- předp., že kromě $y_i$ máme pro každé $i \in \hat{n}$ k dispozici také $m$ nezávislých prom. $x_{i1}, x_{i2}, \ldots, x_{im}$
- model:

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

  kde $e_1, \ldots, e_n$ jsou nezávislé (nekorelované) chyby a $e_i \sim (0, \sigma^2)$
- na základě pozorování $(x_{i1}, \ldots, x_{im}, y_i)$, $i = 1, \ldots n$, chceme odhadnout par. $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^T$
- předpoklad: $n > m + 1$
- maticový zápis: $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T, \quad \boldsymbol{y} = (y_1, \ldots, y_n)^T, \quad \boldsymbol{e} = (e_1, \ldots, e_n)^T,$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \qquad \text{matice modelu (regresní matice)}$$

Model:

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

Nejdříve budeme předpokládat $e_1, \ldots, e_n$ nezávislé, $e_i \sim N(0, \sigma^2)$, tzn.

$$\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n) \qquad \text{a} \qquad \boldsymbol{Y} \sim N_n(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n).$$

Věrohodnostní funkce:

Pro pevné $\sigma^2$ je

$$\max_{\boldsymbol{\beta}} L(\beta, \sigma^2) \qquad \Longleftrightarrow \qquad \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta) = \min_{\boldsymbol{\beta}} g(\beta)\,.$$

### Věta 3.1

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

$$\boldsymbol{X}^T \boldsymbol{X} \beta = \boldsymbol{X}^T \boldsymbol{y} \,. \qquad \text{(soustava normálních rovnic)} \tag{3.1}$$

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

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

Důkaz.

### Lemma 3.1

Soustava lineárních rovnic $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ má řešení $\iff$ $\langle \boldsymbol{y}; \boldsymbol{z} \rangle = 0$ pro všechna $\boldsymbol{z}$ splňující $\boldsymbol{A}^T\boldsymbol{z} = \boldsymbol{0}$.

### Věta 3.2

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

### Důkaz.

$\square$

### Poznámka 3.1

- z vět plyne, že MLE $\widehat{\beta}$ může být nalezen řešením $m+1$ lineárních rovnic o $m+1$ neznámých
- málokdy existuje analytické řešení, je třeba použít numerické metody
- matice $\boldsymbol{X}^T\boldsymbol{X}$ může být špatně podmíněná, což ovlivňuje numerickou přesnost $\widehat{\beta}$
- často se užívají metody jako Choleského rozklad, QR rozklad, singulární rozklad (SVD)

- odvodili jsme pro normální chyby, minimalizace $g(\beta)$ lze ale použít i pro jiné druhy chyb
- nalezené $\widehat{\beta}$ se pak nazývá ordinary least squares estimate (OLS)

Jak poznat, že existuje jednoznačné řešení NR bez nutnosti výpočtu $\boldsymbol{X}^T\boldsymbol{X}$?

**Věta 3.3**

Matice $\boldsymbol{X}^T\boldsymbol{X}$ je nesingulární $\iff$ jsou sloupce matice $\boldsymbol{X}$ lineárně nezávislé.

**Důkaz.**

$\square$

## Poznámka 3.2

- $n > m+1,\ h(\boldsymbol{X}) = m+1 \quad \Longrightarrow \quad$ ex. jednoznačné řešení NR a to $\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$,
- pokud navíc $\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$, je to MLE

## Poznámka 3.3

- pokud jsou sloupce $\boldsymbol{X}$ LZ, je $\boldsymbol{X}^T\boldsymbol{X}$ singulární (většinou detekováno num. metodou výpočtu $\widehat{\beta}$)
- horší situace je, pokud jsou sloupce $\boldsymbol{X}$ „téměř“ LZ, tzv. multikolinearita
- způsobuje problémy při výpočtu $\widehat{\beta}$, protože $\boldsymbol{X}^T\boldsymbol{X}$ je „téměř“ singulární
- jak detekovat a řešit multikolinearitu probereme na konci přednášky

# Odhad parametru $\sigma^2$

- pro normální chyby získáme MLE $\sigma^2$ derivací $\ln L(\boldsymbol{\beta}, \sigma^2)$, tedy

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

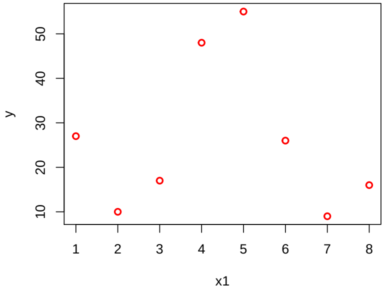
  kde $\widehat{y}_i = (\boldsymbol{X}\widehat{\beta})_i = \boldsymbol{x}_i^T \widehat{\beta}$, $i = 1, \ldots, n$, a $\boldsymbol{x}_i^T$ značí $i$-tý řádek matice $\boldsymbol{X}$
- jedná se o vychýlený odhad, proto se obecně používá nestranný odhad

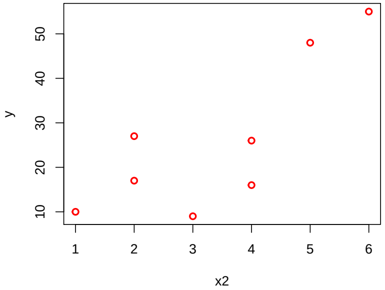
$$s_n^2 = \frac{1}{n - (m+1)} SSE = \frac{1}{n - m - 1} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

  a $s_n = \sqrt{s_n^2}$ jako odhad $\sigma$
- pro $e_i \sim (0, \sigma^2)$ se také používají statistiky $s_n^2, s_n$

## PŘÍKLAD 3.1 (Umělá data)

| x1 | x2 | y |
|---|---|---|
| 2 | 1 | 10 |
| 3 | 2 | 17 |
| 4 | 5 | 48 |
| 1 | 2 | 27 |
| 5 | 6 | 55 |
| 6 | 4 | 26 |
| 7 | 3 | 9 |
| 8 | 4 | 16 |




**Jednorozměrné modely:**

$y = \beta_0 + \beta_1 x_1 + e$: $\quad \widehat{\beta}_0 = 28.5,\ p_{val} = 0.093,\quad \widehat{\beta}_1 = -0.57,\ p_{val} = 0.847,\quad R^2 = 0.00668$

$y = \beta_0 + \beta_1 x_2 + e$: $\quad \widehat{\beta}_0 = -1.34,\ p_{val} = 0.891,\quad \widehat{\beta}_1 = 8.10,\ p_{val} = 0.018,\quad R^2 = 0.6356$

**Dvourozměrný model:**

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$: $\quad \widehat{\beta}_0 = 8,\quad \widehat{\beta}_1 = -5,\quad \widehat{\beta}_2 = 12,\quad p_{val} = 0.000,\quad R^2 = 1$

## Příklad 3.2 (Porodní váha a gestační stáří)

| Sex | Age | Weight |
|---|---|---|
| boy | 40 | 2968 |
| boy | 38 | 2795 |
| girl | 40 | 3163 |
| ⋮ | ⋮ | ⋮ |

Definujeme proměnné:

$$y = \text{Weight}, \quad x_1 = \text{Age}, \quad x_2 = \begin{cases} 0, & \text{Sex} = \text{boy}; \\ 1, & \text{Sex} = \text{girl}. \end{cases}$$ (dummy variable)

**Model:** $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

Uvažujeme vlastně dvě rovnoběžné přímky

$$Y_b = \beta_0 + \beta_1 x_1 + e \quad \text{pro chlapce}, \qquad Y_g = (\beta_0 + \beta_2) + \beta_1 x_1 + e \quad \text{pro dívky}$$

Pro různé směrnice, můžeme přidat interakci $x_1$ a $x_2$, tj. $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$

```
mod <- lm(Weight ~  Age + Sex + Age:Sex)

            Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1268.67    1114.64  -1.138 0.268492
## Age           111.98      29.05   3.855 0.000986 ***
## Sexgirl      -872.99    1611.33  -0.542 0.593952
## Age:Sexgirl    18.42      41.76   0.441 0.663893
## ---
## Residual standard error: 180.6 on 20 degrees of freedom
## Multiple R-squared:  0.6435,        Adjusted R-squared:   0.59
## F-statistic: 12.03 on 3 and 20 DF,  p-value: 0.000101
```

## 3.1 Vlastnosti odhadů $\widehat{\beta}, s_n^2$

VĚTA 3.4

Nechť $\widehat{\beta}$ je OLS odhad parametru $\beta$ v modelu $(**)$, kde $h(\boldsymbol{X}) = m+1$ a $e_1, \ldots, e_n$ jsou nekorelované a $e_i \sim (0, \sigma^2)$. Potom platí

1) $\mathrm{E}(\widehat{\beta}) = \beta$, ($\widehat{\beta}$ je nestranný)
2) $\mathrm{Cov}(\widehat{\beta}) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$,
3) $\mathrm{E}(s_n^2) = \sigma^2$, ($s_n^2$ je nestranný)
4) pokud navíc $e_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, potom $\widehat{\beta} \sim N_{m+1}\left(\beta, \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right)$,

   speciálně $\widehat{\beta}_i \sim N(\beta_i, \sigma^2 v_i)$, kde $v_i = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}_{ii}$ .

Důkaz.

Poznámka 3.4 (Vlastnosti projekční matice $\boldsymbol{H}$)

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

$$\boldsymbol{H}^2 = \boldsymbol{H}, \quad (\boldsymbol{I}_n - \boldsymbol{H})^2 = (\boldsymbol{I}_n - \boldsymbol{H}), \quad (\textit{idempotentnost})$$

$$\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}, \quad \boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{H}) = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{H} = \boldsymbol{0}, \quad \mathrm{tr}(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = m + 1\,.$$