# ŠTATISTICKÉ SKÚMANIE ZÁVISLOSTI V JAZYKU R

*Michal Páleš*

## 1 ÚVOD

Cieľom príspevku je podať základnú informácie o štatistickom skúmaní v závislosti (regresná a korelačná analýza) v jazyku R v kontexte publikácie (Šoltés a kol., 2015, **príklad 5.1**, s. 148). V postupnom slede uvádzame zodpovedajúce riešenie vybraných úloh predmetného príkladu v jazyku R (v komparácii s publikáciou, ktorá obsahuje riešenia v softvéroch *SAS Enterprise Guide* a *Statgraphics Centurion*). Z riešených úloh boli vyňaté tie úlohy, ktoré obsahujú zhodu so zadanou konštantou a komparácie ďalších modelov. Využívame štandardné rozhranie (funkciu `lm`) jazyka R bez inštalácii doplňujúcich balíčkov (*packages*, knižníc), pričom uvádzame jednotlivé príkazy aj výstupy a sa zameriavame najmä na technickú stránku výpočtu. Jednotlivé mierne odchýlky vo výpočtoch sú spôsobené zabudovaným zaokrúhľovaním príslušných softvérov. Tieto poznatky uvádzame v kontexte s využitím jazyka R a zodpovedajúcich štatistických analýz v aktuárskej praxi (Páleš, 2017) najmä ako pomôcku pre študentov k predmetu *Softvérové aplikácie pre aktuárov*.

## 2 VSTUPNÉ ÚDAJE

Vektorovo definujeme vysvetľovanú (závislú) a vysvetľujúcu (nezávislú) premennú

```
x<-c(196,182,207,169,136,151,113,89,158,98,74)
y<-c(4.4,3.9,3.5,5.6,6.0,6.7,7.6,8.1,8.3,9.4,10.2)
```

Prípadne potreby môžeme tieto načítať zo súboru príkazom

```
read.csv2("D:/Udaje.csv")
```

## 3 FORMULÁCIA ÚLOH

***a)*** ***Odhadnime regresnú priamku*** *(resp. lineárny regresný model)* ***charakterizujúcu závislosť vysvetľovanej a vysvetľujúcej premennej.***
***b)*** ***Na hladine významnosti 0,05 overme štatistickú významnosť regresného modelu.***
***c)*** ***Na hladine významnosti 0,05 overme štatistickú významnosť regresného koeficientu.***
***d)*** ***So spoľahlivosťou 0,95 odhadnime priemernú zmenu vysvetľovanej premennej spôsobenú nárastom vysvetľujúcej premennej o jednotku.***
***e)*** **–**
***f)*** ***Tesnosť skúmanej závislosti kvantifikujme korelačnými charakteristikami.***
***g)*** ***Na hladine významnosti 0,05 overme štatistickú významnosť koeficienta korelácie.***
***h)*** **–**
***i)*** ***So spoľahlivosťou 0,95 odhadnime intenzitu lineárnej závislosti medzi vysvetľovanou a vysvetľujúcou premennou.***
***j)*** **–**

## 4 RIEŠENIE

### a)

Lineárny regresný model odhadneme pomocou príkazov

```
model <-lm(y~x) # pre ďalšie informácie pozri ?lm
summary(model)
```

a dostávame výstup

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0463 -0.7319  0.0691  0.3785  2.2745

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.130077   1.078288  12.177  6.8e-07 ***
x           -0.044966   0.007225  -6.223 0.000155 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.023 on 9 degrees of freedom
Multiple R-squared:  0.8114,    Adjusted R-squared:  0.7905
F-statistic: 38.73 on 1 and 9 DF,  p-value: 0.0001545
```

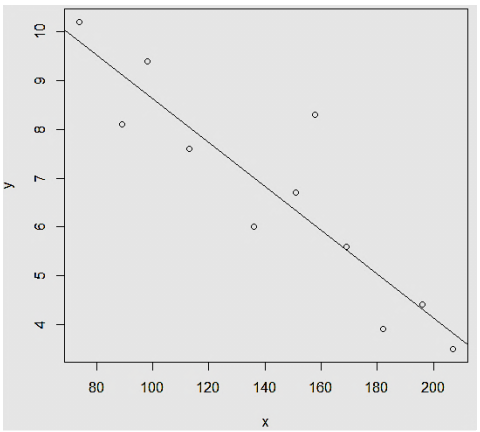
odkiaľ

$$\hat{y}_i = 13{,}130 - 0{,}045 x_i$$

a slovne interpretujeme v závislosti od charakteru jednotlivých premenných.

Pre grafický výstup nižšie volíme príkazy (tento možno vizuálne upraviť pomocou syntaxe funkcie `plot`)

```
plot(x,y)
abline(model)
```



V prípade manuálnych výpočtov v jazyku R je výhodné využiť tieto príkazy

| | | |
|---|---|---|
| $\overline{x}$ | `mean(x)` | 143 |
| $\overline{xy}$ | `mean(x*y)` | 876,1727 |
| $\overline{x^2}$ | `mean(x*x)` | 22271 |
| $\operatorname{cov} xy$, kde $\operatorname{cov} xy = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \overline{x})(y_i - \overline{y})\right]$ | `cov(x,y)`, kde $\operatorname{cov} xy = \frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i - \overline{x})(y_i - \overline{y})\right]$ | – 90,12 [–81.93] |

(* alternatívou je voliť aj funkciu v tvare `cov(x,y,method = c("kendall"))`

**Poznámka.**

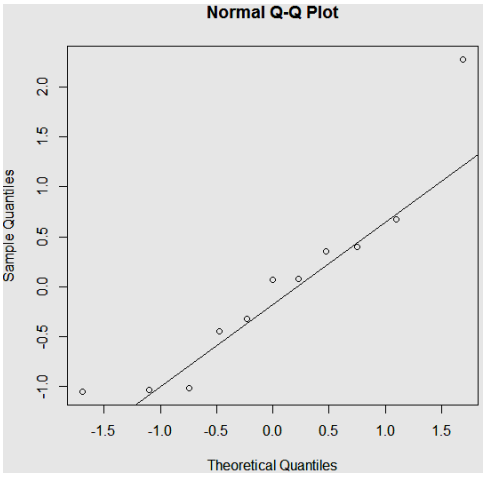
V prípade potreby analyzovať rezíduá tieto získame pomocou príkazu

```
rez<-residuals(model)
```

a overiť predpoklad, že $\varepsilon_i \sim N(0;\sigma^2)$ môžeme napr. pomocou *qq grafu* takto

```
qqnorm(rez)
qqline(rez)
```

s výstupom




**b)**

Štatistickú významnosť regresného modelu zistíme podľa výstupu z úlohy a)

$$p\text{-value} = 0{,}0001545 \;<\; 0{,}05 = \alpha \;\Rightarrow H_1$$

*Regresný model je štatisticky významný.*

**c)**

Štatistickú významnosť regresného koeficienta zistíme podľa výstupu z úlohy a)

$$p\text{-value} = 0{,}000155 \ < \ 0{,}05 = \alpha \ \Rightarrow H_1$$

*Regresný koeficient je štatisticky významný.*

**d)**

Pre intervalový odhad regresného koeficienta volíme

```
confint(model,x,level=0.95)
```

a dostávame výstup

```
x           -0.06131068 -0.02862046
```

a platí

$$P(-0{,}0613 < \beta_1 < -0{,}0286) = 0{,}95$$

**f)**

Korelačné charakteristiky (*výberový Pearsonov koeficient korelácie, koeficient determinácie*) získame pomocou príkazov

| | | |
|---|---|---|
| $r_{xy}$ | `cor(x,y)` | $-0{,}9007959$ |
| $r_{xy}^2$ | `cor(x,y)^2` | $0{,}8114332$ |

**g)**

Štatistickú významnosť koeficienta korelácie získame pomocou príkazu

```
cor.test(x,y)
```

a dostávame výstup

```
Pearson's product-moment correlation

data:  x and y
t = -6.2232, df = 9, p-value = 0.0001545
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9742307 -0.6546952
sample estimates:
       cor
-0.9007959
```

Vidíme, že test štatistickej významnosti koeficienta korelácie je ekvivalentný (v jednoduchej lineárnej regresii) s testom štatistickej významnosti regresného modelu, teda

$$p\text{-value} = 0{,}0001545 \; < \; 0{,}05 = \alpha \; \Rightarrow H_1$$

*Koeficient korelácie je štatisticky významný.*

**i)**

Pre intenzitu lineárnej závislosti medzi vysvetľovanou a vysvetľujúcou premennou využijeme výstup z úlohy g); t. j. (*95 percent confidence inteval*)

$$P(-0{,}9742 < \rho_{xy} < -0{,}6547) = 0{,}95$$

**Poznámka.**

Pre viacnásobnú regresiu treba len upraviť syntax funkcie `lm`, napr. `multiple<-read.csv2("D:/multiple.csv"); lm(y~x1+x2+x3,data=multiple)`.

## Zdroje


[1] PÁLEŠ, M.: Grafická podpora jazyka R pri štatistických analýzach. In *Slovenská štatistika a demografia 1/2016*. Bratislava : Štatistický úrad Slovenskej republiky, 2016.

[2] PÁLEŠ, M.: *Jazyk R v aktuárskych analýzach*. Bratislava : Vydavateľstvo EKONÓM, 2017.

[3] PÁLEŠ, M.: Využitie lineárnych regresných modelov v neživotnom poistení s podporou jazyka R. In *Softvérová podpora v predmetoch študijného programu Aktuárstvo : vedecká konferencia KMA FHI EU v Bratislave (29. júna - 1. júla 2016 Vzdelávacie zariadenie EU, Virt)*. Bratislava : Vydavateľstvo EKONÓM, 2016.

[4] PACÁKOVÁ, V. a kol.: *Štatistické metódy pre ekonómov*. Bratislava: Iura Edition, 2009.

[5] ŠOLTÉS, E.: *Regresná a korelačná analýza s aplikáciami*. Bratislava : Iura Edition, 2008.

[6] ŠOLTÉS, E. a kol.: *Štatistické metódy pre ekonómov (zbierka príkladov)*. Bratislava: Wolters Kluwer, 2015.

[7] R CORE TEAM: R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014 <http://www.R-project.org/>


## Kontaktné údaje


Páleš, Michal, Ing., PhD., Katedra matematiky a aktuárstva, Fakulta hospodárskej informatiky, Ekonomická univerzita v Bratislave, Dolnozemská cesta 1, 852 35 Bratislava, tel. +421 2/672 95 841, e-mail: pales.euba@gmail.com