

Pro řešení lineárního regresního modelu  $Y_i = m(x_i) + \varepsilon_i$ ,  $i = 1, \dots, n$ , v R slouží příkaz `lm` (*linear model*):

```
model <- lm (formule, data = DatováTabulka) , příp.
```

```
model <- lm (formule, data = DatováTabulka, weights = VektorVah) .
```

Pro tzv. *formuli* se používá speciální syntaxe, kde **Y** je název sloupce závisle proměnné, **x** je název sloupce nezávisle proměnné:

$m(x)$	<i>formule</i>
$\beta_0 + \beta_1 x$	$Y \sim x$ nebo $Y \sim 1 + x$ , člen $\beta_0$ je totiž vkládán implicitně
$\beta_1 x$	$Y \sim 0 + x$ , odstranění členu $\beta_0$ nutno zapsat explicitně
$\beta_0 + \beta_1 x + \beta_2 x^2$	$Y \sim x + I(x^2)$
$\beta_2 x^2$	$Y \sim 0 + I(x^2)$
$\beta_1  x $	$Y \sim 0 + I(abs(x))$
$\beta_0 + \beta_1 e^x$	$Y \sim I(exp(x))$
$\beta_0 + \beta_1 \ln x$	$Y \sim I(log(x))$
$\beta_0 + \beta_1 \sqrt{x}$	$Y \sim I(sqrt(x))$

Detailní výsledky a další číselné charakteristiky získáme příkazem

```
prehled <- summary (model) ,
```

příp. `prehled <- summary (model, correlation=TRUE)` pro výběrovou korelační matici parametrů.

$\hat{\beta}$	MNČ-odhady parametrů	<code>model\$coefficients</code> <code>coef(model)</code>
$(\hat{\beta}_j, SD(\hat{\beta}_j), T_j, p_j)$	odhady, směrodatné odchylky, testy významnosti, p-hodnoty	<code>prehled\$coefficients</code> <code>coef(prehled)</code>
$\hat{Y}$	aproximované hodnoty	<code>model\$fitted.values</code> <code>fitted.values(model)</code>
$r$	rezidua	<code>model\$residuals</code> <code>residuals(model)</code>
$n - k$	stupně volnosti modelu: # měření = $n > k$ = # parametrů	<code>model\$df.residual</code>
$X$	matice plánu	<code>model.matrix(model)</code>
$w$	váhy	<code>model\$weights</code>
$s$	odhad sm. odchylky chyb $\varepsilon_i$	<code>prehled\$sigma</code>
$R^2$	index determinace	<code>prehled\$r.squared</code>
$\bar{R}^2$	korigovaný index determinace	<code>prehled\$adj.r.squared</code>
$(F, k-1, n-k)$	celkový F-test	<code>prehled\$fstatistic</code>
$(k, n-k, k)$	stupně volnosti	<code>prehled\$df</code>
$R(\hat{\beta})$	korelační matice odhadů $\hat{\beta}$	<code>prehled\$correlation</code>

V následujících příkladech určete (pro jednotlivé regresní funkce):

- 🧩 MNČ-odhady parametrů  $\hat{\beta}$  regresní funkce  $m(x)$ , sledujte jejich významnost
- 🧩 zapište matematický tvar regresní funkce  $m(x)$
- 🧩 reziduální součet čtverců  $S_e$  a odhad směrodatné odchylky s náhodných chyb
- 🧩 index determinace  $R^2$ , proveďte celkový F-test
- 🧩 vykreslete data a grafy regresních funkcí ( **predict** ), příp. s pásy spolehlivosti
- 🧩 vykreslete boxploty reziduí
- 🧩 modely porovnejte (mj. **anova** ), zvolte z nich nejvhodnější

- 1 Datový soubor **cv11-01.csv** : zkoumejte závislost množství kyseliny mléčné u novorozence na množství stejné látky u matky-prvorodičky (v mg ve 100 ml krve) pomocí regresní přímky a paraboly.
- 2 Datový soubor **cv11-02.csv** : zkoumejte závislost prodloužení měděné trubky v závislosti teplotním rozdílu  $\Delta t$  od referenční hodnoty  $t_0 = 20^\circ\text{C}$  pomocí vhodné regresní přímky a paraboly. Dle fyzikálních zákonů by při  $\Delta t = 0$  prodloužení mělo být nulové.

- 3 Datový soubor **cv11-03.csv** : zkoumejte závislost spotřeby paliva motorového vozidla (v l/100 km) na rychlosti (v km/h) pomocí regresní přímky a paraboly.
- 4 Datový soubor **cv11-04.csv** : zkoumejte závislost koncentrace  $\text{CO}_2$  (v ppm) v atmosféře v letech 1764–1995 pomocí několika polynomických regresních funkcí.
- 5 Datový soubor **cv11-05.csv** : zkoumejte závislost uhlíkových emisí (v milionech tun) v letech 1950–1995 pomocí několika polynomických regresních funkcí.
- 6 Datový soubor **cv11-06.csv** : zkoumejte závislost průměrné teploty (ve  $^{\circ}\text{C}$ ) v letech 1866–1996 pomocí několika polynomických regresních funkcí.
- 7 Datový soubor **cv11-07.csv** : zkoumejte závislost **logaritmu** objemu vytěžené ropy (v tisících barelů) v letech 1880–1988 pomocí několika polynomických regresních funkcí, grafy vykreslete i pro nelogaritmované hodnoty.
- 8 Vyřešte Příklady zadané na přednášce.







