M6120 Lineární statistické modely II

Projekt/Domací úloha

Michal Červenka (518670)

MUNI SCI Ústav matematiky a statistiky

B-MAT Matematika

Přírodovědecká fakulta, Masarykova Univerzita

27. května 2024

Obsah

Řešení příkladu 1

Uvažujme lineární model ve tvaru

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 \mathbb{I}_M\{i\} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

pro $i = 1, 2, \dots, n$, kde $\mathbb{I}_M\{i\}$ označuje indikátorovou funkci nějaké pevně dané množiny M, tj.

$$\mathbb{I}_M\{i\} = \begin{cases} 1 & i \in M, \\ 0 & i \notin M. \end{cases}$$

Podle Scheffého věty platí

$$\Pr\left(\left\{ \mathbf{b}^\top (\mathbf{A}\hat{\beta} - \mathbf{A}\beta) \right\}^2 \leq m F_{1-\alpha}(m, n-p) \hat{\sigma}^2 \mathbf{b}^\top \mathbf{A} \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{A}^\top \mathbf{b} \right) = 1 - \alpha$$

pro všechna $\mathbf{b} \in \mathbb{R}^m$, kde $\mathbf{A} \in \mathbb{R}^{m \times p}$. V případě našeho modelu je $m = 3,\ p = 4$ a $\mathbf{b} = (1, x, x^2)^T$. Odvození tvaru pásu spolehlivosti rozdělíme do dvou částí, a to na případ, kdy $i \in M$ a kdy $i \notin M$. V závislosti na této situaci se bude měnit matice $\mathbf{A}$.

Začneme se situací, kdy $i \in M$. V tomto případě je matice $\mathbf{A}$ ve tvaru

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Nyní dosadíme do Scheffého věty a začneme odvozovat pás spolehlivosti. Po dosazení dostáváme

$$\Pr(\{\hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 - (\beta_0 + \beta_3 + \beta_1 x + \beta_2 x^2)\}^2 \leq$$
$$3 \cdot F_{1-\alpha}(3, n-4) \hat{\sigma}^2 (1, x, x^2, 1)(X^T X)^{-1} (1, x, x^2, 1)^T) = 1 - \alpha.$$

Odmocníme výraz uvnitř Pr, čímž dostaneme

$$\Pr(\hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 - (\beta_0 + \beta_3 + \beta_1 x + \beta_2 x^2) \leq$$
$$\pm \sqrt{3 \cdot F_{1-\alpha}(3, n-4) \hat{\sigma}^2 (1, x, x^2, 1)(X^T X)^{-1} (1, x, x^2, 1)^T}) = 1 - \alpha.$$

Označme $y = \beta_0 + \beta_3 + \beta_1 x + \beta_2 x^2$. Tvar pásu spolehlivosti odvozeného z Scheffého věty pro situaci $i \in M$ s pravděpodobností pokrytí $100 \cdot (1 - \alpha)\,\%$ je

$$y \geq \hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 - \sqrt{3 \cdot F_{1-\alpha}(3, n-4) \hat{\sigma}^2 (1, x, x^2, 1)(X^T X)^{-1} (1, x, x^2, 1)^T},$$
$$y \leq \hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \sqrt{3 \cdot F_{1-\alpha}(3, n-4) \hat{\sigma}^2 (1, x, x^2, 1)(X^T X)^{-1} (1, x, x^2, 1)^T}.$$

Situaci, kdy $i \in M$ máme vyřešenou. Zaměřme se nyní na situaci, kdy $i \notin M$. Matice $\mathbf{A}$ bude v tomto případě v následujícím tvaru

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Odvození pásu spolehlivosti pro tuto situaci provedeme analogicky jako v předchozím případě. Prvně dosaďme do Scheffého věty, čímž dostaneme

$$\Pr(\{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 - (\beta_0 + \beta_1 x + \beta_2 x^2)\}^2 \leq$$
$$3 \cdot F_{1-\alpha}(3, n-4) \hat{\sigma}^2 (1, x, x^2, 0)(X^T X)^{-1} (1, x, x^2, 0)^T) = 1 - \alpha.$$

Odmocníme výraz uvnitř Pr, čímž dostaneme

$$\Pr(\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 - (\beta_0 + \beta_1 x + \beta_2 x^2) \leq$$
$$\pm\sqrt{3 \cdot F_{1-\alpha}(3, n-4)\hat{\sigma}^2(1, x, x^2, 0)(X^T X)^{-1}(1, x, x^2, 0)^T}) = 1 - \alpha.$$

Označme $y = \beta_0 + \beta_1 x + \beta_2 x^2$. Tvar pásu spolehlehlivosti odvozeného z Scheffého věty pro situaci $i \notin M$ s pravděpodobností pokrytí $100 \cdot (1-\alpha)\,\%$ je

$$y \geq \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 - \sqrt{3 \cdot F_{1-\alpha}(3, n-4)\hat{\sigma}^2(1, x, x^2, 0)(X^T X)^{-1}(1, x, x^2, 0)^T},$$
$$y \leq \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \sqrt{3 \cdot F_{1-\alpha}(3, n-4)\hat{\sigma}^2(1, x, x^2, 0)(X^T X)^{-1}(1, x, x^2, 0)^T}.$$

Řešení příkladu 2

(a) Najděte vhodný lineární regresní model pro popis závislosti úspěšnosti studentů na délce přípravy a zařazení skupiny. V následujících bodech pracujte s tímto modelem

Budeme uvažovat dva modely. Jeden z nich bude obsahovat interkaci a druhý nikoliv a následně rozhodneme, který je více vhodný. Podívejme se na jednotlivé modely.

```r
# Model bez interakce
mod1 <- lm(data = data, uspesnost ~ hodiny + skupina)
# Model s interkací
mod2 <- lm(data = data, uspesnost ~ hodiny + skupina + skupina:hodiny)
```

Nyní se podívejme podrobněji na model bez interkace.

```
Call:
lm(formula = uspesnost ~ hodiny + skupina, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2601  -4.4194  -0.9404   4.3164  17.0555

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.1396     1.3841  -3.713 0.000342 ***
hodiny        4.7798     0.1075  44.458  < 2e-16 ***
skupinaB     -4.9704     1.2372  -4.017 0.000116 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.186 on 97 degrees of freedom
Multiple R-squared:  0.9536,    Adjusted R-squared:  0.9526
F-statistic:   996 on 2 and 97 DF,  p-value: < 2.2e-16
```

Můžeme si povšimnout, že všechny koeficienty modelu bez interkace jsou statisticky významné na hladině významnosti $\alpha = 0.05$, a tedy tento model vypadá slibně. Zaměřme se nyní na model s interkací.

```
Call:
lm(formula = uspesnost ~ hodiny + skupina + skupina:hodiny, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3059  -4.4559  -0.9882   4.6161  16.4528

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.2796     1.8542  -2.308   0.0231 *
hodiny            4.6936     0.1637  28.664   <2e-16 ***
skupinaB         -6.4894     2.5013  -2.594   0.0110 *
```

```
hodiny:skupinaB   0.1521     0.2175   0.699   0.4860
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.203 on 96 degrees of freedom
Multiple R-squared:  0.9538,    Adjusted R-squared:  0.9524
F-statistic: 660.7 on 3 and 96 DF,  p-value: < 2.2e-16
```

V modelu s interkací je koeficent u interakce proměnné skupina a hodiny statisticky nevýznamný na hladině významnosti $\alpha = 0.05$, a proto volíme jednodušší model, a tedy ten bez interkace, v našem značení mod1. Můžeme si povšimnout, že oba modely mají adjustované R^2 blízké k hodnotě 0.96, a tedy jsme našli velmi vhodné modely, jelikož vysvětlují velkou část variability našich dat.

(b) Pro každou skupinu zvlášť sestrojte konfidenční pás spolehlivosti s pravděpodobností pokrytí 95 %

```r
xx <- seq(0, 20, by = 0.1)
# Matice A pro skupinu A
A <- rbind(c(1, 0, 0), c(0, 1, 0))
# Odhady koeficientů modelu bez interakce
beta_hat <- mod1$coefficients
# Hodnoty parametrů
m  <- 2
p <- 3
n <- length(data$hodiny)
# Odhad kovarianční matice modelu bez interakce
vcov_mod1 <- vcov(mod1)
# Příslušní F-kvantil
F_kvantil <- qf(0.95, m, n-p)

# Výpočet dolní a horní hranice pásu spolehlivosti pro skupinu A
# Skupina A - dolní hranice pásu spolehlivosti
pas_skupina_A_lower <- numeric(length(xx))
for (i in seq_along(xx)) {
  b <- c(1, xx[i])
  pas_skupina_A_lower[i] <- beta_hat[1] + beta_hat[2] * xx[i] -
    sqrt(m * F_kvantil * t(b) %*% A %*% vcov_mod1 %*% t(A) %*% b)
}
# Skupina A - horní hranice pásu spolehlivosti
pas_skupina_A_upper <- numeric(length(xx))
for (i in seq_along(xx)) {
  b <- c(1, xx[i])
  pas_skupina_A_upper[i] <- beta_hat[1] + beta_hat[2] * xx[i] +
    sqrt(m * F_kvantil * t(b) %*% A %*% vcov_mod1 %*% t(A) %*% b)
}

# Výpočet dolní a horní hranice pásu spolehlivosti pro skupinu B
# Změna matice pro skupinu B
A <- rbind(c(1, 0, 1), c(0, 1, 0))

```

```r
# Skupina B - dolní hranice pásu spolehlivosti
pas_skupina_B_lower <- numeric(length(xx))
for (i in seq_along(xx)) {
  b <- c(1, xx[i])
  pas_skupina_B_lower[i] <- (beta_hat[1] + beta_hat[3]) + beta_hat[2] * xx[i] -
    sqrt(m * F_kvantil * t(b) %*% A %*% vcov_mod1 %*% t(A) %*% b)
}
# Skupina B - horní hranice pásu spolehlivosti
pas_skupina_B_upper <- numeric(length(xx))
for (i in seq_along(xx)) {
  b <- c(1, xx[i])
  pas_skupina_B_upper[i] <- (beta_hat[1] + beta_hat[3]) + beta_hat[2] * xx[i] +
    sqrt(m * F_kvantil * t(b) %*% A %*% vcov_mod1 %*% t(A) %*% b)
}
```

(c) Odhadněte (bodově i intervalově) průměrnou úspěšnost studenta skupiny B, který se učil půl hodiny

```
       fit dolni_hranice horni_hranice
1 -7.720058      -10.3885      -5.05162
```

(d) Vypočtěte simultánní intervalové odhad střední hodnoty úspěšnosti pro studenty skupiny A, kteří se učili 5, 10 a 15 hodin. Zvolte hladinu významnosti $\alpha = 0{,}05$ a využijte Bonferroniho adjustaci

```r
# Hodnoty hodiny = 5, skupina = A
IS_1
```

```
[1] 16.26095 21.25757
```

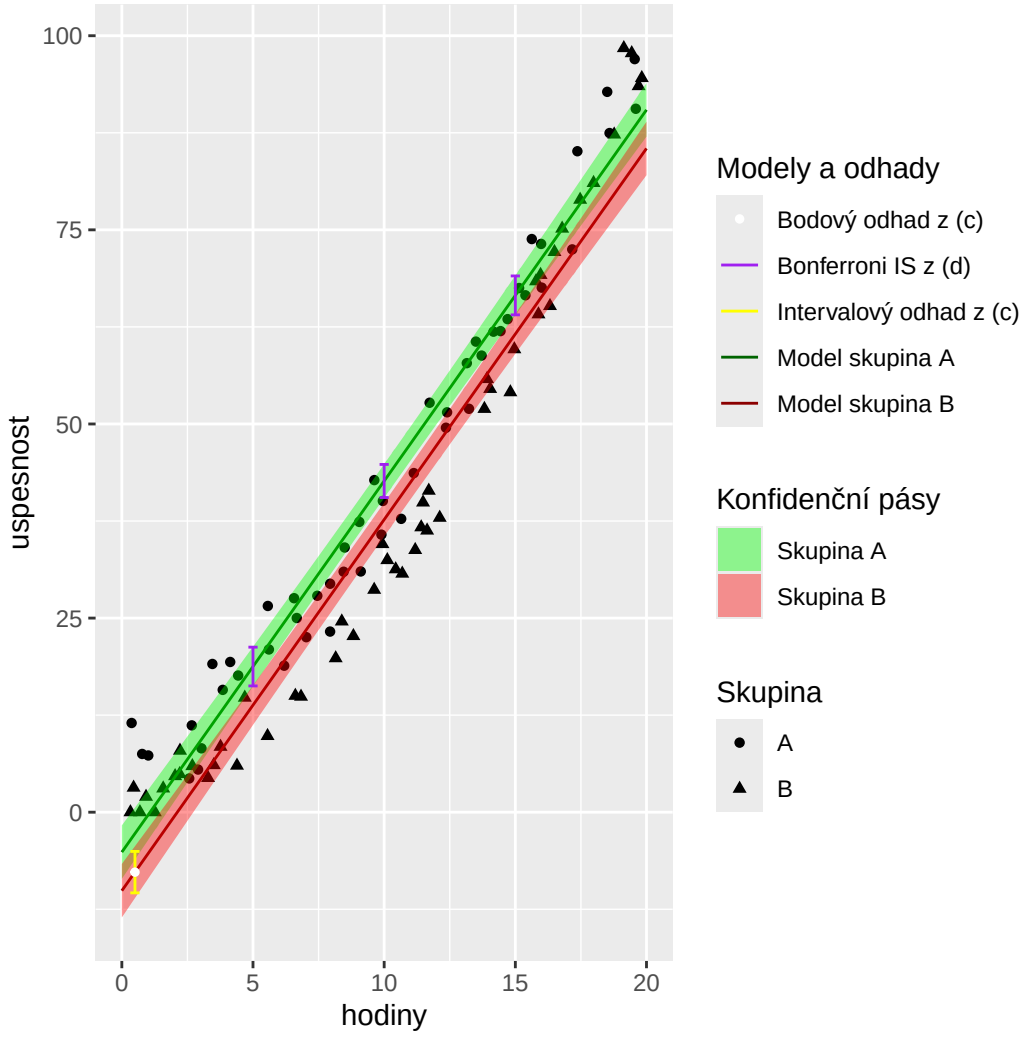
```r
# Hodnoty hodiny = 10, skupina = A
IS_2
```

```
[1] 40.52675 44.78945
```

```r
# Hodnoty hodiny = 15, skupina = A
IS_3
```

```
[1] 64.05215 69.06172
```

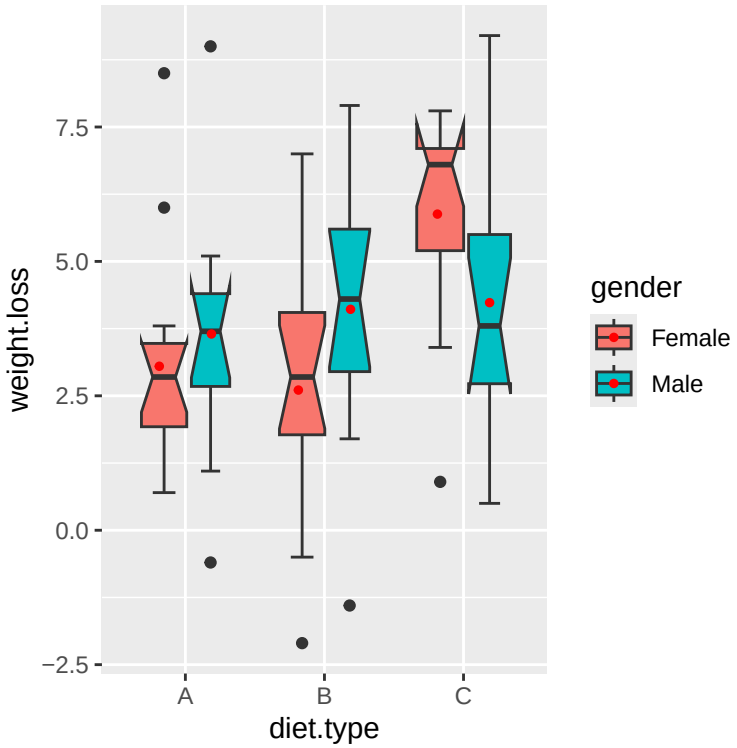
(e) Zakreslete výsledky z předchozích bodů do jednoho obrázku. Tj. vykreslete bodový graf pozorování, přidejte křivky vašeho modelu, zvýrazněte konfidenční pásy spolehlivosti a požadované intervalové odhady i predikce.



Obrázek 1: Regresní přímky modelu mod1 pro obě skupiny

Řešení příkladu 3

(a) Vykreslete krabicové diagramy popisující úbytek hmotnosti v závislosti na typu diety a pohlaví. Vytvořte také krabicové diagramy v závislosti pouze na typu diety. Všechny diagramy vložte vedle sebe do jednoho obrázku, abyste je mohli porovnat mezi sebou. Do „krabic“ přidejte zářezy, nastavte, aby jejich šířka odpovídala proporčně rozsahům a dokreslete do nich aritmetické průměry jako červené body



Obrázek 2: Boxploty úbytku váhy pro jednotlivé typy diet a pohlaví

(b) Modelujte závislost střední hodnoty úbytku hmotnosti na typu diety a pohlaví. Vyzkoušejte různé varianty složitosti modelu: (1) model se vzájemnou interakcí obou faktorů, (2) model bez interakce a (3) model bez vlivu proměnné gender. Vyberte ten nejvhodnější z nich a své rozhodnutí zdůvodněte a podpořte příslušným výstupem.

Prvně sestavme požadované modely (1) až (3).

```
mod1 <- lm(data = diet, weight.loss ~ gender + diet.type + diet.type:gender)
mod2 <- lm(data = diet, weight.loss ~ gender + diet.type)
mod3 <- lm(data = diet, weight.loss ~ diet.type)
```

Nyní provedeme porovnání modelů, tj. budeme zjišťovat, zda se od sebe modely statisticky významně liší na hladině významnosti $\alpha = 0.05$.

```
anova(mod1, mod2)
```

```
Analysis of Variance Table
```

```
Model 1: weight.loss ~ gender + diet.type + diet.type:gender
Model 2: weight.loss ~ gender + diet.type
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     70 376.33
2     72 410.23 -2   -33.904 3.1532 0.04884 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na základě výstupu vidíme, že model s oběma kategoriemi a jejich vzájemnou interakcí se liší od modelu s oběma kategoriemi, ale bez interakce. Vzhledem k tomu, že mod3 je podmodelem modelu mod2, tak zvolíme nejsložitější model, a tedy model mod3 s oběma kategoriemi "gender" a "diet.type" a jejich vzájemnou interakcí. Nezanedbatelnost interakce mezi typem diety a pohlavím nám již naznačily boxploty v (a). Můžeme si povšimnout, že úbytek váhy pro jednotlivé diety se liší v závislosti na pohlaví.

(c) Pro vámi vybraný model model z (b) vypište do tabulky odhady středních hodnot úbytku hmotnosti pro jednotlivé skupiny (tj. v případě modelu (1) skupiny symbolicky zapíšeme jako f.A, f.B, f.C, m.A, m.B a m.C, ale např. v případě modelu (3) uvažujme pouze skupiny A, B a C).

Tabulka 1: Odhady středních hodnot úbytku hmotnosti pro jednotlivé skupiny

Skupina	Odhad střední hodnoty
f.A	3.050000
f.B	2.607143
f.C	5.880000
m.A	3.650000
m.B	4.109091
m.C	4.233333

(d) Pro model (1) z (b) vykreslete do grafu zvlášť pro ženy (červená lomená čára) a zvlášť pro muže (modrá lomená čára) odhadnuté střední hodnoty úbytku hmotnosti pro všechny druhy diet. Na ose x budou úrovně faktoru diet.type a na ose y úbytek hmotnosti. Uveďte také, jak získáte jednotlivé odhady středních hodnot pomocí příslušných koeficientů modelu.

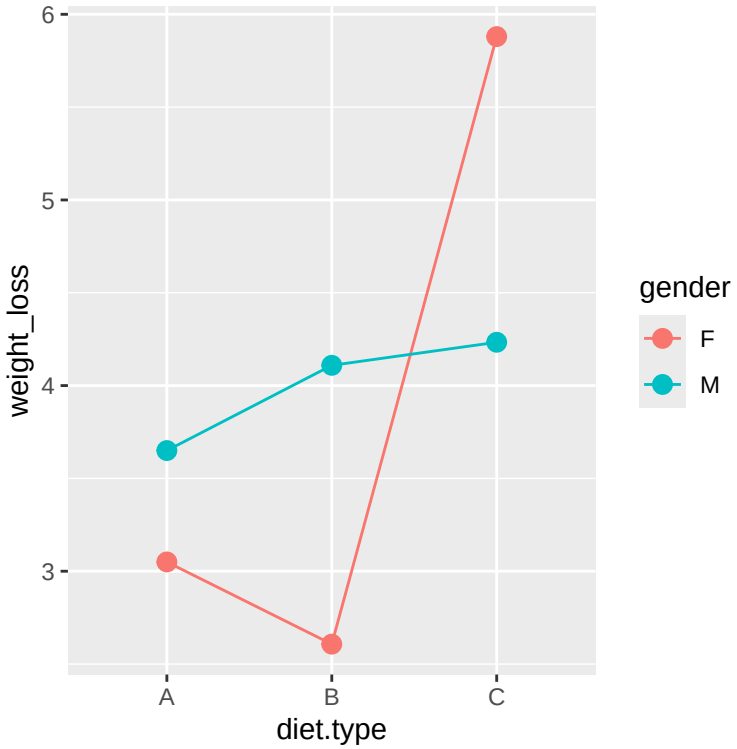
Prvně se podívejme, jak vypočítat jednotlivé odhady úbytku hmotnosti pro všechny diety a jednotlivá pohlaví. Pro přehlednost se podívejme na summary přislušného modelu, respektive na odhady jeho koeficientů.

```
                          Estimate Std. Error    t value     Pr(>|t|)
(Intercept)             3.0500000  0.6196848  4.9218572 5.488725e-06
genderMale              0.6000000  0.9600115  0.6249925 5.340081e-01
diet.typeB             -0.4428571  0.8763666 -0.5053332 6.149123e-01
diet.typeC              2.8300000  0.8616367  3.2844468 1.597680e-03
genderMale:diet.typeB   0.9019481  1.3395411  0.6733261 5.029569e-01
genderMale:diet.typeC -2.2466667  1.3145499 -1.7090767 9.186713e-02
```

Když už známe jednotlivé odhady koeficientů, respektive jejich pořadí, tak provedeme požadované výpočty.

```
# Odhady koeficientů
beta_hat <- mod1$coefficients
weight_loss_F_A <- beta_hat[1] # Žena s dietou A
weight_loss_F_B <- beta_hat[1] + beta_hat[3] # Žena s dietou B
weight_loss_F_C <- beta_hat[1] + beta_hat[4] # Žena s dietou C
weight_loss_M_A <- beta_hat[1] + beta_hat[2] # Muž s dietou A
weight_loss_M_B <- beta_hat[1] + beta_hat[2] + beta_hat[3] + beta_hat[5] # Muž s dietou B
weight_loss_M_C <- beta_hat[1] + beta_hat[2] + beta_hat[4] + beta_hat[6] # Muž s dietou C
```

Nyní budeme výpočty vizualizovat pomocí lomenných čar.



Obrázek 3: Lomenné čáry odhadnutých úbytků váhy v závislosti na pohlaví a typu diety