

Cvičení MV011 Statistika I

11. Lineární regresní model

Monika Kroupová, Ondřej Pokora, Petra Ráboňová

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

jaro 2017



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 = DatovaTabulka), příp.
```

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

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

$m(x)$	formule
$\beta_0 + \beta_1 x$	<code>Y ~ x</code> nebo <code>Y ~ 1 + x</code> , člen β_0 je totiž vkládán implicitně
$\beta_1 x$	<code>Y ~ 0 + x</code> , odstranění členu β_0 nutno zapsat explicitně
$\beta_0 + \beta_1 x + \beta_2 x^2$	<code>Y ~ x + I(x^2)</code>
$\beta_2 x^2$	<code>Y ~ 0 + I(x^2)</code>
$\beta_1 x $	<code>Y ~ 0 + I(abs(x))</code>
$\beta_0 + \beta_1 e^x$	<code>Y ~ I(exp(x))</code>
$\beta_0 + \beta_1 \ln x$	<code>Y ~ I(log(x))</code>
$\beta_0 + \beta_1 \sqrt{x}$	<code>Y ~ I(sqrt(x))</code>

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	<code>model\$df.residual</code>
X	matice plánu	<code>model.matrix(model)</code>
w	váhy	<code>model\$weights</code>
s	odhad sm. odchylky chyb ε_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>

Úkoly v příkladech:

- MNČ-odhady parametrů $\hat{\beta}$ regresní funkce $m(x)$, sledujte i 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.

Příklad 1

*Datový soubor **KysMlecna.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.*

Příklad 2

*Datový soubor **prodlouzeni.csv**: zkoumejte závislost prodloužení měděné trubky v závislosti teplotním rozdílu Δt 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.*

Příklad 3

Datový soubor [spotreba2.csv](#): zkoumejte závislost spotřeby paliva motorového vozidla (v l/100 km) na rychlosti (v km/h) pomocí regresní přímky a paraboly.

Příklad 4

Datový soubor [CO2.csv](#): zkoumejte závislost koncentrace CO₂ (v ppm) v atmosféře v letech 1764–1995 pomocí několika polynomických regresních funkcí.

Příklad 5

Datový soubor [EmiseUhliku.csv](#): zkoumejte závislost uhlíkových emisí (v milionech tun) v letech 1950–1995 pomocí několika polynomických regresních funkcí.

Příklad 6

Datový soubor [teplota.csv](#): zkoumejte závislost průměrné teploty (ve °C) v letech 1866–1996 pomocí několika polynomických regresních funkcí.

Příklad 7

Datový soubor [ropa.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.







