Regresní analýza dat
01REAN - Cvičení 02

Jiri Franc

Czech Technical University
Faculty of Nuclear Sciences and Physical Engineering
Department of Mathematics

Cvičení 02

Osnova dnešního cvičení:

- If, Loops, Funkce, Vektory.
- Pravděpodobnostní rozdělení.
- Data frame.
- Základní grafika.

Simulace hazeni minci (T = tail, H = head)

```
> x <- sample(c(0,1), 100, rep=T)
> table(x)
x
0  1
55 45
> head(ifelse(x==0, "T", "H"))
[1] "T" "H" "H" "T" "T" "H"
> sum(as.logical(x))
[1] 45
```

For cyklus

```
> n <- 500
> v <- w <- z <- numeric(n)
> for (i in 1:n) {
+       t <- runif(1)
+       u <- runif(1)
+       v[i] <- 2 * t - 1
+       w[i] <- 2 * u - 1
+       a <- v[i]^2 + w[i]^2
+       z[i] <- ifelse(a <=1 , a, NA)
+       }
>  round(summary(z), 2)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.00    0.26    0.50    0.50    0.75    1.00     111
> is.na(z)
```

Simulace hazeni minci 2

```
> toss <- function(n=5, p=0.5) sum(rbinom(n, 1, p))
> toss()
[1] 3
> toss(20)
[1] 11
> toss(20, 0.75)
[1] 16
> toss(p=0.75)
[1] 4
> replicate(10, toss(n=100))
[1] 47 53 41 53 59 49 49 53 49 52
> sapply(10:20, toss)
[1]  4  6  6  4  8  6  7  6  8 10 12
> as.numeric(Map("toss", n=10:20))
[1]  7  3  7  3  7  9  4 11 11 10  9
```

Vlastni zadáni dat

```
> pet    = rep(c("cat","dog","fish","other","none"),times=8)
> house  = rep(c("bungalow","villa", "row_house", "apartment"),each=10
> members= rbinom(40,6,0.3)+1
> income = rexp(40,1/19)+11
> area   = abs(rnorm(40,100,90))+20
> dat   <- data.frame(pet,house,members,income,area)
> dim(dat)
[1] 40  5
> summary(dat)
pet           house       members          income            area
cat   :8 apartment:10 Min.   :1.000  Min.   :11.81  Min.   : 35.13
dog   :8 bungalow :10 1st Qu.:2.000  1st Qu.:17.65  1st Qu.: 86.26
fish  :8 row_house:10 Median :3.000  Median :24.94  Median :127.07
none  :8 villa    :10 Mean   :2.825  Mean   :32.60  Mean   :137.79
other:8               3rd Qu.:4.000  3rd Qu.:39.34  3rd Qu.:177.23
Max.   :6.000  Max.   :82.25  Max.   :328.62
```

Další kód v R

Úkoly: Vezměte data `tree` z minulé hodiny.

- Pro vypočtený BMI index stromu přiřaďte faktorovou proměnnou tak, aby 25% stromů s nejnižším indexem bylo označeno jako *tenke*, 50% stromů jako *stredni* a 25% stromu s nejvyšší hodnotou indexu jako *silne*.
- Vykreslete 2 Box ploty do jednoho obrázku nad sebe (nebo vedle sebe), kde zobrazíte rozdělení výšky a obvodu podle spočtených skupin. Použijte k tomu jak klasickou funkci `plot`, popřípadě textttboxplot tak `ggplot`.
- Vykreslete bodově závislosti výšky na obvodu, výšky na objemu, objemu na obvodu. Použijte k tomu jak klasickou funkci `plot` tak `pairs`.
- Pro všechny 3 numerické proměnné vykreslete tzv. qqploty pro normalni rozdělení.
- Pro všechny 3 numerické proměnné vykreslete histogramy a proložte je křivkou hustoty příslušného normálního rozdělení, kde použijete odhad střední hodnoty a rozptylu získaný z dat.

Úkoly

Uvažujte Markovsky proces "Jump and slide".

Nechť X je Markovský řetězec na $\{0, 1, 2, \ldots\}$ s maticí přechodu danou pomocí $p_{0j} = a_j$ pro $j \geq 0$, $p_{ii} = r$ a $p_{i,i-1} = 1 - r$ pro $i \geq 1$.

Například pro matici $n \times n$ **P** vyberte $r = 0.5$ a $a_i = \frac{b_i}{\sum_i b_i}$ kde $b_i = \frac{1}{i^2}$, $i = 1, 2, \ldots, n$.

- Sestrojte matici **P**.
- Spočte stacionární rozdělení.
- Napište funkci, která nasimuluje jednu trajektorii procesu. Jako vstup bude sloužit matice **P**, počáteční stav a počet časových kroků.
- Porovnejte histogram nasimulovaných dat se stacionárním rozdělením.

Nápověda: stacionární rozdělení π je popsáno pomocí rovnice $\pi = \mathbf{P}\pi$, tudíž stačí najít vyřešit tuto rovnici. Použijte k tomu například příkaz `Null` pro nalezení jádra lineárního zobrazení.