

Aplikace jazyka R v biomedicině — Úvod do jazyka R

Lubomír Štěpánek^{1, 2}



¹Oddělení biomedicínské statistiky & výpočetní techniky
Ústav biofyziky a informatiky
1. lékařská fakulta
Univerzita Karlova v Praze



²Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze

(2017) Lubomír Štěpánek, CC BY-NC-ND 3.0 (CZ)



Dílo lze dále svobodně šířit, ovšem s uvedením původního autora a s uvedením původní licence. Dílo není možné šířit komerčně ani s ním jakkoliv jinak nakládat pro účely komerčního zisku. Dílo nesmí být jakkoliv upravováno. Autor neručí za správnost informací uvedených kdekoli v předložené práci, přesto vynaložil nezanedbatelné úsilí, aby byla uvedená fakta správná a aktuální, a práci sepsal podle svého nejlepšího vědomí a svých „nejlepších“ znalostí problematiky.

Obsah

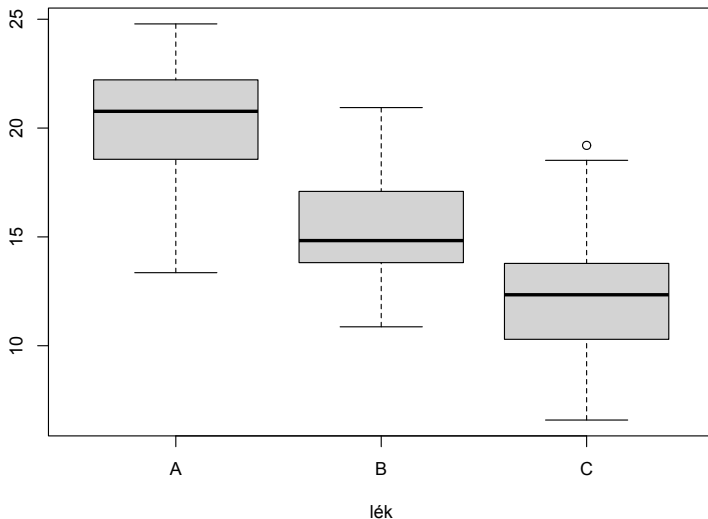
- 1 Lineární model
- 2 Lineární regrese
- 3 Literatura

Jednofaktorová ANOVA

- testuje nulovou hypotézu $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ o statisticky nevýznamném rozdílu ve středních hodnotách k výběrů
- **předpokládá** normalitu výběrů
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α

```
1  set.seed(1)
2  lek_A <- rnorm(30, 20, 3)
3  lek_B <- rnorm(30, 15, 3)
4  lek_C <- rnorm(30, 12, 3)
5
6  my_data <- data.frame(
7    "mira" = c(lek_A, lek_B, lek_C),
8    "lek"  = c(rep("A", 30), rep("B", 30),
9              rep("C", 30))
10 )
```

Jednofaktorová ANOVA



Jednofaktorová ANOVA

- diagnostika

```
1 | shapiro.test(  
2 |   my_data[my_data$lek == "A", "mira"]  
3 | )  
4 | shapiro.test(  
5 |   my_data[my_data$lek == "B", "mira"]  
6 | )  
7 | shapiro.test(  
8 |   my_data[my_data$lek == "C", "mira"]  
9 | )
```

Jednofaktorová ANOVA

- výsledek testu

```
1 | summary(aov(mira ~ lek, my_data))
2 |
3 |           Df Sum Sq Mean Sq F value Pr(>F)
4 | lek         2   955.9    478.0   66.14 <2e-16 ***
5 | Residuals   87   628.7       7.2
```

- lépe formátovaný výstup pro T_EX

```
1 | xtable(summary(aov(mira ~ lek, my_data)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lek	2	955.94	477.97	66.14	< 0.0001
Residuals	87	628.70	7.23		

Vícefaktorová ANOVA

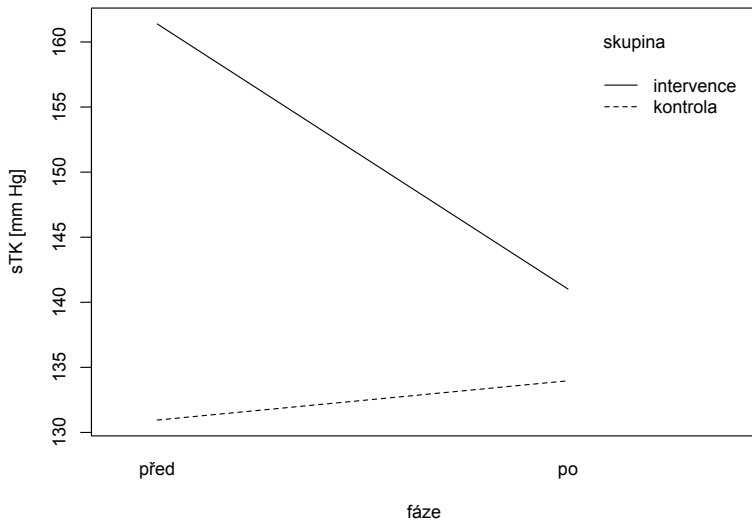
- testuje několik nulových hypotéz typu $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ o statisticky nevýznamném rozdílu ve středních hodnotách mezi k hodnotami některého faktoru
- **předpokládá** normalitu všech (pod)výběrů
- hodnota $p < \alpha$ vede k zamítnutí nulové hypotézy H_0 na hladině významnosti α
- v R opět funkce `aov()`

Vícefaktorová ANOVA

- vytvořme si vhodný dataset

```
1 | set.seed(1); my_data <- data.frame(  
2 |   "sTK" = c(rnorm(20, 130, 5), rnorm(20, 133, 5),  
3 |             rnorm(20, 160, 10), rnorm(20, 140,  
4 |               10)),  
5 |   "faze" = c(rep("před", 20), rep("po", 20),  
6 |             rep("před", 20), rep("po", 20)),  
7 |   "skupina" = c(rep("kontrola", 40),  
8 |                rep("intervence", 40))  
9 | )  
10 | my_data$faze <- factor(  
11 |   my_data$faze, levels = c("před", "po")  
12 | )  
13 | my_data$skupina <- factor(  
14 |   my_data$skupina,  
15 |   levels = c("kontrola", "intervence")  
16 | )
```

Vícefaktorová ANOVA



Vícefaktorová ANOVA

- výsledek testu

```
1 summary(aov(sTK ~ faze + skupina, my_data))
2
3           Df Sum Sq Mean Sq F value    Pr(>F)
4 faze         1   1506     1506   16.97 9.47e-05 *
5 skupina       1   7026     7026   79.17 1.89e-13 *
6 Residuals    77   6833         89
```

- lépe formátovaný výstup pro T_EX

```
1 xtable(summary(aov(sTK ~ faze + skupina, my_data)
2           ))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
faze	1	1506.08	1506.08	16.97	0.0001
skupina	1	7025.65	7025.65	79.17	< 0.0001
Residuals	77	6832.81	88.74		

Vícefaktorová ANOVA

- zaved' me interakci

```
1 summary(aov(sTK ~ faze * skupina, my_data))
2           Df Sum Sq Mean Sq F value    Pr(>F)
3 faze         1   1506      1506    27.93 1.17e-06 *
4 skupina       1   7026      7026   130.28 < 2e-16 *
5 faze:skupina  1   2734      2734    50.71 5.19e-10 *
6 Residuals    76   4098         54
```

- lépe formátovaný výstup pro T_EX

```
1 xtable(summary(aov(sTK ~ faze*skupina, my_data)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
faze	1	1506.08	1506.08	27.93	< 0.0001
skupina	1	7025.65	7025.65	130.28	< 0.0001
faze:skupina	1	2734.44	2734.44	50.71	< 0.0001
Residuals	76	4098.38	53.93		

Vícefaktorová ANOVA

- velikosti efektů

```
1 | summary(lm(sTK ~ faze + skupina, my_data))
2 |           Estimate Std. Err   t val Pr(>|t|)
3 | (Intercept)    136.799    1.824  74.992   < 2e-16 *
4 | faze = po       -8.678    2.106  -4.120  9.47e-05 *
5 | skupina = int    18.743    2.106   8.898  1.89e-13 *
```

- lépe formátovaný výstup pro T_EX

```
1 | xtable(summary(lm(sTK ~ faze + skupina, my_data)))
  | )
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	136.7990	1.8242	74.99	< 0.0001
faze = po	-8.6778	2.1064	-4.12	0.0001
skupina = intervence	18.7425	2.1064	8.90	< 0.0001

Vícefaktorová ANOVA

- velikosti efektů u interakcí

```
1 | summary(lm(sTK ~ faze * skupina, my_data))
2 |           Estimate Std Err t val Pr(>|t|)
3 | (Intercept      130.953   1.642 79.750 < 2e-16
4 | faze = po        3.015   2.322  1.298  0.198
5 | skupina = int    30.435   2.322 13.106 < 2e-16
6 | fazepo:skupinaint -23.386   3.284 -7.121 5.2se-10
```

- lépe formátovaný výstup pro T_EX

```
1 | xtable(summary(lm(sTK ~ faze*skupina, my_data)))
```

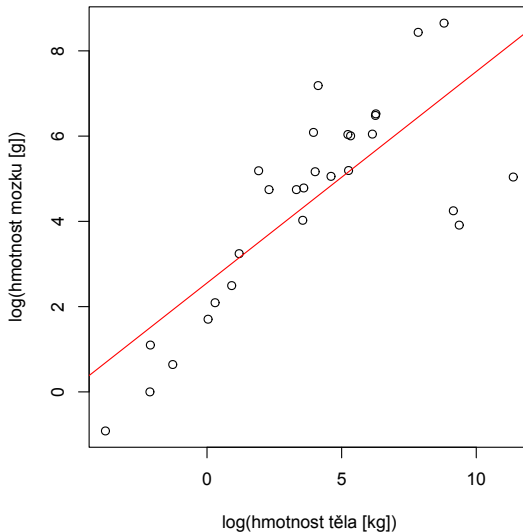
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	130.9526	1.6420	79.75	< 0.0001
faze = po	3.0150	2.3222	1.30	0.1981
skupina = int.	30.4353	2.3222	13.11	< 0.0001
faze = po : skup. = int.	-23.3856	3.2841	-7.12	< 0.0001

Lineární regrese

- obdobná syntaxe lineárnímu modelu
- v R pomocí funkce `lm()`
- odhad parametrů metodou nejmenších čtverců, proto očekáváno splnění tzv. slabé sady předpokladů
 - v R diagnostika pomocí `plot(lm())`

```
1 library(MASS)
2 data(Animals)
3
4 summary(lm(log(brain) ~ log(body), Animals))
5
6           Estimate Std. Err t value Pr(>|t|)
7 (Intercept)  2.55490   0.41314   6.184 1.53e-06 *
8 log(body)    0.49599   0.07817   6.345 1.02e-06 *
```

Lineární regrese



Lineární regrese

- lépe formátovaný výstup pro T_EX

```
1 | xtable(  
2 |   summary(lm(log(brain) ~ log(body), Animals))  
3 | )
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.5549	0.4131	6.18	0.0000
log(body)	0.4960	0.0782	6.35	0.0000

Lineární regrese

- diagnostika modelu

- rezidua vs. vyrovnané hodnoty
- Q-Q diagram
- odmocniny z reziduí vs. vyrovnané hodnoty
- rezidua vs. pákové body (Cookova distance)

```
1 || plot(lm(log(brain) ~ log(body), Animals))
```

Literatura



ZVÁRA, Karel. *Základy statistiky v prostředí R*.
Praha, Česká republika: Karolinum, 2013. ISBN 978-80-246-2245-3.



WICKHAM, Hadley. *Advanced R*.
Boca Raton, FL: CRC Press, 2015. ISBN 978-1466586963.

Děkuji za pozornost!

lubomir.stepanek@lf1.cuni.cz

lubomir.stepanek@fbmi.cvut.cz