SAP – Četvrta auditorna vježba

Case study *Karakteristike klijenata banke*: ANOVA i logistička regresija

Tessa Bauman, Stjepan Begušić, David Bojanić, Tomislav Kovačević, Andro Merćep

12.1.2022.

Podatci o potrošačkim kreditima klijenata banke

Dani su podatci o potrošačkim nenamjenskim kreditima jedne banke.

```
# Ucitavanje podataka
creditdata = read.csv('creditdata.csv')
summary(creditdata)
```

```
##        ID              age          education    marriage     apartment
##  Min.   :  1.0   Min.   :19.00   Min.   :1   Min.   :1.0   Min.   :1.0
##  1st Qu.:150.8   1st Qu.:43.00   1st Qu.:1   1st Qu.:1.0   1st Qu.:1.0
##  Median :300.5   Median :51.00   Median :2   Median :1.5   Median :1.5
##  Mean   :300.5   Mean   :49.96   Mean   :2   Mean   :1.5   Mean   :1.5
##  3rd Qu.:450.2   3rd Qu.:56.00   3rd Qu.:3   3rd Qu.:2.0   3rd Qu.:2.0
##  Max.   :600.0   Max.   :83.00   Max.   :3   Max.   :2.0   Max.   :2.0
##      income          amount         default
##  Min.   : 1030   Min.   : 6010   Min.   :0.000
##  1st Qu.: 4358   1st Qu.:14630   1st Qu.:0.000
##  Median : 5525   Median :17210   Median :0.000
##  Mean   : 5659   Mean   :17427   Mean   :0.155
##  3rd Qu.: 6912   3rd Qu.:19848   3rd Qu.:0.000
##  Max.   :11580   Max.   :29660   Max.   :1.000
```

Svaki redak predstavlja kredit za određenog klijenta, uz neke njegove značajke:

- education – obrazovanje (1 – osnovna škola, 2 – srednja škola, 3 – viša škola ili fakultet)
- marriage – bračno stanje (1 – neudana/neoženjen, 2 – udana/oženjen)
- apartment – vlasništvo stana (1 – podstanar, 2 – vlasnik stana)
- income – prosječna mjesečna plaća
- amount – iznos kredita
- default – je li klijent kasnio s plaćanjem kredita (0/1)

```
# Priprema podataka
creditdata$education = factor(creditdata$education,levels = c(1,2,3),labels = c('elementary','secondary
creditdata$marriage = factor(creditdata$marriage,levels = c(1,2),labels = c('single','married'))
creditdata$apartment = factor(creditdata$apartment,levels = c(1,2),labels = c('rent','own'))
creditdata$default = factor(creditdata$default,levels = c(0,1),labels = c(FALSE,TRUE))
summary(creditdata)
```

```
##        ID              age              education      marriage   apartment
##  Min.   :  1.0   Min.   :19.00   elementary:200   single :300   rent:300
##  1st Qu.:150.8   1st Qu.:43.00   secondary :200   married:300   own :300
##  Median :300.5   Median :51.00   university:200
```

```
##  Mean   :300.5   Mean   :49.96
##  3rd Qu.:450.2   3rd Qu.:56.00
##  Max.   :600.0   Max.   :83.00
##      income          amount       default
##  Min.   : 1030   Min.   : 6010   FALSE:507
##  1st Qu.: 4358   1st Qu.:14630   TRUE : 93
##  Median : 5525   Median :17210
##  Mean   : 5659   Mean   :17427
##  3rd Qu.: 6912   3rd Qu.:19848
##  Max.   :11580   Max.   :29660
```

Neka od ključnih pitanja koja zanimaju banke i kreditne institucije su:

- Kako varira plaća u ovisnosti o nekim značajkama klijenata (npr. obrazovanje)?
- Postoje li interakcijski efekti u više pojedinih značajki klijenata koji određuju visinu plaće?
- Kako, koristeći dane podatke, najbolje predvidjeti vjerojatnost da će klijent kasniti s otplatom?

ANOVA

ANOVA (engl. *ANalysis Of VAriance*) je metoda kojom testiramo sredine više populacija. U analizi varijance pretpostavlja se da je ukupna varijabilnost u podatcima posljedica varijabilnosti podataka unutar svakog pojedine grupe (populacije) i varijabilnosti između različitih grupa. Varijabilnost unutar pojedinog uzorka je rezultat slučajnosti, a ako postoje razlike u sredinama populacija, one će biti odražene u varijabilnosti među grupama. Jedan od glavnih ciljeva analize varijance je ustanoviti jesu li upravo te razlike između grupa samo posljedica slučajnosti ili je statistički značajna.

Jednofaktorska ANOVA

U jednofaktorskom ANOVA modelu razmatra se utjecaj jednog faktora koji ima k razina. Neka su:

$$\begin{aligned} X_{11}, X_{12}, \ldots, X_{1n_1} &\sim N(\mu_1, \sigma^2) \\ X_{21}, X_{22}, \ldots, X_{2n_2} &\sim N(\mu_2, \sigma^2) \\ &\vdots \\ X_{k1}, X_{k2}, \ldots, X_{kn_k} &\sim N(\mu_k, \sigma^2) \end{aligned}$$

nezavisni uzorci iz k različitih populacija (populacije se razlikuju upravo po razini faktora od interesa). Jednofaktorski ANOVA model glasi:

$$X_{ij} = \mu_j + \epsilon_{ij},$$

gdje je μ_j sredina svake populacije $j = 1, .., k$. Analizom varijance testiramo:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \ldots = \mu_k \\ H_1 &: \neg H_0. \end{aligned}$$

Pretpostavke ANOVA-e su:

- nezavisnost pojedinih podataka u uzorcima,
- normalna razdioba podataka,
- homogenost varijanci među populacijama.

Kad su veličine grupa podjednake, ANOVA je relativno robusna metoda na blaga odstupanja od pretpostavke normalnosti i homogenosti varijanci. Ipak, dobro je provjeriti koliko su ta odstupanja velika.

Provjera normalnosti može se za svaku pojedinu grupu napraviti KS testom ili Lillieforsovom inačicom KS testa. U ovom slučaju razmatrat ćemo zaposlenje kao varijablu koja određuje grupe (populacije) i plaću kao zavisnu varijablu.

```
require(nortest)
```

```
## Loading required package: nortest
```

```
lillie.test(creditdata$income)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  creditdata$income
## D = 0.041548, p-value = 0.01539
```

```
lillie.test(creditdata$income[creditdata$education=='elementary'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  creditdata$income[creditdata$education == "elementary"]
## D = 0.05522, p-value = 0.1438
```

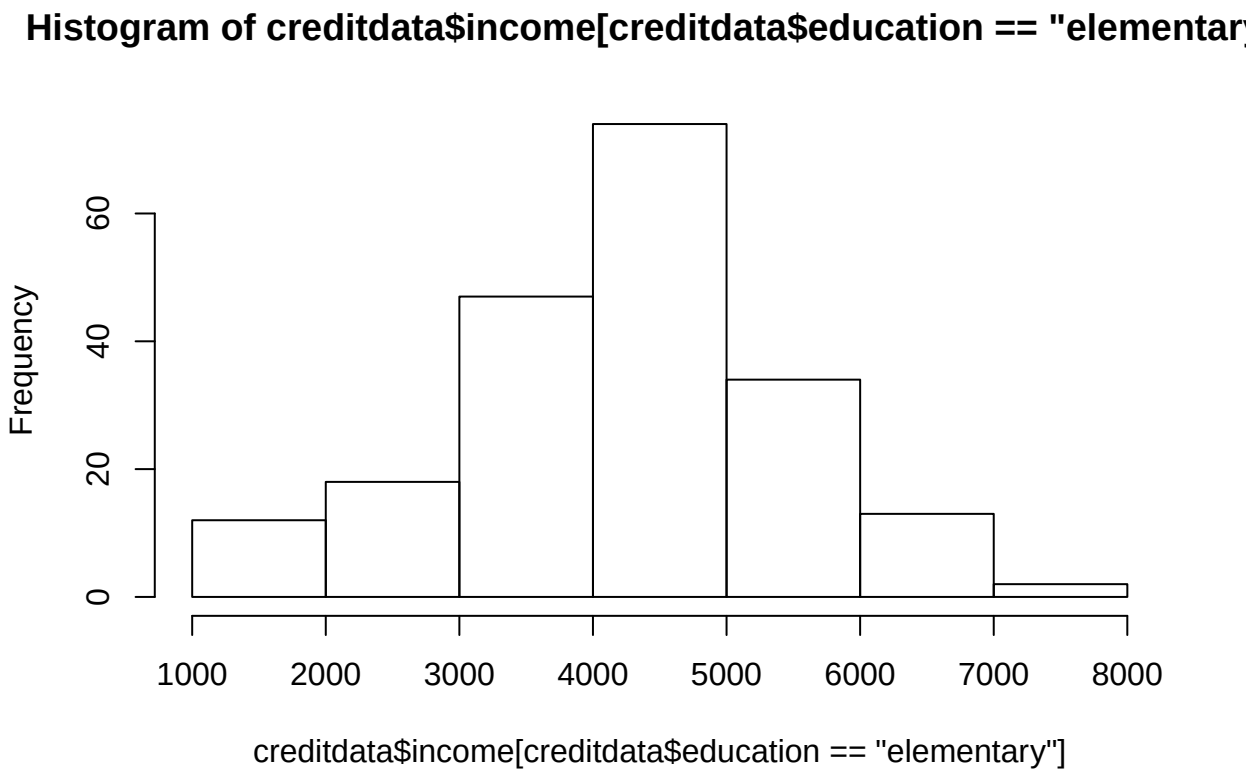
```
lillie.test(creditdata$income[creditdata$education=='secondary'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  creditdata$income[creditdata$education == "secondary"]
## D = 0.035209, p-value = 0.7889
```

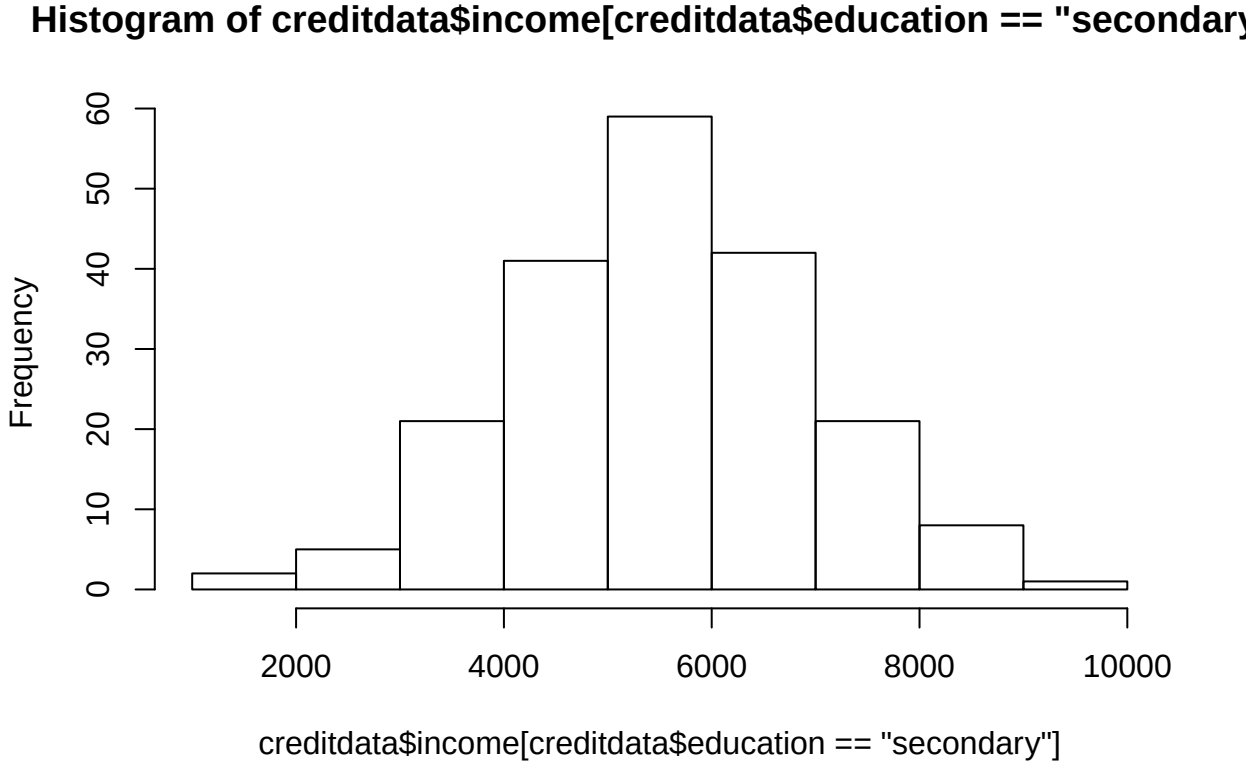
```
lillie.test(creditdata$income[creditdata$education=='university'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  creditdata$income[creditdata$education == "university"]
## D = 0.030324, p-value = 0.9251
```

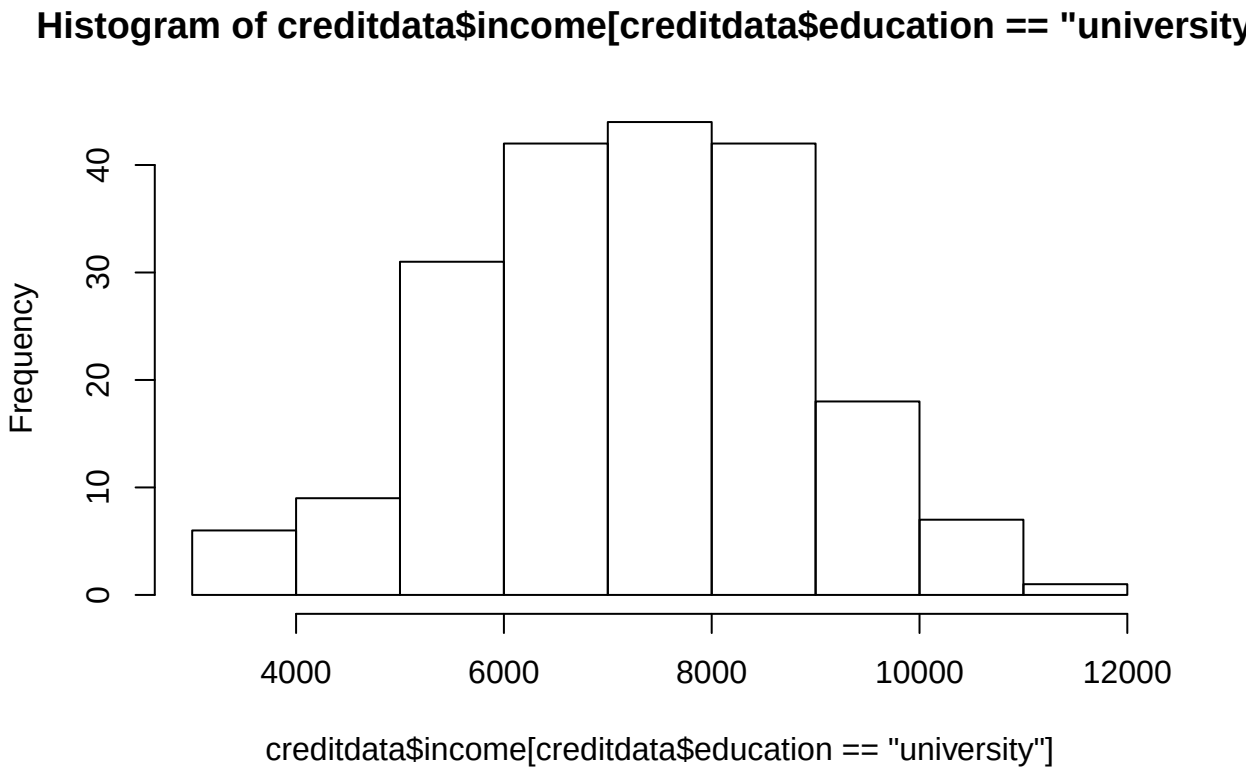
```
hist(creditdata$income[creditdata$education=='elementary'])
```




```
hist(creditdata$income[creditdata$education=='secondary'])
```




```
hist(creditdata$income[creditdata$education=='university'])
```



Što se tiče homogenosti varijanci različitih populacija, potrebno je testirati:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$
$$H_1 : \neg H_0.$$

Navedenu hipotezu možemo testirati Bartlettovim testom. Bartlettov test u R-u implementiran je naredbom `bartlett.test()`.

```
# Testiranje homogenosti varijance uzoraka Bartlettovim testom
bartlett.test(creditdata$income ~ creditdata$education)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  creditdata$income by creditdata$education
## Bartlett's K-squared = 11.464, df = 2, p-value = 0.00324
```

```
var((creditdata$income[creditdata$education=='elementary']))
```

```
## [1] 1605464
```

```
var((creditdata$income[creditdata$education=='secondary']))
```

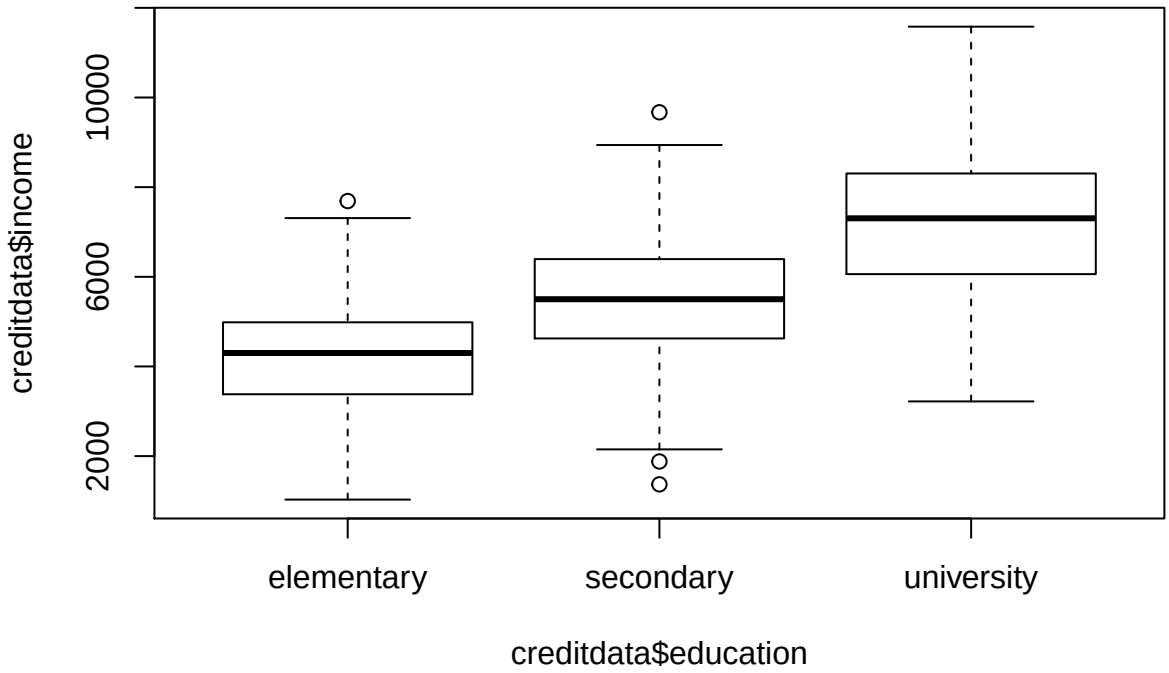
```
## [1] 2036133
```

```
var((creditdata$income[creditdata$education=='university']))
```

```
## [1] 2598369
```

Provjerimo postoje li razlike u prihodima za različite razine školovanja klijenata.

```
# Graficki prikaz podataka
boxplot(creditdata$income ~ creditdata$education)
```




```
# Test
a = aov(creditdata$income ~ creditdata$education)
summary(a)
```

```
##                         Df    Sum Sq   Mean Sq F value Pr(>F)
## creditdata$education   2 9.215e+08 460725212   221.5 <2e-16 ***
## Residuals            597 1.242e+09   2079989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Grafički prikaz sugerira da postoji jasna razlika između grupa, što potvrđuje i ANOVA. Kako bismo procijenili model koji pomoću varijable o školovanju klijenata objašnjava njihov prihod?

```
# Linearni model
model = lm(income ~ education, data = creditdata)
summary(model)
```

```
##
## Call:
## lm(formula = income ~ education, data = creditdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4133.5  -959.9    48.9   914.9  4331.3
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4225.1      102.0  41.431   <2e-16 ***
## educationsecondary    1278.4      144.2   8.864   <2e-16 ***
## educationuniversity   3023.5      144.2  20.965   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1442 on 597 degrees of freedom
```

```
## Multiple R-squared:  0.426,  Adjusted R-squared:  0.424
## F-statistic: 221.5 on 2 and 597 DF,  p-value: < 2.2e-16
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: income
##            Df     Sum Sq   Mean Sq F value    Pr(>F)
## education   2  921450424 460725212   221.5 < 2.2e-16 ***
## Residuals 597 1241753253   2079989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linearni model koji ima samo kategorijsku varijablu grupe (populacije) kao prediktor istovjetan je ANOVA modelu – statistički zaključci su u oba slučaja isti.

Dvofaktorska ANOVA

Kod dvofaktorske analize varijance promatra se utjecaj dvaju faktora, pri čemu prvi faktor ima a razina, a drugi faktor b razina. Dakle, promatramo ukupno $a \cdot b$ populacija. Pretpostavimo da iz svake populacije uzimamo nezavisne slučajne uzorke jednake duljine n, svaki za obilježje X reprezentirano sa $X_{ij} \sim N(\mu_{ij}, \sigma^2)$ u populaciji ij, gdje je $i \in \{1, 2, \ldots, a\}$, a $j \in \{1, 2, \ldots, b\}$.

Potrebno je testirati hipoteze:

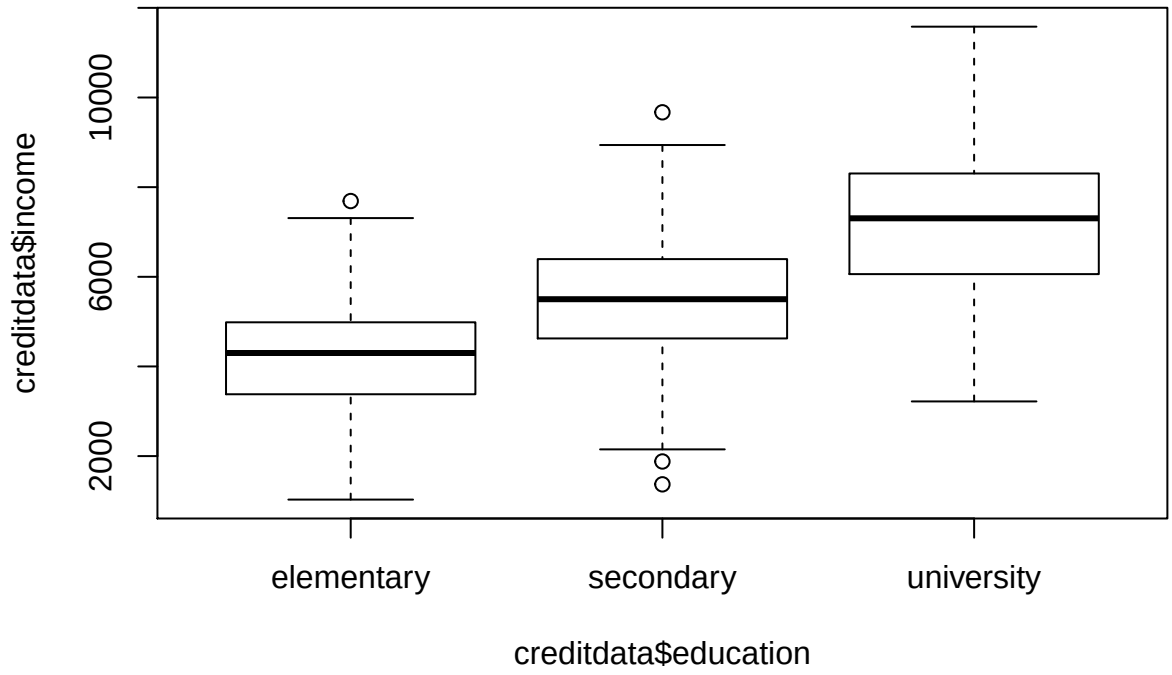
- H_0': prvi faktor je beznačajan
- H_0'': drugi faktor je beznačajan
- H_0''': nema interakcije među faktorima

Sve tri hipoteze testiraju se dvofaktorskim ANOVA testom, koji pretpostavlja model:

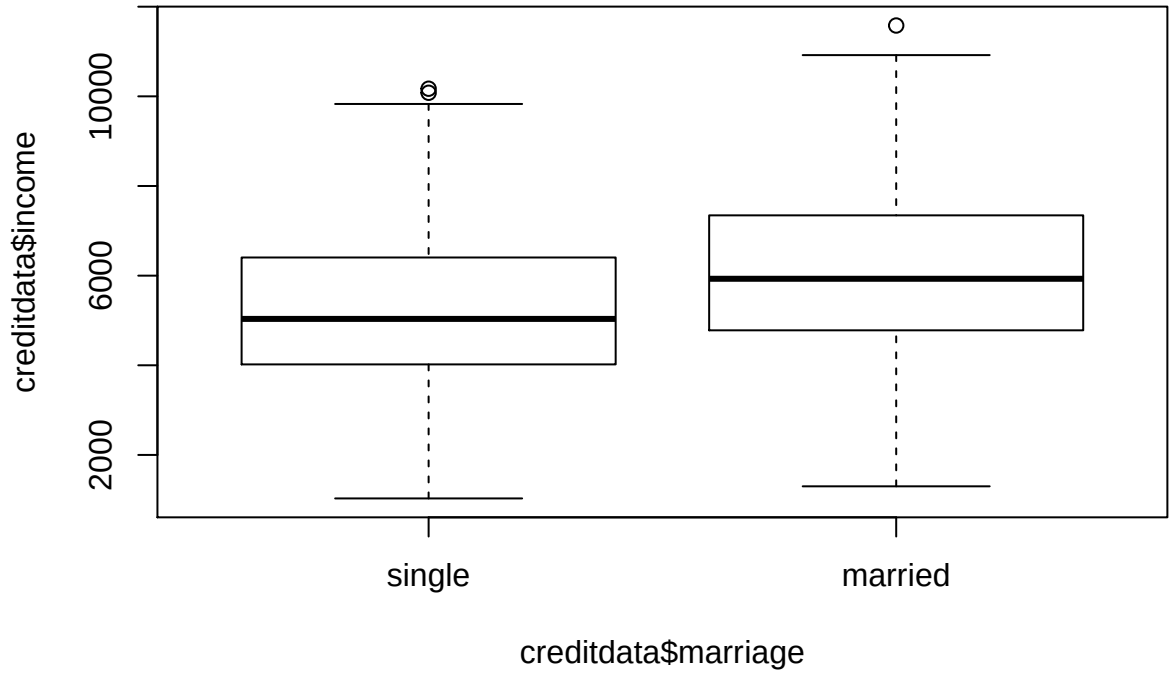
$$X_{ijk} = \mu_{ij} + \epsilon_{ijk},$$

gdje se sredine μ_{ij} mogu zapisati kao: $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, koje odgovaraju sredinama prvog faktora α_i, drugog faktora β_j, i interakcije $(\alpha\beta)_{ij}$. U standardnoj proceduri će dvofaktorski ANOVA test imati iste pretpostavke kao jednofaktorski, uz zahtjev na jednake veličine uzoraka pojedinih grupa (populacija). To u praksi najčešće nije slučaj, pa se koriste verzije s otežanim srednjim vrijednostima – u R-u je upravo takav pristup defaultni u funkciji `aov()`.

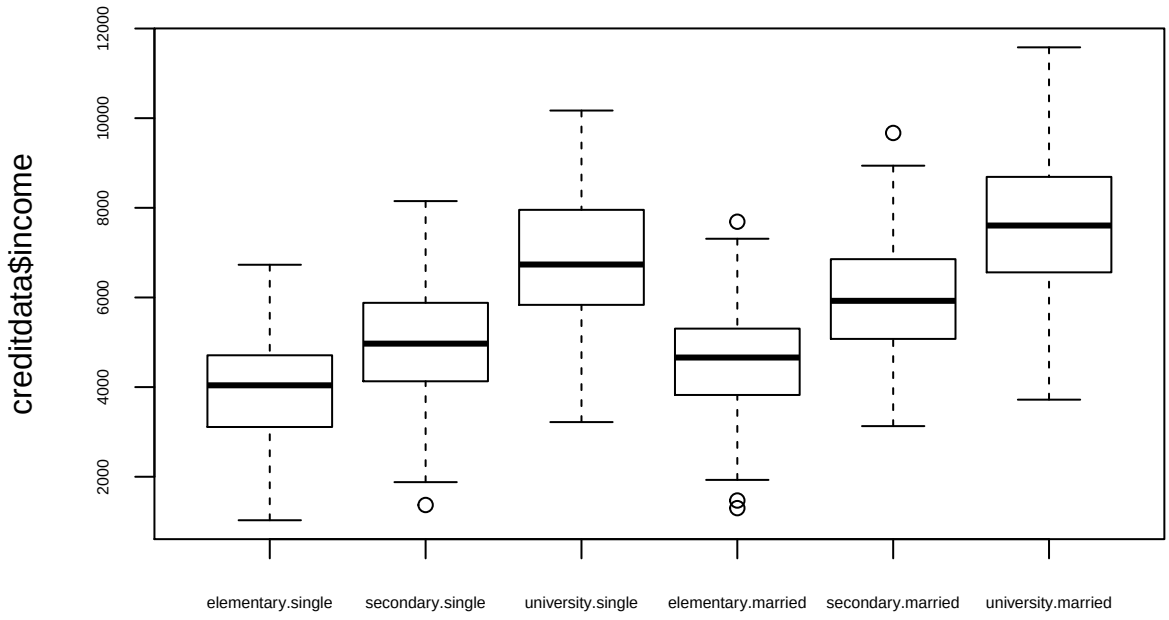
```
# Graficki prikaz podataka
boxplot(creditdata$income ~ creditdata$education)
```




```
boxplot(creditdata$income ~ creditdata$marriage)
```




```
inter = interaction(creditdata$education,creditdata$marriage)
boxplot(creditdata$income ~ inter,cex.axis=0.5)
```




```
# Bartlettov test za jednakost varijanci između pojedinih grupa
bartlett.test(creditdata$income ~ inter)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  creditdata$income by inter
## Bartlett's K-squared = 12.821, df = 5, p-value = 0.02511
```

```
aggregate(creditdata$income, by=list(inter), FUN=var)
```

```
##              Group.1       x
## 1  elementary.single 1454731
## 2   secondary.single 1795602
## 3  university.single 2378584
## 4 elementary.married 1536952
## 5  secondary.married 1756313
## 6 university.married 2525275
```

```
# ANOVA test
a = aov(income ~ education * marriage, data = creditdata)
summary(a)
```

```
##                     Df    Sum Sq   Mean Sq F value  Pr(>F)
## education            2 9.215e+08 460725212 241.482 < 2e-16 ***
## marriage             1 1.052e+08 105219313  55.149 3.9e-13 ***
## education:marriage   2 3.236e+06   1617840   0.848   0.429
## Residuals          594 1.133e+09   1907910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Linearni model
model = lm(income ~ education * marriage, data = creditdata)
summary(model)
```

```
##
## Call:
## lm(formula = income ~ education * marriage, data = creditdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3926.1  -887.3    31.5   888.9  3933.9
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          3883.7      138.1  28.117  < 2e-16 ***
## educationsecondary                   1102.3      195.3   5.643 2.59e-08 ***
## educationuniversity                  2967.5      195.3  15.191  < 2e-16 ***
## marriagemarried                       682.8      195.3   3.495 0.000509 ***
## educationsecondary:marriagemarried    352.1      276.3   1.275 0.202967
## educationuniversity:marriagemarried   112.1      276.3   0.406 0.685046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1381 on 594 degrees of freedom
## Multiple R-squared:  0.4761, Adjusted R-squared:  0.4717
## F-statistic:   108 on 5 and 594 DF,  p-value: < 2.2e-16
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: income
##                     Df     Sum Sq   Mean Sq F value  Pr(>F)
## education            2  921450424 460725212 241.482 < 2e-16 ***
## marriage             1  105219313 105219313  55.149 3.9e-13 ***
## education:marriage   2    3235680   1617840   0.848  0.4288
## Residuals          594 1133298260   1907910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rezultati sugeriraju da interakcije nema, ali da se pojedine populacije (podijeljene po kategorijama edukacije ili braka) razlikuju po srednjim vrijednostima prihoda. Štoviše, iz linearnog modela možemo zaključiti koje pojedine grupe imaju viša očekivanja.

Logistička regresija

Kad bismo htjeli koristiti postojeće podatke za predvidjeti hoće li koji klijent zakasniti s otplatom kredita, moguće je procijeniti regresijski model s podatcima o klijentima kao nezavisnim varijablama. Zavisna varijabla u tom slučaju nije kontinuirana. Koje su pretpostavke linearne regresije onda (jako) prekršene i ne možemo je koristiti u ovom slučaju?

Imamo na raspolaganju skup podataka $D = \{X_1, ..., X_N\}$ gdje je svaki X_i vektor vrijednosti prediktorskih varijabli, one mogu biti diskretne (uz prikladno dummy-kodiranje) ili kontinuirane. Imamo i skup očekivanih izlaza $\{y_1, ..., y_n\}$ gdje je svaki y_i binarna varijabla tj. 0 ili 1. Želimo dobiti kao izlaz modela skup izlaza $\{\hat{y_1}, ..., \hat{y_N}\}$. Idealno bismo od dobrog modela očekivali da bude (što je češće moguće) $\hat{y_i} = y_i$, tj. da radi dobre predikcije. Također, želimo imati vjerojatnost $P(\hat{Y_i} = 1|x_i)$ koja bi nam dala mjeru koliko je model

"siguran" u svoju odluku i omogućavala da izračunamo predikcije na sljedeći način

$$\hat{y}_i = \begin{cases} 1 & \text{ako } P(\hat{Y}_i = 1|\vec{x_i}) \geq 0.5 \\ 0, & \text{inače} \end{cases}$$

Glavni problem zbog kojeg ne možemo koristiti linearnu regresiju za ovaj zadatak je što $\beta^T X$ može poprimiti vrijednosti van intervala $[0, 1]$ pa izlaz linearne regresije ne možemo interpretirati kao vjerojatnost.

Logistička regresija rješava taj problem tako što transformira $\beta^T X$ koristeći logističku (sigmoidalnu) funkciju:

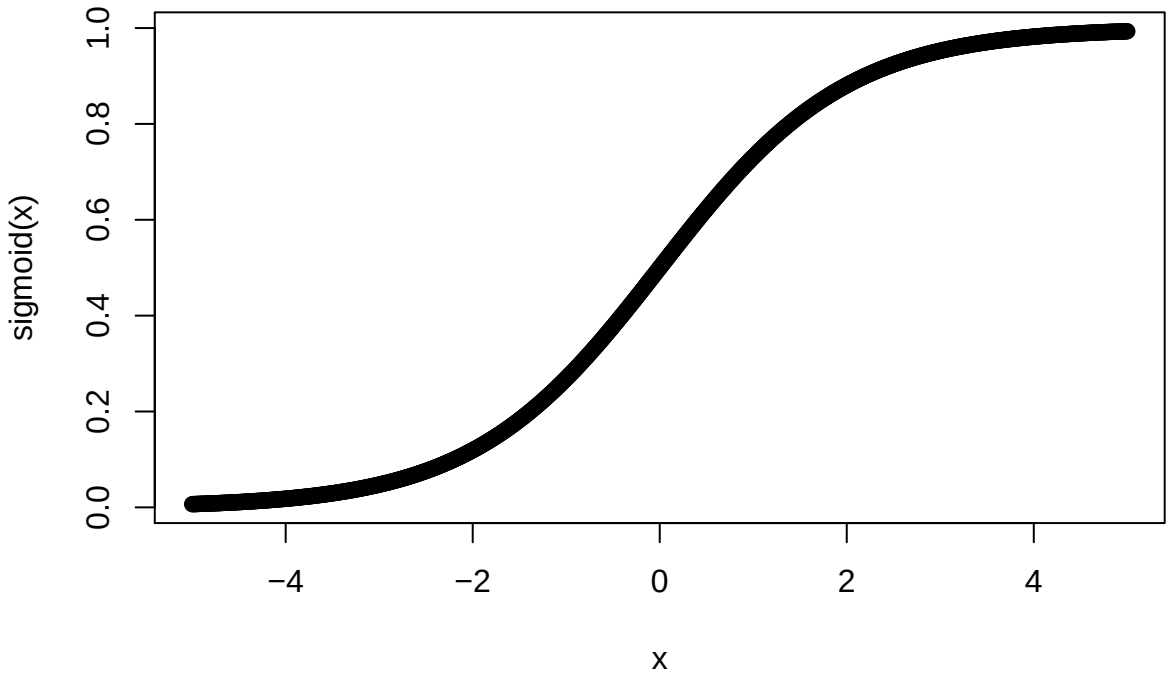
$$\sigma(\alpha) = \frac{1}{1 + e^{-\alpha}}$$

Koja je prikazana na sljedećem grafu:

```
sigmoid = function(x) {
   1 / (1 + exp(-x))
}

x <- seq(-5, 5, 0.01)

plot(x,sigmoid(x))
```



Postoji više razloga zašto koristimo baš ovu funkciju:

1. Ona ima upravo željeno djelovanje – ograničava izlaz linearnog modela između 0 i 1
2. Ima svojstvo da je njena derivacija $\sigma'(\alpha) = \sigma(\alpha)(1 - \sigma(\alpha))$ što olakšava implementaciju algoritma učenja.
3. Omogućuje lakše interpretiranje koeficijenata β ovog modela (više o ovom kasnije).

Model dakle prikazuje gore traženu vjerojatnost na sljedeći način:

$$P(\hat{Y}_i = 1|X_i) = \frac{1}{1 + e^{-\beta^T X_i}}$$

Uz to što za svaki x_i možemo dobiti vjerojatnost da je pripadni y_i jednak 1, možemo i donijeti binarne odluke na temelju usporedbe dobivene vjerojatnosti s pragom od 0.5 kao što je opisano gore.

Učenje

Kako bismo naučili dobre vrijednosti za β koristimo postupak procjene najveće izglednosti (vjerodostojnosti) (engl. *Maximum Likelihood Estimation*). Za neki fiksni vektor težina β možemo izračunati vjerojatnost koju model daje našem cijelom skupu podataka. Npr. ako je $D = \{X_1, X_2, X_3\}$ i skup točnih izlaza je $1, 1, 0$ tada je vjerojatnost podataka uz model logističke regresije koji koristi te konkretne težine jednaka

$$P(D|\beta) = P(Y_1 = 1|X_1)P(Y_2 = 1|X_2)(1 - P(Y_3 = 1|X_3).$$

Ova veličina se još zove izglednost (vjerodostojnost) $L(\vec{\beta})$ parametara uz dane podatke. Da smo uzeli neki drugi skup težina β', dobili bismo neku drugu vjerodostojnost $L(\beta')$. Algoritam učenja radi tako pronađe onaj skup težina β koji maksimizira ovu veličinu. Upravo taj skup težina najbolje opisuje podatke.

Interpretacija i testiranje koeficijenata β

Kao kod linearne regresije i ovdje možemo odrediti koje značajke su statistički značajne. U `summary` naredbi modela logističke regresije R će nam također ispisati i devijancu (engl. *deviance*). To je mjera zasnovana na izglednosti i opisuje nam koliko je model dobar, u smislu koliko dobro se prilagodio podacima (veći broj znači da je prilagodba gora). R će nam izbaciti dvije vrste devijance (1) `null deviance` – koja opisuje model koji ima samo slobodni član i (2) `residual deviance` koja uključuje sve prediktorske varijable. Koristeći te dvije veličine, moguće je i izračunati R^2 danog modela kao:

$$R^2 = 1 - \frac{D_{mdl}}{D_0}.$$

Važna napomena: ovaj R^2 nema istu interpretaciju kao R^2 modela linearne regresije:

- nije vezan uz koeficijent korelacije,
- ne govori o udjelu opisane varijance.

No, može se koristiti kao mjera koja govori koliko je procijenjeni model blizu/daleko od null modela (0-1).

```
require(caret)
```

```
## Loading required package: caret
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
logreg.mdl = glm(default ~ age + education + marriage + apartment + income + amount, data = creditdata,
summary(logreg.mdl)
```

```
##
## Call:
## glm(formula = default ~ age + education + marriage + apartment +
##     income + amount, family = binomial(), data = creditdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7737  -0.5628  -0.3161  -0.1470   2.9208
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.876e-01  9.072e-01  -0.537  0.59092
## age                 -2.621e-03  1.277e-02  -0.205  0.83733
## educationsecondary  -1.039e+00  3.166e-01  -3.280  0.00104 **
## educationuniversity -1.080e+00  4.474e-01  -2.414  0.01579 *
## marriagemarried     -6.485e-01  2.753e-01  -2.355  0.01850 *
```

```
## apartmentown        -2.721e-01  2.602e-01  -1.046  0.29574
## income              -7.441e-04  1.120e-04  -6.643 3.07e-11 ***
## amount               2.085e-04  4.385e-05   4.756 1.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 517.54  on 599  degrees of freedom
## Residual deviance: 400.98  on 592  degrees of freedom
## AIC: 416.98
##
## Number of Fisher Scoring iterations: 6
```

```
Rsq = 1 - logreg.mdl$deviance/logreg.mdl$null.deviance
Rsq
```

```
## [1] 0.2252137
```

Važno je imati na umu da sam omjer oznaka u izlaznoj varijabli može jako utjecati na neke mjere kvalitete modela. Bolju informaciju moguće je dobiti iz tzv. matrice zabune (engl. *confusion matrix*), koja je zapravo kontingencijska matrica oznaka iz podataka i modela. Matrica će biti oblika:

	$\hat{Y}=0$	$\hat{Y}=1$
$Y=0$	TN	FP
$Y=1$	FN	TP

Mjere koje mogu biti od interesa su:

- točnost (eng. accuracy): $\frac{TP+TN}{TP+FP+TN+FN}$
- preciznost (eng. precision): $\frac{TP}{TP+FP}$ (udio točnih primjera u svim koji su klasificirani kao TRUE)
- odziv (eng. recall): $\frac{TP}{TP+FN}$ (udio točnih primjera u skupu svih koji su stvarno TRUE)
- specifičnost (eng. specificity): $\frac{TN}{TN+FP}$ (udio točnih primjera u svim koji su klasificirani kao FALSE)

Postoje još druge tehnike za ispitivanje kvalitete klasifikacijskih modela, poput F1 ili ROC krivulje, u koje ovaj case study neće ulaziti u detalje, a bit će obrađene se na kasnijim predmetima na diplomskom studiju.

```
yHat <- logreg.mdl$fitted.values > 0.4
tab <- table(creditdata$default, yHat)

tab
```

```
##        yHat
##         FALSE TRUE
##   FALSE   479   28
##   TRUE     59   34
```

```
accuracy = sum(diag(tab)) / sum(tab)
precision = tab[2,2] / sum(tab[,2])
recall = tab[2,2] / sum(tab[2,])
specificity = tab[1,1] / sum(tab[,1])

accuracy
```

```
## [1] 0.855
```

```
precision
```

```
## [1] 0.5483871
```

```
recall
```

```
## [1] 0.3655914
```

```
specificity
```

```
## [1] 0.8903346
```

Test omjera izglednosti (likelihood ratio test)

Pokazuje se da za dva modela logističke regresije M_1 sa N_1 prediktorskih varijabli i M_2 sa N_2 prediktorskih varijabli statistika $-2\ln\frac{L_1}{L_2}$, gdje su L_1 i L_2 izglednosti za oba modela, ima χ^2 distribuciju s $|N_1 - N_2|$ stupnjeva slobode. Tu statistiku možemo iskoristiti za testiranje postoji li značajna razlika u kvaliteti više alternativnih modela. Ovaj test ima sličnu ulogu kao F-test u slučaju linearne regresije.

Na primjer, možemo testirati postoji li razlika između dva modela – originalnog modela i modela s dodanim interakcijskim članom. U tom slučaju ćemo prihvatiti prošireni model ako ima značajno manju devijancu, na što će nam odgovor dati test omjera izglednosti.

```
logreg.mdl = glm(default ~ age + education + marriage + apartment + income + amount, data = creditdata,
summary(logreg.mdl)
```

```
##
## Call:
## glm(formula = default ~ age + education + marriage + apartment +
##     income + amount, family = binomial(), data = creditdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7737  -0.5628  -0.3161  -0.1470   2.9208
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.876e-01  9.072e-01  -0.537  0.59092
## age                 -2.621e-03  1.277e-02  -0.205  0.83733
## educationsecondary  -1.039e+00  3.166e-01  -3.280  0.00104 **
## educationuniversity -1.080e+00  4.474e-01  -2.414  0.01579 *
## marriagemarried     -6.485e-01  2.753e-01  -2.355  0.01850 *
## apartmentown        -2.721e-01  2.602e-01  -1.046  0.29574
## income              -7.441e-04  1.120e-04  -6.643 3.07e-11 ***
## amount               2.085e-04  4.385e-05   4.756 1.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 517.54  on 599  degrees of freedom
## Residual deviance: 400.98  on 592  degrees of freedom
## AIC: 416.98
##
## Number of Fisher Scoring iterations: 6
```

```r
logreg.mdl.2 = glm(default ~ age + education + marriage + apartment + income + amount + I(income/amount)
summary(logreg.mdl.2)
```

```
##
## Call:
## glm(formula = default ~ age + education + marriage + apartment +
##     income + amount + I(income/amount), family = binomial(),
##     data = creditdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8843  -0.5578  -0.3252  -0.1517   2.9644
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.105e+00  2.172e+00   0.969  0.33232
## age                  -2.828e-03  1.277e-02  -0.221  0.82481
## educationsecondary   -1.036e+00  3.182e-01  -3.256  0.00113 **
## educationuniversity -1.102e+00  4.518e-01  -2.439  0.01472 *
## marriagemarried      -6.339e-01  2.756e-01  -2.300  0.02147 *
## apartmentown         -2.689e-01  2.607e-01  -1.032  0.30217
## income               -1.518e-04  4.613e-04  -0.329  0.74216
## amount                5.343e-05  1.248e-04   0.428  0.66851
## I(income/amount)     -9.815e+00  7.540e+00  -1.302  0.19306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 517.54  on 599  degrees of freedom
## Residual deviance: 399.11  on 591  degrees of freedom
## AIC: 417.11
##
## Number of Fisher Scoring iterations: 6
```

```r
anova(logreg.mdl, logreg.mdl.2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: default ~ age + education + marriage + apartment + income + amount
## Model 2: default ~ age + education + marriage + apartment + income + amount +
##     I(income/amount)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       592     400.98
## 2       591     399.11  1   1.8794   0.1704
```

Također možemo testirati i razliku originalnog modela i smanjenog modela koji ne sadrži neke nesignifikantne regresore. U tom slučaju ćemo prihvatiti smanjeni model ukoliko devijanca nije značajno veća.

```r
logreg.mdl.3 = glm(default ~ education + marriage + income + amount, data = creditdata, family = binomia
summary(logreg.mdl.3)
```

```
##
## Call:
## glm(formula = default ~ education + marriage + income + amount,
##     family = binomial(), data = creditdata)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8298  -0.5668  -0.3223  -0.1508   2.8654
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -7.059e-01  6.441e-01  -1.096  0.27316
## educationsecondary  -1.028e+00  3.155e-01  -3.259  0.00112 **
## educationuniversity -1.070e+00  4.466e-01  -2.397  0.01655 *
## marriagemarried     -6.430e-01  2.752e-01  -2.337  0.01946 *
## income              -7.320e-04  1.111e-04  -6.588 4.45e-11 ***
## amount               2.020e-04  4.342e-05   4.651 3.30e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 517.54  on 599  degrees of freedom
## Residual deviance: 402.22  on 594  degrees of freedom
## AIC: 414.22
##
## Number of Fisher Scoring iterations: 6
```

```
anova(logreg.mdl, logreg.mdl.3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: default ~ age + education + marriage + apartment + income + amount
## Model 2: default ~ education + marriage + income + amount
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       592     400.98
## 2       594     402.22 -2   -1.231   0.5404
```

Analiza konačnog modela:

```
yHat <- logreg.mdl.3$fitted.values > 0.5
tab <- table(creditdata$default, yHat)

tab
```

```
##        yHat
##         FALSE TRUE
##   FALSE   494   13
##   TRUE     71   22
```

```
accuracy = sum(diag(tab)) / sum(tab)
precision = tab[2,2] / sum(tab[,2])
recall = tab[2,2] / sum(tab[2,])
specificity = tab[1,1] / sum(tab[,1])

accuracy
```

```
## [1] 0.86
```

```
precision
```

```
## [1] 0.6285714
```

```
recall
```

```
## [1] 0.2365591
```

```
specificity
```

```
## [1] 0.8743363
```