Multivariate Analysis for the Behavioral Sciences,
Second Edition (Chapman and Hall/CRC, 2019)

Exercises of Chapter 5: Generalized Linear Models

Kimmo Vehkalahti and Brian S. Everitt

9 November 2018

Exercises

Exercise 5.1

Use the bladder cancer data (see below) and modify the R code given in the **Examples of Chapter 5** to analyse the data as instructed.

```
bladder <- read.table("data/bladder.txt", header = T)
bladder <- within(bladder,
             X <- factor(X, labels = c("< 3cm", "> 3cm"))
)
bladder
```

```
##    Time     X n
## 1    11 > 3cm 1
## 2     2 < 3cm 1
## 3     3 < 3cm 1
## 4     6 < 3cm 1
## 5     8 < 3cm 1
## 6     9 < 3cm 1
## 7    10 < 3cm 1
## 8    11 < 3cm 1
## 9    13 < 3cm 1
## 10   14 < 3cm 1
## 11   16 < 3cm 1
## 12   21 < 3cm 1
## 13   22 < 3cm 1
## 14   24 < 3cm 1
## 15   26 < 3cm 1
## 16   27 < 3cm 1
## 17    7 < 3cm 2
## 18   13 < 3cm 2
## 19   15 < 3cm 2
## 20   18 < 3cm 2
## 21   23 < 3cm 2
## 22   20 < 3cm 3
## 23   24 < 3cm 4
## 24    1 > 3cm 1
```

```
## 25    5 > 3cm 1
## 26   17 > 3cm 1
## 27   18 > 3cm 1
## 28   25 > 3cm 1
## 29   18 > 3cm 2
## 30   25 > 3cm 2
## 31    4 > 3cm 3
## 32   19 > 3cm 4
```

Helpful hints:

Consider a *Poisson process*, in which the waiting times between successive events of interest (the tumors in this case) are independent and exponentially distributed with common mean, $1/\lambda$ (say).Then the number of events that occur up to time t has a Poisson distribution with mean $\mu = \lambda t$. Here the parameter of real interest is the rate at which events occur, λ, and for a single explanatory variable, x, we can adopt a Poisson regression approach using the model

$$\log(\lambda) = \log \frac{\mu}{t} = \beta_0 + \beta_1 x$$

to examine the dependence of λ on x. Rearranging this model we obtain

$$\log(\mu) = \beta_0 + \beta_1 x + \log t$$

In this form the model can be fitted within the GLM framework with $\log t$ as a variable in the model whose regression coefficient is *fixed* at unity; this is usually known as an *offset*.

Exercise 5.2

Use the CHDrisks data (see below) and modify the R code given in the **Examples of Chapter 5** to analyse the data.

```
CHDrisks <- read.table("data/CHDrisks.txt", header = T)
CHDrisks <- within(CHDrisks,
            {
             Smoking <- factor(Smoking, labels = c("non-smoker", "1-10", "11-20", "20+"))
             Press <- factor(Press, labels = c("< 140", ">= 140"))
             Behavior <- factor(Behavior, labels = c("Type B", "Type A"))
            }
)
str(CHDrisks)
```

```
## 'data.frame':    16 obs. of  5 variables:
##  $ Years   : num  5268 2542 1141 615 4451 ...
##  $ Smoking : Factor w/ 4 levels "non-smoker","1-10",..: 1 2 3 4 1 2 3 4 1 2 ...
##  $ Press   : Factor w/ 2 levels "< 140",">= 140": 1 1 1 1 1 1 1 1 2 2 ...
##  $ Behavior: Factor w/ 2 levels "Type B","Type A": 1 1 1 1 2 2 2 2 1 1 ...
##  $ nCHD    : int  20 16 13 3 41 24 27 17 8 9 ...
```

Helpful hints:

Letting y_i be the number of cases of CHD and T_i be the person years of follow-up (this is defined as the total duration of observed follow-up, from entry into the study until either disease detection or end of follow-up) where i indexes the risk group and takes values 1 to 16. To fit a GLM with only smoking behavior as the single risk factor assume that the values of this variable are quantitative although this is not strictly the case and an alternative would be to use a factor (or three dummy variables) to code the four categories of smoking. The model used is the same as for the bladder cancer data (in Exercise 5.1):

$$\log \frac{\mu_i}{T_i} = \beta_0 + \beta_1 \text{Smoking}_i$$

,

where $\mu_i = E(y_i)$ and T_i is the number of person years for category i. Remember that $\log T_i$ has to be included as an offset (see previous exercise).