Poznámky k předmětu Aplikovaná statistika, 12. téma

Lineární regrese

Z teorie často víme, že naměřená data vykazují funkcionální závislost, ovšem zatíženou chybou měření. V našem případě je tato závislost lineární. Chtěli bychom najít takovou přímku, která „nejlépe“ (v jakém smyslu?) vystihuje naše data měřená s náhodnou chybou. Tuto úlohu lze popsat pomocí tzv. **lineární regrese**

$$Y(x) = \beta_1 + \beta_2 x + \epsilon(x),$$

kde

- pro každé x (nenáhodné!) z podmnožiny M reálných čísel je $Y(x)$ náhodná veličina,
- $\beta_1, \beta_2 \in \mathbb{R}$ jsou parametry,
- pro každé $x \in M$ je $\epsilon(x)$ náhodná veličina (chyba měření).

Ona funcionální závislost se nazývá **regresní funkce** $\eta(x)$ a v lineárním případě je definovaná vztahem

$$\eta(x) = \mathbb{E}\left[Y(x)\right] = \beta_1 + \beta_2 x.$$

Sestavme tzv. **lineární regresní model**:

$$Y_j = \beta_1 + \beta_2 x_1 + \epsilon_j,\ j = 1, \ldots, n,$$

kde pro hodnoty nezávislé proměnné $x_j, j = 1, \ldots, n$, jsme označili

$$Y_j \equiv Y(x_j),\ \epsilon_j \equiv \epsilon(x_j),\ \eta(x_j) \equiv \eta_j,\ j = 1, \ldots, n.$$

Jeho **předpoklady** jsou

(i) $n > 2$ a existují $i, j \in \{1, \ldots, n\}$ taková, že $x_i \neq x_j$,

(ii) $\mathbb{E}\, Y_j = \eta_j = \beta_1 + \beta_2 x_j,\ j = 1, \ldots, n$,

(iii) $\operatorname{var} Y_j = \sigma^2 > 0,\ j = 1, \ldots, n$,

(iv) $\operatorname{cov}(Y_i, Y_j) = 0,\ i, j = 1, \ldots, n, i \neq j$.

Poznámka 1. *Podmínky (ii)–(iv) lze přepsat v řeči chyb $\epsilon(x_j)$ následovně:*

$$\mathbb{E}\,\epsilon_j = 0, \operatorname{var}\epsilon_j = \sigma^2,\ j = 1, \ldots, n,\ \operatorname{cov}(\epsilon_i, \epsilon_j) = 0,\ i, j = 1, \ldots, n, i \neq j.$$

Bodové odhady parametrů modelu

Model má tři parametry $\beta_1, \beta_2, \sigma^2$, pro něž bychom chtěli nalézt bodové odhady

$$\hat{\beta}_1 = b_1,\ \hat{\beta}_2 = b_2,\ \hat{\sigma}^2 = s^2.$$

- Pro odhady β_1, β_2 se standardně používá **metoda nejmenších čtverců** – hledáme b_1, b_2 tak, aby platilo

$$\sum_{j=1}^{n} \big(Y_j - (b_1 + b_2 x_j)\big)^2 = \min_{\beta_1, \beta_2 \in M} \sum_{j=1}^{n} \big(Y_j - (\beta_1 + \beta_2 x_j)\big)^2.$$

  Je to aplikační úloha na lokální extrémy funkcí dvou proměnných (viz Matematiku II) a nalezená řešení b_1, b_2 jsou nejlepšími nestrannými lineárními odhady parametrů β_1, β_2.

Poznámka 2. *Buď $X_1, \dots, X_n$ náhodný výběr z rozdělení náhodné veličiny X. Bodový odhad $\hat{\theta}_N \equiv \hat{\theta}_N(X_1, \dots, X_n)$ parametru θ tohoto rozdělení je* ***nestranný****, jestliže*

$$\mathbb{E}\,\hat{\theta}_N = \theta.$$

Nestranný bodový odhad $\hat{\theta}_{NN}$ parametru θ je ***nejlepší nestranný odhad****, jestliže*

$$\operatorname{var}\hat{\theta}_{NN} = \min_{\hat{\theta}_N} \operatorname{var}\hat{\theta}_N.$$

Bodový odhad $\hat{\theta}$ je ***lineární****, pokud existují konstanty $a_0, a_1, \dots, a_n \in \mathbb{R}$ tak, že*

$$\hat{\theta} = a_0 + \sum_{i=1}^{n} a_i X_i.$$

Ona řešení β_1, β_2 lze vyjádřit ve tvaru

$$b_1 = \bar{Y} - b_2 \bar{x}, \tag{1}$$

$$b_2 = \frac{n \sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \tag{2}$$

kde

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

- Tudíž bodovým odhadem regresní funkce η_j je

$$\hat{\eta}_j = b_1 + b_2 x_j,\ j = 1, \dots, n.$$

- Bodové odhady chyb $\hat{\epsilon}_j,\ j = 1, \dots, n$, získané na základě metody nejmenších čtverců

$$\hat{\epsilon}_j = Y_j - b_1 - b_2 x_j$$

  se nazývají **rezidua** a součet jejich kvadrátů (značí se S_e)

$$S_e = \sum_{j=1}^{n} \hat{\epsilon}_j^2 = \sum_{j=1}^{n} (Y_j - \hat{\eta}_j)^2 = \sum_{j=1}^{n} (Y_j - b_1 - b_2 x_j)^2$$

  pak **reziduální součet čtverců**. Snadnou úpravou lze získat tvar vhodný pro výpočty

$$S_e = \sum_{j=1}^{n} Y_j^2 - b_1 \sum_{j=1}^{n} Y_j - b_2 \sum_{j=1}^{n} x_j Y_j.$$

Využitím reziduálního součtu čtverců obdržíme nestranný odhad rozptylu

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-2} S_e,$$

který se nazývá **reziduální rozptyl**.

Intervaly spolehlivosti – pás spolehlivosti a predikční pás

Pokud bychom chtěli konstruovat intervaly spolehlivosti pro parametry regresního modelu, potřebujeme dodatečné předpoklady na chyby:

(i) $\epsilon_j,\ j = 1, \ldots, n$, jsou vzájemně nezávislé náhodné veličiny,

(ii) $\epsilon_j,\ j = 1, \ldots, n$ mají normální rozdělení $\mathcal{N}(0, \sigma^2)$.

Pak $(1-\alpha)100\%$ oboustranný interval spolehlivosti pro

- β_1 je
$$\left[b_1 - t_{1-\frac{\alpha}{2}}(n-2) s_{b_1}, b_1 + t_{1-\frac{\alpha}{2}}(n-2) s_{b_1}\right],$$
- β_2 je
$$\left[b_2 - t_{1-\frac{\alpha}{2}}(n-2) s_{b_2}, b_2 + t_{1-\frac{\alpha}{2}}(n-2) s_{b_2}\right], \tag{3}$$
- regresní funkci $\eta(x)$ je
$$\left[b_1 + b_2 x - t_{1-\frac{\alpha}{2}}(n-2) s_{\hat{\eta}(x)}, b_1 + b_2 x + t_{1-\frac{\alpha}{2}}(n-2) s_{\hat{\eta}(x)}\right],$$

kde

$$\begin{aligned}
s_{b_1}^2 &= s^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}, \\
s_{b_2}^2 &= s^2 \frac{n}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}, \\
s_{\hat{\eta}(x)}^2 &= s^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right),
\end{aligned}$$

a $t_{1-\frac{\alpha}{2}}(n-2)$ je $\left(1-\frac{\alpha}{2}\right)100\%$ kvantil t–rozdělení s $n-2$ stupni volnosti.

Poznámka 3. *(a) Vzhledem k nápadné podobnosti těchto intervalů spolehlivosti s intervaly spolehlivosti pro střední hodnotu (při neznámém rozptylu) též založených na t–rozdělení (s $n-1$ stupni volnosti), si zkuste napsat jednostranné intervaly spolehlivosti pro parametry regresní funkce.*

(b) Pomocí intervalů spolehlivosti lze testovat nulovost parametrů regresní funkce. Jak testovat hypotézy pomocí intervalů spolehlivosti jsme se naučili v tématu o jednovýběrových testech.

Jestliže se díváme na horní a dolní meze intervalu spolehlivosti pro regresní funkci jako na funkce proměnné x, pak plocha ohraničená grafy těchto funkcí se nazývá **pás spolehlivosti kolem regresní přímky**. Obdobně lze zkonstruovat tzv. **predikční pás kolem regresní přímky**, což je plocha ohraničená mezemi $(1-\alpha)100\%$ intervalů zkonstruovaných přímo pro náhodnou veličinu $Y(x)$

$$\left[b_1 + b_2 x - t_{1-\frac{\alpha}{2}}(n-2) s_{\hat{Y}(x)}, b_1 + b_2 x + t_{1-\frac{\alpha}{2}}(n-2) s_{\hat{Y}(x)}\right],$$

kde

$$s^2_{\hat{Y}(x)} = s^2 \left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right) = s^2 + s^2_{\hat{\eta}(x)}.$$

Kvalitu modelu, tj. jak dobře lineární regresní model vystihuje naše data, lze posoudit např. pomocí tzv. **koeficientu determinace**

$$R^2 = 1 - \frac{S_e}{S_t},$$

kde S_e je výše definovaný reziduální součet čtverců a S_t je tzv. **celkový součet čtverců**

$$S_t = \sum_{j=1}^n (Y_j - \bar{Y})^2.$$

Koeficient determinace udává, jakou část celkové variability závisle proměnné Y se podařilo regresním modelem vysvětlit. Nabývá hodnot z intervalu $[0,1]$ a čím je blíže k jedné, tím lépe model naše data vystihuje.

Příklad: Experimentálně byla zjišťována závislost koncentrace nasyceného roztoku hydroxidu vápenatého Y (v %) na teplotě x (ve °C)

x_j	20	30	40	50	60	70	80
y_j	16,1	16,2	14,0	13,3	11,9	10,2	10,1

(a) Odhadněte bodově parametry lineární regrese a regresní funkci.

(b) Určete 95% intervaly spolehlivosti pro parametry regrese.

(c) Určete 95% interval spolehlivosti pro střední hodnotu koncentrace při teplotě 40 °C.

(d) Odhadněte bodově koncentraci nasyceného roztoku při teplotě 25 °C.