M6120 – 7. CVIČENÍ : **M6120cv07** (*Klasický lineární regresní model*)

A. Klasický lineární regresní model, modely neúplné hodnosti, rozšířený lineární regresní model a vážená metoda nejmenších čtverců.

Mějme **regresní model** plné hodnosti:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \wedge \quad h(\mathbf{X}) = h(\mathbf{X}'\mathbf{X}) = p+1 \quad \wedge \quad n > p+1 \quad \wedge \quad \boldsymbol{\varepsilon} \sim \mathcal{L}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

vektor závisle proměnných $\mathbf{Y} = (Y_1, \ldots, Y_n)'$
matice plánu $\mathbf{X} = (x_{ij}) \quad i = 1, \ldots, n; \quad j = 0, \ldots, p$
vektor chyb $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)', \quad E\boldsymbol{\varepsilon} = \mathbf{0}, \quad D\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}_n;$

Odhad neznámých parametrů $\boldsymbol{\beta}$ provedený *metodou nejmenších čtverců* je řešením normálních rovnic $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$ a platí: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

Označme

$$\widehat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \underbrace{\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'}_{\mathbf{H}}\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\underbrace{\mathbf{I}-\mathbf{H}}_{\mathbf{M}})\mathbf{Y} = \mathbf{M}\mathbf{Y} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) = \underbrace{\mathbf{M}\mathbf{X}}_{=\mathbf{0}}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\varepsilon} = (\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}$$

$$s^2 = \frac{S_e}{n-p-1} = \frac{1}{n-p-1}(\mathbf{Y}-\widehat{\mathbf{Y}})'(\mathbf{Y}-\widehat{\mathbf{Y}}) = \frac{1}{n-p-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{n-p-1}\widehat{\mathbf{Y}}'(\mathbf{I}-\mathbf{H})\widehat{\mathbf{Y}} = \frac{1}{n-p-1}\boldsymbol{\varepsilon}'(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}$$

Platí

- $E\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$
- $Es^2 = \frac{E(S_e)}{n-p-1} = \sigma^2$, tj. s^2 je nestranným odhadem rozptylu
- $D\widehat{\boldsymbol{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Platí–li navíc $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, pak

- $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$
- $\widehat{\boldsymbol{\varepsilon}} \sim N_n(\mathbf{O}, \sigma^2(\mathbf{I}-\mathbf{H}))$
- $\widehat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- $\frac{S_e}{\sigma^2} \sim \chi^2(n-p-1)$
- $\widehat{\boldsymbol{\beta}}$ a s^2 jsou stochasticky nezávislé
- $T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 v_{jj}}} \sim t(n-p-1)$, kde $(\mathbf{X}'\mathbf{X})^{-1} = (v_{ij})_{i,j=0,\ldots,p}$
- $F = \frac{1}{qs^2}(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)'\mathbf{W}^{-1}(\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \sim F(q, n-p-1)$,

  kde $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \mathbf{V} & \mathbf{U} \\ \mathbf{U} & \mathbf{W} \end{pmatrix}$, $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$ $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$ a $h(\mathbf{W}) = q$
- $T = \frac{\mathbf{c}'\widehat{\boldsymbol{\beta}} - \mathbf{c}'\boldsymbol{\beta}}{\sqrt{s^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim t(n-p-1)$, kde $\mathbf{c} = (c_0, c_1, \ldots, c_p)'$ a $E(\mathbf{c}'\widehat{\boldsymbol{\beta}}) = \mathbf{c}'\boldsymbol{\beta}$
- $\left.\begin{array}{l} Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \\ \hat{Y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i) \end{array}\right\} \Rightarrow \quad Y_i - \hat{Y}_i \sim N(0, \sigma^2(1 + \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i))$

  kde $\mathbf{x}_i' = (x_{i0}, \ldots, x_{ip})$ *je* i-tý řádek matice plánu $\mathbf{X}$

Intervaly spolehlivosti	
pro parametry β_j $j=0,\dots,p$	$\left(\beta_j - t_{1-\frac{\alpha}{2}}(n-p-1)s\sqrt{v_{jj}},\ \beta_j + t_{1-\frac{\alpha}{2}}(n-p-1)s\sqrt{v_{jj}}\right)$
pro střední hodnotu predikce $E\hat{Y}_i = E\mathbf{x}_i'\widehat{\boldsymbol{\beta}} = \mathbf{x}_i'\boldsymbol{\beta}$	$\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}} - t_{1-\frac{\alpha}{2}}(n-p-1)s\sqrt{\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i},\ \mathbf{x}_i'\widehat{\boldsymbol{\beta}} + t_{1-\frac{\alpha}{2}}(n-p-1)s\sqrt{\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}\right)$
pro predikci $\hat{Y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ $i=1,\dots,n$	$\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}} - t_{1-\frac{\alpha}{2}}(n-p-1)s\sqrt{1+\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}, \mathbf{x}_i'\widehat{\boldsymbol{\beta}} + t_{1-\frac{\alpha}{2}}(n-p-1)s\sqrt{1+\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}\right)$

kde $t_{1-\frac{\alpha}{2}}(n-p-1)$ je $1-\frac{\alpha}{2}$ kvantil Studentova rozdělení o $n-p-1$ stupních volnosti

Až doposud jsme uvažovali lineární regresní model plné hodnosti. V některých situacích je však vhodné použít model s **neúplnou hodností**, tj. $h(\mathbf{X}) = r < k \leq n$. V tom případě systém normálních rovnic má nekonečně mnoho řešení, takže žádný vektor středních hodnot $E\mathbf{Y} = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ neurčuje jednoznačně vektor $\boldsymbol{\beta}$. Není však vyloučeno, že existují nějaké **lineární kombinace vektoru $\boldsymbol{\beta}$**, jejichž hodnoty jsou vektorem středních hodnot $\boldsymbol{\mu} \in \mathcal{M}(\mathbf{X})$ určeny jednoznačně. Ukazuje se (viz Anděl, 1978), že těmito hledanými vektory jsou **(nestranně lineárně) odhadnutelné** parametrické funkce $\theta = \mathbf{c}'\boldsymbol{\beta}$. Jejich důležitou vlastností je, že jsou to právě lineární kombinace řádků matice $\mathbf{X}$, tj. $\mathbf{c} \in \mathcal{M}(\mathbf{X}')$. Pokud máme vektor $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$, $m \in \mathbb{N}$, jehož složky jsou odhadnutelné, jde o odhadnutelný vektor parametrů.

Dá se ukázat (viz Anděl, 1978), že nejlepším nestranným lineárním odhadem odhadnutelné parametrické funkce $\theta = \mathbf{c}'\boldsymbol{\beta}$ je $\hat{\theta} = \mathbf{c}'\widehat{\boldsymbol{\beta}}$, kde $\widehat{\boldsymbol{\beta}}$ je libovolné řešení normálních rovnic. Odtud je ihned vidět, že vektor středních hodnot $\boldsymbol{\mu} = E\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ je vždy odhadnutelný a jeho nejlepší nestranný lineární odhad je tvaru

$$\widehat{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}.$$

Platí–li navíc $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, pak (viz Anděl, 1978)

(1) Statistika $S_e/\sigma^2 = \frac{1}{\sigma^2}(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) = \frac{1}{\sigma^2}\mathbf{Y}'[\mathbf{I}_n-\mathbf{H}]\mathbf{Y} \sim \chi^2(n-r)$.

(2) Statistika $s^2 = \frac{S_e}{n-r}$ je nestranným odhadem parametru σ^2.

(3) Vektor $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ a s^2 jsou **nezávislé**.

(4) Statistika $T = \dfrac{\mathbf{c}'\widehat{\boldsymbol{\beta}} - \mathbf{c}'\boldsymbol{\beta}}{s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}}} \sim t(n-r)$.

Někdy musíme vzít současně se základním lineárním modelem v úvahu i několik speciálních případů tohoto modelu, kterým se říká **podmodely** nebo **submodely**. Mějme náhodný vektor $\mathbf{Y} = (Y_1, \dots, Y_n)'$ a předpokládejme, že platí model M a jsou dány další dva submodely M_1 a M_2, přičemž pro $n \geq k \geq r \geq r_1 \geq r_2$ máme

$$\boxed{M}\ \mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),\quad \mathbf{X} \text{ je typu } n \times k,\quad h(\mathbf{X}) = r,\quad \boldsymbol{\beta} \text{ je typu } k \times 1$$

$$\boxed{M_1}\ \mathbf{Y} \sim N_n(\mathbf{U}\boldsymbol{\beta}_1, \sigma^2\mathbf{I}_n),\ \mathbf{U} \text{ je typu } n \times k_1,\ h(\mathbf{U}) = r_1,\ \boldsymbol{\beta}_1 \text{ je typu } k_1 \times 1$$

$$\boxed{M_2}\ \mathbf{Y} \sim N_n(\mathbf{T}\boldsymbol{\beta}_2, \sigma^2\mathbf{I}_n),\ \mathbf{T} \text{ je typu } n \times k_2,\ h(\mathbf{T}) = r_2,\ \boldsymbol{\beta}_2 \text{ je typu } k_2 \times 1$$

Položme $\widehat{\boldsymbol{\mu}}_1 = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-}\mathbf{U}'\mathbf{Y}$ a $\widehat{\boldsymbol{\mu}}_2 = \mathbf{T}(\mathbf{T}'\mathbf{T})^{-}\mathbf{T}'\mathbf{Y}$, pak (viz Anděl, 1978)

(5) platí-li model $\boxed{M_1} \Rightarrow F_1 = \frac{(\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_1)'(\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_1)}{r-r_1}\frac{1}{s^2} \sim F(r-r_1, n-r)$,

(6) platí-li model $\boxed{M_2} \Rightarrow F_2 = \frac{(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)'(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)}{r_1-r_2}\frac{1}{s^2} \sim F(r_1-r_2, n-r)$.

Podmodel vzniklý vypuštěním sloupců matice plánu. Podmodel může být dán požadavkem vynechat z matice plánu $\mathbf{X}$ některé sloupce. Bez újmy na obecnosti předpokládejme, že matice, které určují model a podmodel se liší právě posledními sloupci matice $\mathbf{X}$, takže $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1)$.

Mějme náhodný vektor $\mathbf{Y} = (Y_1, \dots, Y_n)'$ a předpokládejme, že platí model M a je dán submodel M_0, přičemž

$$\begin{array}{llllll} \boxed{M} & \mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) & \text{kde } \mathbf{X} & \text{je typu } n \times k, & h(\mathbf{X}) = r, & \boldsymbol{\beta} \quad \text{je typu } k \times 1 \\ \boxed{M_0} & \mathbf{Y} \sim N_n(\mathbf{X}_0\boldsymbol{\beta}_0, \sigma^2 \mathbf{I}_n) & \text{kde } \mathbf{X}_0 & \text{je typu } n \times k_0, & h(\mathbf{X}_0) = r_0, & \boldsymbol{\beta}_0 \quad \text{je typu } k_0 \times 1 \end{array}$$

$$\text{kde} \qquad n \geq k \geq r \geq r_0$$

Podle definice model M_0 je podmodelem M pokud $\mathbf{X}_0 = \mathbf{X}\mathbf{K}$, v našem případě matice $\mathbf{K} = \binom{\mathbf{I}_{k_0}}{\mathbf{0}}$ je typu $k \times k_0$.

Položme $\widehat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{Y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$ a $\widehat{\boldsymbol{\mu}}_0 = \mathbf{H}_0\mathbf{Y} = \mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-}\mathbf{X}_0'\mathbf{Y}$,

$$S_e = (\mathbf{Y} - \widehat{\boldsymbol{\mu}})'(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) \qquad S_{e_0} = (\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0)'(\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0)$$

pak

$$S_{\Delta_0} = (\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0)'(\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0) \qquad S_e = S_{e_0} - S_{\Delta_0}$$

Pokud platí model $\boxed{M_0}$, pak statistika

$$F_0 = \frac{(S_{e_0} - S_e)/(r - r_0)}{S_e/(n - r)} \sim F(r - r_0, n - r).$$

Rozšířený lineární regresní model a vážená metoda nejmenších čtverců. Mějme regresní model, ve kterém $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{L}_n(\mathbf{0}, \sigma^2\mathbf{V})$, $\mathbf{V} > 0$, a hodnost matice $h(\mathbf{X}) = k$ (tj. $\mathbf{V}$ je pozitivně definitní), pak odhad pomocí metody nejmenších čtverců je roven

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

což lze snadno dokázat. Vzhledem k předpokadu $\mathbf{V} > 0$ (tj. $\mathbf{V}$ je pozitivně definitní) existuje $\mathbf{V}^{-\frac{1}{2}}$, která je symetrická a regulární. Proto

$$h(\mathbf{V}^{-\frac{1}{2}}\mathbf{X}) = h(\mathbf{X}) = k = h(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}) = h(\mathbf{X}'\mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}\mathbf{X})$$

takže $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ je regulární. Položme

$$\mathbf{Z} = \mathbf{V}^{-\frac{1}{2}}\mathbf{Y}, \qquad \mathbf{F} = \mathbf{V}^{-\frac{1}{2}}\mathbf{X}, \qquad \eta = \mathbf{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon}.$$

Pak z $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ plyne, že $\mathbf{V}^{-\frac{1}{2}}\mathbf{Y} = \mathbf{V}^{-\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon}$, tj. $\mathbf{Z} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\eta}$.

Pak

$$E\boldsymbol{\eta} = E\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon} = \mathbf{V}^{-\frac{1}{2}}\underbrace{E\boldsymbol{\varepsilon}}_{=0} = 0$$

a

$$D\boldsymbol{\eta} = D(\mathbf{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}^{-\frac{1}{2}}\mathbf{V}\mathbf{V}^{-\frac{1}{2}} = \sigma^2\mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{\frac{1}{2}}\mathbf{V}^{\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}} = \sigma^2\mathbf{I}_n$$

a tento model již splňuje předpoklady klasického lineárního regresního modelu, ve kterém odhad vektoru neznámých parametrů metodou nejmenších čtverců je roven

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{Z} = (\mathbf{X}'\mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}\mathbf{Y} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

Nejčastěji se matice $\mathbf{V}$ uvažuje jako diagonální matice ve tvaru $\mathbf{V} = \mathrm{diag}\{v_1, \dots, v_n\}$.
Položíme-li
$$\mathbf{W} = \mathbf{V}^{-1} = diag\{\tfrac{1}{v_1}, \dots, \tfrac{1}{v_n}\} = diag\{w_1, \dots, w_n\},$$
přičemž prvky $w_1, \dots, w_n$ se nazývají **váhami** (tedy čím je rozptyl větší, tím je váha pozorování menší). Pak odhad neznámých parametrů metodou nejmenších čtverců:
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$$
se nazývá VÁŽENÁ METODA NEJMENŠÍCH ČTVERCŮ.

Poznámka: V prostředí R se ve funkci `lm()` přidá parametr `weights` .

B. Testování rovnoběžnosti a shodnosti dvou regresních přímek

Mějme dva nezávislé náhodné výběry
$$Y_{11}, \dots, Y_{1n_1} \text{ (resp. } Y_{21}, \dots, Y_{2n_1})$$
a k tomu odpovídající hodnoty regresorů
$$x_{11}, \dots, x_{1n_1} \text{ (resp. } x_{21}, \dots, x_{2n_2}).$$

Předpokládejme, že
$$\begin{aligned} Y_{1i} &= a_1 + b_1 x_{1i} + \varepsilon_{1i}, & \varepsilon_{1i} &\sim N(0, \sigma^2) & i &= 1, \dots, n_1 \\ Y_{2i} &= a_2 + b_2 x_{2i} + \varepsilon_{2i} & \varepsilon_{2i} &\sim N(0, \sigma^2) & i &= 1, \dots, n_2 \end{aligned}$$

Vytvořme společný regresní model:
$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \hline Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n_1} & 0 & 0 \\ \hline 0 & 0 & 1 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_{2n_2} \end{pmatrix} \begin{pmatrix} a_1 \\ b_1 \\ \hline a_2 \\ b_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \hline \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \end{pmatrix}.$$

Vyjádřeno blokově:
$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}$$

Pak
$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \mathbf{X}_1'\mathbf{Y}_1 \\ \mathbf{X}_2'\mathbf{Y}_2 \end{pmatrix} \quad \text{a} \quad \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} \mathbf{X}_1'\mathbf{Y}_1 \\ (\mathbf{X}_2'\mathbf{X}_2)^{-1} \mathbf{X}_2'\mathbf{Y}_2 \end{pmatrix}.$$

Označme
$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\varepsilon}}_1 \\ \hat{\boldsymbol{\varepsilon}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 - \widehat{\mathbf{Y}}_1 \\ \mathbf{Y}_2 - \widehat{\mathbf{Y}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 \\ \mathbf{Y}_2 - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 \end{pmatrix}$$

$$S_e = \widehat{\boldsymbol{\varepsilon}}\,'\,\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\varepsilon}}_1\,'\,\widehat{\boldsymbol{\varepsilon}}_1 + \widehat{\boldsymbol{\varepsilon}}_2\,'\,\widehat{\boldsymbol{\varepsilon}}_2 = S_{e_1} + S_{e_2}$$

$$s_1^2 = \tfrac{S_{e_1}}{n_1-2} = \tfrac{\widehat{\boldsymbol{\varepsilon}}_1\,'\,\widehat{\boldsymbol{\varepsilon}}_1}{n_1-2}, \quad s_2^2 = \tfrac{S_{e_1}}{n_2-2} = \tfrac{\widehat{\boldsymbol{\varepsilon}}_2\,'\,\widehat{\boldsymbol{\varepsilon}}_2}{n_2-2} \qquad \Rightarrow \qquad s^2 = \tfrac{S_e}{n_1+n_2-4} = \tfrac{(n_1-2)s_1^2+(n_2-2)s_2^2}{n_1+n_2-4}$$

Testování rovnoběžnosti dvou regresních přímek

Při testování hypotézy $\boxed{H_0 : b_1 = b_2}$ proti alternativě $\boxed{H_1 : b_1 \neq b_2}$ využijeme toho, že statistika

$$T = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}} - \mathbf{c}'\boldsymbol{\beta}}{\sqrt{s^2 \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim t(n-p-1).$$

Položme $\mathbf{c} = (0, 1, 0, -1)$, pak

$$\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} = v_{22} + v_{44}, \quad \text{přičemž} \quad (\mathbf{X}'\mathbf{X})^{-1} = (v_{ij})_{i,j=1,\dots,4}.$$

Za platnosti nulové hypotézy statistika $\boxed{T_0 = \frac{\hat{b}_1 - \hat{b}_2}{s\sqrt{v_{22}+v_{44}}} \sim t(n_1 + n_2 - 4).}$

Nulovou hypotézu zamítáme na hladině významnosti α

pokud ■ $|t_0| > t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 4)$
nebo ■ *p-value* $p_0 = P(T_0 > |t_0|) < \frac{\alpha}{2}$.

Testování shodnosti dvou regresních přímek

Budeme testovat hypotézu $\boxed{H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2}$ proti alternativě $\boxed{H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2}$.

Využijeme vztahů $K_2 = \frac{S_e}{\sigma^2} = \frac{(n_1+n_2-4)s^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 4)$

$$\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2 \sim N\Big(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2, \sigma^2 \underbrace{\big((\mathbf{X}_1'\mathbf{X}_1)^{-1} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}\big)}_{\mathbf{W}}\Big)$$

a za platnosti H_0 $K_1 = \frac{1}{\sigma^2}\big(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\big)' W^{-1}\big(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\big) \sim \chi^2(2)$

pak

$$\boxed{F_0 = \frac{K_1/2}{K_2/(n_1+n_2-4)} = \frac{1}{2s^2}\big(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\big)'\mathbf{W}^{-1}\big(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\big) \sim F(2, n_1+n_2-4)}$$

Nulovou hypotézu zamítáme na hladině významnosti α

pokud ■ $f_0 < F_{\frac{\alpha}{2}}(2, n_1 + n_2 - 4)$ nebo $f_0 > F_{1-\frac{\alpha}{2}}(2, n_1 + n_2 - 4)$
nebo ■ *p-value* $p_0 = P(F > f_0) < \frac{\alpha}{2}$ nebo $1 - p_0 < \frac{\alpha}{2}$.

Ověřování shodnosti rozptylů

Při testování hypotézy $\boxed{H_0 : \sigma_1^2 = \sigma_2^2}$ proti alternativě $\boxed{H_1 : \sigma_1^2 \neq \sigma_2^2}$ využijeme toho, že statistika F_0 za platnosti H_0 má F-rozdělení

$$\boxed{F_0 = \frac{\frac{S_{e_1}}{(n_1-2)\sigma^2}}{\frac{S_{e_2}}{(n_2-2)\sigma^2}} = \frac{s_1^2}{s_2^2} \sim F(n_1 - 2, n_2 - 2)}$$

Nulovou hypotézu zamítáme na hladině významnosti α

pokud ■ $f_0 < F_{\frac{\alpha}{2}}(n_1 - 2, n_2 - 2)$ nebo $f_0 > F_{1-\frac{\alpha}{2}}(n_1 - 2, n_2 - 2)$
nebo ■ *p-value* $p_0 = P(F_0 > f_0) < \frac{\alpha}{2}$ nebo $1 - p_0 < \frac{\alpha}{2}$.

Reference: Anděl, J. *Matematická statistika.* SNTL, ALFA Praha 1978

Příklad 1: Rozvodovost v České a Slovenské republice (1960-1970)

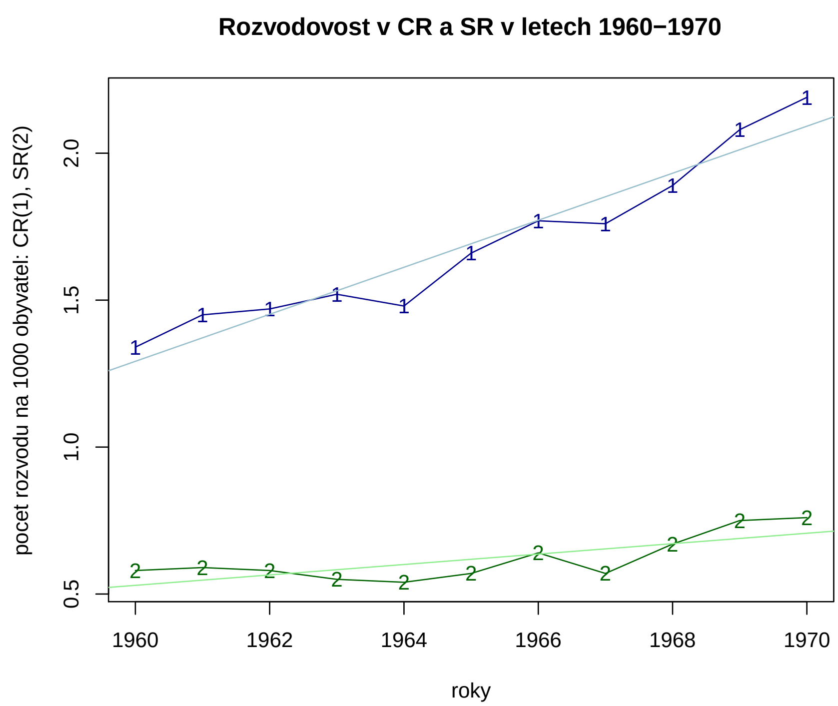
V celostátních statistikách byl v letech 1960–1970 sledován počet rozvodů na 1000 obyvatel zvlášť pro Českou a Slovenskou republiku.

Do proměnných `y1` a `y2` vložíme počty rozvodů na 1000 obyvatel pro jednotlivé republiky.

```
> TXT <- "Rozvodovost v CR a SR v letech 1960-1970"
> TxtY <- "pocet rozvodu na 1000 obyvatel"
> TxtX <- "roky"
> TxtGr <- "republika"
> y1 <- c(1.34, 1.45, 1.47, 1.52, 1.48, 1.66, 1.77, 1.76, 1.89, 2.08,
    2.19)
> y2 <- c(0.58, 0.59, 0.58, 0.55, 0.54, 0.57, 0.64, 0.57, 0.67, 0.75,
    0.76)
> xtime <- 1960:1970
```

Hodnoty obou republik vykreslíme do jediného grafu pomocí funkce `matplot()`. Do grafu zakreslíme také regresní přímky.

```
> matplot(xtime, cbind(y1, y2), type = "o", xlab = TxtX, main = TXT, ylab = paste(TxtY,
    ": CR(1), SR(2)", sep = ""), lty = 1, col = c("darkblue", "darkgreen"))
> abline(lm(y1 ~ xtime), col = "lightblue3", lty = 1)
> abline(lm(y2 ~ xtime), col = "lightgreen", lty = 1)
```



Obrázek 1: `matplot`: pro data *Rozvodovost v České a Slovenské republice v letech 1960-1970*

Pro obě dvě republiky provedeme odhad všech parametrů regresní přímky pomocí funkce `lm()`. Kvůli vysokým x–vým hodnotám se doporučuje proměnnou x (roky) centrovat.

```
> n <- length(xtime)
> xshift <- mean(xtime)
> model.CR <- lm(y1 ~ I(xtime - xshift))
> summary(model.CR)
```

```
Call:
lm(formula = y1 ~ I(xtime - xshift))

Residuals:
      Min        1Q    Median        3Q       Max
-0.131818 -0.036818 -0.001818  0.058182  0.098182

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.691818   0.022866   73.99 7.61e-14 ***
I(xtime - xshift) 0.080000   0.007231   11.06 1.53e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07584 on 9 degrees of freedom
Multiple R-squared: 0.9315,       Adjusted R-squared: 0.9239
F-statistic: 122.4 on 1 and 9 DF,  p-value: 1.534e-06
```

```
> model.SR <- lm(y2 ~ I(xtime - xshift))
> summary(model.SR)
```

```
Call:
lm(formula = y2 ~ I(xtime - xshift))

Residuals:
      Min        1Q    Median        3Q       Max
-0.083636 -0.040455  0.004091  0.046591  0.060909

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.618182   0.015953  38.750 2.52e-11 ***
I(xtime - xshift) 0.017727   0.005045   3.514  0.00658 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05291 on 9 degrees of freedom
Multiple R-squared: 0.5784,       Adjusted R-squared: 0.5316
F-statistic: 12.35 on 1 and 9 DF,  p-value: 0.006577
```

Testování homogenity rozptylu

Chceme–li testovat rovnoběžnost a shodnost obou přímek, musíme nejprve otestovat homogenitu rozptylu pomocí statistiky

$$F_0 = s_1^2/s_2^2 \ \sim \ F(n_1 - 2, n_2 - 2).$$

Pomocí funkce `summary()` získáme odhady statistik s_1^2 a s_2^2 (viz `Residual standard error` v předchozích výpisech) a spočítáme hodnotu statistiky F_0.

```
> (s.CR <- summary(model.CR)$sigma)

[1] 0.07583874

> (s.SR <- summary(model.SR)$sigma)

[1] 0.05291025

> (f0 <- s.CR^2/s.SR^2)

[1] 2.054483
```

Abychom mohli provést testování, potřebuje mít buď kritickou hodnotu testu nebo příslušnou p–hodnotu. K tomu potřebujeme znát stupně volnosti, zbytek spočítáme pomocí distribuční a kvantilové funkce F–rozdělení.

```
> (nu.sigma.CR <- model.CR$df.residual)

[1] 9

> (nu.sigma.SR <- model.SR$df.residual)

[1] 9

> (F.kritH <- qf(0.975, nu.sigma.CR, nu.sigma.SR))

[1] 4.025994

> (F.kritD <- qf(0.025, nu.sigma.CR, nu.sigma.SR))

[1] 0.2483859

> (pValF0 <- 1 - pf(f0, nu.sigma.CR, nu.sigma.SR))

[1] 0.1492161
```

Protože konkrétní hodnota $f_0 = 2.054483$ statistiky F neleží v kritické oblasti testu $W_\alpha = \{(0, 0.2483859) \cup (4.025994, \infty)\}$ a také p–hodnota není menší než 0.05, **nezamítáme** hypotézu o shodě rozptylu.

Testování rovnoběžnosti přímek

Při testování hypotézy $H_0 : b_1 = b_2$ proti alternativě $H_1 : b_1 \neq b_2$ využijeme toho, že statistika

$$T = \frac{\hat{b}_1 - \hat{b}_2}{s\sqrt{v_{22} + v_{44}}} \sim t(n_1 + n_2 - 4).$$

Diagonální prvky matice $(\mathbf{X}'\mathbf{X})^{-1}$ získáme pomocí funkce `summary()`.

```
> (vjj.CR <- diag(summary(model.CR)$cov.unscaled))

(Intercept) I(xtime - xshift)
 0.09090909        0.00909091

> (vjj.SR <- diag(summary(model.SR)$cov.unscaled))

(Intercept) I(xtime - xshift)
 0.09090909        0.00909091
```

Protože obě regresní přímky byly odhadnuty na základě stejného počtu pozorování, tak vážený průměr obou rozptylů je obyčejným aritmetickým průměrem. A pak již snadno spočítáme hodnotu testové statistiky a příslušnou kritickou hodnotu a p–hodnotu.

```
> (s2 <- 0.5 * (s.CR^2 + s.SR^2))

[1] 0.004275505

> (t0 <- (coef(model.CR)[2] - coef(model.SR)[2])/sqrt(s2 * (vjj.CR[2] +
     vjj.SR[2])))

I(xtime - xshift)
          7.06294

> (t.krit <- qt(0.975, nu.sigma.CR + nu.sigma.SR))

[1] 2.100922

> (pValt <- 2 * (1 - pt(t0, nu.sigma.CR + nu.sigma.SR)))

I(xtime - xshift)
     1.377579e-06
```

Protože vypočtená hodnota

$$t_0 = 7.06294 \in W_\alpha = \{(-\infty, -2.100922) \cup (2.100922, \infty)\}$$

a také p–hodnota je menší než 0.05, proto **zamítáme** nulovou hypotézu o rovnoběžnosti obou přímek.

Testování shodnosti přímek

Testování shodnosti přímek provedeme pomocí F–statistiky

$$F_0 = \frac{1}{2s^2}\left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\right)' \mathbf{W}^{-1} \left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\right) \sim F(2, n_1 + n_2 - 4)$$

kde

$$\mathbf{W} = (\mathbf{X}_1'\mathbf{X}_1)^{-1} + (\mathbf{X}_2'\mathbf{X}_2)^{-1}$$

Nejprve spočítáme matici $\mathbf{W}$, pak její inverzní matici pomocí funkce `solve()`.

```
> W <- summary(model.CR)$cov.unscaled + summary(model.SR)$cov.unscaled
> (invW <- solve(W))
```

```
                    (Intercept) I(xtime - xshift)
(Intercept)        5.500000e+00     -7.910962e-16
I(xtime - xshift) -7.910962e-16      5.500000e+01
```

Nakonec dopočítáme hodnotu statistiky F_0, kritické hodnoty a p–hodnotu.

```
> diffBeta <- as.matrix(coef(model.CR) - coef(model.SR))
> F0 <- t(diffBeta) %*% invW %*% diffBeta
> (F0 <- F0/(2 * s2))
```

```
         [,1]
[1,] 766.3547
```

```
> (F0.kritH <- qf(0.975, 2, nu.sigma.CR + nu.sigma.SR))
```

```
[1] 4.559672
```

```
> (F0.kritD <- qf(0.025, 2, nu.sigma.CR + nu.sigma.SR))
```

```
[1] 0.02535345
```

```
> (pValF00 <- 1 - pf(F0, 2, nu.sigma.CR + nu.sigma.SR))
```

```
     [,1]
[1,]    0
```

Protože konkrétní hodnota $f_0 = 766.3547$ statistiky F_0 leží v kritické oblasti testu $W_\alpha = \{(0, 0.2535345) \cup (4.559672, \infty)\}$ a také p–hodnota je menší než 0.05, **zamítáme** hypotézu o shodě regresních přímek.

Testování rovnoběžnosti a shodnosti přímek pomocí podmodelů

Předchozí postup testování rovnoběžnosti a shodnosti dvou přímek je velmi pracný. Naštěstí totéž můžeme provést mnohem jednodušším postupem, a to pomocí podmodelů.

Vytvoříme jediný regresní model s využitím kategoriální proměnné, která bude udávat, zda se jedná o data jedné či druhé republiky. Dostaneme tak tzv. ANCOVA model nebo též mluvíme o **analýze kovariance**.

ANCOVA modely jsou lineární regresní modely, jejichž závisle proměnné (nazývané též **kovariáty** či **regresory**) jsou jak spojité, tak kategoriální.

Ve shodě s implicitním nastavení kontrastů v prostředí R vytvoříme jediný regresní model a to tak, abychom vypuštěním jednoho sloupce mohli testovat i rovnoběžnost přímek. Plný model, který označíme jako $\boxed{M}$, bude mít různé koeficienty α i β, tedy předpokládá, že přímky mohou být různoběžné.

$$\boxed{M}: \quad Y_{ji} = \alpha + \alpha_2 + (\beta + \beta_2)x_{ji} + \varepsilon_{ji} \quad \text{kde} \quad j = 1, 2 \quad i = 1, \ldots, n_j \qquad (n_1 = n_2 = 11)$$

a parametry α_2 a β_2 budou nulové pro Českou republiku a pro Slovenskou republiku budou značit odchylku od parametru α `(Intercept)`, resp. β (směrnice přímky).

Nejprve vytvoříme datový rámec.

```
> data <- data.frame(x = rep(xtime - xshift, 2), y = c(y1, y2), gr = c(rep("CR",
     n), rep("SR", n)))
> str(data)
```

```
'data.frame':       22 obs. of  3 variables:
 $ x : num  -5 -4 -3 -2 -1 0 1 2 3 4 ...
 $ y : num  1.34 1.45 1.47 1.52 1.48 1.66 1.77 1.76 1.89 2.08 ...
 $ gr: Factor w/ 2 levels "CR","SR": 1 1 1 1 1 1 1 1 1 1 ...
```

Nyní pomocí funkce `lm()` provedeme odhad modelu $\boxed{M}$. Všimněme si, jak zadáme různé směrnice i průsečíky (`y ~ x * gr`), což lze také zapsat názornější a tím i delší formou `y ~ x + gr + x:gr`.

```
> model.M <- lm(y ~ x * gr, data)
> summary(model.M)
```

```
Call:
lm(formula = y ~ x * gr, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.131818 -0.039545  0.001364  0.049886  0.098182

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.691818   0.019715  85.814  < 2e-16 ***
x            0.080000   0.006234  12.832 1.70e-10 ***
grSR        -1.073636   0.027881 -38.507  < 2e-16 ***
```

```
x:grSR      -0.062273   0.008817  -7.063 1.38e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06539 on 18 degrees of freedom
Multiple R-squared: 0.9892,       Adjusted R-squared: 0.9875
F-statistic: 551.9 on 3 and 18 DF,  p-value: < 2.2e-16
```

Vypočteme koeficienty obou přímek.

```
> (Coef.M <- coef(model.M))
```

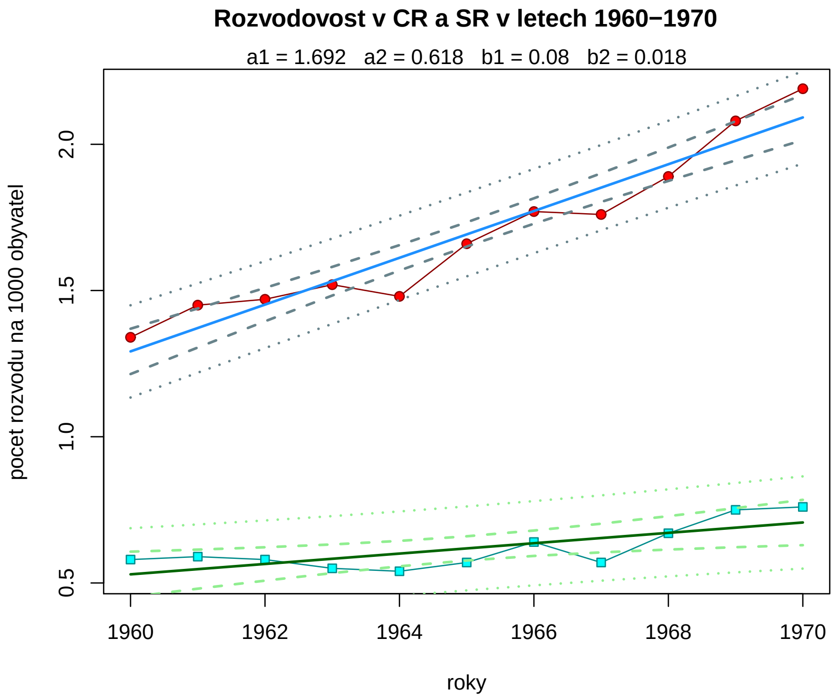
```
(Intercept)           x        grSR      x:grSR
 1.69181818  0.08000000 -1.07363636 -0.06227273
```

```
> a1.M <- Coef.M[1]
> a2.M <- Coef.M[1] + Coef.M[3]
> b1.M <- Coef.M[2]
> b2.M <- Coef.M[2] + Coef.M[4]
> txtCoef.M <- paste("a1 =", round(a1.M, 3), "  a2 =", round(a2.M, 3),
     "  b1 =", round(b1.M, 3), "  b2 =", round(b2.M, 3))
> cat(paste("Model M :", txtCoef.M, "\n"))
```

```
Model M : a1 = 1.692   a2 = 0.618   b1 = 0.08   b2 = 0.018
```

Vykreslíme do jednoho grafu obě dvě přímky, přidáme intervaly spolehlivosti kolem střední hodnoty a prediční intervaly.

```
> pred.M.CR <- predict(model.M, newdata = data.frame(x = xtime - xshift,
     gr = "CR"))
> pred.M.SR <- predict(model.M, newdata = data.frame(x = xtime - xshift,
     gr = "SR"))
> CR.ci.conf.M <- predict(model.M, newdata = data.frame(x = xtime - xshift,
     gr = "CR"), interval = "confidence")
> CR.ci.pred.M <- predict(model.M, newdata = data.frame(x = xtime - xshift,
     gr = "CR"), interval = "prediction")
> SR.ci.conf.M <- predict(model.M, newdata = data.frame(x = xtime - xshift,
     gr = "SR"), interval = "confidence")
> SR.ci.pred.M <- predict(model.M, newdata = data.frame(x = xtime - xshift,
     gr = "SR"), interval = "prediction")
> yrange <- range(c(data$y, pred.M.CR, pred.M.SR))
> plot(xtime[c(1, n)], yrange, type = "n", xlab = TxtX, main = TXT, ylab = TxtY)
> matlines(xtime, cbind(y1, y2), type = "o", pch = c(21, 22), bg = c("red",
     "cyan"), lty = c(1, 1), col = c("darkred", "darkcyan"))
> matlines(xtime, cbind(pred.M.CR, pred.M.SR), lty = c(1, 1), col = c("dodgerblue",
     "darkgreen"), lwd = 2)
> matlines(xtime, cbind(CR.ci.conf.M[, 2:3], SR.ci.conf.M[, 2:3]), lty = c(2,
     2, 2, 2), col = c("lightblue4", "lightblue4", "lightgreen", "lightgreen"),
     lwd = 2)
> matlines(xtime, cbind(CR.ci.pred.M[, 2:3], SR.ci.pred.M[, 2:3]), lty = c(3,
     3, 3, 3), col = c("lightblue4", "lightblue4", "lightgreen", "lightgreen"),
     lwd = 2)
> mtext(txtCoef.M)
```



Obrázek 2: Různoběžné regresní přímky pro data *Rozvodovost v České a Slovenské republice v letech 1960-1970*

Nyní budeme uvažovat podmodel modelu M, kdy v matici plánu vypustíme poslední sloupec. Model bude tvaru

$$\boxed{M_1}: \ Y_{ji} = \alpha + \alpha_2 + \beta x_{ji} + \varepsilon_{ji} \quad \text{kde} \quad j = 1,2 \quad i = 1,\dots,n_j \qquad (n_1 = n_2 = 11)$$

a parametr α_2 bude nulový pro Českou republiku a pro Slovenskou republiku bude značit odchylku od parametru α `(Intercept)`.

Nyní pomocí funkce `lm()` provedeme odhad modelu $\boxed{M_1}$.

```
> model.M1 <- lm(y ~ x + gr, data)
> summary(model.M1)


Call:
lm(formula = y ~ x + gr, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16295 -0.07494 -0.03068  0.04608  0.25386

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.691818   0.037266  45.399  < 2e-16 ***
x            0.048864   0.008333   5.864 1.20e-05 ***
grSR        -1.073636   0.052701 -20.372 2.28e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1236 on 19 degrees of freedom
Multiple R-squared: 0.9594,       Adjusted R-squared: 0.9552
F-statistic: 224.7 on 2 and 19 DF,  p-value: 5.988e-14
```

Vypočteme koeficienty obou přímek.

```
> (Coef.M1 <- coef(model.M1))
```

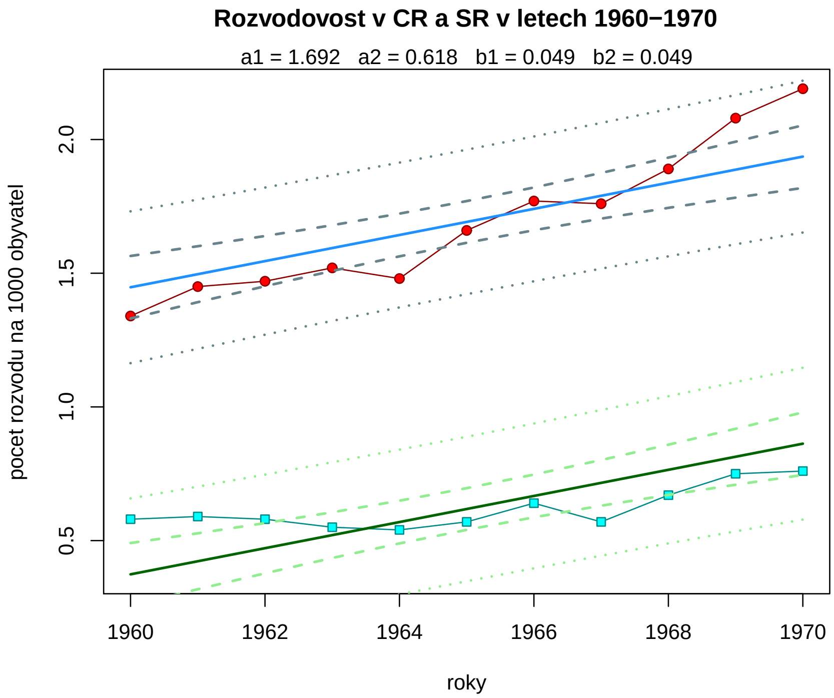
```
(Intercept)           x        grSR
 1.69181818  0.04886364 -1.07363636
```

```
> a1.M1 <- Coef.M1[1]
> a2.M1 <- Coef.M1[1] + Coef.M1[3]
> b1.M1 <- Coef.M1[2]
> b2.M1 <- Coef.M1[2]
> txtCoef.M1 <- paste("a1 =", round(a1.M1, 3), "  a2 =", round(a2.M1,
     3), "  b1 =", round(b1.M1, 3), "  b2 =", round(b2.M1, 3))
> cat(paste("Model M1 :", txtCoef.M1, "\n"))
```

```
Model M1 : a1 = 1.692   a2 = 0.618   b1 = 0.049   b2 = 0.049
```

Vykreslíme do jednoho grafu obě dvě přímky, intervaly spolehlivosti kolem střední hodnoty a prediční intervaly.

```
> pred.M1.CR <- predict(model.M1, newdata = data.frame(x = xtime - xshift,
     gr = "CR"))
> pred.M1.SR <- predict(model.M1, newdata = data.frame(x = xtime - xshift,
     gr = "SR"))
> CR.ci.conf.M1 <- predict(model.M1, newdata = data.frame(x = xtime -
     xshift, gr = "CR"), interval = "confidence")
> CR.ci.pred.M1 <- predict(model.M1, newdata = data.frame(x = xtime -
     xshift, gr = "CR"), interval = "prediction")
> SR.ci.conf.M1 <- predict(model.M1, newdata = data.frame(x = xtime -
     xshift, gr = "SR"), interval = "confidence")
> SR.ci.pred.M1 <- predict(model.M1, newdata = data.frame(x = xtime -
     xshift, gr = "SR"), interval = "prediction")
> yrange <- range(c(data$y, pred.M1.CR, pred.M1.SR))
> plot(xtime[c(1, n)], yrange, type = "n", xlab = TxtX, main = TXT, ylab = TxtY)
> matlines(xtime, cbind(y1, y2), type = "o", pch = c(21, 22), bg = c("red",
     "cyan"), lty = c(1, 1), col = c("darkred", "darkcyan"))
> matlines(xtime, cbind(pred.M1.CR, pred.M1.SR), lty = c(1, 1), col = c("dodgerblue",
     "darkgreen"), lwd = 2)
> matlines(xtime, cbind(CR.ci.conf.M1[, 2:3], SR.ci.conf.M1[, 2:3]), lty = c(2,
     2, 2, 2), col = c("lightblue4", "lightblue4", "lightgreen", "lightgreen"),
     lwd = 2)
> matlines(xtime, cbind(CR.ci.pred.M1[, 2:3], SR.ci.pred.M1[, 2:3]), lty = c(3,
     3, 3, 3), col = c("lightblue4", "lightblue4", "lightgreen", "lightgreen"),
     lwd = 2)
> mtext(txtCoef.M1)
```



Obrázek 3: Rovnoběžné regresní přímky pro data *Rozvodovost v České a Slovenské republice v letech 1960-1970*

Nakonec budeme uvažovat podmodel modelu M_1, kdy v matici plánu vypustíme poslední sloupec. Model bude tvaru

$$\boxed{M_2}: \; Y_{ji} = \alpha + \beta x_{ji} + \varepsilon_{ji} \quad \text{kde} \quad j = 1, 2 \quad i = 1, \dots, n_j \qquad (n_1 = n_2 = 11)$$

a parametry α a β budou společné pro Českou i pro Slovenskou republiku.

Pomocí funkce `lm()` provedeme odhad modelu $\boxed{M_2}$.

```
> model.M2 <- lm(y ~ x, data)
> summary(model.M2)


Call:
lm(formula = y ~ x, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.68273 -0.56557  0.02159  0.50136  0.79068 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.15500    0.12275   9.409 8.72e-09 ***
x            0.04886    0.03882   1.259    0.223    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.5758 on 20 degrees of freedom
```

```
Multiple R-squared: 0.07341,        Adjusted R-squared: 0.02708
F-statistic: 1.585 on 1 and 20 DF,  p-value: 0.2226
```

Vypočteme koeficienty.

```
> (Coef.M2 <- coef(model.M2))
```

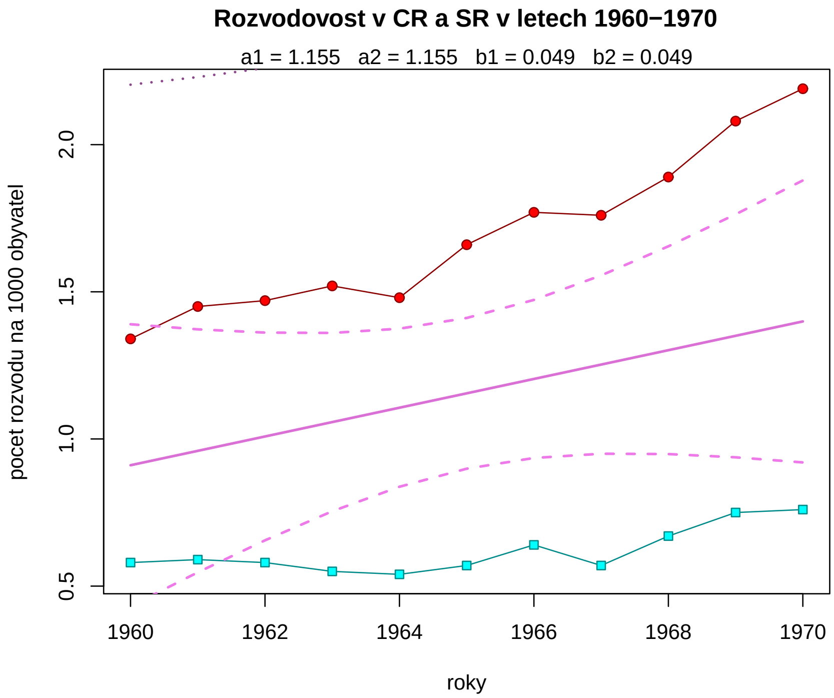
```
(Intercept)           x
 1.15500000  0.04886364
```

```
> a1.M2 <- Coef.M2[1]
> a2.M2 <- Coef.M2[1]
> b1.M2 <- Coef.M2[2]
> b2.M2 <- Coef.M2[2]
> txtCoef.M2 <- paste("a1 =", round(a1.M2, 3), "  a2 =", round(a2.M2,
     3), "  b1 =", round(b1.M2, 3), "  b2 =", round(b2.M2, 3))
> cat(paste("Model M2 :", txtCoef.M2, "\n"))
```

```
Model M2 : a1 = 1.155   a2 = 1.155   b1 = 0.049   b2 = 0.049
```

Vykreslíme do jednoho grafu jedinou regresní přímku, intervaly spolehlivosti kolem střední hodnoty a prediční intervaly.

```
> pred.M2 <- predict(model.M2, newdata = data.frame(x = xtime - xshift,
     gr = "CR"))
> pred.M2 <- predict(model.M2, newdata = data.frame(x = xtime - xshift,
     gr = "SR"))
> ci.conf.M2 <- predict(model.M2, newdata = data.frame(x = xtime - xshift,
     gr = "CR"), interval = "confidence")
> ci.pred.M2 <- predict(model.M2, newdata = data.frame(x = xtime - xshift,
     gr = "CR"), interval = "prediction")
> yrange <- range(c(data$y, pred.M2))
> plot(xtime[c(1, n)], yrange, type = "n", xlab = TxtX, main = TXT, ylab = TxtY)
> matlines(xtime, cbind(y1, y2), type = "o", pch = c(21, 22), bg = c("red",
     "cyan"), lty = c(1, 1), col = c("darkred", "darkcyan"))
> lines(xtime, pred.M2, lty = 1, col = c("orchid"), lwd = 2)
> matlines(xtime, ci.conf.M2[, 2:3], lty = c(2, 2), col = c("orchid2"),
     lwd = 2)
> matlines(xtime, ci.pred.M2[, 2:3], lty = c(3, 3), col = c("orchid4"),
     lwd = 2)
> mtext(txtCoef.M2)
```



Obrázek 4: Jediná regresní přímka pro data *Rozvodovost v České a Slovenské republice v letech 1960-1970*

Nyní pomocí příkazu `anova()` otestujeme pomocí F–testů, zda se modely významně zhoršily, když jsme postupně odstranili poslední sloupec matice plánu.

```
> anova(model.M, model.M1, model.M2)
```

```
Analysis of Variance Table

Model 1: y ~ x * gr
Model 2: y ~ x + gr
Model 3: y ~ x
  Res.Df    RSS Df Sum of Sq        F    Pr(>F)
1     18 0.0770
2     19 0.2902 -1   -0.2133   49.885 1.378e-06 ***
3     20 6.6301 -1   -6.3398 1482.824 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Z výsledků jasně vidíme to, co již bylo patrné z grafů, a to že není možné uvažovat ani model, kde jsou regresní přímky rovnoběžné. Takže musíme zůstat u plného modelu, který jsme označili jako $\boxed{M}$.

Testování homogenity rozptylu

Pro výsledný model $\boxed{M}$ ještě provedeme test homogenity rozptylu, a to jak pomocí F–testu, tak pomocí Bartlettova testu. Testovat samozřejmě musíme rezidua modelu.

```
> var.test(resid(model.M) ~ data$gr)


        F test to compare two variances

data:  resid(model.M) by data$gr
F = 2.0545, num df = 10, denom df = 10, p-value = 0.2717
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5527571 7.6360862
sample estimates:
ratio of variances
          2.054483
```

```
> bartlett.test(resid(model.M) ~ data$gr)


        Bartlett test of homogeneity of variances

data:  resid(model.M) by data$gr
Bartlett's K-squared = 1.2086, df = 1, p-value = 0.2716
```

Oba dva testy ukazují, že homogenitu rozpylu **nezamítáme**, neboť

F–test

- interval spolehlivosti neobsahuje jedničku
- p–hodnota není menší než 0.05

Bartlettův test

- p–hodnota není menší než 0.05

C. Úkol:

U 126 podniků řepařské oblasti v České republice byl sledován hektarový výnos cukrovky ve vztahu ke spotřebě průmyslových hnojiv.

Data jsou uložena v souboru nazvaném `cukrovka.txt` ve 4 sloupcích:

1. sloupec	dolní hranice spotřeby $K_2 0$ (kg/ha)
2. sloupec	horní hranice spotřeby $K_2 0$ (kg/ha)
3. sloupec	četnosti
4. sloupec	průměrné výnosy cukrovky (q/ha)

(a) Načtěte soubor dat `cukrovka.txt`.

(b) Odhadněte parametry regresních funkcí tvaru

$$
\begin{aligned}
y =& \beta_0 + \beta_1 x \\
y =& \beta_0 + \beta_1 x + \beta_2 x^2 \\
y =& \beta_0 + \beta_1 x^{0.5}
\end{aligned}
$$

<u>Poznámka</u>: Za hodnoty nezávisle proměnné volte střed intervalu. Nezapomeňte zohlednit fakt, že v každém intervalu byl jiný počet pozorování (viz 3. sloupec).

(c) Porovnejte vhodnost tří použitých regresních modelů.