Exercise Logistic regression. Day 2

Exercise 1

The file ‘bonemarrow.csv’ contains data on 166 patients with cancer receiving a bone marrow transplantation from a family member. The variables in the data set are:

- The outcome AGVHD, which indicates the presence or absence of acute graft versus host disease (AGVHD =1) respectively (AGVHD=0) . AGVHD=1 indicates that the bone marrow transplant is rejected.
- DIAG: diagnosis of the disease in 3 categories.
- AGEDON: age of the donor.
- AGEREC: age of the recipient (=the patient).
- SEXDON: biological sex of the donor (0=male, 1 = female)
- SEXREC: biological sex of the recipient. (0=male, 1 = female)

a. We would like to predict which patients are at highest risk to develop AGVHD. Start performing a logistic regression analysis with AGVHD as dependent and AGEREC as independent variable. Create a new variable with the predicted probabilities and make a plot with the predicted probabilities on the y-axis and the age of the recipient on the x-axis. What do you see?

b. A 95% confidence bound for the predicted probabilities can be obtained as follow:

- First, calculate for each person the logit with its standard error as follows:
```
logit <- predict(model.lr1, type="response", se.fit = TRUE)
```
- Then use the fitted value and the standard error to calculate the lower and upper bound of the 95% confidence interval
```
logit.lwb <- logit$fit-1.96*logit$se.fit
logit.upb <- logit$fit+1.96*logit$se.fit
```
- Transform the upper and lower bound from the logit scale to the probability scale.

Make a plot of the predicted probabilities with the 95% confidence bounds.

c. Check if the assumption of a linear effect of AGEREC on the logit scale is adequate by adding a quadratic term in AGEREC. Make a plot of the predicted probabilities of the model with the quadratic term and compare the model with only a linear term. What do you conclude?

d. Make a graph of AGEDON against AGEREC. What do you see?

e. Make a model with both AGEDON and AGEREC. Compare the regression coefficient of AGEREC in this model to the coefficient found in part a. What do you observe? Explain.

f. Make a model with AGEDON and the diagnosis (as categorical variable). Again make a plot of the predicted probabilities against age of the donor. Try to link the three curves in the plot with the three categories of the diagnosis variable.

g. Add sex donor and sex recipient to the model. Note the p-values. What is the deviance of this model? How many parameters have been estimated? Check if the formula for the AIC given in the lecture corresponds to the AIC of the output.

h. You can perform a likelihood ratio test to compare this model (alt_model) to the previous model with only AGEDON and DIAG (null_model), using a likelihood ratio test, as follows:

1. Obtain the log likelihoods of the two models, and calculate the Likelihood ratio test statistic:
   lr_stat <- -2 * (logLik(null_model) - logLik(alt_model))
2. Calculate the difference in number of parameters. This can be done as follows:
   df <- df.residual(null_model) - df.residual(alt_model)
3. Calculate the p-value using the chi-square distribution:
   p_value <- 1 - pchisq(lr_stat, df)

i. Add an interaction term between sex donor and sex recipient. What do you see? Calculate for each combination of sex donor and sex recipient the odds ratio and try to give an explanation.

j. Calculate a variable named mismatch which is 1 if donor and recipient are of different sex, and 0 if they are of the same sex. Use this variable instead of the model with sex donor, sex recipient and the interaction. Compare the deviances of the models.

k. perform a stepwise variable selection procedure using

stepwise_model <- step(initial_model, direction = "both")

where the initial model contains the variables DIAG (categorical), AGEDON, AGEREC, SEXDON, SEXREC and MISMATCH, and explore the resulting model.

l. Check if there are any points with high leverage by using the command `plot(hatvalues(stepwise_model)`. Try to identify the observation with the very high leverage value and find an explanation for the high value.

Exercise 2

Consider a cohort study where the relation between infection with the human papillomavirus ((X=1 is infected,X=0 is not) and cervical cancer (Y=1 is cancer, Y=0 no cancer) is studied.

The following code generates (hypothetical) data from a cohort study on this research question

```
nsim<-100000
set.seed(2468)
probX <- 0.5
X <- rbinom(n=nsim, size=1,prob=probT)
probY <- 0.01 + 0.05*X
Y<-rbinom(n=nsim, size=1, prob=probY)
out <- cbind(X,Y)
cohort<-data.frame(out)
```

a. Look at the code. What are the true Risk Difference, Relative Risk and Odds Ratio in the population?
b. Run this code, and calculate the observed risk difference, relative risk and odds ratio in the generated data. Compare with the true values.
c. Now generate data using outcome dependent sampling (a case control study) by selecting all cases (those with Y=1) and a random sample of 1 percent of controls.
d. Calculate the risk difference, relative risk and odds ratio in the generated case-control data. What do you observe?