ldots . n,$$

mají konstantní rozptyl $D(r_i) = 1$, $h_{ii}$ jsou diagonální prvky matice $H$.

# Analýza reziduí

- Vynecháme-li v lineárním regresním modelu $i$-té pozorování $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki})$, dostaneme matici plánu $\boldsymbol{X}_{(i)}$, která vznikne z původní matice $\boldsymbol{X}$ vynecháním $i$-tého řádku, a vektor $\boldsymbol{Y}_{(i)}$. Pro tento model jsou odhady parametrů metodou nejmenších čtverců rovny $\hat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}'_{(i)}\boldsymbol{Y}_{(i)}$. Pomocí takto zvoleného modelu odhadneme hodnotu vysvětlující proměnné pro vynechané $i$-té pozorování $\hat{Y}_{(i)} = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{(i)}$, rozdíly

$$e_{(i)} = Y_i - \hat{Y}_{(i)}$$

se označují jako **predikovaná rezidua**. Takto je možné určit predikovaná rezidua pro $i = 1, 2, \ldots, n$. Nicméně lze ukázat, že platí

$$e_{(i)} = \frac{e_i}{1 - h_{ii}},$$

tedy predikovaná rezidua lze přímo určit z reziduí původního regresního modelu.

# Analýza reziduí

- Standardizovaná rezidua $r_i$ lze použít pro diagnostiku odlehlých pozorování. Pro odhad parametru $\sigma^2$ se obvykle používá reziduální rozptyl $s_e^2$. Tento parametr však lze odhadnout i jinak. Jednou z možností, je určit odhad na základě datového souboru, kterém bylo $i$-té pozorování vynecháno, označme jej $s_{e(i)}^2$. Tento odhad je roven

$$s_{e(i)}^2 = \frac{(\boldsymbol{Y}_{(i)} - \hat{\boldsymbol{Y}}_{(i)})'(\boldsymbol{Y}_{(i)} - \hat{\boldsymbol{Y}}_{(i)})}{n-k-1},$$

lze jej vyjádřit ve tvaru

$$s_{e(i)}^2 = \frac{(n-k)s_e^2 - e_i^2/(1-h_{ii})}{n-k-1}.$$

**Studentizovaná rezidua** potom určíme ze vzorce

$$t_i = \frac{e_i}{\sqrt{s_{e(i)}^2(1-h_{ii})}}, \quad i = 1, 2, \ldots . n.$$

# Analýza reziduí – grafické znázornění

- **histogram reziduí** – posouzení rozdělení reziduí (normalita)
- **QQ plot** – graf porovnávající empirické a teoretické kvantily.
  V případě ověření normality reziduí vyneseme do grafu body jejichž první souřadnicí je hodnota kvantilu $X_p, 0 < p < 1$ normálního rozdělení $N(0, 1)$, druhou souřadnicí je potom odpovídají empirický kvantil určený z reziduí. Rezidua lze potom považovat za normálně rozdělená, leží-li tyto body přibližně na přímce.
- **rezidua a vyrovnané hodnoty**
  Vykreslíme-li graf reziduí $e_i$ (případně $d_i$, $r_i$ či $t_i$) vůči vyrovnaným hodnotám $\hat{Y}_i$, můžeme detekovat řadu porušení předpokladů modelu. Vykreslené body grafu by měly tvořit pás rovnoměrně rozložený kolem nulové hodnoty. Je-li tvar odlišný, lze usuzovat např. na nekonstantní rozptyl (lze stabilizovat pomocí vhodné transformace vysvětlované proměnné), případně nelinearitu (může se upravit přidáním dalšího regresoru případně vhodnou transformací). Graf může také zachytit velké hodnoty reziduí, což může ukazovat na potenciální odlehlé pozorování.
- **rezidua a jednotlivé regresory**
  Graf reziduí vůči hodnotám jednotlivých regresorů by podobně jako v předcházejícím případě měl tvořit pás kolem nulové hodnoty. Při porušení tohoto tvaru ze opět usuzovat na nestacionární rozptyl, či jiný něž předpokládaná typ závislosti mezi vysvětlovanou proměnnou a daným regresorem.

# Analýza reziduí

Ovevění nekorelovanosti reziduí – autokorelační a parciální autokorelační funkce, Durbin-Watsonův test, Ljung-Boxův test (příkaz `Box.test`)

Ověření normality reziduí – testy normality (Shapiro-Wilkovým testem (`shapiro. test`) nebo Lillieforsovým testem (balíček `nortest`, příkaz `lillie.test`).)

# Detekce neočekávaných pozorování

Při zpracování dat se můžeme setkat s případy, kdy se v datovém souboru vyskytují hodnoty, které se výrazně odlišují od hodnot ostatních. V zásadě může jít o

- neočekávané hodnoty vysvětlované proměnné, tzv. **odlehlá pozorování – „outliers“**,
- neočekávané hodnoty vektoru vysvětlujících proměnných, tzv. **vlivné body – „leverage points“**.

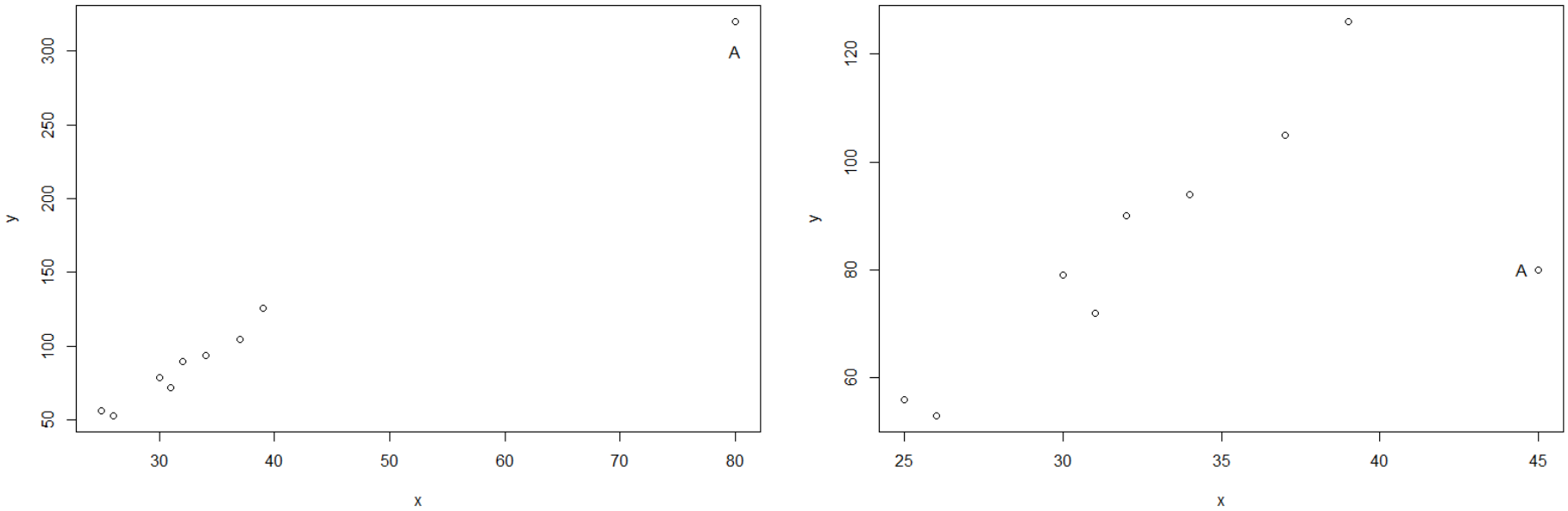
# Detekce neočekávaných pozorování

- Odlehlá pozorování (outliers) – detekce pomocí analýzy reziduí
- Vlivné body (leverage points) – body které mají vliv na hodnoty odhadů regresních parametrů. Důležitou roli při jejich detekci hrají diagonální prvky $h_{ii}$ matice

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'.$$

Velké hodnoty $h_{ii}$ ukazují na potenciální vlivné body. Vzhledem k tomu, že $\sum_{i=1}^{n} h_{ii} = k$, je průměrná hodnota diagonálních prvků $h_{ii}$ rovna $\bar{h} = k/n$. Pozorování, pro která je $h_{ii} > 2k/n$ se považují za tzv. „leverage points“. Ne všechny leverage points ale musejí být body vlivnými

# Detekce neočekávaných pozorování




Obrázek: Příklady leverage a vlivných bodů

# Detekce neočekávaných pozorování

Pro měření vlivu $i$-tého pozorování na hodnoty odhadů regresních parametrů lze použít tzv. **Cookovu vzdálenost**

$$D_i = \frac{(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)})'(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(i)})}{k s_e^2} = \frac{r_i^2}{k} \frac{h_{ii}}{1 - h_{ii}}.$$

Je-li hodnota Cookovy vzdálenosti $D_i > 1$, je $i$-té pozorování považováno za vlivné. Na Cookovu vzdálenost lze pohlížet jako na eukleidovskou vzdálenost (až na člen $ks_e^2$) mezi vektorem predikce $\hat{\boldsymbol{Y}}$ a vektorem predikce $\hat{\boldsymbol{Y}}_{(i)}$, který odpovídá odhadům při vynechání $i$-tého pozorování.

# Lineární model

Lineárním modelem budeme rozumět model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

kde

- $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$ je náhodný vektor
- $\boldsymbol{X}$ je matice typu $n \times k$, $k < n$
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$ je vektor parametrů
- $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_k)'$ je vektor náhodných chyb, pro nějž platí $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ a $\mathrm{var}(\boldsymbol{\epsilon}) = \boldsymbol{V}$ a nezávisí na $\boldsymbol{\beta}$

Vektor $\hat{\boldsymbol{\beta}}$ nazveme **lineárním odhadem** vektoru $\boldsymbol{\beta}$, existuje-li taková matice $\boldsymbol{U}$ typu $k \times n$, že $\hat{\boldsymbol{\beta}} = \boldsymbol{U}\boldsymbol{Y}$.
Lineární odhad je **nestranný** právě když

$$\boldsymbol{U}\boldsymbol{X} = \boldsymbol{I}.$$

O vektoru $\hat{\boldsymbol{\beta}}$ řekneme, že je **nejlepším nestranným lineárním odhadem** vektoru $\boldsymbol{\beta}$, jestliže platí

- $\hat{\boldsymbol{\beta}}$ je nestranným odhadem parametru $\boldsymbol{\beta}$,
- je-li $\hat{\boldsymbol{\beta}}^*$ jiný nestranný odhad $\boldsymbol{\beta}$, pak $\mathrm{var}(\hat{\boldsymbol{\beta}}^*) - \mathrm{var}(\hat{\boldsymbol{\beta}}^*) \geq 0$.

# Lineární model s plnou hodností

Předpokládejme, že hodnost matice $h(\boldsymbol{X}) = k$ a že $\boldsymbol{V} = \text{var}(\boldsymbol{\epsilon})$ je regulární ($\det(\boldsymbol{V}) \neq 0$). Pak nejlepší nestranný lineární odhad parametru $\boldsymbol{\beta}$ je roven

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}$$

a má varianční matici

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}.$$

Nejlepší nestranný odhad parametru $\theta = \boldsymbol{c}'\boldsymbol{\beta}$ je $\hat{\theta} = \boldsymbol{c}'\hat{\boldsymbol{\beta}}$.

# Lineární model s neúplnou hodností

Nechť matice $\boldsymbol{X}$ má hodnost menší než $k$ a pro varianční matici chybové složky platí $\text{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$, kde $\sigma^2 > 0$ je neznámý parametr. K získání odhadů se opět použije metoda nejmenších čtverců.

Výraz

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{6}$$

nabývá vzhledem k $\boldsymbol{\beta}$ nejmenší možné hodnoty, pokud je $\boldsymbol{\beta}$ řešením soustavy rovnic

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}. \tag{7}$$

Hodnota výrazu (6) je stejná pro všechna $\boldsymbol{\beta}$, která jsou řešení soustavy (7).

# Lineární model s neúplnou hodností

Soustava (7) je ekvivalentní se soustavou (**soustava normálních rovnic**)

$$\frac{\partial}{\partial \beta_j}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = 0, \quad j = 1, 2, \ldots, k$$

resp.

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \left( Y_i - \sum_{j=1}^{k} x_{ij}\beta_j \right)^2 = 0, \quad j = 1, 2, \ldots, k$$

Mějme vektor $\boldsymbol{c} = (c_1, c_2, \ldots, c_k)'$. Řekneme, že parametr $\theta = c_1\beta_1 + c_2\beta_2 + \cdots + c_k\beta_k$ je **odhadnutelný**, existuje-li pro $\theta$ alespoň jeden vektor $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)'$ takový, že $E(\boldsymbol{u}'\boldsymbol{Y}) = \theta$, což lze zapsat ve tvaru

$$\boldsymbol{u}'E(\boldsymbol{Y}) = \theta \quad \text{nebo} \quad \boldsymbol{u}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{c}'\boldsymbol{\beta},$$

odkud $\boldsymbol{u}'\boldsymbol{X} = \boldsymbol{c}'$ neboli $\boldsymbol{c} = \boldsymbol{X}'\boldsymbol{u}$.

Parametr $\theta = \boldsymbol{c}'\boldsymbol{\beta}$ je odhadnutelný právě tehdy,

- je-li $\boldsymbol{c}$ nějakou lineární kombinací řádků matice $\boldsymbol{X}$,
- je-li $\theta$ nějakou lineární kombinací složek vektoru $E(\boldsymbol{Y})$.

# Lineární model s neúplnou hodností

Předpokládejme, že $\theta = \boldsymbol{c}'\boldsymbol{\beta}$ je odhadnutelný parametr. Pak nejlepší nestranný lineární odhad tohoto parametru je dán vzorcem $\hat{\theta} = \boldsymbol{c}'\hat{\boldsymbol{\beta}}$, kde $\hat{\boldsymbol{\beta}}$ je libovolné řešení soustavy normálních rovnic (7). Přitom hodnota $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ je stejná pro všechna řešení této soustavy.

Máme-li vektor parametrů $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ se složkami typu $\theta_j = \boldsymbol{c}_j'\boldsymbol{\beta}$, pak řekneme, že tento vektor je **odhadnutelný**, je-li odhadnutelná každá jeho složka.

Vektor $\boldsymbol{\theta} = E(\boldsymbol{Y})$ **je vždy odhadnutelný** a nejlepší nestranný lineární odhad je roven

$$\hat{\boldsymbol{\theta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}.$$

Pozn. $(.)^{-}$ značí pseudoinverzi. Pseudoinverze matice $\boldsymbol{A}$ typy $n \times k$ je matice $\boldsymbol{A}^{-}$ splňující

$$\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{A} = \boldsymbol{A}.$$

Pro **reziduální součet čtverců** $S_e = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ platí

$$S_e = \boldsymbol{Y}'[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}']\boldsymbol{Y}.$$

# Analýza rozptylu – jeden faktor

Máme k dispozici dvě aditiva, která by měla zvýšit dojezd automobilu na jeden litr (nebo galon) benzínu. Střední hodnota dojezdu bez aditiv je $\mu$. Přidáním prvního aditiva očekáváme zvýšení dojezdu o $\alpha_1$ kilometrů, po přidání druhého aditiva zvýšení o $\alpha_2$ kilometrů. Model můžeme zapsat

$$Y_1 = \mu + \alpha_1 + \epsilon_1, \quad Y_2 = \mu + \alpha_2 + \epsilon_2.$$

Cílem bude odhadnout parametry $\mu, \alpha_1$ a $\alpha_2$ a testovat hypotézu $\alpha_1 = \alpha_2$.

Předpokládejme následující uspořádání experimentu: máme 6 identických aut, 3 budou mít natankovaný benzín s aditivem 1, 3 s aditivem 2. Pro jednotlivá pozorování dostáváme

$$\begin{aligned} Y_{11} &= \mu + \alpha_1 + \epsilon_{11}, & Y_{12} &= \mu + \alpha_1 + \epsilon_{12}, & Y_{13} &= \mu + \alpha_1 + \epsilon_{13}, \\ Y_{21} &= \mu + \alpha_2 + \epsilon_{21}, & Y_{22} &= \mu + \alpha_2 + \epsilon_{22}, & Y_{23} &= \mu + \alpha_2 + \epsilon_{23}, \end{aligned}$$

nebo

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \ i = 1, 2, \ j = 1, 2, 3.$$

# Analýza rozptylu – jeden faktor

V maticové podobě dostáváme

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}$$

neboli

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Hodnost matice $\boldsymbol{X}$ je rovna 2 (první sloupec je součtem 2. a 3. sloupce), počet parametrů, které je třeba odhadnou je 3 (model s neúplnou hodností).

# Analýza rozptylu – jeden faktor

Možné přístupy k řešení

1. Redukce počtu parametrů – označme $\mu_1 = \mu + \alpha_1$ (střední hodnota dojezdu pro první typ aditiva), $\mu_2 = \mu + \alpha_2$ (střední hodnota dojezdu pro první typ aditiva)

$$Y_{ij} = \mu_i + \epsilon_{ij}, \ i = 1, 2, \ j = 1, 2, 3.$$

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}$$

Dostáváme model plné hodnosti.

# Analýza rozptylu – jeden faktor

1. Přidání podmínek – předpokládejme, že $\alpha_1 + \alpha_2 = 0$, dostáváme model $Y_{ij} = \mu^* + \alpha_i^* + \epsilon_{ij}$, který lze vyjádřit jako $Y_{1j} = \mu^* + \alpha_1^* + \epsilon_{1j}$ a $Y_{2j} = \mu^* - \alpha_1^* + \epsilon_{2j}$

$$Y_{ij} = \mu_i + \epsilon_{ij}, \ i = 1, 2, \ j = 1, 2, 3.$$

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \mu^* \\ \alpha_1^* \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}$$

   Dostáváme model plné hodnosti.
2. Odhadování lineárních kombinací parametrů, kontrasty.

# Analýza rozptylu – jeden faktor

Mějme nezávislé výběry z rozdělení $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2), \ldots, N(\mu_k, \sigma^2)$ o rozsazích $n_1, n_2, \ldots, n_k, \sum_{i=1}^{k} n_i = n$. Model lze zapsat ve tvaru

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, 2, \ldots k, \; j = 1, 2, \ldots, n_i, \tag{8}$$

kde $\epsilon_{ij}$ jsou nezávislé náhodné veličiny s rozdělením $N(0, \sigma^2)$. Model je přeparametrizovaný, obsahuje o jeden parametr více než je třeba. (Lze nahradit $\mu + \alpha_i = \mu_i$.) Formulace modelu (8) je nicméně názorná u složitějších modelů. Normální rovnice dostaneme parciálním derivováním výrazu

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \mu - \alpha_i)^2$$

podle parametrů $\mu$ a $\alpha_i$. Vzniklá soustava má singulární matici. K modelu přidáme podmínku

$$\sum_{i=1}^{k} n_i \alpha_i = 0.$$

Nejlepším nestranným lineárním odhadem

- $E(Y_{ij}) = \mu + \alpha_i$ je $\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$,
- $\alpha_i - \alpha_t$ je $\overline{y}_i - \overline{y}_t = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} - \frac{1}{n_t} \sum_{j=1}^{n_t} Y_{ij}$

# Analýza rozptylu – jeden faktor

Hypotézu $H : \mu_1 = \mu_2 = \cdots = \mu_k$ lze vyjádřit ve tvaru $H : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, při platnosti $H$ dostáváme submodel

$$Y_{ij} = \mu + \epsilon_{ij}, \quad i = 1, 2, \ldots k, \; j = 1, 2, \ldots, n_i, \tag{9}$$

| Zdroj variability | Součet čtverců | Stupně volnosti | Testová statistika |
|---|---|---|---|
| Faktor | $S_m$ | $f = k - 1$ | $F = \frac{S_m/f}{S_e/f_e}$ |
| Reziduální | $S_e$ | $f_e = n - k$ | – |
| Celkový | $S_c$ | $f_c = n - 1$ | – |

Tabulka: Tabulka jednofaktorové analýzy rozptylu

Označme $\overline{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$, $\overline{y_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$

Variabilita způsobená faktorem (meziskupinová) $S_m = \sum_{i=1}^{k}(\overline{y_i} - \overline{y})^2 n_i$

Variabilita reziduální (vnitroskupinová) $S_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2$

Variabilita celková $S_c = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$

$$S_c = S_m + S_e$$

# Analýza rozptylu – dva faktory bez interakcí

Často je třeba zkoumat závislost kvantitativní proměnné na více faktorech. Omezíme se na případ dvou faktorů.

- Budeme se zabývat vlivem dvou vysvětlujících proměnných (faktorů $A$, $B$) na proměnnou vysvětlovanou $Y$.
- Označme $a$ počet úrovní faktoru $A$, podobně $b$ bude označovat počet úrovní faktoru $B$.
- Předpokládejme, že pro každou dvojici hodnot faktorů máme $r \geq 2$ pozorování.
- Pro pozorování s $i$-tou hodnotou faktoru $A$ a $j$-tou hodnotou faktoru $B$ platí

$$Y_{ij1}, \ldots, Y_{ijr} \sim N(\mu_{ij}, \sigma^2).$$

Předpokládejme, že náhodné veličiny $Y_{ijp}$ se řídí modelem

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \quad i = 1, \ldots, a, \; j = 1, \ldots, b, \; k = 1, \ldots, r, \tag{10}$$

kde $\epsilon_{ijk}$ jsou nezávislé náhodné veličiny s rozdělením $N(0, \sigma^2)$, parametry $\alpha_i$ jsou tzv. **řádkové efekty**, parametry $\beta_j$ jsou tzv. **sloupcové efekty**.

# Analýza rozptylu – dva faktory bez interakcí

Označme:
$\overline{y}_{ij.} = \frac{1}{r}\sum_{k=1}^{r} Y_{ijk}$,
$\overline{y}_{i..} = \frac{1}{br}\sum_{j=1}^{b}\sum_{k=1}^{r} Y_{ijk}$,
$\overline{y}_{.j.} = \frac{1}{ar}\sum_{i=1}^{a}\sum_{k=1}^{r} Y_{ijk}$,
$\overline{y}_{...} = \frac{1}{abr}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} Y_{ijk}$
Normální rovnice dostaneme tak, že postupně parciálně derivujeme výraz

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(Y_{ijk} - \mu - \alpha_i - \beta_j)^2$$

podle parametrů $\mu$, $\alpha_i$ a $\beta_j$. Vzniklá soustava má singulární matici. K modelu přidáme podmínky

$$\sum_{i=1}^{a}\alpha_i = 0, \sum_{j=1}^{b}\beta_j = 0.$$

Získáme řešení $\mu_0 = \overline{y}_{...}$, $\alpha_i^0 = \overline{y}_{i..} - \overline{y}_{...}$, $\beta_j^0 = \overline{y}_{.j.} - \overline{y}_{...}$

# Analýza rozptylu – dva faktory bez interakcí

Nejlepší nestranný lineární odhad pro $E(Y_{ijk}) = \mu + \alpha_i + \beta_j$ je

$$\mu^0 + \alpha_i^0 + \beta_j^0 = \overline{y}_{i..} + \overline{y}_{.j.} - \overline{y}_{...}.$$

Nezáleží-li na sloupcové klasifikaci, můžeme položit $\beta_1 = \beta_1 = \cdots = \beta_b = 0$, dostaneme submodel

$$Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}.$$

Ten odpovídá jednofaktorové analýze rozptylu pro faktor $A$. Jestliže položíme $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$, obdržíme submodel

$$Y_{ijk} = \mu + \beta_j + \epsilon_{ijk}.$$

Nezáleží-li ani na sloupcové a ani na řádkové klasifikaci, dostaneme submodel

$$Y_{ijk} = \mu + \epsilon_{ijk}.$$

# Analýza rozptylu – dva faktory bez interakcí

| Zdroj variability | Součet čtverců | Stupně volnosti | Testová statistika |
|---|---|---|---|
| Faktor $A$ | $S_A$ | $f_A = a - 1$ | $F_A = \frac{S_A/f_A}{S_e/f_e}$ |
| Faktor $B$ | $S_B$ | $f_B = b - 1$ | $F_B = \frac{S_B/f_B}{S_e/f_e}$ |
| Reziduální | $S_e$ | $f_e = n - a - b + 1$ | – |
| Celkový | $S_c$ | $f_c = n - 1$ | – |

Tabulka: Tabulka dvoufaktorové analýzy rozptylu bez interakce

Pro celkovou variabilitu lze psát

$$S_c = S_A + S_B + S_e,$$

kde $S_c = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{r} (Y_{ijk} - \overline{y}_{...})^2$,
$S_A = br \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2$,
$S_B = ar \sum_{j=1}^{b} (\overline{y}_{.j.} - \overline{y}_{...})^2$,
$S_e = S_c - S_A - S_B$.

# Analýza rozptylu – dva faktory s interakcemi

Při dvoufaktorové analýze rozptylu nás může kromě vlivu faktorů $A$ a $B$ na vysvětlovanou proměnnou $Y$ zajímat také vliv interakce obou faktorů. Danou situaci lze popsat modelem

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots, r, \tag{11}$$

Soustavu normálních rovnic získáme derivováním výrazu

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(Y_{ijp} - \mu - \alpha_i - \beta_j - \lambda_{iij})^2$$

podle parametrů $\mu$, $\alpha_i$ a $\beta_j$. Tuto soustavu doplníme o podmínky

$$\sum_{i=1}^{a}\alpha_i = 0, \sum_{j=1}^{b}\beta_j = 0,\ \sum_{i=1}^{a}\lambda_{ij} = 0,\ \sum_{j=1}^{b}\lambda_{ij} = 0.$$

Získáme řešení $\mu_0 = \overline{y}_{...}$, $\alpha_i^0 = \overline{y}_{i..} - \overline{y}_{...}$, $\beta_j^0 = \overline{y}_{.j.} - \overline{y}_{...}$, $\lambda_{ij}^0 = \overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...}$
Nejlepší nestranný lineární odhad pro $E(Y_{ijk})$ je

$$\mu^0 + \alpha_i^0 + \beta_j^0 + \lambda_{ij}^0 = \overline{y}_{ij.}.$$

# Analýza rozptylu – dva faktory s interakcemi

Zpravidla nás zajímají tři různé hypotézy
$H_A$: $\alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$,
$H_B$: $\beta_1 = \beta_2 = \cdots = \beta_b = 0$
$H_{AB}$: $\lambda_{ij} = 0$ pro $i = 1, \ldots, a$, $j = 1, \ldots, b$

Pro celkovou variabilitu lze psát

$$S_c = S_A + S_B + S_{AB} + S_e,$$

kde $S_c = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{r} (y_{ijk} - \overline{y}_{...})^2$,
$S_A = br \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2$,
$S_B = ar \sum_{j=1}^{b} (\overline{y}_{.j.} - \overline{y}_{...})^2$,
$S_{AB} = r \sum_{i=1}^{a} \sum_{j=1}^{b} (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2$,
$S_e = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{r} (y_{ijk} - \overline{y}_{ij.})^2 = S_c - S_A - S_B - S_{AB}$.

# Analýza rozptylu – dva faktory s interakcemi

| Zdroj variability | Součet čtverců | Stupně volnosti | Testová statistika |
|---|---|---|---|
| Faktor $A$ | $S_A$ | $f_A = a - 1$ | $F_A = \frac{S_A / f_A}{S_e / f_e}$ |
| Faktor $B$ | $S_B$ | $f_B = b - 1$ | $F_B = \frac{S_B / f_B}{S_e / f_e}$ |
| Interakce | $S_{AB}$ | $f_{AB} = (a-1)(b-1)$ | $F_{AB} = \frac{S_{AB} / f_{AB}}{S_e / f_e}$ |
| Reziduální | $S_e$ | $f_e = n - ab$ | – |
| Celkový | $S_c$ | $f_c = n - 1$ | – |

Tabulka: Tabulka dvoufaktorové analýzy rozptylu s interakcemi

Pozn.: Součet $S_e + S_{AB}$, resp. $f_e + f_{AB}$ dá hodnotu $S_e$ resp. $f_e$ v tabulce bez interakcí.

# Analýza rozptylu – dva faktory

**Příklad**: Cílem experimentu je zkoumat vliv dvou typů benzínu a tří různých aditiv na spotřebu automobilu. Výsledky jsou uvedeny v tabulce.

| Typ | Aditivum | | |
|---|---|---|---|
| | A1 | A2 | A3 |
| B1 | 8,58 | 7,13 | 7,02 |
| | 8,22 | 7,35 | 7,28 |
| B2 | 7,06 | 6,61 | 7,04 |
| | 6,82 | 6,84 | 7,11 |