Assignment3

Cameron Todd (5611230)

7 April 2017

Question 1

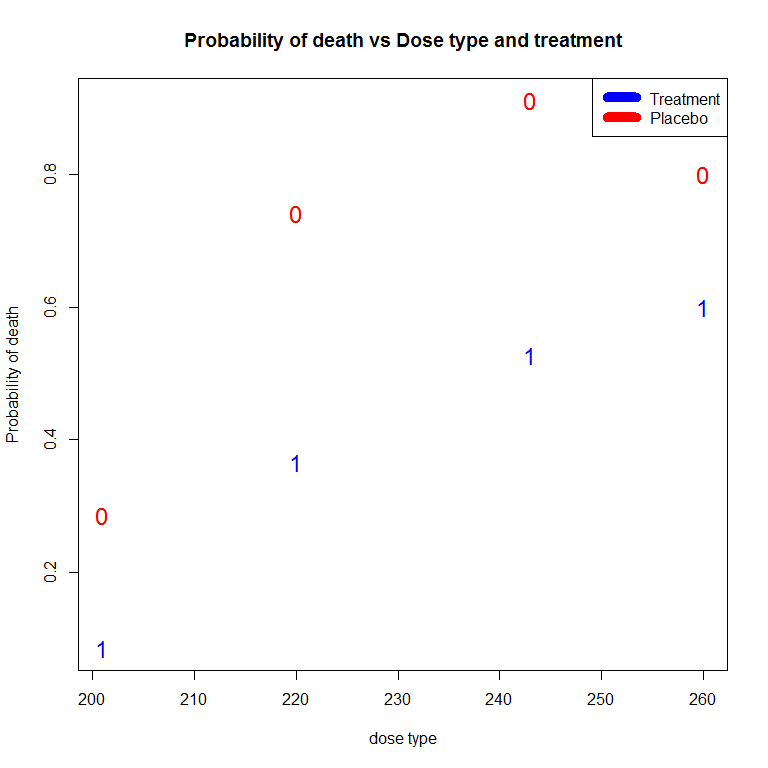
library("dplyr",warn.conflicts=FALSE, quietly=TRUE)  
rats.df <- read.csv("rats\_radiation.csv", header = TRUE)  
  
ratsg.df = group\_by(rats.df,treatment,dose)  
grouped\_rats.df = summarise(ratsg.df, dead = sum(death), alive = length(death) - sum(death))  
grouped\_rats.df$treatment <- factor(grouped\_rats.df$treatment)  
print(grouped\_rats.df)

## Source: local data frame [8 x 4]  
## Groups: treatment [?]  
##   
## treatment dose dead alive  
## <fctr> <int> <int> <int>  
## 1 0 201 10 25  
## 2 0 220 89 31  
## 3 0 243 83 8  
## 4 0 260 12 3  
## 5 1 201 3 32  
## 6 1 220 44 76  
## 7 1 243 48 43  
## 8 1 260 9 6

1. Grouped data is data that has been organised into groups known as classes, the grouped data has been classified and thus come level of data analysis has taken place already. We group data when their are certain observations that all have similar or related properties, in this case the level of doseage and treatment. In most cases, data operations are most useful when done on groups defined by variables in the dataset, this is because the data is generally cleaner and less of it e.g. the sort algorithm can take very long on a huge dataset but when it is grouped it saves alot more time. The downside of course is we get less plots in the scattter plots and therefore harder to see trends or diagnostic outliers.

Question 2 (A)

plot(I(dead/(dead+alive)) ~ dose, pch = substr(treatment,1,1), col=ifelse(treatment == "0","red","blue"), cex=1.5, data = grouped\_rats.df, xlab = "dose type", ylab = "Probability of death", main = "Probability of death vs Dose type and treatment")  
legend("topright", c("Treatment", "Placebo"), col=c("blue", "red"), lwd=10)



#text(grouped\_rats.df$dose,grouped\_rats.df$dead, 1:8, col = ifelse(grouped\_rats.df$treatment == 1, "orange","blue"))

1. As you can see from the plot, we have two continuous variables that probability of death, the dose type (should be a factor, will change later) and treatment as a categorical variable of 0 or 1, 1 being the mouse got treated. We can firstly obviously see that the mice that got treated had a less likely chance of death compared to mice that weren't treated as expected. We can also see as the dose type of radiation increases, the probability of death increases, their also appears to be curvatures in both placebo and treated groups, this might mean we will need a quadratic in our model.

Question 3

rats.df$treatment <- factor(rats.df$treatment)  
rats.fit\_Interaction = glm(cbind(dead,alive) ~ dose \* treatment, family = binomial, data = grouped\_rats.df)  
rats.fitNI = glm(cbind(dead,alive) ~ dose + treatment, family = binomial, data = grouped\_rats.df)  
summary(rats.fit\_Interaction)

##   
## Call:  
## glm(formula = cbind(dead, alive) ~ dose \* treatment, family = binomial,   
## data = grouped\_rats.df)  
##   
## Deviance Residuals:   
## 1 2 3 4 5 6 7 8   
## -1.5016 1.2015 0.2732 -2.4209 -1.6674 1.0487 -0.1270 -0.7182   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -12.87784 2.47210 -5.209 1.90e-07 \*\*\*  
## dose 0.06221 0.01116 5.572 2.52e-08 \*\*\*  
## treatment1 3.67700 3.15463 1.166 0.2438   
## dose:treatment1 -0.02379 0.01404 -1.694 0.0903 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 140.676 on 7 degrees of freedom  
## Residual deviance: 14.046 on 4 degrees of freedom  
## AIC: 53.657  
##   
## Number of Fisher Scoring iterations: 5

summary(rats.fitNI)

##   
## Call:  
## glm(formula = cbind(dead, alive) ~ dose + treatment, family = binomial,   
## data = grouped\_rats.df)  
##   
## Deviance Residuals:   
## 1 2 3 4 5 6 7 8   
## -2.3449 1.0901 1.1577 -1.8288 -1.1114 1.5268 -0.7269 -1.2380   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.718699 1.499494 -6.481 9.09e-11 \*\*\*  
## dose 0.047954 0.006715 7.141 9.24e-13 \*\*\*  
## treatment1 -1.671224 0.207685 -8.047 8.49e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 140.676 on 7 degrees of freedom  
## Residual deviance: 16.999 on 5 degrees of freedom  
## AIC: 54.609  
##   
## Number of Fisher Scoring iterations: 4

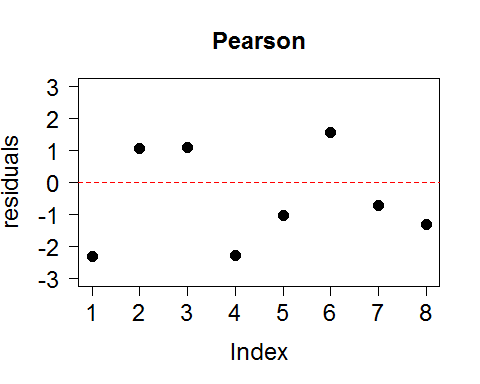
Q3 (B) Analysis of coefficients show an increase in dosage has a positive effect on the probability of death for both models. The interaction term in the first model appears to be insignificant and therefore should be taken out as in the 2nd model the the treatment appears to be a significant factor to explaining the variances in the probability of death, the treatment has a negative effect on probability of deat. Analysis of residuals - In both models the we have a residual deviance for the null hypothesis of 140 and 7 degrees of freedom. When we include the independent variables of dose and treatment, the interaction model decreases it's deviance by 126 and 3 degrees of freedom to 14 residual deviance with 4 degrees of freedom, the 2nd model with no interaction, its residual deviances reduce by 124 to 17 and drops degrees of freedom by 2 to 5. The interaction model also appears to have a lower AIC, which is ok to measure when comparing against 2 models, obviously the lower the better.

It appears the interaction model has the best fit for the data as it has a lower residual deviance, degrees of freedom, however this is normal as we add more terms to the model, the R^2 or in this case the Residual deviances will always decrease but i think the proportional decrease in residual deviances by adding the extra interaction term is not huge and therefore i would prefer to use the more basice model, model 2 with no interaction is actually the better fit.

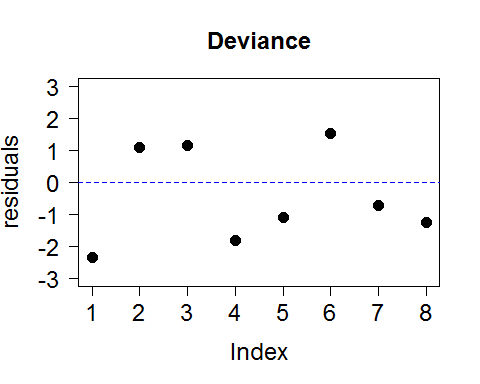
However both models are still not a great model as their is quite a big difference between the residual deviance and degrees of freedom.

q3 (c)

pearson.residuals = residuals(rats.fitNI, type = "pearson")  
deviance.residuals = residuals(rats.fitNI, type = "deviance")  
  
plot(pearson.residuals, ylab = "residuals", main = "Pearson", cex = 1.5, cex.main = 1.5, cex.lab = 1.5, cex.axis = 1.5, ylim = c(-3,3), pch = 16, las = 1)  
abline(h = 0, lty = 2, col = "red")



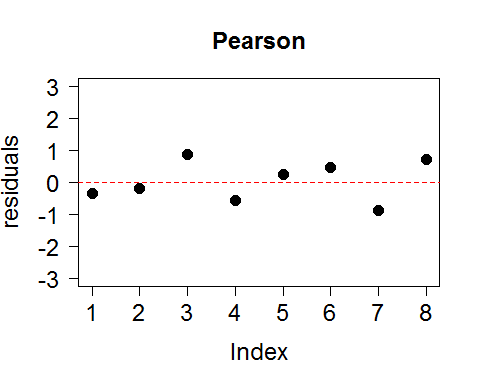
plot(deviance.residuals, ylab = "residuals", main = "Deviance", cex = 1.5, cex.main = 1.5, cex.lab = 1.5, cex.axis = 1.5, ylim = c(-3,3), pch = 16, las = 1)  
abline(h = 0, lty = 2, col = "blue")



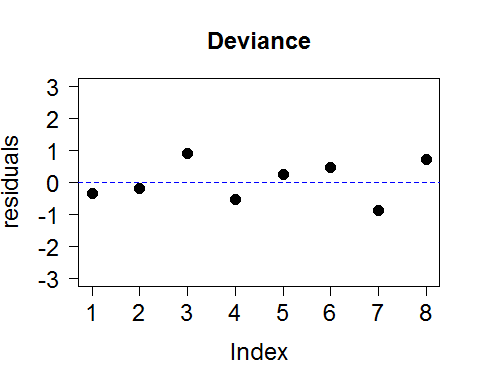
Both residual plots are fairly similar. Becuase these are grouped observations, any points such as groups 1 and 4 hat are outside the -2 and 2 bracket will mean within those 2 groups their will be many observations outside this bracket. With these 2 groups outside the bracket, it is very strange and the model should be refitted. It appears their may be curvature for both treated and non treated groups, this can help to refit the model.

Question 4 (A)

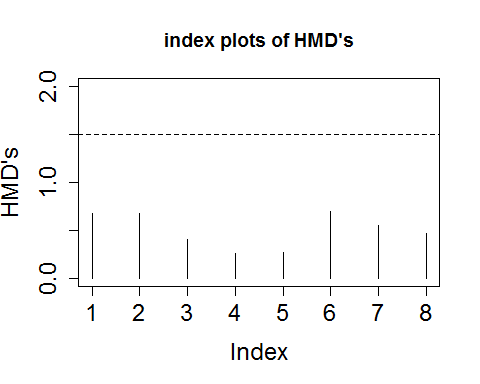
#grouped\_rats.df$dose = as.factor(grouped\_rats.df$dose, levels = c(0,4))  
rats.fitSQ = glm(cbind(dead,alive) ~ dose + I(dose^2) + treatment, family = binomial, data = grouped\_rats.df)  
  
####Residuals  
pearson.residuals = residuals(rats.fitSQ, type = "pearson")  
deviance.residuals = residuals(rats.fitSQ, type = "deviance")  
  
plot(pearson.residuals, ylab = "residuals", main = "Pearson", cex = 1.5, cex.main = 1.5, cex.lab = 1.5, cex.axis = 1.5, ylim = c(-3,3), pch = 16, las = 1)  
abline(h = 0, lty = 2, col = "red")



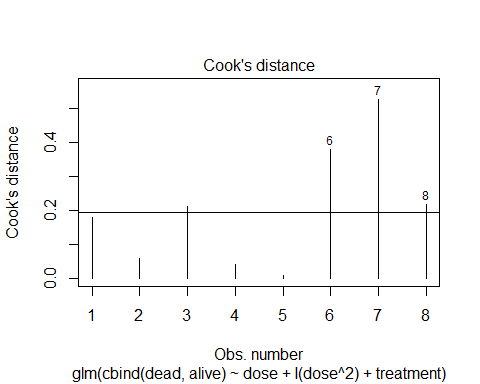
plot(deviance.residuals, ylab = "residuals", main = "Deviance", cex = 1.5, cex.main = 1.5, cex.lab = 1.5, cex.axis = 1.5, ylim = c(-3,3), pch = 16, las = 1)  
abline(h = 0, lty = 2, col = "blue")



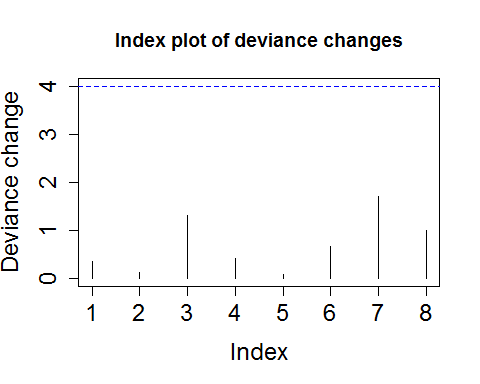
#Hat matrix diagonals  
 HMD = hatvalues(rats.fitSQ)  
 plot(HMD, ylab = "HMD's", type = "h", cex = 1.5, cex.axis = 1.5, cex.lab = 1.5, main = "index plots of HMD's",ylim = c(0,2))  
 abline(h = 3\*length(coef(rats.fitSQ))/nrow(grouped\_rats.df), lty =2)



#Cooks  
hat.extreme <- 3\*length(coef(rats.fitSQ))/nrow(grouped\_rats.df)  
plot(rats.fitSQ, which = 4)  
abline(h = qchisq(0.1, 3)/3)



#Deviance change diagnostic  
dev.change = deviance.residuals^2 + pearson.residuals^2 \* HMD / (1-HMD)  
plot(dev.change, ylab = "Deviance change", type = "h", cex = 1.5, cex.axis = 1.5, cex.lab = 1.5, main = "Index plot of deviance changes",ylim = c(0,4))  
abline(h = 4, lty = 2, col = "blue")

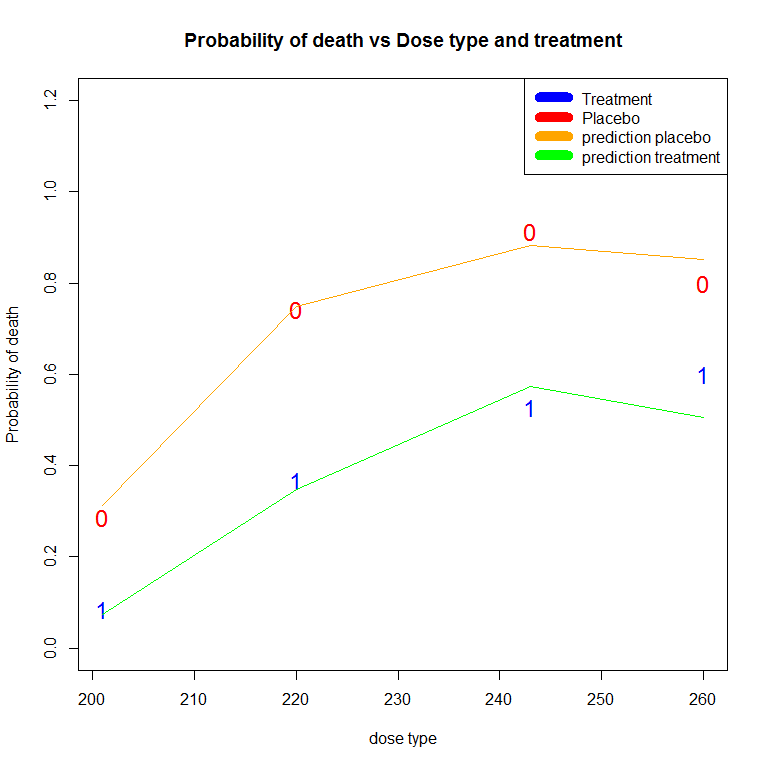


#summary(rats.fitSQ)

By adding the quadratic term to the model, it has fixed the residual deviance problems, with the plots being alot closer to 0 for Pearson and Deviance plots. I have applied the diagnostics for any influential points by using the Hat matrix diagonals, cooks distance and leave-one-out deviance change. The diagnostics seem fine with the hat matrix showing they are all below the 1.5, however the cooks show groups 6,7 and 8 can be all influential but this can remove alot of data and plus the deviance change plot shows that the groups are all below the limit of 4 and therefore we are not too worried, diagnostics ok and we can move forward.

Q4 (B)

grouped\_rats.df$treatment <- factor(grouped\_rats.df$treatment)  
rats.fitSQ = glm(cbind(dead,alive) ~ dose + I(dose^2) + treatment, family = binomial, data = grouped\_rats.df)  
plot(I(dead/(dead+alive)) ~ dose, pch = substr(treatment,1,1), col=ifelse(treatment == "0","red","blue"), cex=1.5, data = grouped\_rats.df, xlab = "dose type", ylab = "Probability of death", main = "Probability of death vs Dose type and treatment",ylim = c(0,1.2))  
legend("topright", c("Treatment", "Placebo", "prediction placebo","prediction treatment"), col=c("blue", "red","orange","green"), lwd=10)  
  
xx = seq(200,260,length.out = 100)  
fit = rats.fitSQ  
pred.prob = predict(rats.fitSQ, newdata = data.frame(dose = c(201, 220, 243, 260) , treatment = "1"), type = "response")  
lines(x = c(201, 220, 243, 260) , y = pred.prob, col = "green")  
  
pred.prob1 = predict(rats.fitSQ, newdata = data.frame(dose = c(201, 220, 243, 260) , treatment = "0"), type = "response")  
lines(x = c(201, 220, 243, 260) , y = pred.prob1, col = "orange")



Q4 (C)

1 - pchisq(deviance(rats.fitSQ),df.residual(rats.fitSQ))

## [1] 0.5889968

anova(rats.fitSQ, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: cbind(dead, alive)  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 7 140.676   
## dose 1 51.217 6 89.459 8.27e-13 \*\*\*  
## I(dose^2) 1 12.069 5 77.390 0.0005128 \*\*\*  
## treatment 1 74.574 4 2.816 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(rats.fitSQ)

##   
## Call:  
## glm(formula = cbind(dead, alive) ~ dose + I(dose^2) + treatment,   
## family = binomial, data = grouped\_rats.df)  
##   
## Deviance Residuals:   
## 1 2 3 4 5 6 7 8   
## -0.3385 -0.1954 0.8945 -0.5466 0.2455 0.4505 -0.8774 0.7292   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.284e+01 1.985e+01 -4.172 3.01e-05 \*\*\*  
## dose 6.905e-01 1.733e-01 3.984 6.79e-05 \*\*\*  
## I(dose^2) -1.404e-03 3.768e-04 -3.727 0.000194 \*\*\*  
## treatment1 -1.728e+00 2.131e-01 -8.106 5.23e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 140.6758 on 7 degrees of freedom  
## Residual deviance: 2.8164 on 4 degrees of freedom  
## AIC: 42.427  
##   
## Number of Fisher Scoring iterations: 4

exp(confint.default(rats.fitSQ))

## 2.5 % 97.5 %  
## (Intercept) 1.327724e-53 8.366366e-20  
## dose 1.420131e+00 2.801534e+00  
## I(dose^2) 9.978596e-01 9.993344e-01  
## treatment1 1.170097e-01 2.698229e-01

After adding a quadratic for does on the model, we can see straight away the fit is alot better then previous models, with the residual deviance of 2.8 and 4 degrees of freedom as almost the same. When looking at the coefficients, you can see that does of radiation has a positive effect on the probability of death but you can also see the quadratic has a negative number meaning that as dose of radiation increases, their is diminishing effect on the probability of death (this is the curvature that you can see in the plots for both treated and non-treated groups). The treatment coefficient is also having a negative effects on the probability of death as expected. All these coefficients are significant to explaining the variability in the model. You can also see with a 95% confidence interval that for every 1 unit increase in dose, we can expect a 1.4 - 2.8 times increase in the probability of death. We can also say that if a mouse has no treatment they were between 1.17 - 2.69 times more likely to die then a mouse that was treated.