# M5201 – CVIČENÍ 2B: *Regresní analýza v R – pokračování*

## 1 Trigonometrická regrese

PŘÍKLAD 1:

Máme k dispozici údaje o průměrných měsíčních teplotách v New Yorku. Zjištěné teploty ve stupních Celsia za období leden 1949 – prosinec 1959 jsou uloženy v souboru `nytemps.dat`.

Soubor obsahuje jediný sloupec se zjištěnými teplotami a není v něm uvedena hlavička. Data načteme obvyklým způsobem pomocí funkce `read.table()`, zobrazíme strukturu vytvořeného datového rámce a vypíšeme začátek datového souboru. Proměnnou v datovém rámci pojmenujeme a zpřístupníme pomocí příkazu `attach()`.

```
> fileDat <- paste(data.library, "nytemps.dat", sep = "")
> data <- read.table(fileDat)
> str(data)
```

```
'data.frame':        168 obs. of  1 variable:
 $ V1: num  11.5 11 14.4 14.4 16.5 ...
```

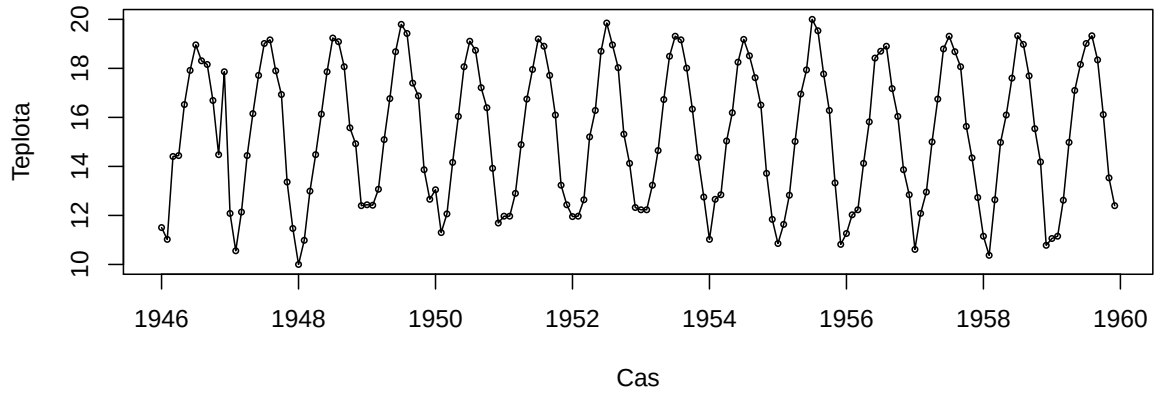
```
> head(data)
```

```
      V1
1 11.506
2 11.022
3 14.405
4 14.442
5 16.524
6 17.918
```

```
> names(data) <- "teploty"
> attach(data)
```

Vývoj teplot v čase znázorníme graficky.

```
> TXT <- "Prumerne mesicni teploty v New Yorku"
> n <- length(teploty)
> Time <- 1:n
> par(mar = c(4, 4, 1, 0) + 0.5)
> plot(1946 + (Time - 1)/12, teploty, type = "o", xlab = "Cas", ylab = "Teplota",
     cex = 0.5)
```




Obrázek 1: Průměrné měsíční teploty v New Yorku v letech 1946–1959.

Z grafu je patrné, že vývoj teplot v čase je periodický. Z povahy věci můžeme předpokládat, že perioda má délku 12 měsíců. Proto jako regresní funkci použijeme takovou funkci, která je periodická s periodou 12. K tomu jsou nejvhodnější funkce sinus a cosinus. Jednotlivá pozorování popíšeme rovnicí

$$Y_i = \beta_0 + \beta_1 \sin\left(\frac{2\pi}{12} x_i\right) + \cos\left(\frac{2\pi}{12} x_i\right) + \varepsilon_i,$$

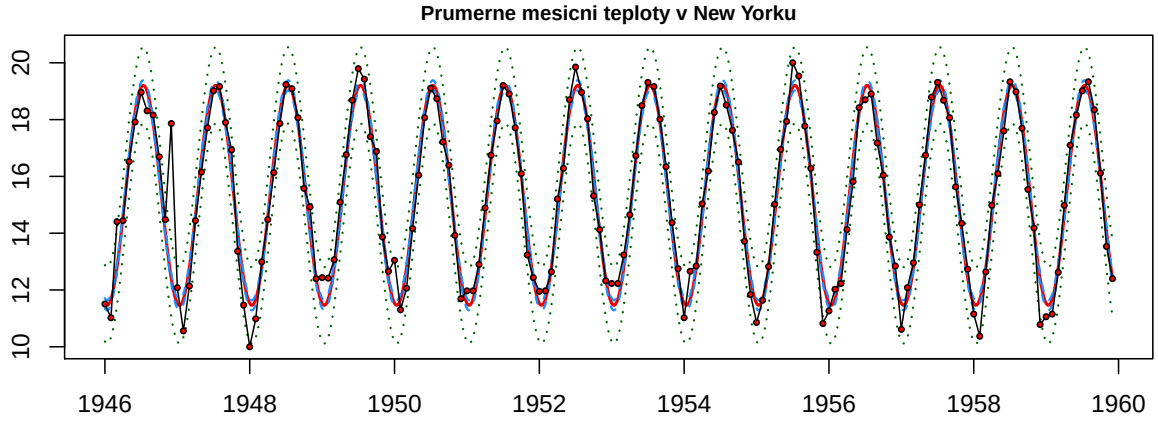
kde $Y_i$ označuje teplotu odpovídající $i$-tému pozorování a $x_i$ je odpovídající čas.

K odhadu parametrů modelu opět použijeme funkci `lm()`.

```
> modelPeriod <- lm(teploty ~ 1 + I(sin(2 * pi/12 * Time)) + I(cos(2 *
    pi/12 * Time)))
> summary(modelPeriod)


Call:
lm(formula = teploty ~ 1 + I(sin(2 * pi/12 * Time)) + I(cos(2 *
    pi/12 * Time)))

Residuals:
    Min      1Q  Median      3Q     Max
-1.6248 -0.4007  0.0111  0.3264  5.4562

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           15.33289    0.05203  294.70   <2e-16 ***
I(sin(2 * pi/12 * Time)) -2.53766    0.07358  -34.49   <2e-16 ***
I(cos(2 * pi/12 * Time)) -2.92704    0.07358  -39.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6744 on 165 degrees of freedom
Multiple R-squared: 0.9438,        Adjusted R-squared: 0.9431
F-statistic:  1386 on 2 and 165 DF,  p-value: < 2.2e-16
```

Nakonec zobrazíme data spolu s odhadnutou regresní křivkou a s intervaly spolehlivosti.

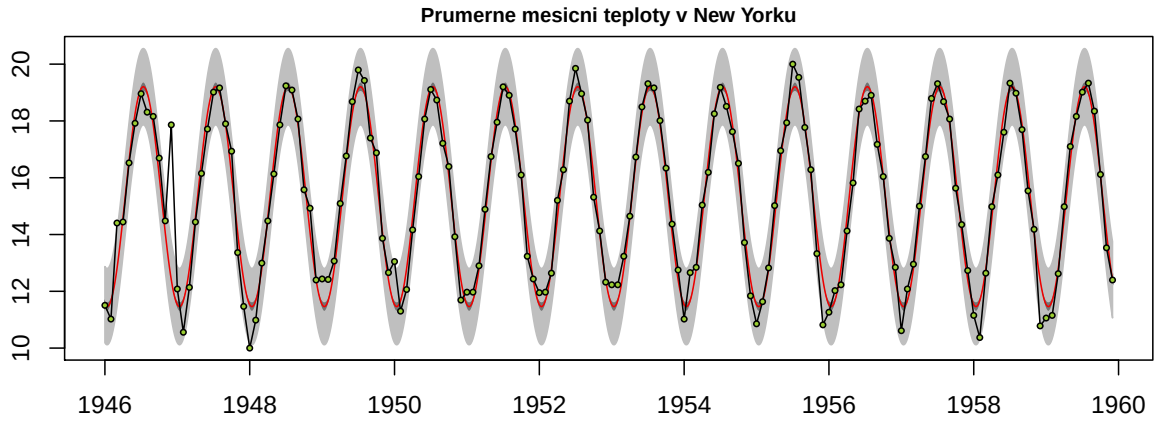
```
> n <- length(data$teploty)
> gridt <- seq(Time[1], Time[n], length.out = 1000)
> Gridt <- 1946 + (gridt - 1)/12
> tt <- 1946 + (Time - 1)/12
> par(mar = c(2, 2, 1, 0) + 0.5)
> ci.conf <- predict(modelPeriod, newdata = data.frame(Time = gridt),
    interval = "confidence")
> ci.pred <- predict(modelPeriod, newdata = data.frame(Time = gridt),
    interval = "prediction")
> yrange <- range(c(data$teploty, ci.pred[, 2], ci.pred[, 3]))
> plot(c(tt[1], tt[n]), yrange, type = "n", main = TXT, cex.main = 0.85)
> matlines(Gridt, ci.conf[, 1], lty = 1, lwd = 2, col = "red2")
> matlines(Gridt, ci.conf[, 2:3], lty = 2, lwd = 1.5, col = "dodgerblue")
> matlines(Gridt, ci.pred[, 2:3], lty = 3, lwd = 1.5, col = "darkgreen")
> points(tt, data$teploty, type = "o", pch = 21, cex = 0.5, bg = "red1")
```




Obrázek 2: *Průměrné měsíční teploty v New Yorku* s trigonometrickou regresní funkcí a s intervaly spolehlivosti.

Ukážeme, že obdobný graf s vyplněnými intervaly spolehlivosti je názornější.

```
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(c(tt[1], tt[n]), yrange, type = "n", main = TXT, cex.main = 0.85)
> xx <- c(Gridt, rev(Gridt))
> yy <- c(ci.conf[, 2], rev(ci.conf[, 3]))
> polygon(xx, yy, col = "gray45", border = "gray45")
> yy <- c(ci.conf[, 2], rev(ci.pred[, 2]))
> polygon(xx, yy, col = "gray75", border = "gray75")
> yy <- c(ci.conf[, 3], rev(ci.pred[, 3]))
> polygon(xx, yy, col = "gray75", border = "gray75")
> matlines(Gridt, ci.conf[, 1], lty = 1, lwd = 1, col = "red2")
> points(tt, data$teploty, type = "o", pch = 21, cex = 0.5, bg = "yellowgreen")
```




Obrázek 3: *Průměrné měsíční teploty v New Yorku* s trigonometrickou regresní funkcí – intervaly spolehlivosti pomocí ploch.

Na závěr příkladu nesníme zapomenout zrušit zpřístupnění jména proměnných v datovém rámci

```
> detach(data)
```

## 2 Analýza rozptylu

PŘÍKLAD 2: JEDNOFAKTOROVÁ ANOVA

Byl proveden experiment, v němž se do pastí odchytával jistý druh můry. Pasti obsahovaly vždy jeden ze tří různých typů návnady a byly umístěny v různých výškách stromů. Data z experimentu jsou uložena v souboru `moths.csv` a informace o datovém souboru jsou k dispozici v souboru `moths_info.txt`.

Nejprve pomocí funkce `readLines()` načteme soubor `moths_info.txt`.

```
> fileTxt <- paste(data.library, "moths_info.txt", sep = "")
> con <- file(fileTxt)
> (popis <- readLines(con))
```

```
[1] "Spruce Moth Trap"
[2] ""
[3] "Response: number of spruce moths found in trap after 48 hours"
[4] "Factor 1: Location of trap in tree (top branches, middle branches, lower branches, ground)"
[5] "Factor 2: Type of lure in trap (scent, sugar, chemical) "
[6] ""
[7] ""
```

```
> close(con)
```

V souboru `moths.csv` jsou data uložená ve třech proměnných, jejichž názvy jsou uvedeny na prvním řádku a jsou odděleny středníkem. Proto použijeme funkci `read.table()` s argumenty `sep=";"` a `header=TRUE`.

```
> fileDat <- paste(data.library, "moths.csv", sep = "")
> data <- read.table(fileDat, header = TRUE, , sep = ";")
> str(data)
```

```
'data.frame':        60 obs. of  3 variables:
 $ Location: Factor w/ 4 levels "Ground","Lower",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ Lure    : Factor w/ 3 levels "Chemical","Scent",..: 2 2 2 2 2 3 3 3 3 3 ...
 $ Number  : int  28 19 32 15 13 35 22 33 21 17 ...
```

Zajímá nás, jaký vliv má umístění pasti na počet chycených můr. Konkrétní úrovně proměnné `Location` typu faktor zjistíme příkazem `levels()`.

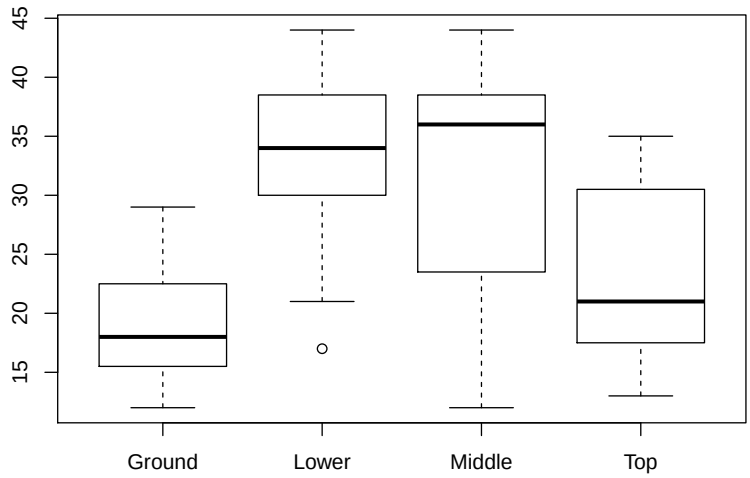
```
> levels(data$Location)
```

```
[1] "Ground" "Lower"  "Middle" "Top"
```

Umístění pasti je v našem případě kategoriální proměnná se čtyřmi úrovněmi *Ground*, *Lower*, *Middle* a *Top*.

Přibližnou představu o rozdílech mezi zjištěnými údaji pro různá umístění pastí můžeme získat z krabicového grafu.

```
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(data$Location, data$Number, ylab = "Number")
```




Obrázek 4: Počet chycených můr v závislosti na umístění pasti.

Než začneme vytvářet lineární regresní model, je třeba vedle grafického vyjádření provést i Bartletův test, který ověřuje předpoklad o shodnosti rozptylu jednotlivých subpopulací.

```
> with(data, bartlett.test(Number ~ Location))
```

```
        Bartlett test of homogeneity of variances

data:  Number by Location
Bartlett's K-squared = 5.498, df = 3, p-value = 0.1388
```

Protože *p–hodnota* testu není menší než 0.05, hypotézu o shodnosti rozptylu **nezamítáme**.

Nyní budeme předpokládat, že střední hodnota počtu odchycených můr se liší v závislosti na tom, v jaké výšce byla past umístěna, což můžeme zapsat

$$\begin{aligned} E(Y_{ground}) &= \mu + \alpha_{ground} \\ E(Y_{lower}) &= \mu + \alpha_{lower} \\ E(Y_{middle}) &= \mu + \alpha_{middle} \\ E(Y_{top}) &= \mu + \alpha_{top} \end{aligned}$$

kde proměnná $Y$ označuje počet odchycených můr a index označuje umístění pasti.

Tomu odpovídá model, v němž jednotlivá pozorování popíšeme rovnicí

$$Y_{jk} = \mu + \alpha_j + \varepsilon_{jk}, \tag{1}$$

kde index $j = 1, 2, 3, 4$ označuje umístění pasti odpovídající danému údaji a index $k$ označuje pořadí pozorování mezi pozorováními se stejným umístěním pasti. Tento model se nazývá *analýza rozptylu jednoduchého třídění*, zkráceně *jednofaktorová ANOVA*.

K odhadu koeficientů $\mu$ a $\alpha_j$ použijeme opět funkci `lm()`.

```
> anova1 <- lm(Number ~ Location, data)
> summary(anova1)


Call:
lm(formula = Number ~ Location, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-19.000  -4.517  -1.200   5.950  13.000

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      19.067      1.970   9.676 1.49e-13 ***
LocationLower    14.267      2.787   5.120 3.90e-06 ***
LocationMiddle   11.933      2.787   4.282 7.32e-05 ***
LocationTop       4.267      2.787   1.531    0.131
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.632 on 56 degrees of freedom
Multiple R-squared: 0.3779,        Adjusted R-squared: 0.3446
F-statistic: 11.34 on 3 and 56 DF,  p-value: 6.459e-06
```

Systém R ale pracuje s jinou parametrizací, než jaká je uvedena v rovnici (1). První skupina pozorování je zvolena jako referenční (zde jsou to pozorování, u nichž má proměnná `Location` úroveň `Ground`) a je zavedena následující omezující podminka

$$\alpha_1 = \alpha_{ground} = 0.$$

Odhadnuté parametry pak mají tento význam:
$\widehat{\mu}$ vyjadřuje střední hodnotu $Y$ pro první (referenční) úroveň kategoriální proměnné,

$\widehat{\alpha}_j$ vyjadřuje rozdíl ve střední hodnotě $Y$ mezi první a $j$-tou úrovní kategoriální proměnné.

Z výpisu informací o modelu `anova1` zjistíme tyto údaje:

$$\begin{aligned}
(\texttt{Intercept}) &= \widehat{E}(Y_{ground}) = 19.067 \\
\texttt{LocationLower} &= \widehat{E}(Y_{lower}) - \widehat{E}(Y_{ground}) = 14.267 \\
\texttt{LocationMiddle} &= \widehat{E}(Y_{middle}) - \widehat{E}(Y_{ground}) = 11.933 \\
\texttt{LocationTop} &= \widehat{E}(Y_{top}) - \widehat{E}(Y_{ground}) = 4.267
\end{aligned}$$

A odtud můžeme dopočítat střední hodnoty pro jednotlivé úrovně proměnné `Location`:

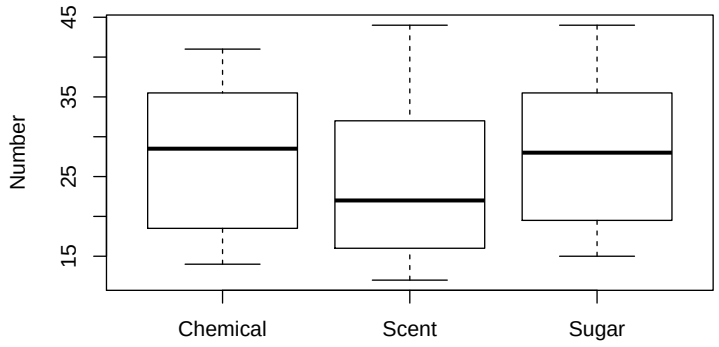
$$\begin{aligned}
\widehat{E}(Y_{ground,k}) &= (\texttt{Intercept}) = 19.067 \\
\widehat{E}(Y_{lower,k}) &= (\texttt{Intercept}) + \texttt{LocationLower} = 19.067 + 14.267 = 33.334 \\
\widehat{E}(Y_{middle,k}) &= (\texttt{Intercept}) + \texttt{LocationMiddle} = 19.067 + 11.933 = 31 \\
\widehat{E}(Y_{top,k}) &= (\texttt{Intercept}) + \texttt{LocationTop} = 19.067 + 4.267 = 23.334
\end{aligned}$$

Příklad 3: Dvoufaktorová ANOVA

Pro stejná data týkající se odchycených můr zkusíme odhadnout model, v němž bychom zohlednili jak vliv umístění pasti, tak vliv zvolené návnady na počet můr, které se podařilo lapit do pasti.

Abychom si udělali představu, jak počet chycených můr ovlivňuje návnada, vykreslíme krabicové graf pro zjištěná dat zvlášť pro jednotlivé typy návnady.

```
> plot(data$Lure, data$Number, ylab = "Number")
```




Obrázek 5: Počet chycených můr v závislosti na typu návnady.

Situaci se pokusíme popsat modelem, do něhož zahrneme vliv obou kategorických proměnných. Nejprve zkusíme odhadnout jednodušší model

$$Y_{jkl} = \mu + \alpha_j + \beta_k + \varepsilon_{jkl}, \qquad \text{kde} \quad \begin{aligned} j &= 1,2,3,4 \\ k &= 1,2,3 \\ l &= 1,\ldots,n_{jk} \end{aligned}$$

kde $\alpha_j$ zachycuje vliv umístění pasti a $\beta_k$ zachycuje vliv $k$-tého typu návnady a pro počty pozorování platí $\sum_{j=1}^{4}\sum_{k=1}^{3} n_{jk} = n = 60$.

```
> anova2 <- lm(Number ~ Lure + Location, data)
> summary(anova2)
```

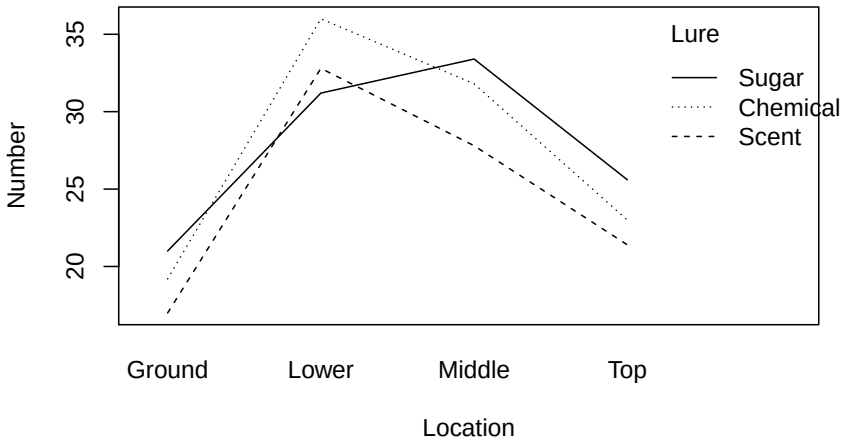
```
Call:
lm(formula = Number ~ Lure + Location, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-17.4500  -5.1208  -0.5167   6.0625  12.9333

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      19.883      2.415   8.234 4.14e-11 ***
LureScent        -2.750      2.415  -1.139    0.260
LureSugar         0.300      2.415   0.124    0.902
LocationLower    14.267      2.788   5.117 4.23e-06 ***
LocationMiddle   11.933      2.788   4.280 7.70e-05 ***
LocationTop       4.267      2.788   1.530    0.132
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.636 on 54 degrees of freedom
Multiple R-squared: 0.3995,        Adjusted R-squared: 0.3439
F-statistic: 7.184 on 5 and 54 DF,  p-value: 3.236e-05
```

Systém R opět zvolil první úrovně faktorů jako referenční. Proto koeficient označený v tabulce jako `(Intercept)` představuje odhadnutou střední hodnotu chycených můr pro $Location = Ground$ a $Lure = Chemical$. Zbylé koeficienty představují odchylky středních hodnot od referenční v případě, že se změní úroveň jedné z proměnných.

V modelu `anova2` jsme mlčky předpokládali, že vliv proměnných `Location` a `Lure` se sčítá, nebrali jsme v úvahu možnost, že by mohlo docházet ke složitějším interakcím mezi nimi.

Jak se mění průměrné hodnoty chycených můr pro různé kombinace úrovní obou vysvětlujících proměnných si můžeme prohlédnout v následujícím grafu.

```
> interaction.plot(data$Location, data$Lure, data$Number, xlab = "Location",
    ylab = "Number", trace.label = "Lure")
```




Obrázek 6: Porovnání průměrného počtu chycených můr v závislosti na umístění pasti a typu návnady.

Nyní odhadneme složitější model, který bude mít tvar

$$Y_{jkl} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{jkl}, \qquad \text{kde} \quad \begin{array}{l} j = 1,2,3,4 \\ k = 1,2,3 \\ l = 1,\ldots,n_{jk} \end{array}$$

kde $\alpha_j$ zachycuje vliv umístění pasti, $\beta_k$ zachycuje vliv $k$-tého typu návnady a $(\alpha\beta)_{jk}$ zachycuje vliv interakce mezi $j$-tou úrovní proměnné `Location` a $k$-tou úrovní proměnné `Lure`.

```
> anova3 <- lm(Number ~ Lure * Location, data)
> summary(anova3)


Call:
lm(formula = Number ~ Lure * Location, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-15.80  -5.00  -0.90   5.85  14.20

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               19.200      3.555   5.400 2.04e-06 ***
LureScent                 -2.200      5.028  -0.438  0.66367
LureSugar                  1.800      5.028   0.358  0.72191
LocationLower             16.800      5.028   3.341  0.00162 **
LocationMiddle            12.600      5.028   2.506  0.01565 *
LocationTop                3.800      5.028   0.756  0.45347
LureScent:LocationLower   -1.000      7.111  -0.141  0.88875
LureSugar:LocationLower   -6.600      7.111  -0.928  0.35795
LureScent:LocationMiddle  -1.800      7.111  -0.253  0.80124
LureSugar:LocationMiddle  -0.200      7.111  -0.028  0.97768
```

```
LureScent:LocationTop       0.600      7.111   0.084  0.93310
LureSugar:LocationTop       0.800      7.111   0.113  0.91089
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.95 on 48 degrees of freedom
Multiple R-squared: 0.4214,        Adjusted R-squared: 0.2888
F-statistic: 3.178 on 11 and 48 DF,  p-value: 0.002653
```

# 3 Analýza kovariance

PŘÍKLAD 6: ANCOVA

Při experimentu byla zkoumána závislost rychlosti růstu populace studenokrevných organismů na teplotě prostředí. V laboratoři byla sledována rychlost populačního růstu (proměnná `rate`) u dvou druhů roztočů (kategoriální proměnná `genus` se dvěma úrovněmi `genA` a `genB`) při různých teplotách (proměnná `temp`).

Data jsou obsahem souboru `mite.txt`, v němž je na prvním řádku uvedena hlavička a hodnoty pro jednotlivé proměnné jsou odděleny tabelátorem. Proto k načtení použijeme funkci `read.table()` s argumentem `sep = "\t"`.

```
> fileDat <- paste(data.library, "mite.txt", sep = "")
> roztoci <- read.table(fileDat, sep = "\t", header = TRUE)
> str(roztoci)
```

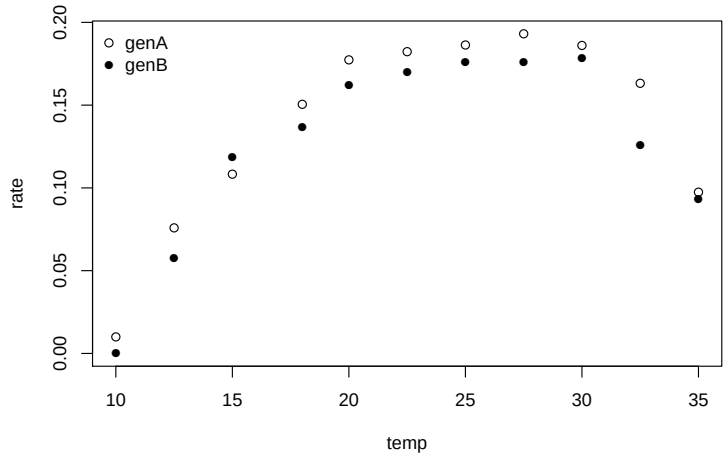
```
'data.frame':        22 obs. of  3 variables:
 $ temp : num  10 12.5 15 18 20 22.5 25 27.5 30 32.5 ...
 $ rate : num  0.01 0.0759 0.1083 0.1505 0.1774 ...
 $ genus: Factor w/ 2 levels "genA","genB": 1 1 1 1 1 1 1 1 1 1 ...
```

```
> attach(roztoci)
```

Načtená data znázorníme graficky. Abychom si udělali představu o tom, jak se výsledky liší pro různé druhy roztočů, znázorníme jinou barvou údaje získané pro roztoče druhu A (bílá barva) a pro roztoče druhu B (černá barva).

Toho dosáhneme tak, že nejprve nakreslíme „prázdný“ graf a pak do něj zvlášť vykreslíme body reprezentující data pro druh A a pro druh B.

```
> plot(temp, rate, type = "n")
> points(temp[genus == "genA"], rate[genus == "genA"])
> points(temp[genus == "genB"], rate[genus == "genB"], pch = 16)
> legend("topleft", legend = c("genA", "genB"), pch = c(1, 16), bty = "n")
```




Obrázek 7: Závislost rychlosti růstu na teplotě pro dva druhy roztočů: ”genA” (bíle), ”genB” (černě).

Podle obrázku bychom mohli usoudit, že závislost rychlosti růstu populace na teplotě není lineární a navíc pro druh A je většinou o něco vyšší. To nás vede k tomu, že model popisující vztah proměnných, by mohl mít následující podobu

$$Y_{jk} = \mu + \alpha_j + \beta x_{jk} + \gamma x_{jk}^2 + \varepsilon_{jk},$$

kde $Y$ označuje rychlost růstu populace, $x$ je teplota, $j$ je index označující druh roztoče a $k$ označuje konkrétní pozorování v rámci daného druhu.

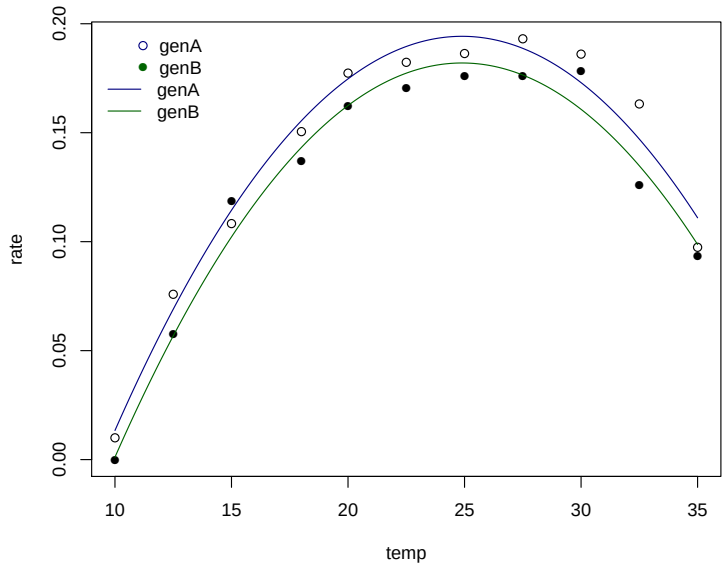
Modely, v nichž figuruje alespoň jedna spojitá a alespoň jedna kategoriální proměnná, se nazývají *analýza kovariance* (ANCOVA).

```
> ancova <- lm(rate ~ temp + I(temp^2) + genus)
> summary(ancova)


Call:
lm(formula = rate ~ temp + I(temp^2) + genus)

Residuals:
      Min        1Q    Median        3Q       Max
-0.013530 -0.006213 -0.002241  0.003962  0.017629

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.111e-01  1.779e-02 -17.489 9.63e-13 ***
temp         4.060e-02  1.694e-03  23.966 4.15e-15 ***
I(temp^2)   -8.154e-04  3.719e-05 -21.927 1.96e-14 ***
genusgenB   -1.224e-02  4.128e-03  -2.966  0.00828 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009682 on 18 degrees of freedom
Multiple R-squared: 0.9753,        Adjusted R-squared: 0.9712
F-statistic: 237.1 on 3 and 18 DF,  p-value: 1.186e-14
```

Podobně jako v předchozích příkladech byl druh A vybrán jako referenční, a proto odhadnutý parametr `genusgenB` vyjadřuje odchylku druhu B od druhu A.

Výsledky odhadu modelu zobrazíme do grafu.

```
> grid <- seq(from = min(temp), to = max(temp), length.out = 300)
> predA <- predict(ancova, data.frame(temp = grid, genus = rep("genA",
    300)))
> predB <- predict(ancova, data.frame(temp = grid, genus = rep("genB",
    300)))
> plot(temp, rate, type = "n")
> points(temp[genus == "genA"], rate[genus == "genA"])
> points(temp[genus == "genB"], rate[genus == "genB"], pch = 16)
> matlines(grid, predA, col = "navy")
> matlines(grid, predB, col = "darkgreen")
> legend(x = 10.5, y = 0.2, legend = c("genA", "genB"), pch = c(1, 16),
    col = c("navy", "darkgreen"), bty = "n")
> legend(x = 9, y = 0.18, legend = c("genA", "genB"), lty = 1, col = c("navy",
    "darkgreen"), bty = "n")
```




Obrázek 8: Závislost rychlosti růstu na teplotě pro dva druhy roztočů: ”genA” (bílé body, modrá čára), ”genB” (černé body, zelená čára).

# 4 Úkoly:

1. Třiceti ženám ve dvou amerických státech, Iowě a Nebrasce, byla zjištěna hladina cholesterolu (v mg/ml). Zároveň byl u každé z nich zaznamenán věk. Zjištěné údaje jsou v datovém souboru `cholesterol.dat`.

   – Datový soubor si prohlédněte a načtěte ho do R.

   – Do grafu vyneste naměřené hodnoty cholesterolu v závislosti na věku.

- Vykreslete tentýž graf, v němž odlišíte ženy z Iowy a ženy z Nebrasky. Vidíte nějaký rozdíl?
- Proložte daty regresní přímku, tj. odhadněte parametry modelu

$$\text{Model 1:} \qquad Y_i = \mu + \beta x_i + \varepsilon_i,$$

kde $Y_i$ je hladina cholesterolu a $x_i$ je věk. Výsledek znázorněte graficky.

- Odhadněte parametry modelu, v němž bude absolutní člen záviset na státě, tj.

$$\text{Model 2:} \qquad Y_{ij} = \mu + \alpha_j + \beta x_i + \varepsilon_i,$$

kde index $j$ označuje příslušnost k jednomu ze dvou států. Výsledky vyneste do grafu.

- Porovnejte odhadnuté přímky v obou modelech.

2. Datový soubor `CrashTest.dat` obsahuje výsledky crash testů. Při nich auta s figurínami na místě řidiče a spolujezdce narazila v rychlosti 35 mil v hodině do překážky a bylo sledováno několik kritérií vyjadřujících, jaký měl náraz dopad na hlavu, hrudník, levou a pravou nohu.

   - Datový soubor načtěte.
   - Prozkoumejte závislost poranění hlavy (proměnná `Head.IC`) na faktu, zda se jednalo o řidiče či spolujezdce (proměnná `D.P`). Odhadněte parametry modelu

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

kde $i$ označuje, zda se jedná o řidiče nebo spolujezdce.

   - Do modelu zahrňte navíc vliv proměnné Protection (vyjadřuje použité ochranné prvky – pásy, airbagy apod.). Odhadněte parametry v modelech

$$\begin{aligned} &\text{Model 1:} && Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\ &\text{Model 2:} && Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \end{aligned}$$

   - Vykreslete graf závislosti proměnné `Chest.decel` (poranění hrudníku) na proměnné `Wt` (hmotnost auta). V dalším grafu vykreslete totéž, ale rozlište údaje pro řidiče a pro spolujezdce.
   - Odhadněte parametry v modelech

$$\begin{aligned} &\text{Model 1:} && Y_{ij} = \mu + \beta x_j + \varepsilon_{ij} && \text{společná regresní přímka} \\ &\text{Model 2:} && Y_{ij} = \mu + \alpha_i + \beta x_j + \varepsilon_{ij} && \text{dvě (rovnoběžné) regresní přímky} \\ &\text{Model 3:} && Y_{ij} = \mu + \alpha_i + \beta_i x_j + \varepsilon_{ij} && \text{dvě regresní přímky (s různými směrnicemi)} \end{aligned}$$

   Výsledky znázorněte graficky a porovnejte.