Regresní analýza dat - 01RAD

ZS 2025/26, 2+2 z,zk

Tomáš Hobza

Katedra matematiky, FJFI, Trojanova 13, 107c

tomas.hobza@fjfi.cvut.cz

Literatura

- Golberg, M. Cho, H.A.: Introduction to Regression Analysis. WITpress, Southampton 2010.
- Víšek, J. Á.: Statistická analýza dat. Vydavatelství ČVUT v Praze, Praha 1998.
- Zvára, K.: Regrese. Matfyzpress, Praha 2008.
- Olive, D.: Linear Regression. Springer, 2017.

Stručný obsah přednášky

1. Úvod - regresní analýza
2. Jednorozměrná lineární regrese
3. Vícerozměrná regrese
4. Rezidua, diagnostika a transformace
5. Výběr regresního modelu
6. Kolinearita (multikolinearita)

1. Úvod - regresní analýza

- jedna z nejužívanějších statistických metod pro analýzu vztahu mezi proměnnými
- pro svou flexibilitu, užitečnost, interpretovatelnost $\longrightarrow$ základní statistický nástroj
- pro úspěšnou a efektivní aplikaci je třeba získat náhled a pochopení

  a) příslušné teorie, b) její praktické aplikace.

  ad a) základy teorie lineární regrese (bude navazovat ZLMA)

  ad b) ilustrace teorie na příkladech - cvičení v R

Historie:

- slovo "regrese": sir *Francis Galton* (1822-1911), studie dědičnosti (1885)
- základní matematický nástroj: metoda nejmenších čtverců

  Carl Friedrich Gauss (1777 - 1855) *Adrien-Marie Legendre* (1752 - 1833)

  myšlenka: minimalizace součtu čtverců deviací pozorovaných hodnot a hodnot predikovaných modelem

  odůvodnění: Gauss - Markov theorem

Použití regresní analýzy:

a) Popis dat — zkoumání případně vyvrácení vztahů mezi proměnnými

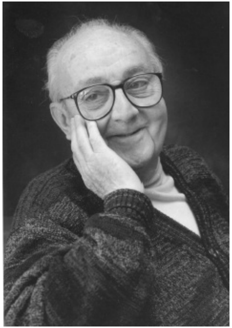
b) Interpretace — získání souhrnu nebo interpretace dat pomocí modelu prokládajícího data křivkou/plochou

c) Inference — hledání nebo vylepšení teoretických modelů
statistické techniky: odhady parametrů, testy hypotéz, predikce

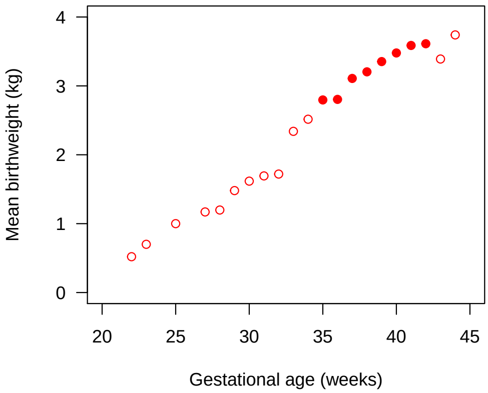
DATA: základní součást regresní analýzy

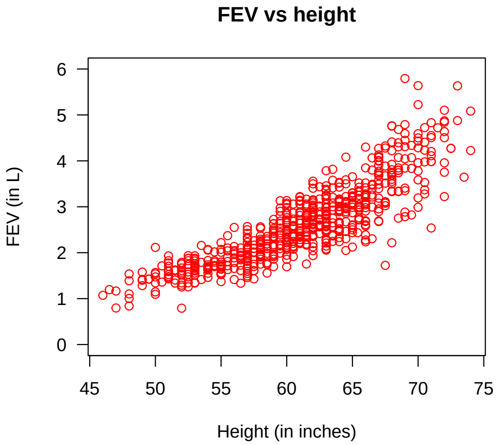
Essentially, all models are wrong, but some are useful. The practical question is how wrong do they have to be to not be useful.

George E. P. Box (1919 - 2013)



2. Jednorozměrná lineární regrese





Model jednorozměrné regrese

Sledujeme dvě fyzikální veličiny x a y, mezi kterými existuje lineární závislost

$$y = \beta_0 + \beta_1 x, \qquad \text{kde } \beta_0, \beta_1 \text{ nejsou známy.}$$

Experiment $\longrightarrow$ hodnoty dvojic (x, y)

- měření hodnot x často probíhá prakticky zcela přesně (například x se nastavuje na předem dané úrovně)
- y se měří s určitou chybou, chyba může být náhodná, y budeme chápat jako náhodnou veličinu (zn. Y).

Pro dvojice $(x_1, Y_1), \ldots, (x_n, Y_n)$ se zavádí model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, \ldots, n, \qquad (*)$$

kde

- Y_i se nazývá vysvětlovaná (závislá) proměnná
- x_i se nazývá vysvětlující (nezávislá) proměnná, někdy také prediktor nebo regresor
- β_0, β_1 jsou neznámé regresní parametry
- e_i je tzv. náhodný šum (náhodná chyba), předpoklad: $e_1, \ldots, e_n$ nezávislé a $e_i \sim (0, \sigma^2)$.

Model jednorozměrné regrese

- měřením se získají data $(x_1, y_1), \ldots, (x_n, y_n)$
- cíl statistické analýzy: určit, zda model $(*)$ dobře popisuje pozorovanou variabilitu v y

První krok: odhad neznámých parametrů β_0, β_1 a σ^2

Proložení dat přímkou - několik způsobů, zásadní bude znalost rozdělení e_i a tedy Y_i

Dvě možnosti:

1. odhadnout β_0, β_1 pomocí metody nezávisející na rozdělení chyb
2. udělat věrohodný předpoklad o rozdělení chyb, odhadnout β_0, β_1 a potom ověřit předpoklad

Poznámka 2.1

Speciální případ $e_i \sim N(0, \sigma^2)$: MLE vede na LSE, LSE může být použito i pro jiný druh chyb

Odhady parametrů pro normální chyby

A) předpokládáme, že $e_1, \ldots, e_n$ jsou *i.i.d.* $N(0, \sigma^2)$, tzn

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad \text{a} \quad Y_1, \ldots, Y_n \text{ nezávislé}$$

MLE odhady:

Odhady parametrů pro normální chyby

Odvodili jsme MLE:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}, \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \qquad \text{a} \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{y}_i)^2 \tag{2.1}$$

kde

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad \text{je predikce modelu (odhad } EY_i)$$

a

$$\widehat{e}_i = y_i - \widehat{y}_i \qquad \text{je } i - \text{té reziduum}$$

POZNÁMKA 2.2

Pro odhad σ^2 se častěji používá

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

což je nestranný odhad σ^2 (pro lib. rozdělení chyb)

Poznámka 2.3

Odhad směrodatné odchylky σ:

$$s_n = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \qquad \text{standardní chyba (standard error)}$$

už není nestranný !

Obecná vlastnost odhadů rozptylu:

$$s^2 \text{ nestranný odhad } \sigma^2 \quad \Rightarrow \quad Es \leq \sigma$$

Odhady parametrů

B) bez předpokladu normality chyb, tzn.

$$e_1, \ldots, e_n \quad \text{nezávislé (nekorelované)} \qquad \text{a} \qquad \mathrm{E}e_i = 0,\ \mathrm{Var}(e_i) = \sigma^2$$

Pro odhad β_0, β_1 lze použít minimalizaci

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

geometrická interpretace:

- $y_i - \beta_0 - \beta_1 x_i$ je vertikální vzdálenost bodu (x_i, y_i) od přímky $y = \beta_0 + \beta_1 x$
- S - "míra"jak dobře přímka prokládá data

Minimalizací S získáme $\widehat{\beta}_0, \widehat{\beta}_1$

- stejné jako u MLE pro normální data
- nazývají se ale odhady metodou nejmenších čtverců (least squares estimators - LSE)

Poznámka 2.4

Existuje více měr vhodnosti přímky, použití LSE pro lib. rozdělení chyb má dvě zdůvodnění

1. pro normální chyby LSE splývá s MLE
2. LSE odhad je navíc Best Linear Unbiased Estimator (BLUE)