EX6

Logistic Regression

Aplikací logistické regrese můžeme zjistit, které faktory nejvíce přispěly k nevěře v manželství.

```
data(Affairs, package="AER")
summary(Affairs)
```

```
##     affairs          gender         age         yearsmarried    children
##  Min.   : 0.000   female:315   Min.   :17.50   Min.   : 0.125   no :171
##  1st Qu.: 0.000   male  :286   1st Qu.:27.00   1st Qu.: 4.000   yes:430
##  Median : 0.000                Median :32.00   Median : 7.000
##  Mean   : 1.456                Mean   :32.49   Mean   : 8.178
##  3rd Qu.: 0.000                3rd Qu.:37.00   3rd Qu.:15.000
##  Max.   :12.000                Max.   :57.00   Max.   :15.000
##  religiousness     education       occupation        rating
##  Min.   :1.000   Min.   : 9.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:14.00   1st Qu.:3.000   1st Qu.:3.000
##  Median :3.000   Median :16.00   Median :5.000   Median :4.000
##  Mean   :3.116   Mean   :16.17   Mean   :4.195   Mean   :3.932
##  3rd Qu.:4.000   3rd Qu.:18.00   3rd Qu.:6.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :20.00   Max.   :7.000   Max.   :5.000
```

```
table(Affairs$affairs)
```

```
##
##   0   1   2   3   7  12
## 451  34  17  19  42  38
```

Transformace do binární proměnné podle toho zda respondent měl/neměl aféru.

```
Affairs$binaffair[Affairs$affairs >  0] <- 1
Affairs$binaffair[Affairs$affairs == 0] <- 0
Affairs$binaffair <- factor(Affairs$binaffair,levels=c(0,1), labels=c("No","Yes"))
table(Affairs$binaffair)
```

```
##
##  No Yes
## 451 150
```

Model logistické regrese k pozorování vztahu mezi proměnnou aféra a vysvětlujícími proměnnými (věk, pohlaví, vzdělání, povolání, děti, atd.)

```
fit.full <- glm(binaffair ~ gender + age + yearsmarried + children + religiousness
+ education + occupation + rating, data=Affairs, family=binomial())
summary(fit.full)
```

```
##
## Call:
## glm(formula = binaffair ~ gender + age + yearsmarried + children +
##     religiousness + education + occupation + rating, family = binomial(),
##     data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5713  -0.7499  -0.5690  -0.2539   2.5191
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.37726    0.88776   1.551 0.120807
## gendermale     0.28029    0.23909   1.172 0.241083
## age           -0.04426    0.01825  -2.425 0.015301 *
## yearsmarried   0.09477    0.03221   2.942 0.003262 **
## childrenyes    0.39767    0.29151   1.364 0.172508
## religiousness -0.32472    0.08975  -3.618 0.000297 ***
## education      0.02105    0.05051   0.417 0.676851
## occupation     0.03092    0.07178   0.431 0.666630
## rating        -0.46845    0.09091  -5.153 2.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 609.51  on 592  degrees of freedom
## AIC: 627.51
##
## Number of Fisher Scoring iterations: 4
```

Pokud se zaměříme na p-hodnoty pro regresní koeficienty, pak zjistíme, že pohlaví,děti, vzdělání a povolání nemají významný vliv. Druhý model zahrnuje pouze významné proměnné, jako je věk, počet manželství, náboženské vyznání.

```
fit.reduced   <- glm(binaffair ~ age + yearsmarried + religiousness + rating, data=
Affairs, family=binomial())
summary(fit.reduced)
```

```
##
## Call:
## glm(formula = binaffair ~ age + yearsmarried + religiousness +
##     rating, family = binomial(), data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6278  -0.7550  -0.5701  -0.2624   2.3998
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.93083    0.61032   3.164 0.001558 **
## age           -0.03527    0.01736  -2.032 0.042127 *
## yearsmarried   0.10062    0.02921   3.445 0.000571 ***
## religiousness -0.32902    0.08945  -3.678 0.000235 ***
## rating        -0.46136    0.08884  -5.193 2.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 615.36  on 596  degrees of freedom
## AIC: 625.36
##
## Number of Fisher Scoring iterations: 4
```

U druhého modelu jsou p-hodnoty statisticky významné. Potom můžeme spustit chí-kvadrát test pro porovnání mezi prvním a druhým modelem.

```
anova(fit.reduced, fit.full, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: binaffair ~ age + yearsmarried + religiousness + rating
## Model 2: binaffair ~ gender + age + yearsmarried + children + religiousness +
##     education + occupation + rating
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       596     615.36
## 2       592     609.51  4   5.8474   0.2108
```

Výsledek potvrzuje původní teorii, že pohlaví, děti, vzdělání a povolání nepřispívají k nevěře.

```
coef(fit.reduced)
```

```
##   (Intercept)           age  yearsmarried religiousness        rating
##    1.93083017   -0.03527112    0.10062274   -0.32902386   -0.46136144
```

```
exp(coef(fit.reduced))
```

```
##   (Intercept)           age  yearsmarried religiousness        rating 
##     6.8952321     0.9653437     1.1058594     0.7196258     0.6304248 
```