Vzorový příklad na binární logistickou regresi

V souboru head.txt máme k dispozici antropometrické údaje 175 mladých dospělých lidí (převážně studentů vysokých škol z Brna a Ostravy). Známe také pohlaví zaznamenaných jedinců (proměnná sex). Sestrojte model, který na základě tělesné výšky (proměnná body.H), délky hlavy (head.L), šířky hlavy (head.W), šířky dolní čelisti (bigo.W) a šířky obličeje (bizyg.W) odhadne pravděpodobnost, že neznámý případ je muž. Všechny rozměry byly měřeny v milimetrech.
Načteme datový soubor a podíváme se na jeho číselné charakteristiky:

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

Zkontrolujeme, že R pracuje s proměnnou pohlaví jako s faktorem. Pokud by byla v datovém souboru kódována například 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().

```
is.factor(head$sex)
[1] TRUE
```

Než budeme sestavovat model, je vhodné se podívat na vztah mezi vysvětlovanou veličinou (v našem případě je to pohlaví) a každou vysvětlující veličinou zvlášť. Vypočítáme rozsahy, výběrové průměry a výběrové směrodatné odchylky všech veličin pro každé pohlaví zvlášť. Abychom nemuseli vše psát ručně, vytvoříme si funkci, která nám tyto hodnoty poskytne. Zároveň si vykreslíme krabicové diagramy.

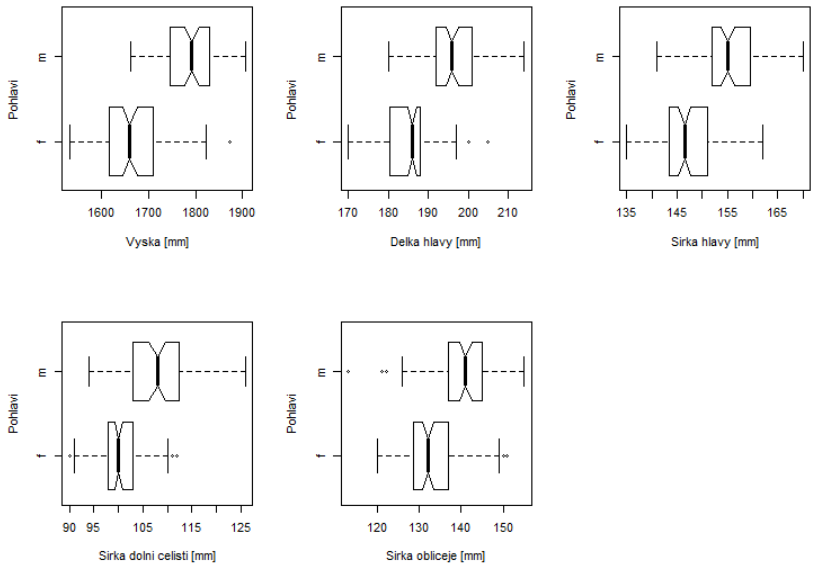
```
charakteristiky <- function(x){
+     # funkce pocitajici pocet pozorovani, prumer a smerodatnou odchylku
+     # argument: x ... vektor
+     # vraci: vektor (pocet pozorovani, prumer, smerodatna odchylka)
+     x <- na.omit(x) #odstrani chybejici hodnoty
+     n <- length(x)
+     m <- mean(x)
+     s <- sd(x)
+     return(c(n, m, s))
+ }

charakteristiky(head$body.H[head$sex=='f'])
[1]  100.00000 1667.33000   67.20811
charakteristiky(head$body.H[head$sex=='m'])
[1]   75.00000 1789.72000   59.70639
charakteristiky(head$head.L[head$sex=='f'])
[1] 100.000000 185.010000   6.545096
charakteristiky(head$head.L[head$sex=='m'])
[1]  75.000000 195.946667   6.970776
charakteristiky(head$head.W[head$sex=='f'])
[1] 100.000000 146.920000   5.336514
charakteristiky(head$head.W[head$sex=='m'])
[1]  75.000000 155.653333   6.081637
charakteristiky(head$bigo.W[head$sex=='f'])
[1] 100.00000 100.57000   4.69957
charakteristiky(head$bigo.W[head$sex=='m'])
[1]  75.000000 107.813333   6.872769
charakteristiky(head$bizyg.W[head$sex=='f'])
[1] 100.000000 133.460000   6.110795
charakteristiky(head$bizyg.W[head$sex=='m'])
[1]  75.000000 140.293333   7.714103
```

		Tělesná výška (mm)		Délka hlavy (mm)		Šířka hlavy (mm)	
	rozsah	průměr	sm. odchylka	průměr	sm. odchylka	průměr	sm. odchylka
Ženy							
Muži							

		Šířka dolní čelisti (mm)		Šířka obličeje (mm)	
	rozsah	průměr	sm. odchylka	průměr	sm. odchylka
Ženy					
Muži					

Vykreslíme krabicové diagramy:

```
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)
```



S-W testem ověříme, že všechny proměnné se ve skupinách mužů a žen řídí normálním rozložením. Např. pro tělesnou výšku žen použijeme příkaz

```
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
```

Zjistíme, že normalita je lehce porušena u proměnné bizyg.W u žen i mužů a u žen navíc ještě u proměnné bigo.W. Vzhledem k dostatečně velkým rozsahům výběrů budeme i s těmito proměnnými zacházet jako s normálně rozloženými proměnnými.

Před provedením dvouvýběrových t-testů, kterými budeme ověřovat, zda se skupiny mužů a žen liší ve středních hodnotách sledovaných pěti proměnných, musíme ještě ve všech pěti případech ověřit shodu rozptylů pomocí F-testu. Provedení F-testu pro proměnnou body.H:

```
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
```

Zjistíme, že u proměnných bigo.W a bizyg.W je na hladině významnosti 0,05 zamítnuta hypotéza o shodě rozptylů. Použijeme proto dvouvýběrový t-test s Welchovou aproximací. Provedení dvouvýběrového t-testu pro proměnnou body.H:

```
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 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
```

Zjistíme, že u všech proměnných je na hladině významnosti 0,05 zamítnuta hypotéza o shodě středních hodnot. Proto do modelu logistické regrese zahrneme na začátku všechny proměnné.

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

Vypíšeme si 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 nevýznamnosti modelu M1, potřebujeme sestavit model konstanty M0, který s ním budeme srovnávat.

```
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
```

Deviance nulového modelu ……., deviance modelu M1 ……

Hodnota testové statistiky pro test, že model konstanty je dostatečný =
p-hodnota =
závěr:

Proměnná bizyg.W není v našem modelu významná na hladině významnosti 0,05 (její p-hodnota je 0,0528). Zkusíme ji tedy z modelu vynechat a posoudíme, zda model M2 je lepší než model M1.

```
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 testové statistiky pro test, že model M2 je dostatečný =
p-hodnota =
závěr:

Budeme tedy dále pracovat s modelem M1. Podíváme se na odhady jeho parametrů a sestrojíme pro ně 95% intervaly spolehlivosti:

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

Výpočet mezí intervalů spolehlivosti:

```
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^{β_j}, tj. podíly šancí. Při interpretaci je potřeba mít na paměti, kterou kategorii bere R jako referenční; v našem případě jsou referenční skupinou ženy.

```
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
```

Výpočet mezí intervalů spolehlivosti pro podíly šancí:

```
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
```

Interpretace podílu šancí u proměnné head.W: pokud se o 1 mm zvětší šířka hlavy (a přitom ostatní rozměry se nezmění), tak šance, že pozorování patří muži, se zvýší 1,31-krát.

Pro hodnocení kvality modelu si vypíšeme hodnoty koeficientů determinace:

```
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 objektů. Nejprve musíme na základě odhadnutých pravděpodobností odhadnout, která pozorování patří mužům a která ženám. Jako dělicí bod zvolíme hodnotu 0,5.

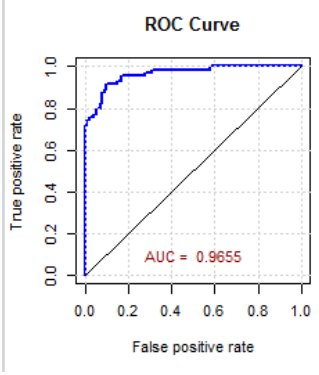
```
fitted <- predict(m.head, newdata=head, type="response")
fitted.cat <- ifelse(fitted < 0.5, "f", "m")
tab <- table(fitted.cat,head$sex)
tab
```

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

Procento správně zařazených žen =
Procento správně zařazených mužů =
Celkové procento správné klasifikace =

Pro hodnocení kvality modelu můžeme použít i ROC křivku a hodnotu AUC (area under the curve - plocha pod křivkou).

```
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")
```



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

Upozornění: Pro tvorbu modelu můžeme použít i STEPWISE proceduru (zpětnou, dopřednou, obousměrnou), 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')

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

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

Výsledky všech tří procedur jsou stejné a shodné s výsledkem předešlého postupu.