Seminář Matematické metody v praxi

LS 2021/2022

Volitelný zápočtový úkol: Lineárna regresia

(Submission Deadline: 30.06.2022 | Koniec skúškového obdobia)

ℹ Všeobecné informácie

- ❏ Vypracovanie aspoň jedného úkolu (výber konkrétneho úkolu z niektorého tématického okruhu je na princípe individuálnej voľby) je v prípade nekompletnej účasti na seminároch nutnou podmienkou k získaniu zápočtu.
- ❏ V prípade úkolu z *lineárni regrese* je potrebné tento úkol vypracovať v programe R. Program R (dostupný pod GNU GPL licenciou) je k dispozícii (free of charge) na adrese `https://www.r-project.org`. K dispozícii sú distribúcie pre Windows, Unix, aj Macintosh.
- ❏ Po inštalácii a spustení programu R sa zobrazí príkazový riadok—konzola. V závislosti na zvolenom datovom súbore (viď podrobnosti nižšie) môže byť potrebné doinštalovať a zprístupniť príslušnú knižnicu (napr. `carData`, alebo `MASS`). Ak je počítač pripojený na internet, postačí do konzoly zadať postupne príkazy

```
> install.packages("carData")
> library(carData)
```

kde v argumente oboch príkazov je názov príslušnej knižnice—v tomto prípade `carData`.

- ❏ Konkrétny datový súbor sa zobrazí po zadaní príslušného názvu suboru do konzoly. Každý datový súbor (viď nižšie) má určitý počet pozorovaní (n riadkov) a niekoľko sledovaných premenných (p stĺpcov). Napríklad:

```
> Angell
                moral hetero mobility region
Rochester        19.0   20.6     15.0      E
Syracuse         17.0   15.6     20.2      E
---
---
```

- ❏ Pre vypracovanie úkolu z *lineárnej regresie* je k dispozícii niekoľko rôznych datových súborov (zoznam v tabuľke nižšie) z ktorých si študent/študentka vyberie jeden a na tomto súbore vypracuje zadanie.

#	Datový súbor	Potrebná knižnica	n/p	Premenná Y	Premenná X
1	`Angell`	`carData`	43/4	`moral`	`hetero`
2	`Anscombe`	`carData`	51/4	`income`	`education`
3	`Bforx`	`carData`	30/6	`partic`	`womwage`
4	`Ericksen`	`carData`	66/9	`rate`	`minority`
5	`Highway1`	`carData`	39/12	`rate`	`trks`
6	`Soils`	`carData`	48/14	`pH`	`Ca`
7	`mtcars`	–	32/11	`mpg`	`disp`
8	`trees`	–	31/3	`Volume`	`Girth`
9	`road`	`MASS`	26/6	`deaths`	`drivers`
10	`hills`	`MASS`	35/3	`time`	`climb`

- ❏ Stručný popis ku každému datovému súboru a jednotlivým premenným je možné získať pomocou helpu (t.j., do konzoly zadať otazník a meno príslušného datového súboru)—napríklad pre help k datovému súboru `Angell`:

```
> ?Angell
```

✍ Zadanie úkolu

1. Uvažujte jednoduchý lineárny regresný model (regresnú přímku) pre závislu premennú Y a nezávislú premennú X, t.j., použije model

$$Y_i \approx a + bX_i, \tag{1}$$

kde $i = 1, \ldots, n$ su jednotlivé pozorovania a $a, b \in \mathbb{R}$ sú neznáme parametre (absolútny člen a a smernica regresnej priamky b), ktoré odhadnete metódou najmenších štvorcov.

2. Pomocou programu R nájdite odhady neznamých parametrov $a, b \in \mathbb{R}$, ktoré získate riešením systému lineárnych rovníc, definovaných ako

$$\mathbb{X}^\top \mathbb{X} \begin{pmatrix} a \\ b \end{pmatrix} = \mathbb{X}\boldsymbol{Y}, \tag{2}$$

kde $\mathbb{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \in \mathbb{R}^{n\times 2}$ je tzv. dizajnová matica lineárneho regresného modelu (1) a $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ je vektor s hodnotami závislej premennej. V programe R vytvoríme maticu $\mathbb{X}$ a vektor Y (pre prípad prvého datového súboru `Angel`) pomocou prikazov:

```
> X <- cbind(rep(1, length(Angell$hetero)), Angell$hetero)
> Y <- Angell$moral
```

Transponovaná matica sa v programe získa pomocou príkazu `t()` a maticove násobenie získame pomocou trojkombinácie symbolov `%*%`, napr. pravá strana rovnice (2) je:

```
> X %*% Y
```

Riešenie systému lineárných rovníc získame v programe R pomocou príkazu `solve()`. Help a jednoduchý návod k ľubovolnej funkcii v programe R získame zadaním otazníku a názvu funkcie do konzoly, napr. `?solve`.

3. Porovnajte odhady, ktoré ste dostali riešením systému lineárných rovníc s výstupom funkcie (čiernej skrinky pre lineárny regresný model) `lm()`, ktorá je v programe R explicitne určená k odhadovaniu parametrov v lineárnom regresnom modeli:

```
> summary(model <- lm(moral ~ hetero, data = Angell))
```

4. Pomocou funkcie `plot()` a funkcie `abline()` vykreslite model a odhadnutú regresnú přímku:

```
> plot(moral ~ hetero, data = Angell)
> abline(model, col = "red")
```

5. Pokúste sa vlastnými slovami interpretovať/vysvetliť regresnú přímku (asociatívnu závislosť, resp. vzťah medzi závislou premennou Y a nezávislou premennou X), ktorú ste dostali.