Lineární statistické modely II

Domácí úkol

Vojtěch Matulík

505487

Obor Statistika a analýza dat

Přírodovědecká fakulta
Masarykova univerzita

26. května 2023

Příklad 1

Mějme lineární regresní model tvaru

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \;\; i = 1,\, 2,\, \ldots,\, n,$$

kde pro náhodné chyby platí $\varepsilon_i \overset{\text{iid}}{\sim} N(0,\, \sigma^2)$. Podle Scheffého věty platí

$$P\left(\{\boldsymbol{b}^T(A\hat{\boldsymbol{\beta}} - A\boldsymbol{\beta})\}^2 \leq mF_{1-\alpha}(m,\, n-p)\hat{\sigma}^2\boldsymbol{b}^T A(\boldsymbol{X}^T\boldsymbol{X})^{-1}A^T\boldsymbol{b};\, \forall \boldsymbol{b} \in \mathbb{R}^m\right) = 1 - \alpha,$$

kde A je rozměrů $m \times p$ a b je rozměrů $m \times 1$. Pro případ kdy $b = (1,\, x,\, x^2)^T$ a $A = I_3$, kde I_3 je jednotková matice rozměrů 3×3, pak platí

$$P\left(\{(1,\, x,\, x^2)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^2 \leq 3F_{1-\alpha}(3,\, n-3)\hat{\sigma}^2(1,\, x,\, x^2)(\boldsymbol{X}^T\boldsymbol{X})^{-1}(1,\, x,\, x^2)^T\right) = 1 - \alpha.$$

Pás spolehlivosti odvozený z Scheffého věty s pravděpodobností pokrytí $100(1-\alpha)\,\%$ je pro tento speciální případ tvaru

$$P\Bigg(|(\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2) - (\beta_0 + \beta_1 x + \beta_2 x^2)| \leq \\ \leq \sqrt{3F_{1-\alpha}(3,\, n-3)\hat{\sigma}^2(1,\, x,\, x^2)(\boldsymbol{X}^T\boldsymbol{X})^{-1}(1,\, x,\, x^2)^T}\Bigg) = 1 - \alpha.$$

Příklad 2

V datovém souboru baseball_hit.Rdata jsou zaznamenány údaje o baseballových odpalech. Máme k dispozici proměnnou vzdálenost (distance), která udává horizontální vzdálenost (v metrech) mezi pálkařem a dopadem míčku, a proměnnou úhel (angle), jež zachycuje velikost úhlu (ve stupních) mezi trajektorií míčku a zemí při odpalu.

(a) Lineární model

Na základě vizuálního posouzení, koeficientu determinace a hodnot Akaikova informačního kritéria (AIC) jsme vybrali kvadratický model, který nejlépe popisuje data baseball_hit.Rdata. Tento kvadratický model vypadá následujícím způsobem

```
mod <- lm(distance ~ angle + I(angle^2), data = data)
```

(b) Scheffého pásy spolehlivosti

Definujeme si funkci ScheffePS1, která slouží k výpočtu Scheffého pásu spolehlivosti lineárního regresního modelu s pravděpodobností pokrytí 95 % v daném bodě.

```
ScheffePS1 <- function(b, A, mod, alpha){
  X <- model.matrix(mod)
  Y.hat <- predict(mod, newdata = data.frame(angle = b[2]))
  n <- dim(X)[1]
  m <- dim(A)[1]
  p <- dim(A)[2]
  M.alpha <- sqrt(m * qf(df1 = m, df2 = n - p, p = 1 - alpha))
  s <- summary(mod)$sigma
  result1 <- c(Y.hat - M.alpha * s *
                 sqrt(t(b) %*% A %*% solve(t(X) %*% X) %*% t(A) %*% b),
               Y.hat + M.alpha * s *
                 sqrt(t(b) %*% A %*% solve(t(X) %*% X) %*% t(A) %*% b))
  return(result1)
}
```

Tato funkce přijímá vektor regresních koeficientů b, maticovou reprezentaci kontrastů A, lineární regresní model mod a hladinu významnosti alpha. Výstupem funkce je dolní a horní mez Scheffého pásu spolehlivosti pro predikované hodnoty Y.

Dále si vytváříme vektor xx, který obsahuje 200 hodnot, které se rovnoměrně rozkládají mezi minimální a maximální hodnotou proměnné angle. Aplikujeme funkci ScheffePS1 na každou hodnotu vektoru xx a výsledek ukládáme do proměnné MatScheffe.

```
xx <- seq(from = min(data$angle), to = max(data$angle),
          length.out = 200)

MatScheffe <- sapply(1:length(xx), function(i){
  result2 <- ScheffePS1(mod = mod, A = diag(3),
                        b = c(1, xx[i], xx[i]^2),
                        alpha = 0.05)
  return(result2)
})
```

Tím jsme tedy získali hodnoty Scheffého pásu spolehlivosti pro náš model mod s pravděpodobností pokrytí 95 % (tyto hodnoty jsou uloženy v proměnné MatScheffe).

(c) Bodová a intervalová predikce

Predikujme (bodově i intervalově) vzdálenost dopadu míčku od pálkaře pokud odpálí míček pod úhlem 55° v R příkazem:

```
prediction <- predict(mod, newdata = data.frame(angle = 55),
                      interval = "prediction", level = 0.95)
```

Intervalová predikce nám vychází (76.5906, 82.47312), bodová pak 70.70808.

(d) Simultánní intervalový odhad

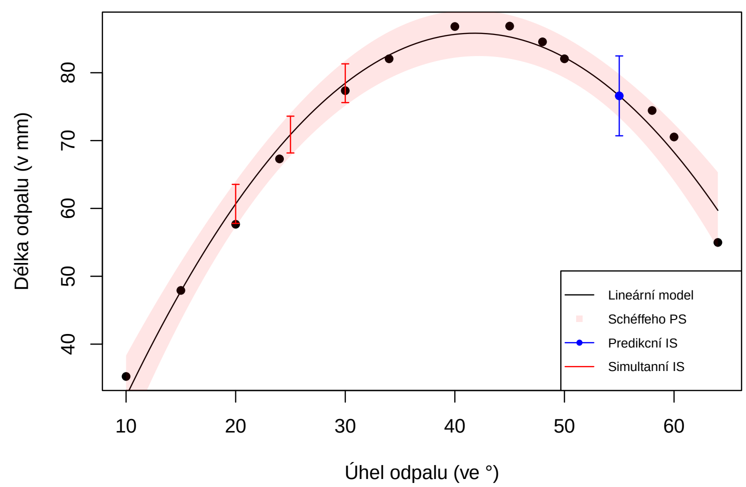
Pomocí následujících řádků kódu vypočteme simultánní intervalový odhad střední hodnoty délky odpalu pro úhly odpalu 20°, 25° a 30°. Přičemž volíme hladinu významnosti $\alpha = 0.05$ a využíváme Bonferroniho adjustace.

```
angles <- c(20, 25, 30)
alpha <- 0.05
bonferroni.alpha <- alpha / length(angles)
simultaneous.interval <- predict(mod, newdata =
                                   data.frame(angle = angles),
                                 interval = "confidence",
                                 level = 1 - bonferroni.alpha)
lower.bound.simultaneous <- simultaneous.interval[, "lwr"]
upper.bound.simultaneous <- simultaneous.interval[, "upr"]
```

Výsledné intervalové odhady nám vycházejí (1587.934, 1628.435) pro úhel odpalu 20°, (1688.657, 1711.751) pro úhel odpalu 25° a (1775.560, 1808.885) pro úhel odpalu 30°.

(e) Vykreslení výsledků

Nyní zakreslíme výsledky z předchozích bodů do jednoho obrázku. Tj. vykreslíme bodový graf pozorování, přidáme křivku našeho modelu, zvýrazníme pás spolehlivosti a požadované intervalové odhady i predikce a nakonec vykreslíme legendu.



Obrázek 1: Vykreslení výsledků v jednom grafu

Příklad 3

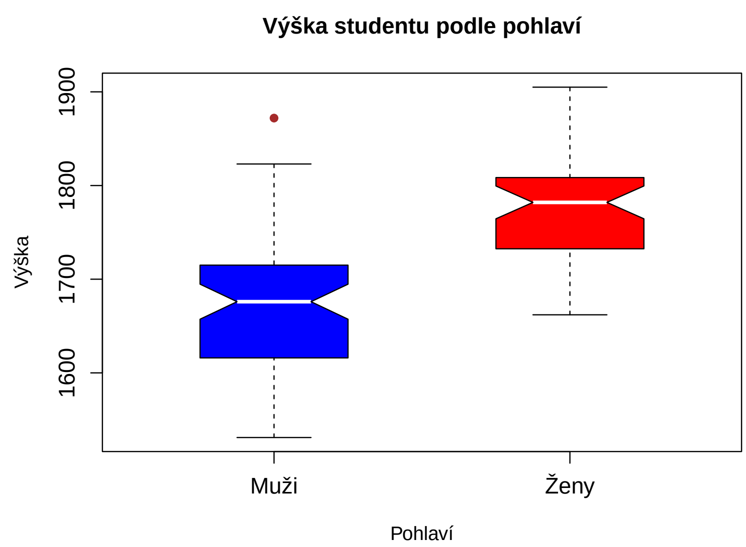
V tomto příkladě budeme pracovat s antropometrickými údaji studentů vysokých škol uloženými v souboru lrmfoot.txt. Data obsahují proměnné: pohlaví (sex), délku chodidla v milimetrech (foot.L) a tělesnou výšku v milimetrech (body.H). Bude nás zajímat efekt pohlaví na tělesnou výšku adjustovaný na délku chodidla.

(a) Čištění dat a vykreslení krabicových grafů

Provedeme čištění dat.

```
any(is.na(data)) # Chybejici hodnoty?
unique(data$sex) # Prvky mimo obor hodnot?
sort(unique(data$foot.L)) # Prvky mimo obor hodnot?
sort(unique(data$body.H)) # Prvky mimo obor hodnot?
```

Zjišťujeme, že data jsou čistá. Pokračujeme vykreslením krabicových diagramů popisujících výšku studentů v závislosti na pohlaví.



Obrázek 2: Vykreslení krabicových diagramů

(b) Výběr modelu

Budeme modelovat závislost střední hodnoty tělesné výšky na pohlaví a délce chodidla. Vyzkoušíme si různé varianty složitosti modelu: model se vzájemnou interakcí pohlaví a délky chodidla (1 – všeobecný, různé sklony přímek), model bez interakce (2 – ANCOVA, stejné sklony přímek) a model bez vlivu proměnné pohlaví (3 – jedna přímka). Modely jsme si v R definovali následovně

```
# Model s interakci
model1 <- lm(body.H ~ sex * foot.L, data = data)
summary(model1)

# Model bez interakce (ANCOVA)
model2 <- lm(body.H ~ sex + foot.L, data = data)
summary(model2)

# Model bez vlivu promenne foot.L
model3 <- lm(body.H ~ foot.L, data = data)
summary(model3)
```

Výstup z R posledních dvou řádků funkce summary aplikované na model1:

```
Multiple R-squared:  0.7252,    Adjusted R-squared:  0.7179
F-statistic: 99.42 on 3 and 113 DF,  p-value: < 2.2e-16
```

Výstup z R posledních dvou řádků funkce summary aplikované na model2:

```
Multiple R-squared:  0.7199,    Adjusted R-squared:  0.715
F-statistic: 146.5 on 2 and 114 DF,  p-value: < 2.2e-16
```

Výstup z R posledních dvou řádků funkce summary aplikované na model3:

```
Multiple R-squared:  0.715,     Adjusted R-squared:  0.7126
F-statistic: 288.6 on 1 and 115 DF,  p-value: < 2.2e-16
```

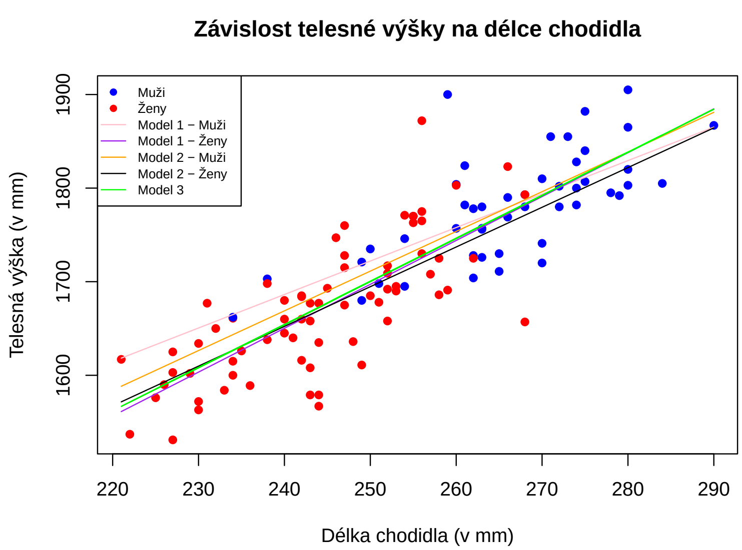
Pokud bychom měli vybrat nejlepší model pouze na základě těchto metrik, model 1 by mohl být mírně preferován, protože má nejvyšší hodnotu koeficientu determinace R^2. Nicméně je důležité zvážit také interpretovatelnost a relevantnost jednotlivých koeficientů, stejně jako další diagnostické testy.

Celkově lze říci, že model 1, model 2 a model 3 jsou v tomto případě velmi podobné a nelze jednoznačně rozhodnout, který z nich je nejlepší. Avšak kvůli interpretovatelnosti volíme model 3. Odhad vektoru parametrů β pro model 3 je:

```
> coefficients(model3)
(Intercept)      foot.L
 549.966107    4.600951
```

(c) Vykreslení modelů

Vykreslíme modely 1–3.



Obrázek 3: Vykreslení modelů v jednom grafu

(d) Vykreslení grafu a simultánní odhad

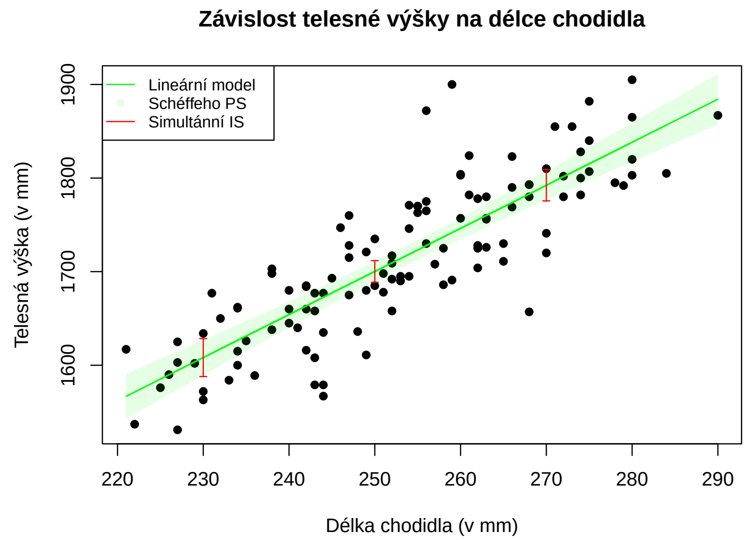
Vypočteme z dat minimální a maximální délku chodidla a v rámci tohoto rozsahu zkonstruujeme 95% Scheffého pás spolehlivosti pro model 3. V R provedeno následovně

```
xx <- seq(from = min.foot, to = max.foot, length.out = 200)

ScheffePS2 <- function(b, A, mod, alpha) {
  X <- model.matrix(mod)
  Y.hat <- predict(mod, newdata = data.frame(foot.L = b[2]))
  n <- dim(X)[1]
  m <- dim(A)[1]
  p <- dim(A)[2]
  M.alpha <- sqrt(m * qf(df1 = m, df2 = n - p, p = 1 - alpha))
  s~<- summary(mod)$sigma
  result1 <- c(Y.hat - M.alpha * s *
                 sqrt(t(b) %*% A %*% solve(t(X) %*% X) %*% t(A) %*% b),
               Y.hat + M.alpha * s *
                 sqrt(t(b) %*% A %*% solve(t(X) %*% X) %*% t(A) %*% b))
  return(result1)
}

MatScheffe <- sapply(1:length(xx), function(i) {
  result2 <- ScheffePS2(mod = model3, A = diag(2), b = c(1, xx[i]),
                        alpha = 0.05)
  return(result2)
})
```

Dále odhadněme střední hodnotu tělesné výšky pro jedince s délkou chodidla 230 mm, 250 mm a 270 mm pomocí 95% simultánních oboustranných intervalů spolehlivosti se Šidákovou adjustací. V R provedeno následovně

```
foot.lengths <- c(230, 250, 270)
alpha <- 0.05
alpha.adjusted <- (1 - (1 - alpha) ^ (1 / length(foot.lengths)))/2
simultaneous.interval <- predict(model3,
                                 newdata = data.frame(foot.L = foot.lengths),
                                 interval = "confidence",
                                 level = 1 - alpha.adjusted)
lower.bound.simultaneous <- simultaneous.interval[, "lwr"]
upper.bound.simultaneous <- simultaneous.interval[, "upr"]
```

Nakonec vytvoříme graf, který bude obsahovat všechna pozorování spolu s regresní přímkou odpovídající modelu 3. Do tohoto grafu zakreslíme námi vypočtený 95% Scheffého pás spolehlivosti a vypočtené 95% simultánní oboustranné intervaly spolehlivosti se Šidákovou adjustací.



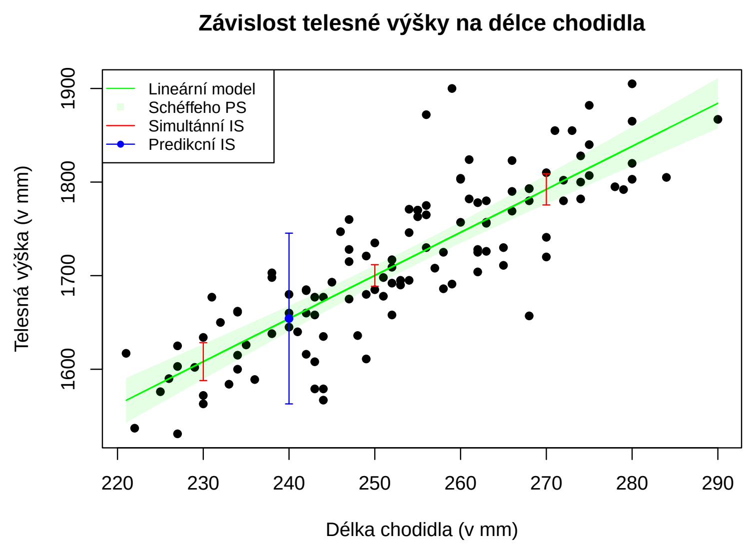
Obrázek 4: Vykreslení modelu, dat, Schefféo pásu a intervalů spolehlivosti

(e) Predikce výšky

Pomocí 95% intervalu spolehlivosti predikujeme výšku jedince, jenž má délku chodidla 240 mm. To uděláme v R příkazem

```
prediction <- predict(model3, newdata = data.frame(foot.L = 240),
                       interval = "prediction", level = 0.95)
```

Tento interval je: (1654.194, 1745.393). Nakonec znázorníme tento interval do obrázku 4.



Obrázek 5: Vykreslení predikce