Exercise Generalized linear models

Exercise 1.

a. Open the bone marrow transplant dataset which you analysed yesterday. Perform a linear regression with age donor as dependent variable and age recipient as independent variable. Use summary to look at the output and write down the estimated coeffient for agerec with standard error, plus the residual standard deviation (σ).

b. To fit a generalized linear model you have to specify the link function, the distribution of the outcome variable and the predictor variables in the model. A generalized linear model in R is specified by

```
glm(formula, family = , data, , ...)
```

Here, the family indicates the distribution of the response variable. There are different options: `binomial,gaussian,Gamma,inverse.gaussian,Poisson, quasi, quasibinomial,quasipoisson`. Each distribution has its own 'canonical' link function. The glm function uses the canonical link function, unless specified otherwise.

Perform a linear regression with the same variables, using the glm function, specifying the correct family. Compare the parameter estimates, standard errors. Also find an estimate of σ. Are the results exactly the same as in 1.a?

c. Perform a logistic regression with AGVHD as dependent and agedon as independent variable. What link function is used?

d. It is possible to use a different link function. For example use a identity link by specifying `family = binomial(link = "identity")`. What is the model for π_i used here? How do you interpret the regression coefficient for agerec.

Exercise 2

In the following exercise, we consider a study on oral contraceptives (the "pill") . A rare complication of birth control pills is thrombosis, which occurs when blood clots block veins or arteries. Prescription data from Dutch pharmacies were linked to the Dutch national disease registry. The aim of the study was to explore if the type of pill (second versus third generation pill) was related to the risk of thrombosis.

The data are aggregated per age at start, type of pill used and month after start of using the pill. For each combination the number of thrombosis cases occurring and the total number of users were calculated.

In this exercise we only use data from the first year after starting the pill.The data are in the file pill.csv. The variables are

- month: Month after start of using the pill
- age: age of the women at start using the pill

- type_pil: second (type_pill=2) or third (type_pill=3) generation pill
- thrombosis: number of thrombosis cases occurring
- users: total number of pill users in the specific category.

a. Explore the data. Calculate the total number of women with thrombosis and the total number of "person-months" (sum of users over the months) , separately for second and third generation pill.

b. Calculate the expected number of thrombosis cases per person-month separately for second and third generation pill.

c. We want to model the rate of thrombosis (counts per person-month), using a Poisson model, with type of pill as independent variable and number of thrombosis cases as dependent variable. In this model log(users) should be specified as an offset.

Fit a glm , with `family = poisson` and specify the formula as
`thrombosis~as.factor(type_pill)+offset(log(users))`

Write down the model and calculate the rate of thromboses cases per person per month for the second and third generation pill. Calculate also the rate ratio for a third generation pill versus a second generation pill.

d. The thrombosis rate may also depend on age and month after starting the pill. This may have confounded the effect of pill on thrombosis, as older women more often use the third generation pill. Add age and month to the model. Explore the effect of these two variables to the model, and the effect of type of pill. Again calculate the rate ratio for a third generation pill versus a second generation pill.

e. Check the goodness of fit of the model using the likelihood ratio test.

f. Check if there is overdispersion in the data using `family = quasipoisson` . Conclusion?