Regresní analýza dat - 01RAD

ZS 2024/25, 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

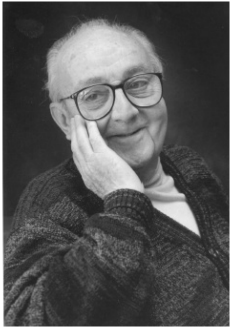
b) Interpretace — získání souhrnu nebo interpretace dat pomocí modelu prokládajícího data křivkou/plochou

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

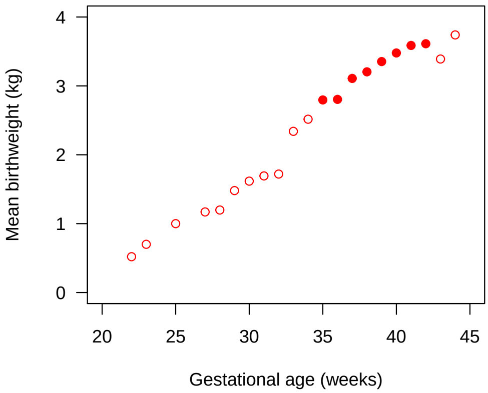
DATA: základní součást regresní analýzy

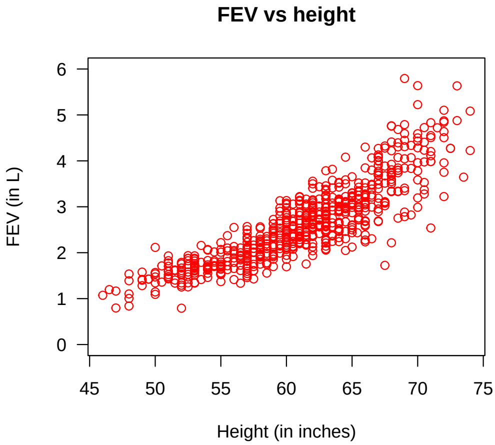
Essentially, all models are wrong, but some are useful. The practical question is how wrong do they have to be to not be useful.

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 $(x_1, y_1), \dots, (x_n, y_n)$
- 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.

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

Pro odhad β_0, β_1 lze použít minimalizaci

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

<u>geometrická interpretace:</u>

- $y_i - \beta_0 - \beta_1 x_i$ 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)

01RAD - přednáška 2, 24.9.2024

Vlastnosti odhadů $\widehat{\beta}_0, \widehat{\beta}_1, s_n^2$

VĚTA 2.1

Nechť $\widehat{\beta}_0, \widehat{\beta}_1$ jsou LSE odhady parametrů β_0, β_1 v lineárním modelu

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, \dots, n,$$

kde $e_i \sim (0, \sigma^2)$ jsou nezávislé náhodné veličiny (postačí i nekorelovanost). Potom platí:

1. $\mathrm{E}[\widehat{\beta}_0] = \beta_0, \qquad \mathrm{E}[\widehat{\beta}_1] = \beta_1, \quad$ (nestranné odhady)
2. $\mathrm{Var}[\widehat{\beta}_1] = \frac{\sigma^2}{S_{xx}}, \quad$ kde $\quad S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2,$
3. $\mathrm{Var}[\widehat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right),$
4. Pokud navíc platí, že $e_i \sim N(0, \sigma^2)$, $i = 1, \dots, n$ potom $\widehat{\beta}_j \sim N(\beta_j, \mathrm{Var}[\widehat{\beta}_j]) \quad j = 0, 1.$

Důkaz:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}, \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

Věta 2.2

Za předpokladů věty 2.1 platí

$$E(s_n^2) = \sigma^2. \qquad (s_n^2 \text{ je nestranný odhad } \sigma^2)$$

Důkaz:

Tvrzení 2.1

Nechť platí předpoklady věty 2.1 a nechť $e_1, \dots, e_n$ jsou *i.i.d.* $N(0, \sigma^2)$. Potom platí:

a) $\dfrac{(n-2)s_n^2}{\sigma^2} \sim \chi^2(n-2)$

b) odhad s_n^2 je nezávislý na $\widehat{\beta}_0$ a $\widehat{\beta}_1$.

Poznámka 2.5

Spočetli jsme

$$\sigma^2(\widehat{\beta}_0) \triangleq \mathrm{Var}(\widehat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right] \quad \text{a} \quad \sigma^2(\widehat{\beta}_1) \triangleq \mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Nestranné odhady jsou:

$$\widehat{\sigma}^2(\widehat{\beta}_0) = s_n^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right] = s_n^2 \delta_0, \qquad \widehat{\sigma}^2(\widehat{\beta}_1) = \frac{s_n^2}{S_{xx}} = s_n^2 \delta_1,$$

Odhady směrodatné odchylky veličin $\widehat{\beta}_0$ a $\widehat{\beta}_1$ pak jsou

$$\hat{\sigma}(\widehat{\beta}_0) = s_n \sqrt{\delta_0} \quad \text{a} \quad \hat{\sigma}(\widehat{\beta}_1) = s_n \sqrt{\delta_1}, \qquad \text{(standardní chyby odhadů } \widehat{\beta}_0, \widehat{\beta}_1\text{)}$$

Gauss-Markov theorem

- Chyby normální $\Rightarrow$ LSE pro $\widehat{\beta}_0, \widehat{\beta}_1$ je MLE (eficientní odhad)
- Pokud nejsou chyby normální, jaké je opodstatnění použít LSE?
  Ukážeme, že LSE jsou BLUE (best linear unbiased estimators), tedy lineární nestranné odhady s minimálním rozptylem
- můžou ale existovat nelineární nebo vychýlené odhady parametrů β_0, β_1, které jsou eficientnější než LSE (pokud se rozdělení chyb liší výrazně od normálního)

Definice 2.2

Lineární odhad parametru β, založený na pozorováních $Y_1, \dots, Y_n$, je statistika tvaru

$$\widehat{\beta} = \sum_{i=1}^{n} c_i Y_i,$$

kde c_i jsou dané reálné konstanty, $i = 1, \dots, n$.

Uvažujme model

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

Věta 2.3 (Gauss-Markov theorem)

Nechť $e_1, \ldots, e_n$ v modelu $(*)$ jsou nekorelované, mají stejný rozptyl $\mathrm{Var}(e_i) = \sigma^2$ a $\mathrm{E}e_i = 0$, $i = 1, \ldots, n$. Potom LSE $\widehat{\beta}_j$ $(j = 0, 1)$ je BLUE parametru β_j.

Důkaz:

Intervaly spolehlivosti pro β_0, β_1

- IS poskytují jistou "míru přesnosti" bodových odhadů
- pro jejich konstrukci potřebujeme znát rozdělení pravděpodobnosti bodového odhadu
- budeme tedy uvažovat normalitu chyb
- spočtené IS se ale často používají, i když rozdělení chyb není normální
  zdůvodnění: LSE odhady par. β jsou lineární funkcí $Y_i, i = 1, \dots, n$, aplikace CLT vede na asymptotickou normalitu $\widehat{\beta}_0, \widehat{\beta}_1$

V modelu $(*)$ platí:

$$\widehat{\beta}_i \sim N(\beta_i, \sigma^2(\widehat{\beta}_i)), \quad \frac{(n-2)s_n^2}{\sigma^2} \sim \chi^2(n-2) \quad \text{(a nezávisí na } \widehat{\beta}_0, \widehat{\beta}_1).$$

Tedy

$$T_i = \frac{\dfrac{\widehat{\beta}_i - \beta_i}{\sigma(\widehat{\beta}_i)}}{\dfrac{s_n}{\sigma}} = \frac{\widehat{\beta}_i - \beta_i}{\hat{\sigma}(\widehat{\beta}_i)} \sim t(n-2), \quad i = 0, 1.$$

To znamená

$$P\left[-t_{1-\alpha/2}(n-2) \leq \frac{\widehat{\beta}_i - \beta_i}{\hat{\sigma}(\widehat{\beta}_i)} \leq t_{1-\alpha/2}(n-2)\right] = 1-\alpha$$

a vyjádřením β_i dostaneme

$$P\left[\widehat{\beta}_i - t_{1-\alpha/2}(n-2)\,\hat{\sigma}(\widehat{\beta}_i) \leq \beta_i \leq \widehat{\beta}_i + t_{1-\alpha/2}(n-2)\,\hat{\sigma}(\widehat{\beta}_i)\right] = 1-\alpha$$

a tedy $(\widehat{\beta}_i \pm t_{1-\alpha/2}(n-2)\,\hat{\sigma}(\widehat{\beta}_i))$ je $100(1-\alpha)\%$ IS pro $\beta_i, i = 0, 1.$

Dosazením za $\hat{\sigma}(\widehat{\beta}_i)$ dostaneme

$100(1-\alpha)\%$ IS pro β_0: $\left(\widehat{\beta}_0 \pm t_{1-\alpha/2}(n-2) \cdot s_n \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}\right)$

$100(1-\alpha)\%$ IS pro β_1: $\left(\widehat{\beta}_1 \pm t_{1-\alpha/2}(n-2) \cdot s_n \frac{1}{\sqrt{S_{xx}}}\right)$

Poznámka 2.6

- IS pro β_0 bude ve většině praktických případů širší než IS pro β_1
- Někdy se konstruují simultánní IS pro oba parametry. Zmíníme podrobněji u vícerozměrné regrese.

Testy hypotéz o parametrech β_0, β_1

- chtěli bychom ověřit platnost předpokladu lineárního vztahu mezi x a y
- předpokládejme, že model je lineární a že x je jediná dostupná vysvětlující proměnná
- otázkou je, zda je x užitečná ve vysvětlení variability v y
- chceme tedy rozhodnout mezi dvěma modely:

$$Y_i = \beta_0 + e_i \qquad \text{a} \qquad Y_i = \beta_0 + \beta_1 x_i + e_i$$

  tzn. otestovat hypotézu $H_0: \beta_1 = 0 \quad \times \quad H_1: \beta_1 \neq 0$.

- $$\begin{array}{lcl} \text{nezamítnutí } H_0 & \rightarrow & x \text{ nevysvětluje nic z variability } y \text{ a není v modelu významné} \\ \text{zamítnutí } H_0 & \rightarrow & x \text{ je v modelu významné.} \end{array}$$

Poznámka 2.7

Tyto závěry jsou správné pouze za předpokladu, že je model lineární!

- nezamítnutí H_0 nemusí znamenat, že x není užitečná (vztah mezi y a x nemusí být lineární)
- zamítnutí H_0 naopak říká, že existuje lineární trend mezi x a y (mohou tam ale být i jiné typy závislosti)

Pro konstrukci testů využijeme odvozené IS.

Opakování: mějme testovat $H_0 : \theta = \theta_0 \quad \times \quad H_1 : \theta \neq \theta_0$ a nechť $(\underline{\theta}, \overline{\theta})$ je $100(1-\alpha)\%$ IS pro θ

Pak $W = \{x | \theta_0 \notin (\underline{\theta}, \overline{\theta})\}$ je kritický obor testu na hladině α.

$H_0 : \beta_1 = 0$ zamítneme, pokud $0 \notin \left(\widehat{\beta}_1 \pm t_{1-\alpha/2}(n-2) \cdot \frac{s_n}{\sqrt{S_{xx}}}\right)$, tzn.

$$\text{buď} \qquad \widehat{\beta}_1 + t_{1-\alpha/2}(n-2) \cdot \frac{s_n}{\sqrt{S_{xx}}} < 0 \quad \Longleftrightarrow \quad \widehat{\beta}_1 \frac{\sqrt{S_{xx}}}{s_n} < -t_{1-\alpha/2}(n-2)$$

$$\text{nebo} \qquad \widehat{\beta}_1 - t_{1-\alpha/2}(n-2) \cdot \frac{s_n}{\sqrt{S_{xx}}} > 0 \quad \Longleftrightarrow \quad \widehat{\beta}_1 \frac{\sqrt{S_{xx}}}{s_n} > t_{1-\alpha/2}(n-2)$$

zapsáno dohromady

$$|T_1| = |\widehat{\beta}_1| \frac{\sqrt{S_{xx}}}{s_n} > t_{1-\alpha/2}(n-2).$$

Poznámka 2.8 (Intuitivní interpretace)

$|T_1| = |\widehat{\beta}_1| \frac{\sqrt{S_{xx}}}{s_n} = \frac{|\widehat{\beta}_1|}{\widehat{\sigma}(\widehat{\beta}_1)}$ je převrácená hodnota relativní chyby $\frac{\widehat{\sigma}(\widehat{\beta}_1)}{|\widehat{\beta}_1|}$.

Poznámka 2.9

Někdy dopředu známe kandidáta b_1 jako hodnotu parametru β_1 a chtěli bychom testovat $H_0 : \beta_1 = b_1 \quad \times \quad H_1 : \beta_1 \neq b_1$. Test bude: zamítnout H_0, pokud

$$\left|\widehat{\beta}_1 - b_1\right| \cdot \frac{\sqrt{S_{xx}}}{s_n} > t_{1-\alpha/2}(n-2).$$

Test významnosti interceptu:

Otázka je, zda přímka prochází počátkem $(0, 0)$, tedy $H_0 : \beta_0 = 0 \quad \times \quad H_1 : \beta_0 \neq 0$.

Nezamítnutí H_0 znamená, že jednodušší model $y = \beta_1 x + e$ lépe popisuje data, než $y = \beta_0 + \beta_1 x + e$.

H_0 zamítneme, pokud

$$|T_0| = \frac{|\widehat{\beta}_0|}{\widehat{\sigma}(\widehat{\beta}_0)} = |\widehat{\beta}_0| \frac{1}{s_n\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}} > t_{1-\alpha/2}(n-2).$$

Příklad 2.1 (Měření rychlosti zvuku v závislosti na teplotě)

teplota	-20	0	20	50	100
rychlost (m/s)	323	327	340	364	386

$$\overline{x} = 30, \quad \overline{y} = 348, \quad \sum_{i=1}^{n} x_i y_i = 57140, \quad \sum_{i=1}^{n} x_i^2 = 13300,$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = 8800, \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - 5\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - 5\overline{x}^2} = 0.561, \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = 331.16,$$

$$s_n^2 = \frac{1}{5-2} \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = 18.95, \quad s_n = 4.35, \quad \hat{\sigma}(\widehat{\beta}_0) = 2.394, \quad \hat{\sigma}(\widehat{\beta}_1) = 0.046$$

$$t_{0.975}(5-2) = 3.18 \Rightarrow \text{ 95\% IS pro } \beta_0 : (323.5, 338.8), \quad \text{95\% IS pro } \beta_1 : (0.414, 0.709)$$

Testy hypotéz:

$H_0 : \beta_0 = 0 \quad |T_0| = \frac{331.16}{2.394} = 138.3 \ > \ 3.18 \quad \Rightarrow \quad$ zamítáme H_0

$H_0 : \beta_1 = 0 \quad |T_1| = \frac{0.561}{0.046} = 12.1 \ > \ 3.18 \quad \Rightarrow \quad$ zamítáme H_0

```r
mod <- lm(rychlost~teplota)

summary(mod)

## Call:
## lm(formula = rychlost ~ teplota)
##
## Residuals:
##      1      2      3      4      5
## 3.068 -4.159 -2.386  4.773 -1.295
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 331.15909    2.39363   138.3 8.33e-07 ***
## teplota       0.56136    0.04641    12.1  0.00122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.354 on 3 degrees of freedom
## Multiple R-squared:  0.9799,     Adjusted R-squared:  0.9732
## F-statistic: 146.3 on 1 and 3 DF,  p-value: 0.001216

confint(mod)

##                  2.5 %      97.5 %
## (Intercept) 323.541507 338.7766749
## teplota       0.413665   0.7090623
```