

## 7. Binární logistická regrese

Zdeňka Geršlová, Markéta Janošová

2023-03-23

### Příklad 1

V souboru `head.txt` máme k dispozici antropometrické údaje mladých dospělých lidí (převážně studentů VŠ z Brna a Ostravy). Známe také pohlaví zaznamenaných jedinců (proměnná `sex`). Sestrojte model, který na základě tělesné výšky (proměnná `body.H`), délky hlavy (proměnná `head.L`), šířky hlavy (proměnná `head.W`), šířky dolní čelisti (proměnná `bigo.W`) a šířky obličeje (proměnná `bizyg.W`) určí pravděpodobnost, že neznámý případ je muž. Všechny rozměry byly měřeny v milimetrech.

Načteme datový soubor a zkontrolujeme, že R pracuje s proměnnou pohlaví jako s faktorem (popř. použijeme argument `stringsAsFactor = T`). *Pozn: Pokud by byla proměnná v datovém souboru kódována např. pomocí 0 a 1, tak by s ní R pracovalo jako s numerickou proměnnou, nikoli kategoriální. V takovém případě bychom ji museli změnit na kategoriální pomocí funkce `factor()`.*

```
head <- read.table("head.txt", header = T, stringsAsFactors = T)
summary(head)
```

```
## sex          body.H          head.L          head.W          bigo.W
## f:100  Min.   :1531  Min.   :170.0  Min.   :135.0  Min.   : 90.0
## m: 75   1st Qu.:1650  1st Qu.:183.5  1st Qu.:145.0  1st Qu.: 99.0
##          Median :1717  Median :189.0  Median :151.0  Median :102.0
##          Mean   :1720  Mean   :189.7  Mean   :150.7  Mean   :103.7
##          3rd Qu.:1788  3rd Qu.:195.5  3rd Qu.:155.0  3rd Qu.:108.0
##          Max.   :1906  Max.   :214.0  Max.   :170.0  Max.   :126.0
##      bizyg.W
## Min.   :113.0
## 1st Qu.:131.0
## Median :136.0
## Mean   :136.4
## 3rd Qu.:142.0
## Max.   :155.0
```

### Výpočet základních charakteristik, ověřování předpokladů

Než budeme sestavovat model, je vhodné se podívat na vztah mezi vysvětlovanou (závislou) veličinou (v našem případě je to pohlaví) a každou vysvětlující veličinou zvlášť. Nejprve zkontrolujeme, že datový soubor neobsahuje chybějící hodnoty (případně inkriminované řádky odstraníme). Vypočítáme rozsahy, výběrové průměry a výběrové směrodatné odchylky všech uvažovaných veličin zvlášť pro každé pohlaví. Abychom nemuseli vše psát ručně, vytvoříme si funkci, která nám tyto hodnoty vypočítá, a použijeme ji jako vstup do funkce `tapply()` pro současný výpočet charakteristik vybrané veličiny rozdělený podle pohlaví. Zároveň si vykreslíme krabicové diagramy.

```
charakteristiky <- function(x){
  # funkce počítající počet pozorování, prumer a smerodatnou odchylku
```

```

# argument: x ... vektor
# vraci: vektor (pocet pozorovani, prumer, smerodatna odchylka)
n <- length(x)
m <- mean(x)
s <- sd(x)
return(c(n, m, s))
}
tapply(head$body.H, head$sex, charakteristiky)

## $f
## [1] 100.00000 1667.33000 67.20811
##
## $m
## [1] 75.00000 1789.72000 59.70639
tapply(head$head.L, head$sex, charakteristiky)

## $f
## [1] 100.00000 185.01000 6.545096
##
## $m
## [1] 75.00000 195.946667 6.970776
tapply(head$head.W, head$sex, charakteristiky)

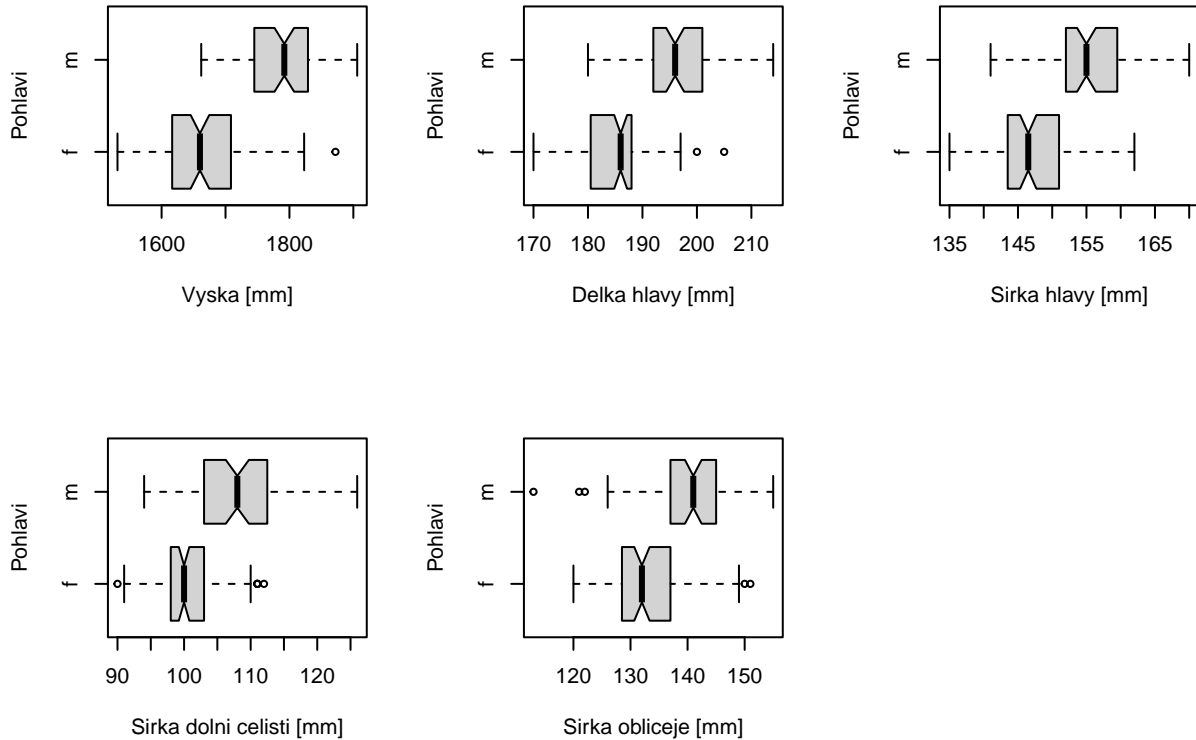
## $f
## [1] 100.00000 146.92000 5.336514
##
## $m
## [1] 75.00000 155.653333 6.081637
tapply(head$bigo.W, head$sex, charakteristiky)

## $f
## [1] 100.00000 100.57000 4.69957
##
## $m
## [1] 75.00000 107.813333 6.872769
tapply(head$bizyg.W, head$sex, charakteristiky)

## $f
## [1] 100.00000 133.46000 6.110795
##
## $m
## [1] 75.00000 140.293333 7.714103
par(mfrow=c(2,3))
boxplot(head$body.H ~ head$sex, varwidth=T, notch=T, xlab="Vyska [mm]",
        ylab="Pohlavi", horizontal=T)
boxplot(head$head.L ~ head$sex, varwidth=T, notch=T, xlab="Delka hlavy [mm]",
        ylab="Pohlavi", horizontal=T)
boxplot(head$head.W ~ head$sex, varwidth=T, notch=T, xlab="Sirka hlavy [mm]",
        ylab="Pohlavi", horizontal=T)
boxplot(head$bigo.W ~ head$sex, varwidth=T, notch=T, xlab="Sirka dolni celisti [mm]",
        ylab="Pohlavi", horizontal=T)
boxplot(head$bizyg.W ~ head$sex, varwidth=T, notch=T, xlab="Sirka obliceje [mm]",

```

```
ylab="Pohlavi", horizontal=T)
```

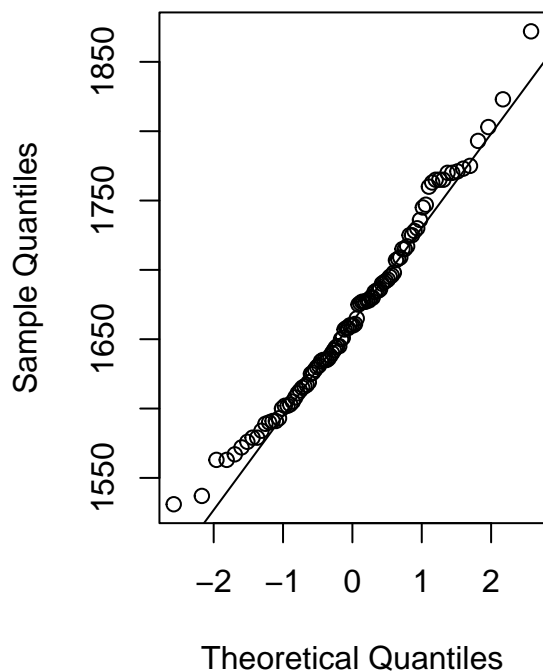


Pomocí t-testu otestujeme hypotézy, že muži a ženy se v jednotlivých mírách liší. Nezapomeneme ověřit předpoklady. *Pozn.: Jaké výsledky testu očekáváte na základě krabicových diagramů?*

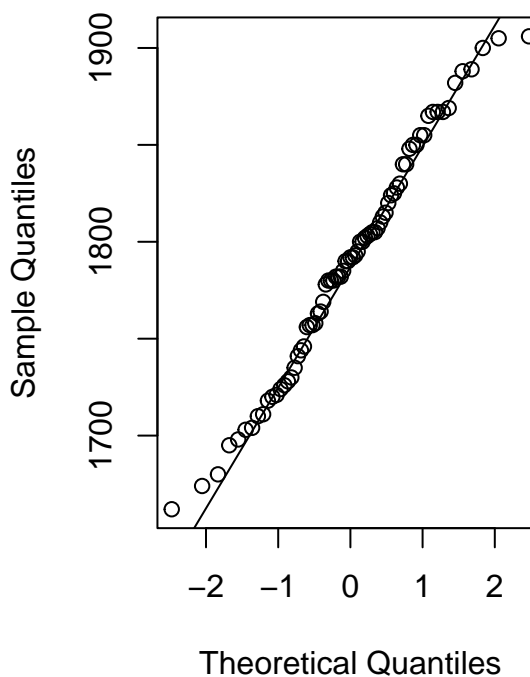
```
shapiro.test(head$body.H[head$sex=='f'])
##
## Shapiro-Wilk normality test
##
## data: head$body.H[head$sex == "f"]
## W = 0.98171, p-value = 0.1803
shapiro.test(head$body.H[head$sex=='m'])
##
## Shapiro-Wilk normality test
##
## data: head$body.H[head$sex == "m"]
## W = 0.98364, p-value = 0.445
```

```
par(mfrow=c(1,2))
qqnorm(head$body.H[head$sex=='f'], main='Telesna vyska - zeny')
qqline(head$body.H[head$sex=='f'])
qqnorm(head$body.H[head$sex=='m'], main='Telesna vyska - muzi')
qqline(head$body.H[head$sex=='m'])
```

Telesna vyska – zeny



Telesna vyska – muzi



```
var.test(head$body.H ~ head$sex)
```

```
##  
## F test to compare two variances  
##  
## data: head$body.H by head$sex  
## F = 1.2671, num df = 99, denom df = 74, p-value = 0.2854  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.819648 1.932581  
## sample estimates:  
## ratio of variances  
## 1.267073
```

```
t.test(head$body.H ~ head$sex)
```

```
##  
## Welch Two Sample t-test  
##  
## data: head$body.H by head$sex  
## t = -12.712, df = 168.04, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group f and group m is not equal to 0  
## 95 percent confidence interval:  
## -141.3977 -103.3823  
## sample estimates:  
## mean in group f mean in group m  
## 1667.33 1789.72
```

Vidíme, že pro tělesnou výšku Shapiro-Wilkův test nezamítá normalitu ani pro muže, ani pro ženy, což nám potvrzují i kvantil-kvantilové grafy. F-test nezamítl hypotézu o rovnosti rozptylů, zatímco výsledky t-testu ukazují, že existuje rozdíl v průměrné tělesné výšce mezi muži a ženami. Podobně sestavíme a vyhodnotíme testy také pro délku a šířku hlavy (proměnné head.L a head.W). U šířky čelisti žen Shapiro-Wilkův test zamítl hypotézu, že data pocházejí z normálního rozdělení. Máme ale dost pozorování a v kvantil-kvantilovém grafu není odchýlení od přímky velké. Budeme tedy tento předpoklad považovat za splněný. Dále u šířky čelisti F-test zamítá hypotézu o rovnosti rozptylů žen a mužů, použijeme proto variantu t-testu pro nestejně rozptyly (nastavíme argument `var.equal=F`).

```
shapiro.test(head$bigo.W[head$sex=='f'])
```

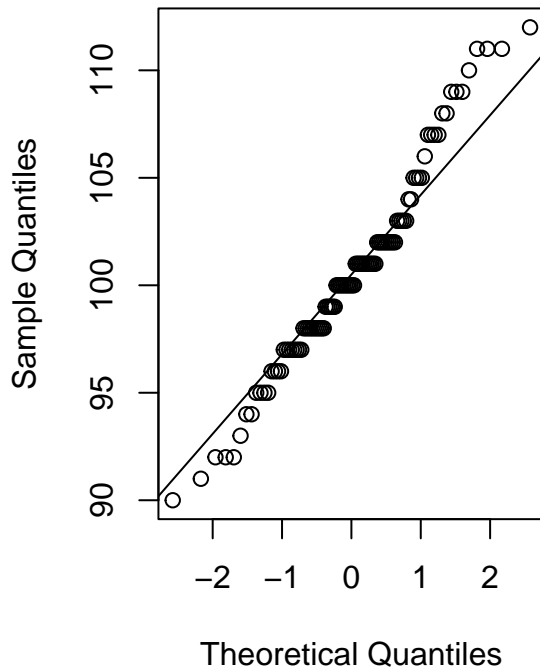
```
##  
## Shapiro-Wilk normality test  
##  
## data: head$bigo.W[head$sex == "f"]  
## W = 0.97377, p-value = 0.04327
```

```
shapiro.test(head$bigo.W[head$sex=='m'])
```

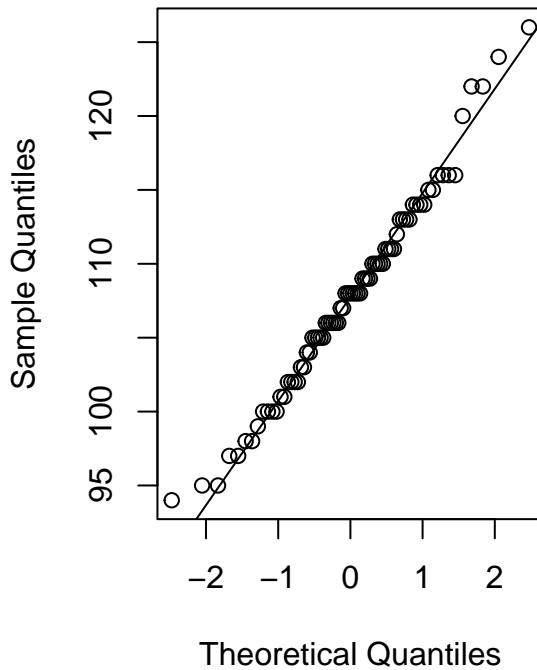
```
##  
## Shapiro-Wilk normality test  
##  
## data: head$bigo.W[head$sex == "m"]  
## W = 0.98478, p-value = 0.5078
```

```
par(mfrow=c(1,2))  
qqnorm(head$bigo.W[head$sex=='f'], main='Sirka celisti - zeny')  
qqline(head$bigo.W[head$sex=='f'])  
qqnorm(head$bigo.W[head$sex=='m'], main='Sirka celisti - muzi')  
qqline(head$bigo.W[head$sex=='m'])
```

**Sirka celisti – zeny**



**Sirka celisti – muzi**



```
var.test(head$bigo.W ~ head$sex)
```

```
##  
## F test to compare two variances  
##  
## data: head$bigo.W by head$sex  
## F = 0.46758, num df = 99, denom df = 74, p-value = 0.0004304  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.3024672 0.7131630  
## sample estimates:  
## ratio of variances  
## 0.4675766
```

```
t.test(head$bigo.W ~ head$sex, var.equal=F)
```

```
##  
## Welch Two Sample t-test  
##  
## data: head$bigo.W by head$sex  
## t = -7.8535, df = 123.64, p-value = 1.667e-12  
## alternative hypothesis: true difference in means between group f and group m is not equal to 0  
## 95 percent confidence interval:  
## -9.068900 -5.417767  
## sample estimates:  
## mean in group f mean in group m  
## 100.5700 107.8133
```

U šířky obličejů žen i mužů Shapiro-Wilkův test zamítl hypotézu, že data pocházejí z normálního rozdělení. Na základě množství dat a grafického posouzení budeme i zde tento předpoklad považovat za splněný. Opět máme problém i s homogenitou rozptylu, proto použijeme variantu t-testu pro nestejné rozptyly (nastavíme argument `var.equal=F`).

```
par(mfrow=c(1,2))
shapiro.test(head$bizyg.W[head$sex=='f'])

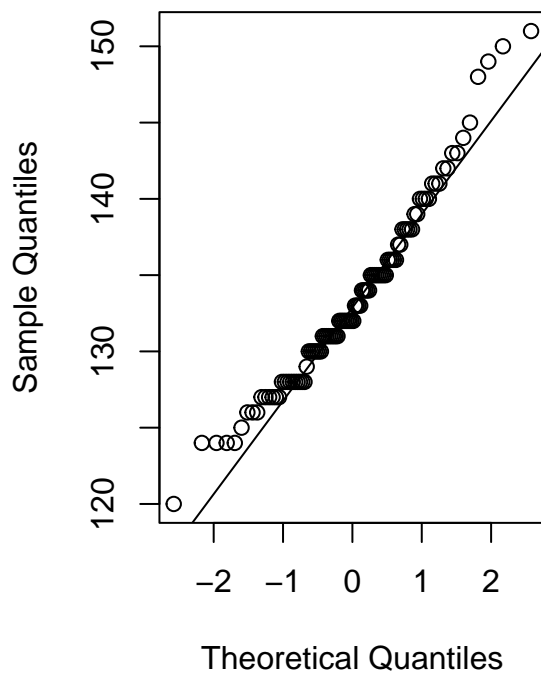
##
## Shapiro-Wilk normality test
##
## data: head$bizyg.W[head$sex == "f"]
## W = 0.96672, p-value = 0.01247

shapiro.test(head$bizyg.W[head$sex=='m'])

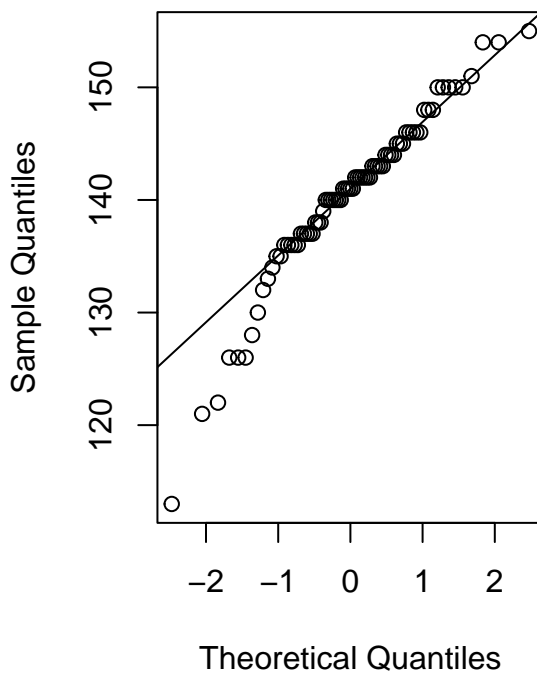
##
## Shapiro-Wilk normality test
##
## data: head$bizyg.W[head$sex == "m"]
## W = 0.9457, p-value = 0.002916

qqnorm(head$bizyg.W[head$sex=='f'], main='Sirka obliceje - zeny')
qqline(head$bizyg.W[head$sex=='f'])
qqnorm(head$bizyg.W[head$sex=='m'], main='Sirka obliceje - muzi')
qqline(head$bizyg.W[head$sex=='m'])
```

**Sirka obliceje – ženy**



**Sirka obliceje – muži**



```
var.test(head$bizyg.W ~ head$sex)
```

```
##  
## F test to compare two variances  
##  
## data: head$bizyg.W by head$sex  
## F = 0.62752, num df = 99, denom df = 74, p-value = 0.03054  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.4059291 0.9571072  
## sample estimates:  
## ratio of variances  
## 0.6275157
```

```
t.test(head$bizyg.W ~ head$sex, var.equal=F)
```

```
##  
## Welch Two Sample t-test  
##  
## data: head$bizyg.W by head$sex  
## t = -6.3259, df = 137.31, p-value = 3.309e-09  
## alternative hypothesis: true difference in means between group f and group m is not equal to 0  
## 95 percent confidence interval:  
## -8.969328 -4.697339  
## sample estimates:  
## mean in group f mean in group m  
## 133.4600 140.2933
```

## Sestavení modelu

U všech vysvětlujících proměnných se prokázaly rozdíly mezi muži a ženami, při sestavování modelu logistické regrese použijeme pro začátek všechny. Model vytvoříme pomocí funkce `glm` s nastavením argumentu `family=binomial(logit)`.

```
m.head <- glm(sex ~ body.H + head.L + head.W + bigo.W + bizyg.W, family=binomial(logit), data=head)
```

Vypíšeme si podrobné informace o modelu.

```
summary(m.head)
```

```
##  
## Call:  
## glm(formula = sex ~ body.H + head.L + head.W + bigo.W + bizyg.W,  
## family = binomial(logit), data = head)  
##  
## Deviance Residuals:  
## Min 1Q Median 3Q Max  
## -1.86737 -0.25202 -0.04043 0.19981 2.92685  
##  
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.086e+02 1.797e+01 -6.045 1.49e-09 ***  
## body.H 2.180e-02 5.302e-03 4.112 3.92e-05 ***  
## head.L 1.658e-01 4.802e-02 3.453 0.000554 ***  
## head.W 2.700e-01 8.503e-02 3.175 0.001499 **
```



```
## bigo.W      1.340e-01  5.920e-02  2.264 0.023578 *
## bizyg.W     -1.150e-01  5.939e-02 -1.937 0.052773 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 239.018 on 174 degrees of freedom
## Residual deviance: 81.205 on 169 degrees of freedom
## AIC: 93.205
##
## Number of Fisher Scoring iterations: 7
```

Abychom mohli provést celkový test významnosti modelu, sestavíme model konstanty a tyto dva modely spolu porovnáme s využitím funkce `anova`.

```
m0 <- glm(sex ~ 1, family=binomial(logit), data=head)
anova(m0, m.head, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sex ~ 1
## Model 2: sex ~ body.H + head.L + head.W + bigo.W + bizyg.W
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         174     239.018
## 2         169      81.205  5   157.81 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hodnota testovací statistiky .....  
p-hodnota .....  
Závěr .....

Protože jsme na hladině významnosti 0.05 zamítli hypotézu, že dostačující je model konstanty, zkusíme z maximálního modelu vynechat proměnné, které dílčí testy ukazují jako nevýznamné. Vynecháme tedy šířku obličejů.

```
m.head2 <- glm(sex ~ body.H + head.L + head.W + bigo.W, family=binomial(logit), data=head)
anova(m.head2, m.head, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: sex ~ body.H + head.L + head.W + bigo.W
## Model 2: sex ~ body.H + head.L + head.W + bigo.W + bizyg.W
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         170      85.283
## 2         169      81.205  1   4.0788 0.04343 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hodnota testovací statistiky .....  
p-hodnota .....  
Závěr .....

Vynechání šířky obličejů tedy nevede k lepšímu modelu (potvrzuje to i vyšší hodnota Akaikova informačního kritéria modelu `m.head2`, kterou zjistíme vypsáním `summary`), zůstaneme proto u maximálního modelu. Podíváme se na odhadnuté parametry a vypočítáme pro ně intervaly spolehlivosti.

```
coef(m.head)

## (Intercept)      body.H      head.L      head.W      bigo.W
## -108.63634714  0.02179998  0.16581097  0.26995038  0.13403001
##      bizyg.W
##      -0.11502670
```

```
lower <- coef(m.head) - qnorm(0.975) * summary(m.head)$coefficients[,2]
upper <- coef(m.head) + qnorm(0.975) * summary(m.head)$coefficients[,2]
cbind(lower,upper)
```

```
##              lower      upper
## (Intercept) -143.85962875 -73.413065533
## body.H      0.01140885  0.032191111
## head.L     0.07170056  0.259921384
## head.W     0.10330082  0.436599940
## bigo.W     0.01799600  0.250064019
## bizyg.W    -0.23143071  0.001377304
```

Lépe se ale interpretují hodnoty  $e^{\beta_i}$ , které udávají změnu šancí na úspěch při jednotkové změně i-té vysvětlující proměnné (za předpokladu, že ostatní vysvětlující proměnné zůstávají stejné). Při interpretaci je potřeba mít na paměti, kterou kategorii bere R jako referenční, v našem případě jsou referenční skupinou ženy. *Pozn.: analogicky k hodnotám koeficientů musíme transformovat i konfidenční intervaly.*

```
exp(coef(m.head))
```

```
## (Intercept)      body.H      head.L      head.W      bigo.W      bizyg.W
## 6.604408e-48  1.022039e+00  1.180350e+00  1.309899e+00  1.143427e+00  8.913423e-01
```

```
exp(cbind(lower,upper))
```

```
##              lower      upper
## (Intercept) 3.330865e-63  1.309516e-32
## body.H     1.011474e+00  1.032715e+00
## head.L     1.074334e+00  1.296828e+00
## head.W     1.108825e+00  1.547437e+00
## bigo.W     1.018159e+00  1.284108e+00
## bizyg.W    7.933977e-01  1.001378e+00
```

Například hodnota  $e^{\beta_3}$ , která se vztahuje k šířce hlavy, znamená, že pokud se o 1 mm zvětší šířka hlavy, šance, že pozorování patří muži, se zvýší 1.31-krát.

## STEPWISE procedury

Pro výběr modelu můžeme použít i STEPWISE proceduru, obdobně jako v případě lineárního regresního modelu.

```
step(glm(sex ~ body.H + head.L + head.W + bigo.W + bizyg.W, family=binomial(logit), data=head),
direction='backward')
```

```
## Start:  AIC=93.2
## sex ~ body.H + head.L + head.W + bigo.W + bizyg.W
##
##           Df Deviance    AIC
## <none>      81.205  93.205
## - bizyg.W   1   85.283  95.283
## - bigo.W    1   87.350  97.350
```

```

## - head.W 1 94.218 104.218
## - head.L 1 96.129 106.129
## - body.H 1 108.208 118.208

##
## Call: glm(formula = sex ~ body.H + head.L + head.W + bigo.W + bizyg.W,
## family = binomial(logit), data = head)
##
## Coefficients:
## (Intercept) body.H head.L head.W bigo.W bizyg.W
## -108.6363 0.0218 0.1658 0.2700 0.1340 -0.1150
##
## Degrees of Freedom: 174 Total (i.e. Null); 169 Residual
## Null Deviance: 239
## Residual Deviance: 81.2 AIC: 93.2

step(glm(sex ~ 1, family=binomial(logit), data=head),
scope= ~ body.H + head.L + head.W + bigo.W + bizyg.W, direction='forward')

## Start: AIC=241.02
## sex ~ 1
##
## Df Deviance AIC
## + body.H 1 133.46 137.46
## + head.L 1 151.75 155.75
## + head.W 1 158.70 162.70
## + bigo.W 1 181.64 185.64
## + bizyg.W 1 200.88 204.88
## <none> 239.02 241.02
##
## Step: AIC=137.46
## sex ~ body.H
##
## Df Deviance AIC
## + head.W 1 106.25 112.25
## + head.L 1 109.34 115.34
## + bigo.W 1 110.10 116.10
## + bizyg.W 1 125.36 131.36
## <none> 133.46 137.46
##
## Step: AIC=112.25
## sex ~ body.H + head.W
##
## Df Deviance AIC
## + head.L 1 89.378 97.378
## + bigo.W 1 98.429 106.429
## <none> 106.252 112.252
## + bizyg.W 1 105.258 113.258
##
## Step: AIC=97.38
## sex ~ body.H + head.W + head.L
##
## Df Deviance AIC
## + bigo.W 1 85.283 95.283
## + bizyg.W 1 87.350 97.350

```

```

## <none>          89.378 97.378
##
## Step: AIC=95.28
## sex ~ body.H + head.W + head.L + bigo.W
##
##           Df Deviance   AIC
## + bizyg.W  1   81.205 93.205
## <none>     1   85.283 95.283
##
## Step: AIC=93.2
## sex ~ body.H + head.W + head.L + bigo.W + bizyg.W
##
## Call: glm(formula = sex ~ body.H + head.W + head.L + bigo.W + bizyg.W,
##           family = binomial(logit), data = head)
##
## Coefficients:
## (Intercept)      body.H      head.W      head.L      bigo.W      bizyg.W
## -108.6363      0.0218      0.2700      0.1658      0.1340     -0.1150
##
## Degrees of Freedom: 174 Total (i.e. Null); 169 Residual
## Null Deviance:      239
## Residual Deviance: 81.2 AIC: 93.2
step(glm(sex ~ body.H + head.L + head.W + bigo.W + bizyg.W, family=binomial(logit), data=head),
direction='both')

## Start: AIC=93.2
## sex ~ body.H + head.L + head.W + bigo.W + bizyg.W
##
##           Df Deviance   AIC
## <none>     1   81.205 93.205
## - bizyg.W  1   85.283 95.283
## - bigo.W   1   87.350 97.350
## - head.W   1   94.218 104.218
## - head.L   1   96.129 106.129
## - body.H   1  108.208 118.208
##
## Call: glm(formula = sex ~ body.H + head.L + head.W + bigo.W + bizyg.W,
##           family = binomial(logit), data = head)
##
## Coefficients:
## (Intercept)      body.H      head.L      head.W      bigo.W      bizyg.W
## -108.6363      0.0218      0.1658      0.2700      0.1340     -0.1150
##
## Degrees of Freedom: 174 Total (i.e. Null); 169 Residual
## Null Deviance:      239
## Residual Deviance: 81.2 AIC: 93.2
step(glm(sex ~ 1, family=binomial(logit), data=head),
scope= ~ body.H + head.L + head.W + bigo.W + bizyg.W, direction='both')

## Start: AIC=241.02
## sex ~ 1
##

```

```

##           Df Deviance    AIC
## + body.H   1   133.46 137.46
## + head.L   1   151.75 155.75
## + head.W   1   158.70 162.70
## + bigo.W   1   181.64 185.64
## + bizyg.W  1   200.88 204.88
## <none>     239.02 241.02
##
## Step:  AIC=137.46
## sex ~ body.H
##
##           Df Deviance    AIC
## + head.W   1   106.25 112.25
## + head.L   1   109.34 115.34
## + bigo.W   1   110.10 116.10
## + bizyg.W  1   125.36 131.36
## <none>     133.46 137.46
## - body.H   1   239.02 241.02
##
## Step:  AIC=112.25
## sex ~ body.H + head.W
##
##           Df Deviance    AIC
## + head.L   1   89.378 97.378
## + bigo.W   1   98.429 106.429
## <none>     106.252 112.252
## + bizyg.W  1  105.258 113.258
## - head.W   1  133.457 137.457
## - body.H   1  158.696 162.696
##
## Step:  AIC=97.38
## sex ~ body.H + head.W + head.L
##
##           Df Deviance    AIC
## + bigo.W   1   85.283 95.283
## + bizyg.W  1   87.350 97.350
## <none>     89.378 97.378
## - head.L   1  106.252 112.252
## - head.W   1  109.342 115.342
## - body.H   1  115.986 121.986
##
## Step:  AIC=95.28
## sex ~ body.H + head.W + head.L + bigo.W
##
##           Df Deviance    AIC
## + bizyg.W  1   81.205 93.205
## <none>     85.283 95.283
## - bigo.W   1   89.378 97.378
## - head.W   1   94.274 102.274
## - head.L   1   98.429 106.429
## - body.H   1  111.026 119.026
##
## Step:  AIC=93.2
## sex ~ body.H + head.W + head.L + bigo.W + bizyg.W

```

```
##
##           Df Deviance      AIC
## <none>      81.205  93.205
## - bizyg.W  1  85.283  95.283
## - bigo.W   1  87.350  97.350
## - head.W   1  94.218 104.218
## - head.L   1  96.129 106.129
## - body.H   1 108.208 118.208

##
## Call:  glm(formula = sex ~ body.H + head.W + head.L + bigo.W + bizyg.W,
##           family = binomial(logit), data = head)
##
## Coefficients:
## (Intercept)      body.H      head.W      head.L      bigo.W      bizyg.W
## -108.6363      0.0218      0.2700      0.1658      0.1340     -0.1150
##
## Degrees of Freedom: 174 Total (i.e. Null);  169 Residual
## Null Deviance:      239
## Residual Deviance: 81.2  AIC: 93.2
```

Ve všech případech jsme došli ke stejnému modelu. Pro hodnocení kvality modelu si vypíšeme hodnoty koeficientu determinace, použijeme funkci knihovny `rsq`.

```
library(rsq)
rsq(m.head, type='n') # nagelkerke
## [1] 0.7977113
rsq(m.head, type='kl') # mcfadden
## [1] 0.6602573
rsq(m.head, type='lr') # cox and snell
## [1] 0.5941575
```

Dále sestavíme klasifikační tabulku, která nám ukáže počty správně a nesprávně zařazených subjektů. Nejprve musíme na základě odhadnutých pravděpodobností odhadnout, která pozorování patří mužům a která ženám. Jako dělicí bod zvolíme hodnotu 0.5.

```
fitted <- predict(m.head, newdata=head, type="response")
fitted.cat <- ifelse(fitted < 0.5, "f", "m")
tab <- table(fitted.cat, head$sex)
tab
```

```
##
## fitted.cat  f  m
##           f 91  9
##           m  9 66
```

Z tabulky můžeme vypočítat relativní četnost správně zařazených pozorování.

```
sum(diag(tab))/sum(tab)
```

```
## [1] 0.8971429
```

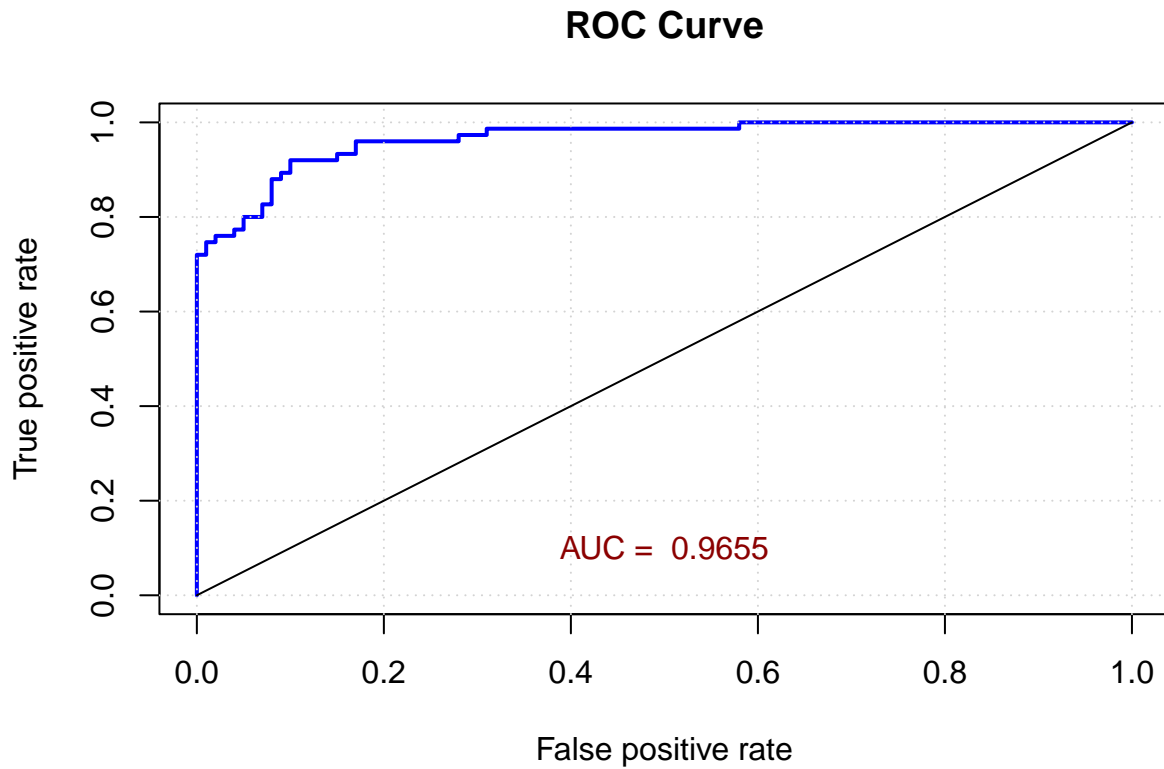
Dalším nástrojem pro hodnocení kvality modelu je ROC křivka a hodnota AUC (area under the curve - plocha pod křivkou). Funkce pro výpočet ROC křivky je součástí knihovny `ROCR`.

```
library(ROCR)
preds <- prediction(fitted, as.numeric(head$sex))
roc <- performance(preds, "tpr", "fpr")
auc <- performance(preds, "auc")
```

```

auc.value <- round(as.numeric(auc@y.values),4)
plot(roc, main="ROC Curve", lwd=2, col="blue")
grid()
lines(c(0,1),c(0,1))
text(0.5, 0.1, paste("AUC = ",auc.value), col="darkred")

```



Negelkerkův koeficient nabývá hodnoty ....., úspěšnost správné klasifikace je ..... a hodnota AUC je ....., můžeme tedy soudit, že .....