M5201 – 2. CVIČENÍ: *Regresní analýza v R*

1 Úvod do regresní analýzy

Regresní analýza je jednou z nejčastěji používaných statistických technik. Zabýváme se při ní vztahy mezi několika veličinami. Snažíme se zjistit, jak hodnoty jedné veličiny (*vysvětlované proměnné, odezvy*) závisí na hodnotách ostatních veličin (*vysvětlujících veličin, regresorů, kovariát*).

Vycházíme z předpokladu, že vysvětlovaná veličina je funkcí vysvětlujících proměnných a náhodné složky, což je náhodná veličina s nulovou střední hodnotou, která reprezentuje náhodné odchylky, chyby v měření apod.

$$Y = f(x_1, x_2, \ldots, x_m, \varepsilon),$$

kde

Y	je vysvětlovaná proměnná
$x_1, \ldots, x_m$	jsou vysvětlující proměnné
ε	je náhodná složka

Funkce f patří do předem zvolené třídy funkcí a závisí na jednom či více parametrech. Ty jsou neznámé a odhadují se na základě zjištěných dat.

Obvykle máme k dispozici n pozorování vysvětlované proměnné $Y_1, Y_2, \ldots, Y_n$ a jim odpovídající hodnoty vysvětlujících proměnných $x_{11}, x_{12}, \ldots, x_{1n}, x_{21}, x_{22}, \ldots, x_{2n}, \ldots, x_{m1}, \ldots, x_{mn}$ a pomocí modelu můžeme popsat jednotlivá pozorování

$$Y_i = f(x_{1i}, X_{2i}, \ldots, x_{mi}, \varepsilon_i), \qquad i = 1, \ldots, n$$

Nejjednodušším případem regresního modelu je lineární regresní model. Je charakteristický tím, že vysvětlovanou proměnnou se v něm snažíme popsat pomocí lineární kombinace známých funkcí vysvětlovaných proměnných s neznámými regresními koeficienty. Pomocí rovnice můžeme lineární regresní model zapsat takto

$$Y_i = \beta_0 + \beta_1 f_1(x_{1i}, x_{2i}, \ldots, x_{mi}) + \beta_2 f_2(x_{1i}, x_{2i}, \ldots, x_{mi}) + \cdots \beta_k f_k(x_{1i}, x_{2i}, \ldots, x_{mi}) + \varepsilon_i,$$

kde

Y	je vysvětlovaná proměnná
$x_1, \ldots, x_m$	jsou vysvětlující proměnné
$f_1, f_2, \ldots, f_k$	jsou známé (předem zvolené) funkce vysvětlujících proměnných
$\beta_0, \beta_1, \ldots, \beta_k$	jsou neznámé regresní koeficienty
ε	je náhodná veličina s nulovou střední hodnotou a konstatním rozptylem

V lineárním regresním modelu tedy vysvětlovanou proměnnou modelujeme jako součet *systematické složky* (což je funkce, která je lineární v neznámých regresních koeficientech) a *náhodné složky*.

V lineárním regresním modelu se dále předpokládá, že pro různé realizace Y_i, $i = 1, \ldots, n$ vysvětlované proměnné jsou odpovídající náhodné veličiny ε_i navzájem nezávislé a mají stejné rozdělení s nulovou střední hodnotou a konstatním rozptylem, což značíme $\varepsilon_i \sim IID(0, \sigma^2)$.

Pro odhad neznámých parametrů regresních modelů existuje více metod. Pro lineární regresní modely se obvykle používá *metoda nejmenších čtverců*. Odhady parametrů se obvykle

označují stříškou, např. $\hat{\beta}_0, \widehat{\beta}_1, \ldots$. Hodnoty vysvětlované proměnné $\widehat{Y}_i$, které získáme po dosazení odhadnutých parametrů do regresní rovnice

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 f_1(x_{1i}, x_{2i}, \ldots, x_{mi}) + \widehat{\beta}_2 f_2(x_{1i}, x_{2i}, \ldots, x_{mi}) + \cdots \widehat{\beta}_k f_k(x_{1i}, x_{2i}, \ldots, x_{mi})$$

se nazývají *vyrovnané hodnoty*. Rozdíly mezi skutečnými a vyrovnanými hodnotami vysvětlované proměnné $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$ se nazývají *rezidua*.

Veličiny v modelu mohou být jak spojité, tak kategorické (sem patří nominální, ordinální a diskrétní veličiny). Kategorické proměnné se také nazývají faktory a jejich hodnoty se nazývají úrovně.

Značení:
Spojité veličiny budeme značit písmeny z konce abecedy: $x, Y, \ldots$
Faktory budeme značit písmeny ze začátku abecedy: $A, B, \ldots$

2 Použití funkce `lm`

K odhadu parametrů lineárního regresního modelu je v systému R k dispozici funkce `lm`. Tato funkce má následující argumenty

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Nejdůležitější je argument `formula`, kterým se zadává, jakou podobu má mít rovnice popisující model. Modelový vztah mezi proměnnými se obecně zadává ve tvaru

vysvětlovaná proměnná $\sim$ *systematická složka*

Zatímco na levé straně můžeme normálně používat obvyklé matematické funkce jako logaritmus nebo matematické operátory $+, -$ apod., na pravé straně mají matematické operátory poněkud odlišný význam. Konkrétně:

`+`	přidat vysvětlující proměnnou
`-`	odstranit vysvětlující proměnnou
`:`	interakce mezi veličinami
`*`	všechny komponenty
`1`	absolutní člen
`-1`	odebrat absolutní člen

Pokud potřebujeme při specifikaci modelu použít matematické operátory v obvyklém smyslu, musíme výraz s těmito operátory napsat jako argument funkce `I()`. Rozdíl můžeme ukázat na příkladu.

`y ~ x1 + x2` definuje model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$,

zatímco

`y ~ I(x1 + x2)` definuje model $Y_i = \beta_0 + \beta_1 (x_{1i} + x_{2i}) + \varepsilon_i$.

Konkrétní příklady zadávání modelů v systému R jsou ukázány v následující tabulce.

Formální zápis	***Syntaxe v R***	***Popis***
$Y_i = \mu + \varepsilon_i$	`y ~ 1`	Model pouze s absolutním členem
$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$	`y ~ x`	Lineární model se spojitou vysvětlující proměnnou x (regresní přímka)
$Y_i = \beta_1 x_i + \varepsilon_i$	`y ~ x - 1`	Lineární model se spojitou vysvětlující x bez absolutního členu
$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$	`log(y) ~ x`	Lineární model se spojitou vysvětlující x a s logaritmicky transformovanou vysvětlovanou proměnnou
$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$	`y ~ x + I(x^2)`, `y ~ poly(x, 2)`	Kvadratický model se spojitou vysvětlující proměnnou x (regresní parabola)
$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$	`y ~ x1+x2`	Model se dvěma spojitými vysvětlujícími proměnnými x_1 a x_2, v každé z nich je lineární (vícenásobná regrese)
$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$	`y ~ x1*x2`, `y ~ x1+x2+I(x1*x2)`	Model se dvěma spojitými vysvětlujícími proměnnými x_1 a x_2 s křížovým členem
$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$	`y ~ A`	Model s jedním faktorem (jednofaktorová analýza rozptylu, ANOVA)
$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$	`y ~ A + B`	Model se dvěma faktory bez interakce (dvoufaktorová analýza rozptylu)
$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$	`y ~ A + B + A:B`, `y ~ A * B`	Model se dvěma faktory s interakcí (dvoufaktorová analýza rozptylu)
$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijkl}$	`y ~ A + B + C`	Model se třemi faktory bez interakcí (třífaktorová analýza rozptylu)
$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}$	`y ~ A+B+C+A:B+A:C+B:C+A:B:C` `y ~ A*B*C`	Model se třemi faktory s interakcemi (třífaktorová analýza rozptylu)
$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijkl}$	`y ~ A+B+C+A:B+A:C+B:C` `y ~ (A+B+C)^2`	Model se třemi faktory obsahující kromě hlavních efektů maximálně dvojné interakce (třífaktorová analýza rozptylu)
$Y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{ij}$	`y ~ x+A`	Model se spojitou proměnnou x a faktorem A bez interakce interakce (ANCOVA)
$Y_{ij} = \mu + \alpha_j + \beta x_{ij} + \delta_j x_{ij} + \varepsilon_{ij}$	`y ~ x*A`	Model se spojitou proměnnou x a faktorem A s interakcí (ANCOVA)

Tabulka 1: Přehled zápisu základních modelů

3 Příklady

PŘÍKLAD 1: REGRESNÍ PŘÍMKA

Máme k dispozici údaje o tom, jaký počet australských domácností měl v letech 1998–2008 připojení na internet:

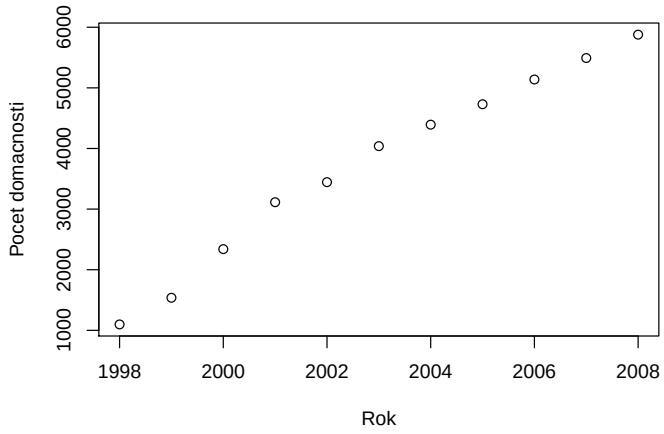
Rok	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Počet domácností	1098	1538	2340	3114	3445	4039	4393	4730	5138	5492	5878

Data nejprve načteme, a to tak, že údaje o počtech domácností zkopírujeme do schránky a použijeme funkci `scan`.

```
> domacnosti <- scan("clipboard", sep = " ")
> time <- 1998:2008
```

Načtená data vykreslíme do grafu.

```
> TxtX <- "Rok"
> TxtY <- "Pocet domacnosti"
> plot(time, domacnosti, type = "p", xlab = TxtX, ylab = TxtY)
```



Obrázek 1: Počet domácností s připojením na internet v letech 1998–2008

Budeme uvažovat regresní model, v němž bude počet domácností majících připojení na internet záviset lineárně na čase. Model můžeme popsat rovnicí

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, 11$$

kde

Y_i	je počet domácností s připojením na internet v i-tém roce
x_i	je rok odpovídající i-tému pozorování
β_0, β_1	jsou neznámé regresní koeficienty, které budeme odhadovat
ε_i	je náhodná odchylka příslušející i-tému pozorování

Parametry tohoto modelu odhadneme v systému R pomocí funkce `lm` a k zobrazení výsledků použijeme funkci `summary`.

```
> model1 <- lm(domacnosti ~ time)
> summary(model1)


Call:
lm(formula = domacnosti ~ time)

Residuals:
    Min      1Q  Median      3Q     Max
-306.45 -200.05   20.18  173.09  318.82

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -948407.45   45085.31  -21.04 5.81e-09 ***
time            475.36      22.51   21.12 5.61e-09 ***
---
Signif. codes:  0 `***´ 0.001 `**´ 0.01 `*´ 0.05 `.´ 0.1 ` ´ 1

Residual standard error: 236.1 on 9 degrees of freedom
Multiple R-squared: 0.9802,        Adjusted R-squared: 0.978
F-statistic:   446 on 1 and 9 DF,  p-value: 5.615e-09
```

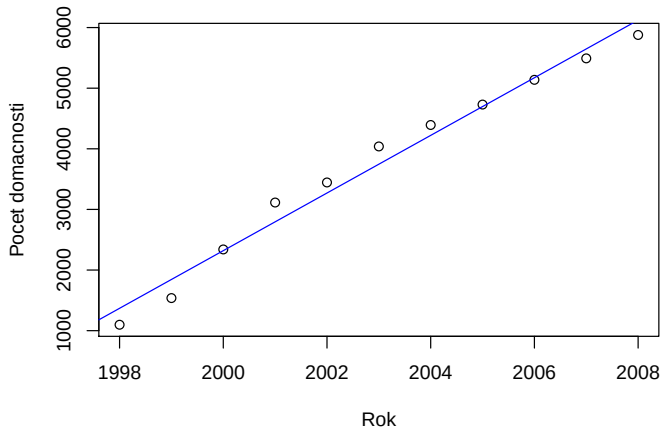
Ve výpisů výsledků odhadu modelu vidíme nejprve zadaný tvar modelu. Dále jsou zde uvedeny informace o reziduích (minimální a maximální reziduum, horní a dolní kvartil a medián). V další tabulce jsou shrnuty odhadnuté parametry

$$\beta_0 = -948407.45 \ (\texttt{Intercept}) \qquad \beta_1 = 475.36 \ (\texttt{time})$$

a v dalších sloupcích jsou směrodatné odchylky odhadnutých parametrů a testové statistiky pro testování některých hypotéz týkajících se parametrů modelu. V závěru jsou uvedeny souhrnné statistiky týkající se celého modelu.

Na závěr zobrazíme data proložená odhadnutou regresní přímkou.

```
> plot(time, domacnosti, type = "p", xlab = TxtX, ylab = TxtY)
> abline(lm(domacnosti ~ time), col = "blue", lty = 1)
```



Obrázek 2: Počet domácností s připojením na internet v letech 1998–2008

Příklad 2: Polynomická regrese

Máme k dispozici průměrné roční průtoky vody v řece Nigeru v Coulicouro (Mali). Údaje se vztahují k období 1907 až 1957 a jsou uvedena v jednotkách cfs.10^{-3}.

Data jsou uložena v souboru `DataNiger.dat`. První řádek obsahuje popis časové řady, pak následuje volný řádek, teprve potom dva sloupce dat: rok a průměrný roční průtok.

Nejprve načteme nadpis.

```
> fileDat <- paste(data.library, "DataNiger.dat", sep = "")
> (TXT <- paste(scan(fileDat, what = "", nlines = 1), collapse = " "))
```

```
[1] "Prumerne rocni prutoky vody v~rece Nigeru v~Coulicoure (Mali) v~letech 1907 az 1957"
```

Pak načteme vlastní data do datového rámce za pomoci funkce `read.table` s argumentem `skip=2`, který způsobí vynechání prvních dvou řádků. Ve vzniklém datovém rámci pojmenujeme proměnné, vypíšeme jeho strukturu a prvních šest řádků. Nakonec data vykreslíme.

```
> dataNiger <- read.table(fileDat, header = F, skip = 2)
> names(dataNiger) <- c("Rok", "Prutok")
> str(dataNiger)
```

```
'data.frame':        51 obs. of  2 variables:
 $ Rok   : int  1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 ...
 $ Prutok: num  39.8 42.9 69.1 43.7 56.7 ...
```

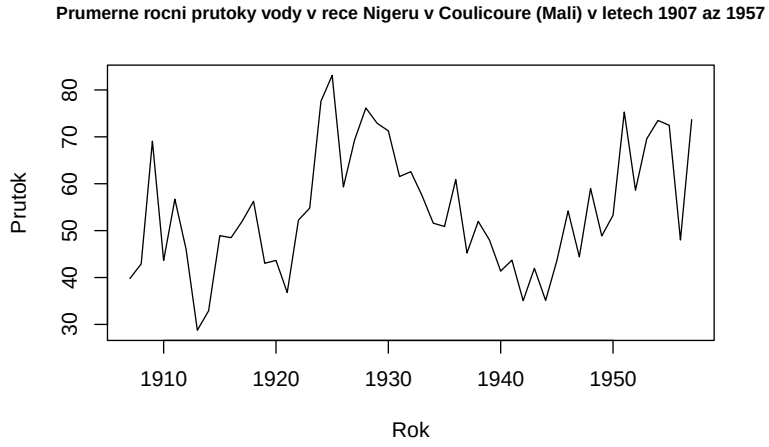
```
> head(dataNiger)
```

```
   Rok Prutok
1 1907 39.793
2 1908 42.892
3 1909 69.070
4 1910 43.653
5 1911 56.700
6 1912 45.990
```

Příkazem `attach` zpřístupníme proměnné v datovém rámci `dataNiger` a data vykreslíme.

```
> attach(dataNiger)
```

```
> plot(Rok, Prutok, main = TXT, cex.main = 0.85, type = "l")
```



Obrázek 3: Průměrně roční průtoky v řece Niger v Coulicoure (Mali).

Z grafu je patrné, že závislost průtoku na čase bude složitější a proložit grafem regresní přímku by mohlo být příliš zjednodušující. Proto zkusíme daty proložit polynom vyššího stupně. Nejprve zkusíme polynom třetího stupně, kdy budeme předpokládat, že jednotlivá pozorování můžeme popsat vztahem

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i,$$

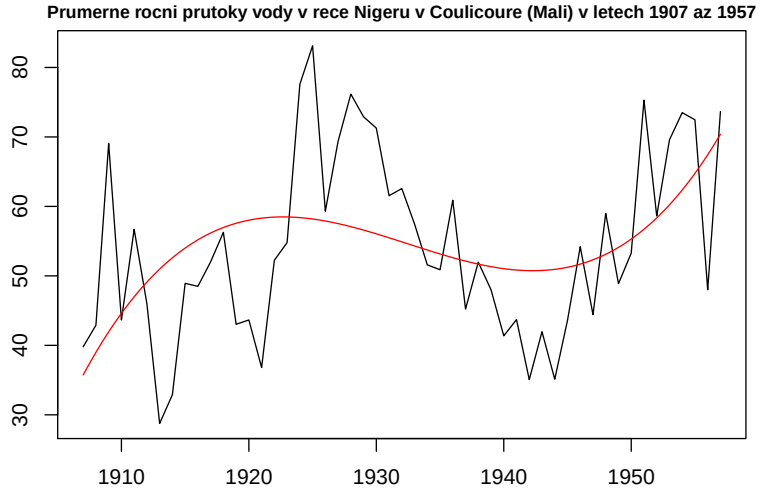
kde Y_i označuje průtok vody a x_i označuje rok odpovídající danému údaji.

Neznámé parametry tohoto modelu nyní odhadneme.

```
> model3 <- lm(Prutok ~ Rok + I(Rok^2) + I(Rok^3))
> summary(model3)


Call:
lm(formula = Prutok ~ Rok + I(Rok^2) + I(Rok^3))

Residuals:
     Min       1Q   Median       3Q      Max
-22.2179  -7.0170  -0.9496   7.6885  27.1292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.483e+07  4.903e+06  -3.024  0.00403 **
Rok          2.302e+04  7.614e+03   3.023  0.00404 **
I(Rok^2)    -1.191e+01  3.941e+00  -3.022  0.00405 **
I(Rok^3)     2.055e-03  6.800e-04   3.022  0.00406 **
---
Signif. codes:  0 `***´ 0.001 `**´ 0.01 `*´ 0.05 `.´ 0.1 ` ´ 1

Residual standard error: 12.14 on 47 degrees of freedom
Multiple R-squared: 0.2108,        Adjusted R-squared: 0.1604
F-statistic: 4.183 on 3 and 47 DF,  p-value: 0.0105
```

Výsledky modelu s polynomem řádu tři zakreslíme do grafu. Odhadnutou regresní křivku zobrazíme do grafu tak, že nejprve vytvoříme vektor `gridt` obsahující 500 hodnot v rozmezí 1907 a 1957. Pak použijeme funkci `predict`, pomocí níž pro všechny prvky `gridt` spočítáme vyrovnané hodnoty na základě odhadnutých parametrů modelu a uložíme je do objektu `Pred3`. Výslednou křivku zobrazíme pomocí funkce `matlines`, kde jako argumenty zadáme `gridt` a `Pred3`.

```
> n <- length(Rok)
> gridt <- seq(Rok[1], Rok[n], length.out = 500)
> Pred3 <- predict(model3, data.frame(Rok = gridt))
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(Rok, Prutok, type = "l", main = TXT, cex.main = 0.85)
> matlines(gridt, Pred3, col = "red")
```



Obrázek 4: Průměrné roční průtoky vody v řece Niger v Coulicoure (Mali) proložené polynomem třetího stupně.

Zkusíme odhadnout jiný model s polynomem šestého stupně, budeme tudíž předpokládat, že jednotlivá pozorování lze popsat vztahem

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5 + \beta_6 x_i^6 + \varepsilon_i.$$

Odhadneme parametry nového modelu.

```
> model6 <- lm(Prutok ~ Rok + I(Rok^2) + I(Rok^3) + I(Rok^4) + I(Rok^5) +
    I(Rok^6))
> summary(model6)


Call:
lm(formula = Prutok ~ Rok + I(Rok^2) + I(Rok^3) + I(Rok^4) +
    I(Rok^5) + I(Rok^6))

Residuals:
     Min       1Q   Median       3Q      Max
-22.2179  -7.0170  -0.9496   7.6885  27.1292
```

```
Coefficients: (3 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.483e+07  4.903e+06  -3.024  0.00403 **
Rok          2.302e+04  7.614e+03   3.023  0.00404 **
I(Rok^2)    -1.191e+01  3.941e+00  -3.022  0.00405 **
I(Rok^3)     2.055e-03  6.800e-04   3.022  0.00406 **
I(Rok^4)            NA         NA      NA       NA
I(Rok^5)            NA         NA      NA       NA
I(Rok^6)            NA         NA      NA       NA
---
Signif. codes:  0 `***´ 0.001 `**´ 0.01 `*´ 0.05 `.´ 0.1 ` ´ 1

Residual standard error: 12.14 on 47 degrees of freedom
Multiple R-squared: 0.2108,        Adjusted R-squared: 0.1604
F-statistic: 4.183 on 3 and 47 DF,  p-value: 0.0105
```

Vidíme, že odhad modelu s vysokým stupněm polynomu je numericky problematický, protože se koeficienty od čtvrtého stupně výše nepodařilo spočítat. V takových případech se doporučuje proměnnou `Rok` centrovat a odhadnout parametry pro model s touto centrovanou proměnnou. Centrování znamená, že od každé hodnoty proměnné `Rok` odečteme průměr všech hodnot. Model se tak změní do podoby

$$Y_i = \beta_0 + \beta_1 z_i + \beta_2 z_i^2 + \beta_3 z_i^3 + \beta_4 z_i^4 + \beta_5 z_i^5 + \beta_6 z_i^6 + \varepsilon_i,$$

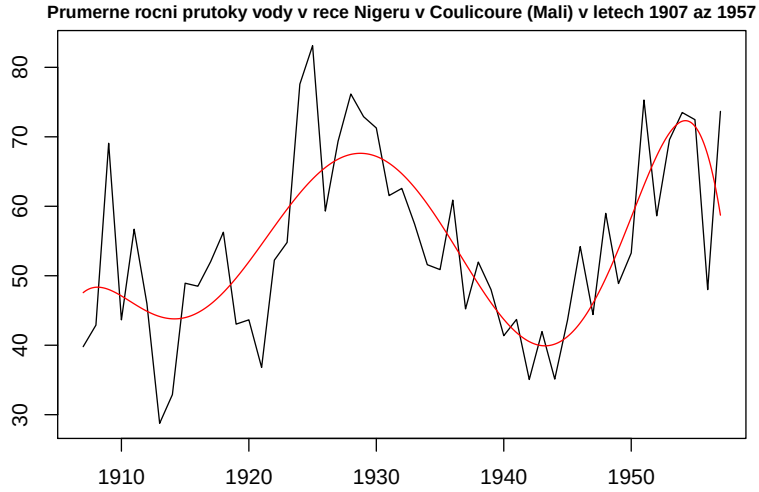
kde $z_i = x_i - \bar{x}$.

Provedeme nový odhad modelu s centrovanou proměnnou.

```
> delta <- mean(Rok)
> RokC <- Rok - delta
> model6C <- lm(Prutok ~ RokC + I(RokC^2) + I(RokC^3) + I(RokC^4) + I(RokC^5) +
     I(RokC^6))
> summary(model6C)


Call:
lm(formula = Prutok ~ RokC + I(RokC^2) + I(RokC^3) + I(RokC^4) +
    I(RokC^5) + I(RokC^6))

Residuals:
     Min       1Q   Median       3Q      Max
-19.3877  -5.4541  -0.9673   5.1213  21.0478

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.469e+01  2.872e+00  22.523  < 2e-16 ***
RokC        -1.757e+00  3.909e-01  -4.494 5.02e-05 ***
I(RokC^2)   -2.385e-01  5.367e-02  -4.444 5.90e-05 ***
I(RokC^3)    1.045e-02  2.371e-03   4.408 6.63e-05 ***
I(RokC^4)    8.809e-04  2.268e-04   3.884 0.000341 ***
I(RokC^5)   -1.165e-05  3.209e-06  -3.631 0.000733 ***
I(RokC^6)   -8.461e-07  2.526e-07  -3.350 0.001666 **
---
Signif. codes:  0 `***´ 0.001 `**´ 0.01 `*´ 0.05 `.´ 0.1 ` ´ 1

Residual standard error: 9.36 on 44 degrees of freedom
Multiple R-squared: 0.5608,        Adjusted R-squared: 0.5009
F-statistic: 9.364 on 6 and 44 DF,  p-value: 1.278e-06
```

Výsledky opět znázorníme graficky.

```
> gridtC <- gridt - delta
> Pred6 <- predict(model6C, data.frame(RokC = gridtC))
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(Rok, Prutok, type = "l", main = TXT, cex.main = 0.85)
> matlines(gridt, Pred6, col = "red")
```



Obrázek 5: Průměrné roční průtoky vody v řece Niger v Coulicoure (Mali) proložené polynomem šestého stupně.

Příklad 3: Trigonometrická regrese

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 ...
```

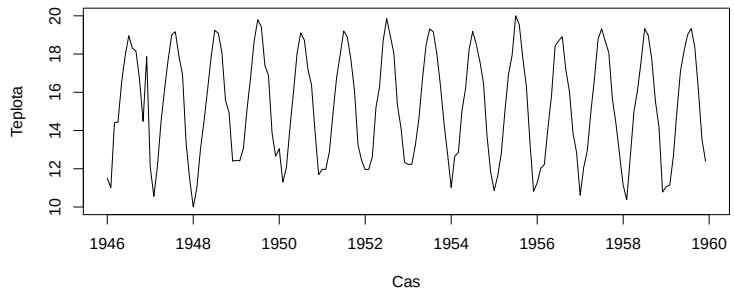
```
> head(data)
```

```
      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
> plot(1946 + (Time - 1)/12, teploty, type = "l", xlab = "Cas", ylab = "Teplota")
```



Obrázek 6: 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.

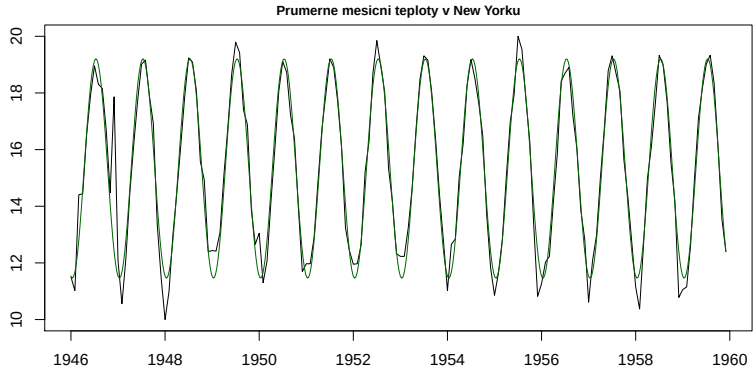
K odhadu parametrů modelu opět použijeme funkci `lm`.

```
> 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.

```
> n <- length(data$teploty)
> gridt <- seq(Time[1], Time[n], length.out = 500)
> PredPeriod <- predict(modelPeriod, data.frame(Time = gridt))
> par(mar = c(2, 2, 1, 0) + 0.5)
> plot(1946 + (Time - 1)/12, data$teploty, type = "l", main = TXT, cex.main = 0.85)
> matlines(1946 + (gridt - 1)/12, PredPeriod, col = "darkgreen")
```



Obrázek 7: Průměrné měsíční teploty v New Yorku s trigonometrickou regresní funkcí.

Příklad 4: 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 na počet můr chycených do pasti má umístění pasti. Úrovně této proměnné typu faktor zjistíme příkazem `levels`.

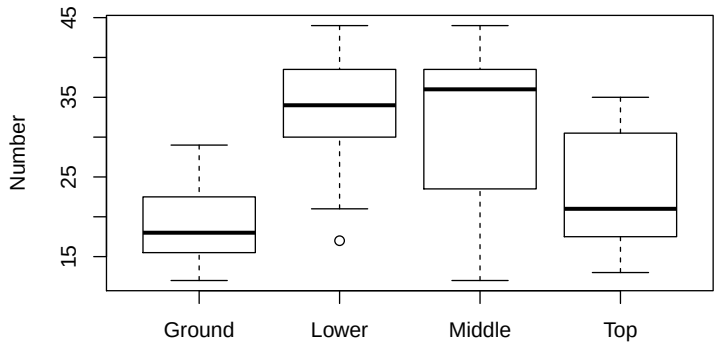
```
> levels(data$Location)
```

```
[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.

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



Obrázek 8: Počet chycených můr v závislosti na umístění pasti.

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 μ a α_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`:

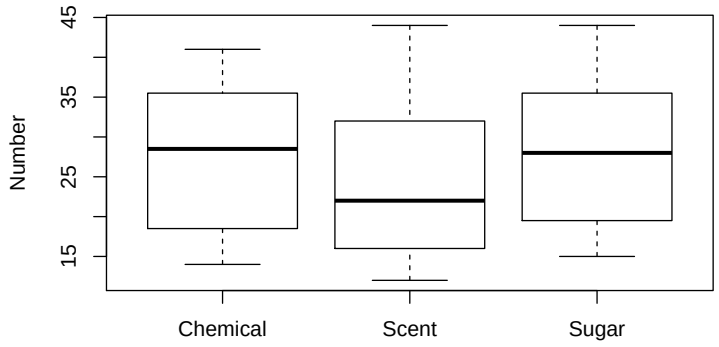
$$\begin{aligned}
\widehat{E}(Y_{ground,k}) &= (\texttt{Intercept}) = 19.067 \\
\widehat{E}(Y_{lower,k}) &= (\texttt{Intercept}) + \texttt{LocationLower} = 19.067 + 14.267 = 33.334 \\
\widehat{E}(Y_{middle,k}) &= (\texttt{Intercept}) + \texttt{LocationMiddle} = 19.067 + 11.933 = 31 \\
\widehat{E}(Y_{top,k})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

Příklad 5: 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 9: 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 α_j zachycuje vliv umístění pasti a β_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)
```

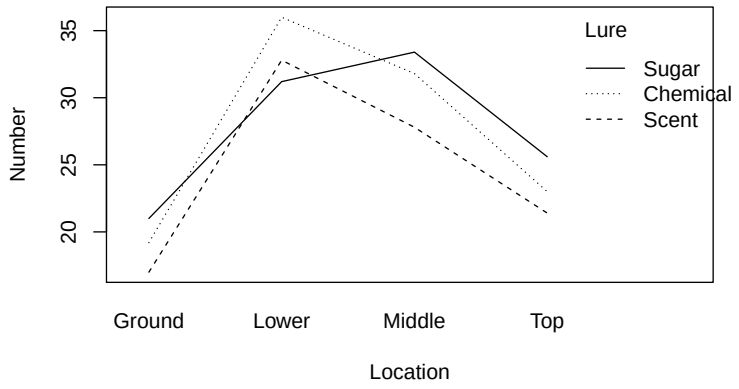
```
Call:
lm(formula = Number ~ Lure + Location, data = data)

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 10: 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 α_j zachycuje vliv umístění pasti, β_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
```

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)
```

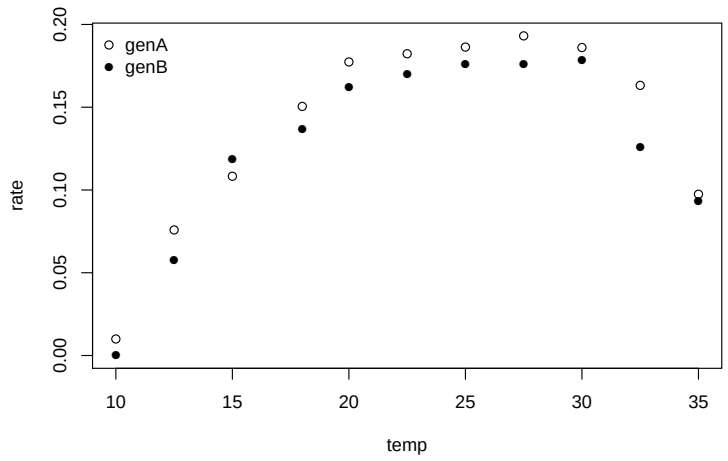
```
'data.frame':        22 obs. of  3 variables:
 $ temp : num  10 12.5 15 18 20 22.5 25 27.5 30 32.5 ...
 $ rate : num  0.01 0.0759 0.1083 0.1505 0.1774 ...
 $ genus: Factor w/ 2 levels "genA","genB": 1 1 1 1 1 1 1 1 1 1 ...
```

```
> 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 11: 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).

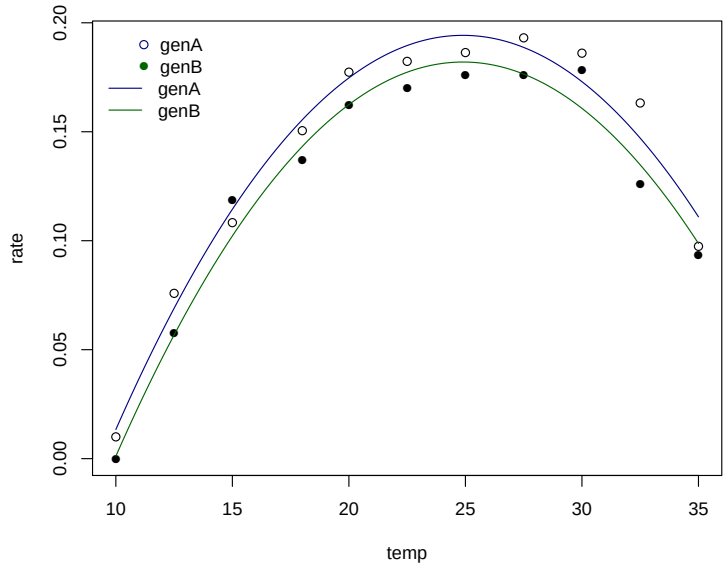
```
> ancova <- lm(rate ~ temp + I(temp^2) + genus)
> summary(ancova)
```

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