Exercise Logistic regression. Day 1

Exercise 1.
We consider data on 358 women with ovarian cancer. Every patient has undergone surgery to remove as much of the tumor as possible. Patients have been followed for at least 4 years after the surgery. The variable death indicates whether the patient died (`death`=1) or was still alive (`death` =0) at 4 years. As explanatory variables we use:

- karn = Karnovsky-index (a measure for the mobility of the patient: the lower the Karnovsky-index, the worse the condition of the patient) at the start of follow-up.
- diameter, the diameter of the residual tumour in cm after the surgery.
- figo staging, an indication of how far the tumour has grown, III (figo = 0) or IV(figo=1).
- ascites, fluid in the abdomen 0 "unknown", 1 "absent", 2 "present".

a. The data are stored in the file ova.csv. Read this file into R and view the data. Make a 2x2 table of death against figo.

b. Use the 2x2 table to estimate the 4-year probability of dying for a patient with figo = 1, and for a patient with figo = 0. Which statistical test is most suited to compare the probability to die within 4 years between the two figo stages? Perform that test and state your conclusions.

c. Compute from the 2x2 table the odds ratio on death for patient with FIGO=1 versus FIGO = 0.

d. Perform a logistic regression analysis with death as dependent variable and figo as independent variable. What is your conclusion regarding the association between figo and death? Look at the estimate of beta for figo, and compare exp(b) with the odds ratio found under c.

e. Write down the logistic model and calculate both for a woman with figo= 0, and for a woman with figo=1, the estimated probability of death within 4 year. Compare these probabilities with the probabilities calculated in b.

e. Calculate by hand a 95 % confidence interval for the regression coefficient for figo, and use this to construct a 95% confidence interval for the odds ratio in the population for figo stage. You can check your answer by using the function `confint.default`, which calculates the Wald confidence intervals.

f. Calculate also profile log likelihood confidence intervals, using the function `confint`. Are the confidence intervals the same?

g. Carry out a logistic regression with diameter as continuous covariate in the model.
How do you interpret the estimated regression coefficient and exp(B) here? What are does the Wald test for the association between diameter and death indicate?

h. Find the deviance of the model and the model with diameter in the model to a model with only an intercept. What would be the value of the likelihood ratio test? Conclusion?

i. Compute the logit and the probability of death within 4 years for a woman with a tumor with a diameter = 5.

j. You can calculate the predicted logit and the predicted probability on death for each women, by using the function `predict`, with and without the option `type="response"`. Make a plot of diameter versus the logit and the predicted probability.

k. Now we study the effect of the categorical variables ascites. First make a cross tabulation of death against ascites and estimate the probability of death in the 3 categories.

l. Perform a logistic regression with ascites as factor in the model. Show that the results of the logistic regression analysis coincide with those of the simple cross tabulation. Try to interpret all regression coefficients. Which category of ascites is used as reference category?

m. Finally make a model with all covariates ascites, karn, figo and diameter. Look at the coefficients for ascites and the corresponding significance and compare them with the previous model. Explain the differences.

Exercise 2.

a. The data are stored in the file data_logreg.csv. Read this file into R and view the data. Make a 2x2 table of y against x. What do you observe? What is the value of the odds ratio?

b. Perform a logistic regression with y as dependent and x as independent variable. What do you observe?