# Series 7

## 1. Poisson Regression

In an experiment one is interested in assessing the concentration of nematodes (a species of worms) in a certain liquid. Three samples of 20 $\mu$l each were generated and thinned with an equal amount of water. From each of the thinned probes the researchers generated 15 subsamples with a concentration of 40 $\mu$l, 20 $\mu$l and 20$\mu$l, respectively. The table shows the counted number of nematodes for each of the 45 subsamples.

| Sample | volume | Number of nematodes in each subsample | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 40 $\mu$l | 31 | 28 | 33 | 38 | 28 | 32 | 39 | 27 | 28 | 39 | 21 | 39 | 45 | 37 | 41 |
| 2 | 20 $\mu$l | 14 | 16 | 18 | 9 | 21 | 21 | 14 | 12 | 13 | 13 | 14 | 20 | 24 | 15 | 24 |
| 3 | 20 $\mu$l | 18 | 13 | 19 | 14 | 15 | 16 | 14 | 19 | 25 | 16 | 16 | 18 | 9 | 10 | 9 |

**a)** Propose a regression model for the dependent variable "number of nematodes" and the independent variable sample.

**b)** Is there any difference among the three samples?

**c)** Could one also use the log-volume instead of the three groups as explanatory variable? How would the corresponding model look like?

**d)** Would $\lambda = c \cdot vol$ also be an appropriate model? Why? Hint: Is the coefficient of log(vol) significantly different from 1?

**e)** Calculate a model in which you fix log(vol) at 1.
**Hint**: With the `offset` comand you can constrain the coefficient of a covariate to 1. E.g., when writing `offset`($x_1$) instead of $x_1$ in the model formula in R the coefficient of $x_1$ is fixed at 1 and not estimated.
Compare the different models.

## 2. Multinomial Logit Model

Do people with the freedom to choose a different investment strategy for their pension make that choice when they have a saying in that investment strategy? The data set `pension.dta` can be read into R with the `read.dta` command from the library `foreign`. The link for the data set is

```
read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge2k/PENSION.DTA")
```

The data set contains observations of 226 subjects on the following variables.

| | |
|---|---|
| id | Identification number of the person |
| pyears | Number of years in pension plan |
| age | Age in years |
| choice | Freedom to choose investment strategy 1=yes, 0=no |
| prftshr | Profit sharing plan 1=yes, 0=no |
| female | Sex 1=female 0=male |
| married | married 1=yes, 0=no |
| black | Ethnic background 1=African-American 0= not African-American |
| educ | Years of education in school |
| finc25 | Income $\leq 25,000$\$ |
| ... | ... |
| wealth89 | Net assets 1989 in 1000 \$ |
| pctstk | Investment strategy 0=mainly obligations, 50= mixed, 100=mainly stocks |

**a)** What is the relationship between freedom to choose the investment strategy and the de-facto chosen investment strategy whilst neglecting all other variables?
**b)** Is there a relationship between income and investment strategy?
**c)** Construct a new income variable with three levels: Levels: up to 25,000, 25,001 up to 50,0000, over 50,000.
**Hint**: Partitioning a continuous variable into several categories is convenient for modelling a non-linear effects.
**d)** Fit a nominal logit model with pctstk=50 as reference category.
**e)** Is the variable choice significant? Interpret the coefficients of `choice` using odds.
**f)** How large is the probability for each of the three investment strategies to be chosen for a 60 year old white male, single, with 13.5 years of school education, an income of over 50,000 $, net assets of 200,000 $ and a profit-sharing plan in the two cases that he either does or does not have the freedom to choose his investment strategy.

## 3. Logistic Regression for Binary Data

A car manufacturer instructed a market research company to analyze which families are going to buy a new car next year using a logistic regession model. Data stems from a random sample of 33 families from an agglomeration area. Assessed variables cover the yearly household income (in 1000 US $) and the age of the oldest car in the family (in years). 12 months later, interviewers assessed which families had bought a new car in the meantime. The data is available in the file `car.dat` and can be read in with following command.

```
read.table("http://stat.ethz.ch/Teaching/Datasets/car.dat",header=T)
```

**a)** Perform a logistic regression. Report the regression equation.
**b)** Estimate $\exp \hat{\beta}_{income}$ and $\exp \hat{\beta}_{age}$ and interpret the values.
**c)** How large is the estimated probability that a family with a yearly household income of 50 000 US $ and whose oldest car is 3 years old will buy a new car?
**d)** Do the residual plots show any abnormalities?
**e)** Is the variable `age` required in the model?
**f)** Is there a non-negligible interaction between `income` and `age`?

## 4. Logistic Regression for Binomial Data

In this task we analyze the example concerning hypertension from Altman (1991). First, we need to enter the data. This is done as follows:

```
> no.yes <- c("No", "Yes")
> smoking <- gl(2,1,8, no.yes)
> obesity <- gl(2,2,8, no.yes)
> snoring <- gl(2,4,8, no.yes)
> n.total <- c(60, 17, 8, 2, 187, 85, 51, 23)
> n.hyper <- c(5, 2, 1, 0, 35, 13, 15, 8)
```

Here, the function `gl` creates a factor with given levels. The factors `smoking, obesity` and `snoring` have an obvious meaning. `n.total` is the number of observations and `n.hyper` is the number of people with hypertension in each group.

**a)** In order to fit a binomial logistic regression model construct a response matrix with two columns containing the number of people with and without hypertension, respectively.
**b)** Fit a binomial logistic regression model to the data.
**c)** Does this model fit well? Assess the goodness-of-fit via the residual deviance.
**d)** Which variables significantly influence the occurence of hypertension?
**e)** Try to find a suitable model. Perform likelihood-ratio tests to achieve this goal.
**f)** Compare the observed and fitted proportions for hypertension under model e). What is striking here? Additionally, calculate the expected and observed counts.

**Preliminary discussion:** Monday, December 12.

**Deadline:** Monday, December 19.