Contents

Exercise 4 (15 Points): Linear regression

Exercise 5 (15 Points): Logistic regression

Submission requirements

- Describe the problem and the methods used
- Discuss and interpret your results
- Include all graphics and code you used to solve the problem

Guidelines for writing code

- *Keep it simple.* Break down complexity into simpler chunks. Avoid implicit or obscure language features. Minimize scope, both logical and visual. Strive for clarity first, then efficiency.
- *Keep it readable.* Use informative variable names. Avoid abbreviations whenever possible. Name variables with noun or adjective noun combinations. The usual way of separating two words in R is the use of a point like in 'linear.fit'.
- *Comment your code.* Clearly comment necessary complexity. Be clear and concise. Do not restate code. Keep code and comments visually separate.

Exercise 4 (15 Points). Consider the data set `kc_house_data.csv` providing the prices and specifications of houses sold between May 2014 to May 2015 in King County (Seattle, USA). For our analysis we consider the following variables (they may have to be recoded):

`price`: Price

`bedrooms`: Number of bedrooms

`bathrooms`: Number of bathrooms per bedroom

`sqft_living`: Square footage of the home

`floors`: Total floors in house

`view`: Has been viewed (1 for viewed, 0 for not viewed / has to be converted)

`condition`: How good is the condition (from 1 to 5)

`grade`: Grade given to the housing unit (from 1 to 13)

`yr_built`: Year the house was built

a) Estimate a linear model with the response variable `price` and all remaining variables as covariates. Are all variables significant? How large is R^2 and how can this be interpreted? Perform the residual analysis to validate the model. Are there any departures from the linear regression model assumptions?

b) Produce a histogram and a QQ-plot of the response variable `price`, as well as of its log-transform `log(price)`. Compare both distributions to the normal one. Fit now a linear model with the response variable `log(price)`. Compare the estimated model with the one from a) in terms of R^2, significance and effect of covariates and model fit via residual analysis. Which model is more adequate?

c) In the model from b) interpret the effect of each covariate on the response. Plot each covariate against `log(price)`. Is the assumption of the linear dependence between covariates and response plausible for all covariates? Add to the model from b) squared terms for `yr_built` and `sqft_living`. Are these terms significant? Does adding these two terms improve the model fit in terms of R^2?

d) Now we would like to compare how well models from b) and c) make prediction. For this divide the dataset into a training and a test set. Sample randomly 10 806 rows to include into the training set (make sure to sample all levels) and test set. Fit both models on the training set and make predictions on the test set. Calculate the mean squared difference between predicted values and values of `log(price)` from the test set for each model. Which prediction error is smaller? Try to extend the model to improve the prediction: my best model gives prediction error of 0.0953.

Exercise 5 (15 Points). Consider the data set `donors.txt` on blood donation. It is available at the **UCI Machine learning repository**. This page contains also the background information on the data. The goal is to build a model that allows to predict best if a donor will donate the blood. The dataset contains the following variables.

`recency`: months since last donation

`frequency`: total number of donations

`amount`: total blood donated in c.c.

`time`: months since first donation

`donation`: 1 stands for donating blood; 0 stands for not donating blood

a) Read the data into R and fit a generalised linear model with the binary response `donation` and covariate `frequency` using the canonical link function. Fit the same model replacing the covariate by `amount`. Compare it to the first model. Plot variable `frequency` against `amount`. Comment on the results. Do you need both of these variables in the model?

b) Fit now the GLM model with the response `donation` and covariate `recency` using all link functions available in the `glm` function. Compare obtained estimators and comment on the differences.

c) Now we would like to build a model that makes the best prediction for the blood donations.

- First divide the dataset into a training and a test set. Sample randomly 374 rows to include into the training set and the rest will be the test set. Fit a GLM model with the response `donation` and canonical link on the training set, choosing appropriate covariates. Predict the model on the test set.
- With the predicted probability perform the classification: set the predicted ith value of `donation` to 0, if the corresponding ith predicted probability is less than 0.5 and to 1 otherwise. Assess the goodness of your classification calculating the classification error

$$CE = \frac{1}{374} \sum_{i=1}^{374} \left| y_i^{test} - \hat{y}_i^{test} \right|,$$

where y_i^{test} is the ith value of `donation` from the test set and $\hat{y}_i^{test}$ is its prediction. Try to extend the model to improve the classification error. Can you beat my classification error of 0.2085561?