Practical 1: Linear and logistic regression

Verena Zuber, Jelena Besevic, Alpha Forna, and Saredo Said

24/1/2019

Part 1: Analysing type 2 diabetes progression using linear regression

Type 2 diabetes is a long-term metabolic disorder that is characterized by high blood sugar, insulin resistance, and relative lack of insulin. The number of people diagnosed with diabetes in the UK has more than doubled in the last twenty years. According to diabetes.org.uk/ figures show that there are now almost 3.7 million people living with a diagnosis of the condition in the UK, an increase of 1.9 million since 1998.

In this practical we consider an observational study that measures progression in type 2 diabetes as outcome and several clinical parameters like age. sex, and bmi, but also common risk factors like map: blood pressure, tc: total cholesterol, ldl: low-density lipoprotein, hdl: high-density lipoprotein, tch: total cholesterol, ltg: triglycerides, and glu: glucose. The dataset includes n=442 cases. It is available in the lars package, which is easy to install using for example the install.packages("lars") command.

As a first step we load the data and assign x as a data.frame of the predictors and y as the quantitative outcome of type 2 diabetes progression.

```r
library(lars)
data(diabetes)
x = as.data.frame.matrix(diabetes$x)
y = diabetes$y
```

From the literature we know that the following 6 predictors are important for type 2 diabetes progression: sex, age, bmi, glu, map and ltg. We consider these 6 predictors as model 1.

Question 1.1 Look at the correlation structure between those 6 predictors and discuss the implications. Use the function corrplot() in the corrplot package to visulalise the correlation structure.

Question 1.2 Does the outcome disease progression follow a Normal-distributed? Look at general summary statistics of y, plot a histogram and a q-q plot against the Normal-distribution.

Question 1.3 Fit a linear model including the predictors of model 1 (sex, age, bmi, glu, map and ltg) using the lm() function and discuss the summary of the model.

Question 1.4 Perform model diagnostics and outlier detection of model 1. Do you think this is a good model fit? Justify your answers.

Question 1.5 Compute the OLS regression coefficient estimate using matrix multiplication $\hat{beta}_{OLS} = (x^t x)^{-1} x^t y$. Use the solve() function to invert a matrix. R distinguishes between scalar multiplication ($\star$) and matrix multiplication (%$\star$%). Make sure to use the matrix multiplication (%$\star$%) for this task and ensure that your matrices have the correct dimensions. Add an intercept by including a row of ones like for example this.

```r
x1 =  cbind(rep(1,nrow(diabetes$x)), x$sex, x$age, x$bmi, x$glu, x$map, x$ltg)
```

Question 1.6 Compute the regression coefficient estimate using the sample covariance based estimate, defined as $\hat{beta}_{COV} = cov(x)^{-1} cov(xy)$. Use the solve() function to invert the covariance matrix cov(x) of dimension 6 x 6 and compute this estimate without the intercept using

```r
x11 =  cbind(x$sex, x$age, x$bmi, x$glu, x$map, x$ltg)
```

Question 1.7 Compare the 3 estimates from questions 1.4-1.6.

Question 1.8 Fit model 2 that only includes glucose and compare how it differs from the multivariable model 1.

Part 2: Predict type 2 diabetes progression using linear regression

Assume we only observed the first 300 cases and need to use the first 300 cases as training data.

```
x_train = data.frame(x[1:300,])
y_train = y[1:300]
```

Now we can consider the remaining 112 cases as new data points for whom we want to predict disease progression.

```
x_new = data.frame(x[301:442,])
y_new = y[301:442]
```

Question 2.1 Use the linear model 1 to predict the disease progression for the 112 cases with predictor information stored in x_new.

Question 2.2 Evaluate the error of your prediction based on linear model 1 by computing the squared difference between the predicted progression and the actual observed progression saved in y_new. Plot a histogram of the squared difference and compute the mean and median.

Question 2.3 Repeat the steps 2.1 and 2.2 using the univariable linear model 2 including only glucose. Contrast the prediction error of linear model 2 with the prediction error of linear model 1.

Question 2.4 Is it good practise to evaluate the prediction performance on a single training data? How appropriate is the split to take the first 300 cases?

Part 3: Distinguishing between severe and light cases of type 2 diabetes using logistic regression

Doctors are particularly concerned with type 2 diabetes cases that have a bad disease progression, in particular cases that have a disease progression score larger than 200. Binarise your outcome like this:

```
y_binary = as.numeric(y>200)
```

Question 3.1 Fit a generalised linear model using the glm() function that can distinguish between bad disease progression and normal progression. Use the 6 predictors as considered in model 1. Look at the summary of the glm output and interpret the findings.

Question 3.2 Consider now model 2 including only glucose. Fit a glm and see if glucose can distinguish between bad disease progression and normal progression.

Question 3.3 Look again at the training data (x_train and ybin_train) based on the first 300 cases, where

```
ybin_train = y_binary[1:300]
```

Build a prediction rule based on model 1 using the training data (x_train and ybin_train) using the glm function. In a second step predict which of the new samples (using x_new as predictor matrix) are at high risk for having a bad diagnosis. How many of the 112 new observations have a probability larger than 0.5 to have bad progression?

PS Use the inverse logit function ($logit^{-1}(eta) = exp(eta)/(eta + 1)$) to transform the linear predictor ($eta = x\beta$) back to a probability which ranges between 0 and 1.

Part 4 (Optional): Which risk factors are important for type 2 diabetes progression?

Look again at the complete dataset including all n=442 cases including all 10 predictors. How would you perform variable selection to decide which variables are important for disease progression in type 2 diabetes?