Assessed practical 2: Investigating the effect of sex, treatment dose and therapy status on mortality rate of rats

Practical Number: P535

June 28, 2021

Abstract

A logistic regression is fitted to the data in order to asses the effect of sex, therapy status and dose level on the mortality rate of rats. The difference in deviance is used to choose which model gives the best fit to the data. The residuals are analyzed and it is argued that the model is valid. It is shown that mortality rate does not depend on therapy status. The significant terms are dose levels and the interaction between sex and dose levels. In a subsequent analysis, more data points should be collected in order to reduce the uncertainty in the fitted values.

Keywords: Deviance; generalized linear model; logistic regression; residuals

1 Introduction

The data analyzed in this report were collected during an experiment conducted on rats with lung cancer. A new treatment for lung cancer was being tested on the rats. The data set consists of 24 batches of 25 rats tested at six different dose levels and split into males and females. The response variable is the number of rats who died whilst undergoing treatment, for each batch. In other words, there are 24 observed responses. The responses will be considered as the proportion of the number of dead rats over the total number of rats (25), in order to fit a logistic models for these proportions. The explanatory variables are dose, sex and standard (which indicates whether the new treatment was in addition to standard therapy (1) or without standard therapy (0)). The 24 batches are equally split between males and females, therapy status and dose (i.e, 12 batches of females out of which 6 have standard therapy and 6 do not, for each treatment dose). Table 1 below gives an overview of the recorded data for the first four batches.

Batch	Number dead	Number alive	Dose	Sex	Therapy status
1	15	10	1	M	1
2	10	15	2	M	1
3	8	17	3	M	1
4	5	20	4	M	1

Table 1: Example of recorded data for the first four batches where M stands for male.

The response for the first batch will then be $15/25 = 0.6$. Note that clearly, sex and therapy status are categorical variables (factors). It is not so obvious whether dose should be treated as a factor or a continuous variable. However, it is said in the settings that the rats are given six different dose levels, hinting that dose should be treated as factor. Indeed, the amount of product for each dose is not specified (or whether the doses are proportional or not, in which case they would be treated as continuous variables). Because we lack such information, the only sensible choice is to treat the doses as factors (i.e, the doses could be any amount of treatment). In the following statistical analysis, dose will thus be taken as a factor. In this report, we are interested in investigating the effect of sex, dose and therapy status on the risk of mortality.

The data will be firstly investigated in a preliminary analysis to have a better overview on how the variables may affect the mortality rate. After which, a logistic regression will be fitted to model the risk of mortality based on dose only, and then based on all three categorical variables. Once the best model is chosen (using the likelihood-ratio test), its validity will be assessed. Finally, results will be presented and the relationships between the variables will be interpreted. The statistical analysis is done using the software R and the code is provided in Appendix A.

2 Preliminary analysis

Figure 1 shows a plot of the observed mortality rates for both sex and therapy status, across the 6 dose levels.

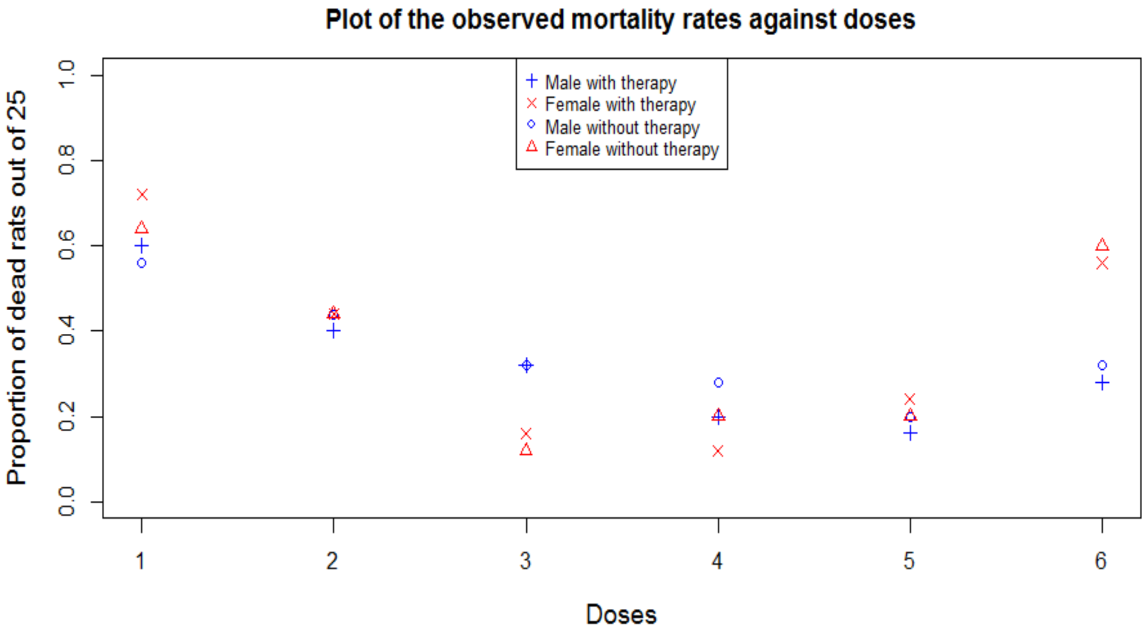

Figure 1: Plot showing the observed data for both sex and therapy status against doses.

One can see in Figure 1 that there does not seem to be a visible difference in mortality rate for the batches with and without therapy. Furthermore, it does not seem like sex have an impact on mortality rates as the red points are sometimes below and sometimes above the blue points, depending on the dose (hinting for interaction between the two). However, there appears to be a consequent difference between males and females for dose 6. Finally, one can assume that the doses have a significant effect on mortality rates. The relationship between mortality rate and dose looks quadratic, but since dose is taken as a factor, it is not possible to interpret as one could simply relabel/reorder the doses. Figure 2 below further confirms that sex and therapy do not seem to have any impact on the mortality rate across all doses.

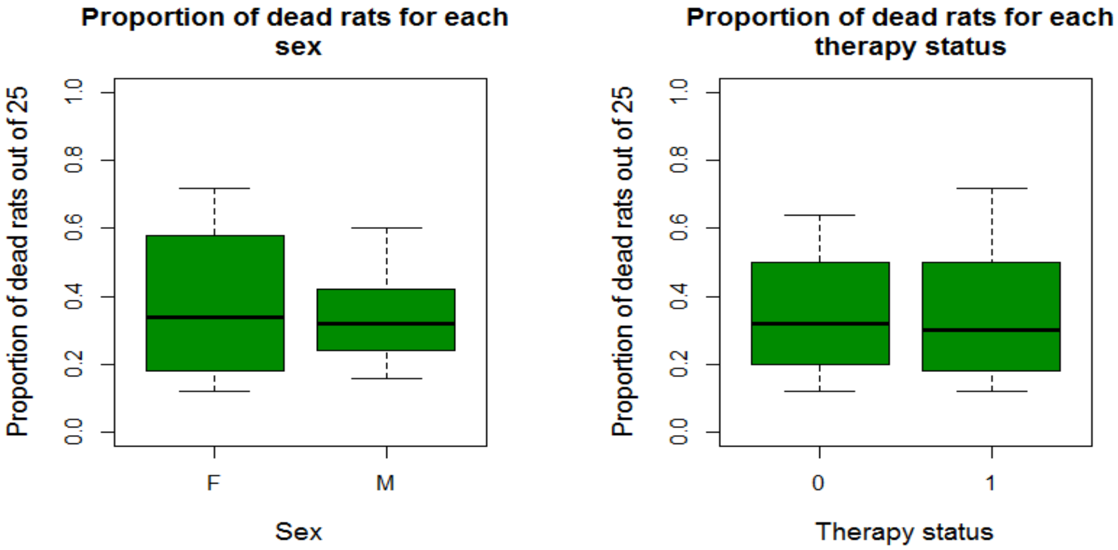

Figure 2: Boxplots showing the effect of sex and therapy status on mortality rate.

It can thus be expected that the two factors will not show any significant effects on mortality rate, later in the analysis. However, it is of interest to investigate whether there might be any interaction between the factors, which is hard to tell looking at Figure 1. Figure 3 below displays all the possible two-way interactions between factors.

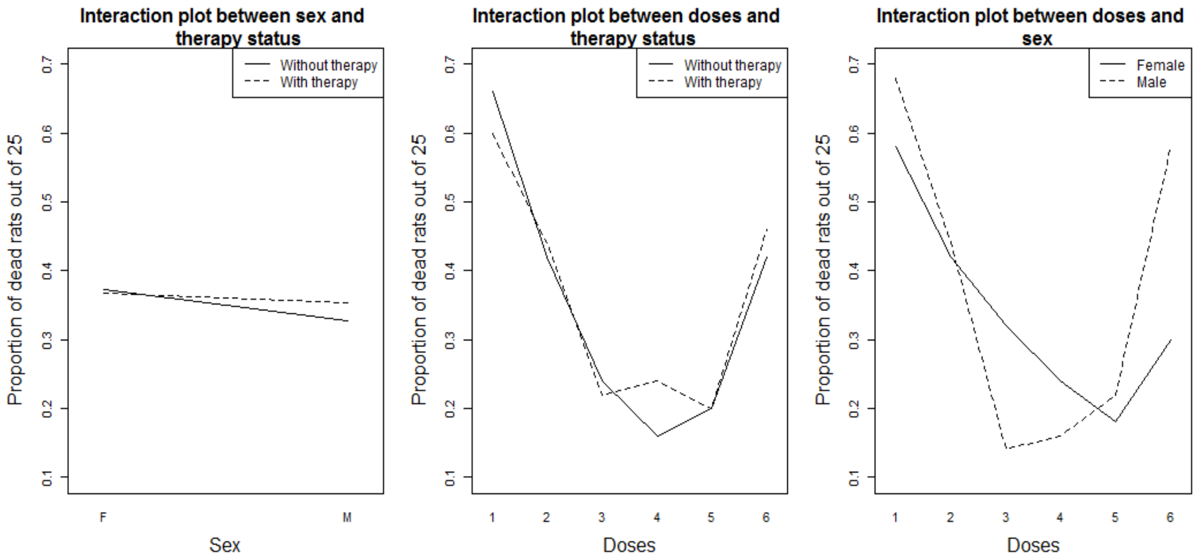

Figure 3: Two way-interactions plots for all factor combinations.

One can see that there seems to be little to no interaction between sex and therapy status as the slopes are very similar. Similarly, there does not appear to be any interaction between therapy status and dose (except perhaps for the fourth dose). Finally, the right-hand side plots shows a possible interaction between sex and dose, in particular for dose 3 and 4. This is important to note because there did not seem to be any difference between males and females, but sex could be interacting with the dose factor. Let us now go into the details of the model in order to rigorously assess the relationship between the response and the explanatory variables.

3 GLM and model fitting

3.1 Theory and method

Generalized linear models (GLMs) are an extension of normal linear models. In GLMs, the response is non-normal and there is a non-identity link function g that relates the expected response to the linear component of the model, as we will see. In addition the variance of the response is a function of the mean, and so the variance is not constant. In GLMs, the link function g can be expressed as

$$g(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j, \tag{1}$$

where $\mu_i = E(Y_i)$.

In particular, in this example, to model proportions, the logistic link function is used and we have

$$\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \sum_{j=1}^{p} x_{ij}\beta_j, \tag{2}$$

which gives

$$\mu_i = \frac{\exp(\sum_{j=1}^{p} x_{ij}\beta_j)}{1+\exp(\sum_{j=1}^{p} x_{ij}\beta_j)}. \tag{3}$$

Alternative and nested GLMs models fitted to the same set of data are compared using the deviance statistic. The deviance D_ω of a model ω is given by

$$D_\omega = -2\{l(\omega) - l(S)\},$$

where l is the log-likelihood at the MLE and S is the saturated model in which the responses μ_i are not restricted (i.e, there is the same number of parameters as there are responses). To test a GLM demoted by Ω that contains p explanatory variables (and thus p parameters) against a simpler alternative model ω containing q variables ($q < p$), we calculate the increase in deviance from omitting the $x_{q+1}, ..., x_p$ explanatory variables. The increase is given by

$$D_\omega - D_\Omega = -2\{l(\omega) - l(\Omega)\}. \tag{4}$$

This statistic has approximately the distribution χ^2_{p-q} if the simpler model ω is true. We are now in a position to try different models and find which one gives the best fit to the data, using the statistic in (4).

3.2 Model fitting and selection

Firstly, it is of interest to fit a model where the mortality rate is only explained by the different dose levels. As Figure 1 hinted, each dose has a significant effect on the mortality rate and must be included in the model. In this case the model is given by

$$\text{logit}(\mu_j) = \beta_0 + \beta_{dose(i)} I_{dose(i)}, \tag{5}$$

where $I_{dose(i)} = 1$ if $i = 2, 3, 4, 5$ and 0 if $i = 1$. For instance, if one wants to estimate a point for dose 2, the model is

$$\text{logit}(\mu_j) = \beta_0 + \beta_{dose(2)} I_{trial(2)}.$$

Note that in this case, since it is assumed that sex and therapy have no effect, the fitted values for each dose are just the average of the 4 observed values at each dose. However, this model might be too simplistic, and we are also interested in investigating the effect of sex and therapy status as well as dose. After fitting the most general model (up to two-way interactions), one obtains that there is no significant interaction between sex and therapy, dose and therapy and that therapy does not have any significant effect on the mortality rate. Instead of removing term by term, we can test the simpler model that none of the above explanatory variables impacts mortality rate by computing the test statistic in (4) against the model including all the two-way interactions. The corresponding p-value after comparing to a χ^2_7 is 0.965. Therefore, there is no evidence that the full model fits the data better and so the simpler model is accepted. The best fitting model is now

$$\text{logit}(\mu_j) = \beta_0 + \beta_{sex} I_{sex} + \beta_{dose(i)} I_{dose(i)} + \beta_{sex,dose(i)} I_{dose(i)} I_{sex}, \tag{6}$$

where, $I_{sex} = 1$ for males and 0 for females. The term $\beta_{sex,trial(i)}$ corresponds to the interaction between males and dose 2,3,4,5,6. Note that from the interaction terms, only the interaction between dose 3 and sex is significant at a 5% level. We thus want to test if the interaction term is needed, and again compute the statistics in (4), where this time the more complex model is (6) and the simpler model does not include the interaction between sex and dose. The p-value obtained after comparing to a χ^2_5 is 0.013, which means that there is strong evidence against the simpler model. Therefore, our final model is given in (6). Before interpreting the model and its parameters, let us look at the goodness of fit of the model to assess its validity.

3.3 Goodness of fit

Residual plots will be used to assess the goodness of fit. The working residuals consider the misfit in the space of the linear predictor. In other words, if there is an apparent pattern when the residuals are plotted against the fitted values, then the link function may not be appropriate. The Pearsons residuals are similar to the raw residuals in a normal linear model but they are standardized since the variances of the response differ as the mean changes. Again, we are looking for a random scatter of the residuals, when plotted against the fitted values. Lastly, the standardized residuals should follow a standard normal for a binomial model with large counts. Figure 4 below displays the required plots to assess the validity of the model. The working residuals are displayed on the plot in the top left corner. There appears to be no pattern (note that it is symmetric since therapy is not significant) and so the logistic link function is appropriate. In addition, the standardized deviance residuals are approximately normal as shown in the Q-Q. The Pearsons residuals are contained within +1 and -1 which means that there are no outliers and that they seem to be approximately standard normal.

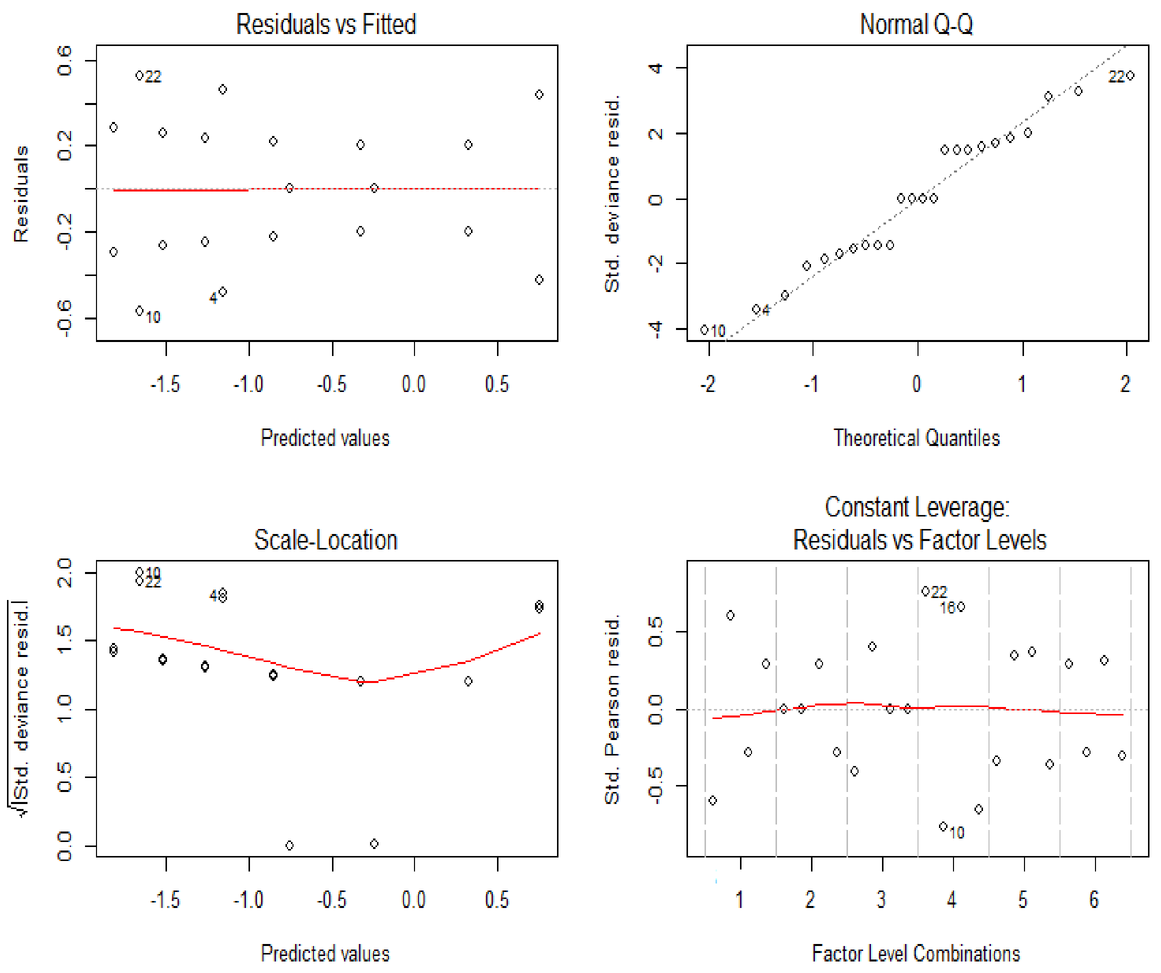

Figure 4: Residual plots for the model in (6).

We can conclude that the model is valid and that it provides a good fit to the data. Let us now look at the estimates of the parameters and at the uncertainty of their estimates.

4 Results and interpretation

Table 2 below shows the estimates of the parameters of model (6) and their associated 95% confidence intervals. Note that the model contains 12 parameters, which means each parameter is estimated with only 2 observed data points (with and without therapy for a given sex and a given dose level). The fitted values are the average of these 2 data points as one can see in Figure 5. One can see that the confidence intervals are very wide, this is because of the fact that only 2 data points are used to estimate a parameter. In turns, the fitted probabilities have large uncertainties as we will see with an example.

Parameters	Estimates	95% CI
β_0	0.75	[0.18,1.37]
β_{sex}	-0.43	[-1.26,0.38]
$\beta_{dose(2)}$	-0.99	[-1.83,-0.19]
$\beta_{dose(3)}$	-2.57	[-3.63,-1.62]
$\beta_{dose(4)}$	-2.41	[-3.43,-1.49]
$\beta_{dose(5)}$	-2.02	[-2.95,-1.15]
$\beta_{dose(6)}$	-0.43	[-1.26,0.38]
$\beta_{sex,dose(2)}$	0.35	[-0.79,1.49]
$\beta_{sex,dose(3)}$	1.49	[0.26,2.82]
$\beta_{sex,dose(4)}$	0.97	[-0.34,2.25]
$\beta_{sex,dose(5)}$	0.18	[-1.10,1.46]
$\beta_{sex,dose(6)}$	-0.74	[-1.91,0.42]

Table 2: Estimates of the parameters with their associated 95% confidence intervals.

For instance, an approximate 95% confidence interval for the fitted value for females and dose 1 is given by

$$\left[\frac{\exp(0.18)}{1+\exp(0.18)}, \frac{\exp(1.37)}{1+\exp(1.37)}\right] = [0.54, 0.80].$$

Proceeding in similar fashion, one can compute approximate confidence intervals for all the fitted values. The intervals will overlap, and so it is not clear what the actual impact of the dose levels and sex on the mortality rate is. In other words, even though dose 2 seems to have a negative effect on the mortality rate relative to dose 1, it may well be that the actual effect is in fact positive, after collecting more data points.

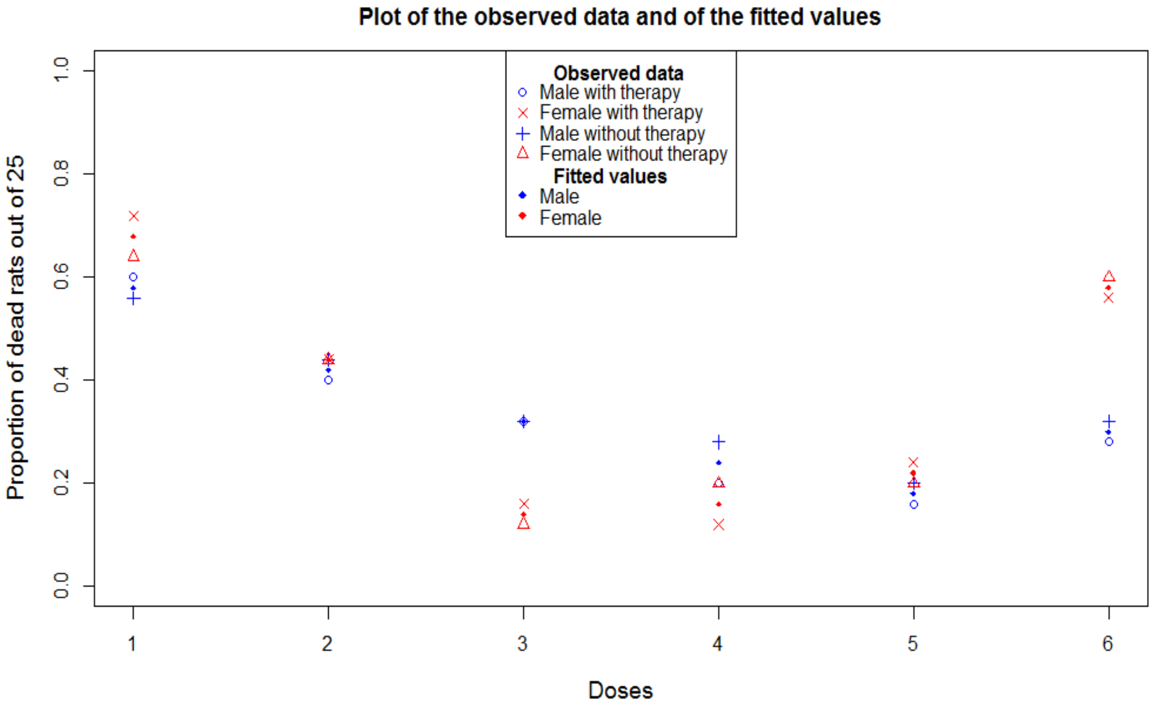

Figure 5: Plot of the observed data and the fitted values of model (6).

Figure 5 shows a plot of the fitted values on top of the observed data points. It can be easily seen that the fitted values are the average between the points with and without standard therapy. It is also easy to see the effect of sex just looking at the plot: females have a higher fitted mortality rate for dose levels 1,2,5 and 6. Whereas males have a higher fitted mortality rate for dose levels 3 and 4.

5 Conclusions

We have seen that dose levels have a significant impact on the mortality rate of rats. Furthermore, there is a significant interaction between sex and dose levels. On the other hand, therapy status does not significantly affect mortality. The fitted model leads to a situation where there are 12 estimated parameters for only 24 observations. We have argued that this leads to big uncertainties in the estimation of the parameters and thus of the fitted values. It would be of interest to collect more data points at each dose and refit a model to have better estimates and smaller uncertainties. In a further analysis, one could also treat the batches as random effects. Finally, we explained in the introduction that dose levels should be treated as factor. In a further experiment, one should record the actual amount of treatment in each dose in order to treat dose as a continuous variable.

Appendix A

```r
rats<- read.csv(file="C:/Users/pinouche/Downloads/rats.csv",header=TRUE)
fix(rats)
names(rats)
sapply(rats, class)

# Load the libraries and packages

library(lattice)
library(nlme)
library(MASS)
install.packages("Rcpp")
install.packages("ggplot2")
library(Rcpp)
library(grid)
library(ggplot2)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Question 2

# Convert dose and standard as factor variables

rats$dose <- as.factor(rats$dose)
rats$standard <- as.factor(rats$standard)

# Partition the data into 4 subsets

Male1 <- subset(rats,  (sex == "M") & (standard == 1))
Female1 <- subset(rats,  (sex == "F") & (standard == 1))
Male0 <- subset(rats,  (sex == "M") & (standard == 0))
Female0 <- subset(rats,  (sex == "F") & (standard == 0))


# Plot the observed data points

par(mfrow=c(1,1))
par(mar=c(5,4,3,3))
plot(-2, -2, ylim=c(0,1),xlim=c(1,6), xlab='Doses'
     , ylab= 'Proportion of dead rats out of 25'
     ,main='Plot of the observed mortality rates against doses'
     , cex.lab=1.2)
points(Male1$dose,(Male1$numdead/25),col='blue',cex=1,pch=3)
points(Female1$dose,(Female1$numdead/25),col='red',cex=1,pch=4)
points(Male0$dose,(Male0$numdead/25),col='blue',cex=1, pch=1)
points(Female0$dose,(Female0$numdead/25),col='red',cex=1, pch=2)
legend( 'top', inset=0,
        legend=c("Male with therapy","Female with therapy","Male without therapy"
                 ,"Female without therapy"),
        col=c('blue','red','blue','red'),
        pch=c(3,4,1,2), merge=FALSE, cex=0.8)


# Plot boxplots for the effect of sex and therapy status on mortality rate

par(mfrow=c(1,2))
par(mar=c(4.2,4.2,4,4.2))
boxplot((rats$numdead/25)~sex,data=rats,ylim=c(0,1),ylab='Proportion of dead rats out of 25', xlab='Sex'
        ,main='Proportion of dead rats for each
sex', cex.lab=1.6,col="green4")
boxplot((rats$numdead/25)~standard,data=rats,ylim=c(0,1),ylab='Proportion of dead rats out of 25'
        , xlab='Therapy status'
        ,main='Proportion of dead rats for each
  therapy status', cex.lab=1.6,col="green4")


# Plot 3 interaction plots on the same window for all the possible two-way interactions.

par(mfrow=c(1,3))
par(mar=c(5,5,3,1))
interaction.plot(rats$sex, rats$standard, (rats$numdead/25)
                 ,ylab='Proportion of dead rats out of 25',xlab='Sex',ylim=c(0.1,0.7)
                 ,cex.lab=1.5,legend= FALSE)
legend( 'topright', inset=0,legend=c('Without therapy','With therapy'),lwd=1,box.lwd=0
        , lty=c(1,2),cex=1.1)
title(main='Interaction plot between sex and
      therapy status', cex.main=1.5)

interaction.plot(rats$dose, rats$standard, (rats$numdead/25)
                 ,ylab='Proportion of dead rats out of 25',xlab='Doses',ylim=c(0.1,0.7)
                 ,cex.lab=1.5,legend= FALSE)
legend( 'topright', inset=0,legend=c('Without therapy','With therapy'),lwd=1,box.lwd=0
        , lty=c(1,2),cex=1.1)
title(main='Interaction plot between doses and
  therapy status', cex.main=1.5)

interaction.plot(rats$dose, rats$sex, (rats$numdead/25)
                 ,ylab='Proportion of dead rats out of 25',xlab='Doses',ylim=c(0.1,0.7)
                 ,cex.lab=1.5,legend= FALSE)
legend( 'topright', inset=0,legend=c('Female','Male'),lwd=1,box.lwd=0, lty=c(1,2),cex=1.1)
title(main='Interaction plot between doses and
  sex', cex.main=1.5)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Question 3

# Fit a logistic model for mortality rate only explained by the dose levels.

rats.glm1 <- glm(cbind(rats$numdead, 25-rats$numdead) ~ rats$dose, family=binomial)
summary(rats.glm1)


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Question 4

# Fit the most general model up to the two-way interaction terms.

rats.glm1 <- glm(cbind(rats$numdead, 25-rats$numdead) ~ (rats$dose+rats$sex+rats$standard)^2, family=binomial)

# Fit a simpler GLM model and compute the p-value of the deviance statistics when compared to a Chi squared.

rats.glm2 <- glm(cbind(rats$numdead, 25-rats$numdead) ~ (rats$dose+rats$sex+rats$standard)^2
                 -rats$sex:rats$standard -rats$dose:rats$standard- rats$standard   , family=binomial)
summary(rats.glm2)
1-pchisq(rats.glm2$deviance-rats.glm1$deviance,7)

# The simpler model rats.glm2 is accepted. Fit yet a simpler model and compute the deviance statistics again.

rats.glm3 <- glm(cbind(rats$numdead, 25-rats$numdead) ~ (rats$dose+rats$sex+rats$standard)^2-rats$sex:rats$standard
                 -rats$dose:rats$standard- rats$standard- rats$dose:rats$sex   , family=binomial)
1-pchisq(rats.glm3$deviance-rats.glm2$deviance,5)

# rats.glm2 is accepted to be the best model fitting the data. Display a summary of the model, the fitted values and
# 95% confidence intervals for the parameters.

summary(rats.glm2)
fitted.values(rats.glm2)
confint(rats.glm2,level=0.95, trace=FALSE)

# Use the anova command to check our results. We find that rats.glm2 is also the best model

anova(rats.glm1,test='Chisq')


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Question 5

```

```
# Plot of the residuals for the chosen model in order to assess the validity of the model .

par(mfrow = c(2, 2))
plot(rats.glm2)

# Compute the working residuals and plot them against the fitted values

workingres.glm2 <- residuals(rats.glm2, type = c(’working’))
ft.glm2 <-fitted.values(rats.glm2)
plot(ft.glm2,workingres.glm2,xlim=c(0.1,0.8),ylim=c(-0.7,0.8),xlab=’fitted values’,ylab=’working residuals’,cex.lab=1.2)
title(’Working residuals vs fitted values’)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Question 6

# Partition the fitted values into 2 vectors to be plotted. Note that only the 12 first values are taken since therapy status
# does not have a significant impact.

Male1fitted <- ft.glm2[1:6]
Female1fitted <- ft.glm2[7:12]

# Plot the observed data and also the fitted values from the model on the same window.

par(mfrow=c(1,1))
par(mar=c(5,4,3,3))
plot(-2, -2, ylim=c(0,1),xlim=c(1,6), xlab=’Doses’
     , ylab= ’Proportion of dead rats out of 25’,main=’Plot of the observed data and of the fitted values’
     , cex.lab=1.2)
points(Male1$dose,(Male1$numdead/25),col=’blue’,cex=1,pch=1)
points(Female1$dose,(Female1$numdead/25),col=’red’,cex=1,pch=4)
points(Male0$dose,(Male0$numdead/25),col=’blue’,cex=1, pch=3)
points(Female0$dose,(Female0$numdead/25),col=’red’,cex=1, pch=2)
points(Female1$dose,Male1fitted,col=’blue’,cex=0.7, pch=18)
points(Female1$dose,Female1fitted,col=’red’,cex=0.7, pch=18)
legend( ’top’, inset=0,
        legend=c(expression(bold("   Observed data")),"Male with therapy","Female with therapy"
                 ,"Male without therapy","Female without therapy", expression(bold("Fitted values")),’Male’,’Female’),
        col=c(’blue’,’red’,’blue’,’red’,’blue’,’red’),
        pch=c(NA,1,4,3,2,NA,18,18), merge=FALSE, cex=0.8)
```