R instructions for the 12th seminar

File *GermanCredit.RData* was provided to the repository http://archive.ics.uci.edu/ml/ by professor Hans Hoffman (Institut für Statistik und Ökonometrie, Universität Hamburg) in November 1994. This file contains credit scoring data from the Federal Republic of Germany during the 90's. Details about all the variables can be found at http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data). For our purposes, we will only use the following variables:

`Class` - Credit classification (0 = good-paid credit, 1 = bad-unpaid credit). We will be modelling unpaid credits, which we will further consider as a "success";
`A00Amount100` - credit amount in 100 DEM;
`A01Duration` - credit duration in months;
`A02Age` - age in years;
`A03IRP` - installment rate in percentage of disposable income;
`A15Gender` - gender (0=female, 1=male);
`A19Housing` - housing (0=rent, 1=own, 2=for free);
`A20Job` - job(0=unemployed/unskilled-non-resident, 1=unskilled - resident, 2=skilled employee/official, 3=management/self-employed/highly qualified employee/officer);

Firstly, we need to load the data file and a library we're going to work with. We can also explore the dimensions and type of data we're going to deal with.

```
load("GermanCredit.RData")
library(DescTools)
dim(GermanCredit)
```

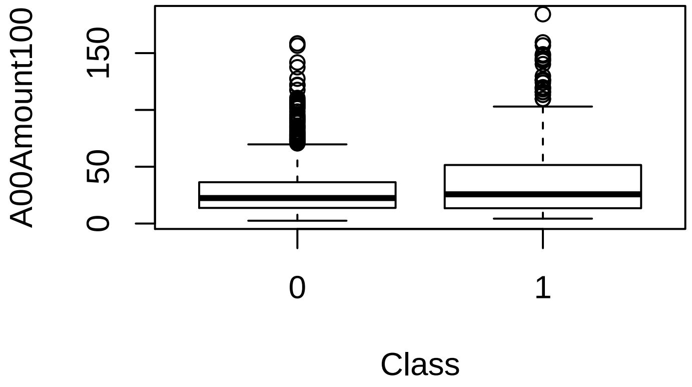
```
## [1] 1000   40
```

R Instructions for Problem 1:

We are going to model the probability of not paying the credit using one quantitative predictor, specifically "credit amount" `A00Amount100`.

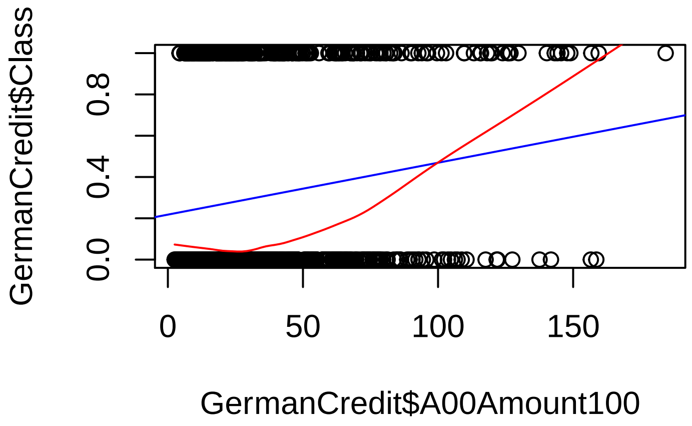
1. Create a boxplot of the variable `A00Amount100` categorized by `Class`. What can be said about the effect of credit amount on the probability of client not paying it back?

   ```
   boxplot(A00Amount100~Class, data = GermanCredit)
   ```



2. Create a scatterplot, with `A00Amount100` on the x-axis and `Class` on the y-axis. Add a straight line or a "Lowess" curve to the plot and iterpret the results.

```
plot(GermanCredit$A00Amount100, GermanCredit$Class)
#add a straight line (blue)
abline(lm(Class~A00Amount100, data = GermanCredit), col = "blue")
#add the Lowess curve (red)
lines(lowess(GermanCredit$A00Amount100, GermanCredit$Class), col = "red")
```



3. Build the model. Let $p(x_1) = P(Y = 1|x_1)$ be the probability of a client not paying back a credit of amount x$_1$100DEM. Then the model will look like $\log(\frac{p(x_1)}{1-p(x_1)} = \beta_0+\beta_1 x_1)$, or odds $= \frac{p(x_1)}{1-p(x_1)} = e^{\beta_0+\beta_1 x_1}$. For our data "not paying back the credit" is a success (we are modelling "not paying back". In `R` philosophy, the first level in any factor variable is treated as "failure". So firstly look into data set at first three variables and check the levels of `ffClass` and `fClass`.

```
levels(GermanCredit$fClass)
```

```
## [1] "Bad"  "Good"
```

```
levels(GermanCredit$ffClass) #This one is appropriate for our model
```

```
## [1] "failure" "success"
```

```
model <- glm(ffClass~A00Amount100, data = GermanCredit, family = binomial(link = "logit"))
```

4. Estimate the model parameters and interpret them. Does the interpretation of β_0 or e^{β_0} make sense? While interpreting β_1, find out how much higher/lower is the chance of "success" (=not paying back) when we compare groups of clients whose credit amounts differ by "one 100DEM", by "ten 100DEM's" and by "hundred 100DEM's".

```
summary(model)
```

```
##
## Call:
## glm(formula = ffClass ~ A00Amount100, family = binomial(link = "logit"),
##     data = GermanCredit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4158  -0.8269  -0.7688   1.3674   1.7022
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.229375   0.108332 -11.348  < 2e-16 ***
## A00Amount100  0.011189   0.002355   4.751 2.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.7  on 999  degrees of freedom
## Residual deviance: 1199.1  on 998  degrees of freedom
## AIC: 1203.1
##
## Number of Fisher Scoring iterations: 4
```

```
coef(model)
```

```
##  (Intercept) A00Amount100
##  -1.22937493   0.01118942
```

To obtain parameter estimater $e^{\hat{\beta}_k}$:

```
exp(coef(model))
```

```
##  (Intercept) A00Amount100
##    0.2924753    1.0112523
```

5. Obtain Wald confidence intervals for parameters β_0 and β_1.
$\beta_k \in (d, h) = (\hat{\beta}_k - u_{1-\alpha/2} se(\hat{\beta}_k), \hat{\beta}_k + u_{1-\alpha/2} se(\hat{\beta}_k))$. Check whether the standard error $se(\hat{\beta}_k)$ isn't much bigger than $\hat{\beta}_k$. In that case, Wald's statistics and subsequent tests are no longer reliable and likelihood ratio test would be more suitable.

```
confint(model, level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
##                     2.5 %      97.5 %
## (Intercept)  -1.444294879 -1.01935617
## A00Amount100  0.006593213  0.01584592
```

6. Obtain confidence intervals for e^{β_0} and e^{β_1}.
   $\mathrm{e}^{\beta_k} \in (\mathrm{e}^{d}, \mathrm{e}^{h})$

```
exp(confint(model, level = 0.95))
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %    97.5 %
## (Intercept)  0.2359124 0.3608272
## A00Amount100 1.0066150 1.0159721
```

7. Test the significance of both parameters by means of Wald's significance test.
   $\mathrm{H}_0 : \beta_k = 0$ vs. $\mathrm{H}_1 : \beta_k \neq 0$

```
summary(model)
```

```
##
## Call:
## glm(formula = ffClass ~ A00Amount100, family = binomial(link = "logit"),
##     data = GermanCredit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4158  -0.8269  -0.7688   1.3674   1.7022
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.229375   0.108332 -11.348  < 2e-16 ***
## A00Amount100  0.011189   0.002355   4.751 2.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1221.7  on 999  degrees of freedom
## Residual deviance: 1199.1  on 998  degrees of freedom
## AIC: 1203.1
##
## Number of Fisher Scoring iterations: 4
```

   We can test this by p-values, which can be found in the column "$Pr(> |z|)$".

8. Test the significance of β_1 using likelihood ratio test.
   `Null model:` $logit(p(x_1)) = \beta_0$ vs `Full model:` $logit(p(x_1)) = \beta_0 + \beta_1 x_1$.
   `Null Deviance:` $= D_0 = -2\log L_0$, where L_0 is the maximum likelihood of a "null" model including nothing but intercept. In our case, $logit(p(x_1)) = \beta_0$.
   `Residual Deviance:` $= D_1 = -2\log L_1$, where L_1 is the maximum likelihood of a "full" model including all predictors. In our case $logit(p(x_1)) = \beta_0 + \beta_1 x_1$.
   Likelihood ratio:

$$LR_{0,1} = D_0 - D_1 = -2\log\frac{L_0}{L_1} \approx \chi^2(df_0 - df_1).$$

   Better model has smaller deviance. If the full model is significantly better than the null model, it will lead to large values of $LR_{0,1}$, which is called the *likelihood-ratio test statistic*. Thus the concerned p-value is on the right tail of χ^2 distribution.

```r
anova(model, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: ffClass
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           999     1221.7
## A00Amount100  1   22.665       998     1199.1 1.929e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For our model, the likelihood-ratio test is in fact a test about the statistical significance of the parameter β_1. $LR_{0,1} = 1221.7 - 1199.1 = 22.6$; $LR_{0,1} \approx \chi^2(999 - 998)$. Thus $p = 0.000002$ and the full model is significantly better than the one including only intercept.

`AIC` $= k - 2\log L_1 = k + D_1$, where $k = 2*$ number of parameters. Better of the two models has smaller value of `AIC`. Compared with deviance, models are penalized for having too many parameters.
In our case, `AIC` $= 2 * 2 + 1199.1 = 1203.1$.

9. Test the model assumption by means of Hosmer-Lemeshow test (the odds ratio derived from the logistic regression model shouldn't be dependent on the credit amount, only on the difference between the credit amounts in different groups.)
   H$_0$: Goodness of fit (we don't want to reject)., vs. H$_1$: Model is badly fitted.
   For this test we have to install the *ResourceSelection* package, which includes `hoslem.test()` function. This will perform Hosmer-Lemeshow test.

```r
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```r
hoslem.test(model$y, fitted(model))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model$y, fitted(model)
## X-squared = 8.6107, df = 8, p-value = 0.3762
```

The first argument, `model$y`, is a binary vector of observations. Second argument, `fitted(model)`, are the fitted values of the model. The statistic of this test follows χ^2 distribution and is included in the output (`X-squared`), along with degrees of freedom and p-value of the test.

The result of Hosmer-Lemeshow test (mainly the p-value, which is equal to 0.3762), suggests that the null hypothesis should not be rejected, which is what we wanted. This means that our model is probably a good fit for the data.

R Instructions for Problem 2:

We are going to model the probability of not paying the credit using one qualitative predictor, `A15Gender`. In this case, boxplot and scatterplot are no longer suitable, as both the outcome and the explanatory variable are categorical. `abline` and Lowess curve also don't make much sense - they are degraded to a simple horizontal line.

However, we can still create a model:

```r
model2 <- glm(ffClass~A15Gender, data = GermanCredit, family = binomial(link = "logit"))
```

Repeat the same exercises as before, but be careful with interpreting the model coefficients.

R Instructions for Problem 3:

We are going to model the probability of not paying the credit using a vector of quantitative and qualitative predictors.