12 Seminar

The dataset *GermanCredit.RData* has been provided by professor Hans Hofmann (Institut für Statistik und Ökonometrie, Universität Hamburg) into the `http://archive.ics.uci.edu/ml/` repository in 1994 (November). The dataset consists of credit scoring data from the Federal Republic of Germany from the nineties. Details about all the variables can be found here: `http://archive.ics.uci.edu/ml/datasets/Stat`
The list of chosen variables follows:

`Class` ... credit classification (0=good-paid credit, 1=bad-unpaid credit) We are modeling bad credit (that is considered 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)

Problem 1

We are going to model the probability of unpaying the credit using one quantitative predictor „credit amount" `A00Amount100`.

1. Draw the box-plot of the `A00Amount100` variable categorized under `Class`. What can we say about the impact of credit amount on the credit unpaying?
2. Draw a plot where there is the `A00Amount100` on the horizontal axis and the `Class` on the vertical axis. Put a straight line or "Lowess"curve through the plots and interpret.
3. Construct a model.
   Let sign $p(x_1) = P(Y = 1|x_1)$; the probability that a client will not pay the credit of $x_1$100 DEM.
   Model: $\log\left(\frac{p(x_1)}{1-p(x_1)}\right) = \beta_0 + \beta_1 x_1$ or $\text{odds} = \frac{p(x_1)}{1-p(x_1)} = e^{\beta_0+\beta_1 x_1}$
4. Estimated the model paramters and interpret them. Does the interpretation of the β_0 or e^{β_0} parameter make sense?
   When interpreting β_1 discover how much higher/lower is the chance of unpaying when we compare groups of clients whose amounts of credit is different by one "hundred Deutche mark", by ten "hundred Deutche mark", or by a hundred "hundred Deutche mark".
5. State the confidence intervals (Wald) for β_0 and β_1 parameters.
   $\beta_k \in (d, h) = (\hat{\beta_k} - u_{1-\alpha/2} se(\hat{\beta_k})\ ,\ \hat{\beta_k} + u_{1-\alpha/2} se(\hat{\beta_k}))$. Check if the standard error $se(\hat{\beta_k})$ is not strongly higher than $\hat{\beta_k}$. In that case the Wald stats and the subsequent tests are not reliable and the ratio of confidence test is more suitable.
6. State the confidence intervals for e^{β_0} and e^{β_1}.
   $e^{\beta_k} \in (e^d, e^h)$
7. Test the significance of both paramters using Wald test of dependence.
   $H_0:\ \beta_k = 0$ vs. $H_1:\ \beta_k \neq 0$
8. We are testing the significance of β_1 using ratio of credibility.
   Null model: $logit(p(x_1)) = \beta_0$ vs. Full model: $logit(p(x_1)) = \beta_0 + \beta_1 x_1$.
9. Test the model assumption by the Hosmer-Lemeshow test (the odds ratio derived from the logistic regression model must not be dependent on the credit amount, but only on the difference of the credit amount during comparing two groups.)
   H_0 : Goodness of fit (I do not want to reject), H_1 : Not goodness of fit.

Problem 2
We are going to model the probability of unpaying the credit using one qualitative predictor. `A15Gender`.

1.

Problem 3
We are going to model the probability of unpaying the credit using vector of quantitative and qualitative predictors.