Generalised Linear Modelling: Assignment 2

Joshua Redolfi

18 October 2019

Question 1

> The probability of a male birth in humans is about 0.51. It has previously been noticed that the lower proportions of male births are observed when offspring are conceived at times of exposure to smog, floods, or earthquakes. Danish researchers hypothesized that sources of stress associated with severe life events may also have some bearing on the sex ratio. To investigate this theory they obtained the sexes of all 3,072 children who were born in Denmark between January 1, 1980 and December 31, 1992, to women who experienced the following kinds of severe life events in the year of the birth or the year prior to the birth: death or admission to hospital for cancer or heart attack of their partner or of their other children. They also obtained sexes on a sample of 20,337 births for mothers who did not experience these life stress episodes.

(a)

We would like to assess whether there was a lower percentage of male births in the exposed group than there were in the control group. We can achieve this by fitting the following logistic regression model for binomial proportions using the logit link function to model the relationship between the response variable μ_i *Percentage of Male Births* and the covariates *Group*:

$$\mu_i \sim Bin(n, p_i),\ \mu_i = p_i$$

$$\eta_i = g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 \times group_i$$

where,

$$group_i = \begin{cases} 1 & if\ exposed \\ 0 & if\ control \end{cases}$$

```
proportion <- PctBoys/100
male.glm1 <- glm(proportion ~ Group, family = binomial, weights = Number)
summary(male.glm1)$coefficients
```

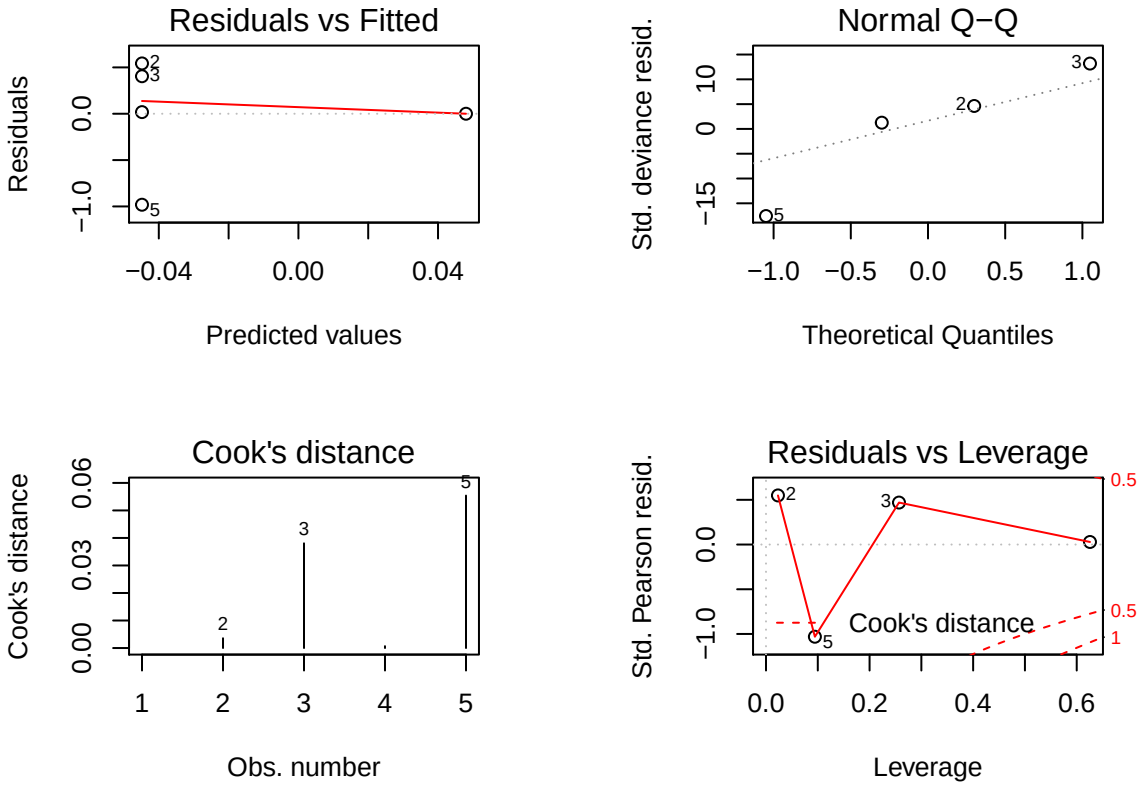
```
##                 Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)   0.04800922 0.01402851  3.422260 0.0006210294
## GroupExposed -0.09281750 0.03872385 -2.396908 0.0165340737
```

From the output in R, $\beta_1 \approx -0.0928$, and so relative to the control group the odds of a male birth for the exposed group is $exp(\beta_1) \approx 0.9114$ times the odds of a male birth for the control group.

We can confirm the significance of this result by conducting the following hypothesis test for β_1 as follows:

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

From the above summary output, we see that the corresponding $p - value \approx 0.0165 < \alpha = 0.05$, hence there is significant evidence to reject the hull hypothesis in favour of the alternative and conclude that *Group* is a significant predictor of probability of birthing a male.



The Residuals vs Fitted plot doesn't appear to show too much of an issue with linearity, as indicated by the lowess curve which remains close to the zero line. Observation 5 may be an outlier however this is difficult to determine on a non-standardised scale.There may be issues with heteroscedasticity however it is important to note that these issues are likely due to there being a very small sample size.

Again from the Normal Quantile-Quantile plot, observation 5 may be an outlier, however this is difficult to determine on a non-standardised scale. However, relative to the data, it may be contributing to a distict left tail thereby affecting the skewness of the distribution although this again is likely due to the limited sample size.

The Cook's Distance and Residuals vs Leverage plots don't raise any concerns about influence. Observation 5 which we identified as a possible outlier, doesn't exceed the 0.5 Cook's distance level (infact it is far from it) and therefore shouldn't have any significant impact on our model fit. No observations have significantly high leverage (none exceed the $\frac{2p}{n} = 0.8$ cut off value).

(b)

We would like to assess whether the probability of male births in the exposed group decreases as the stress event gets closer in time to conception. We can achieve this by fitting the following logistic regression model for binomial proportions using the logit link function to model the relationship between the response variable μ_i *Percentage of Male Births* and the remodeled *avg.time* continuous covariate for our exposed group:

$$\eta_i = g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 \times avg.time_i$$

```
expsd.data <- subset(ex2117, Group=='Exposed')
expsd.data$Time <- as.numeric(c(14,9,3,0))
avg.time <- expsd.data$Time
```

```
expsd.propn <- expsd.data$PctBoys/100
male.glm2 <- glm(expsd.propn ~ avg.time, family = binomial, weights = expsd.data$Number)
summary(male.glm2)$coefficients
```

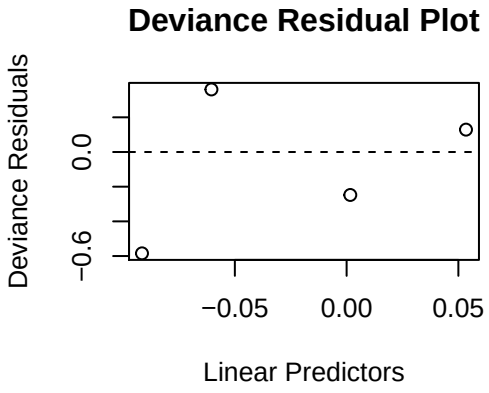
```
##                Estimate Std. Error    z value  Pr(>|z|)
## (Intercept) -0.09152963 0.06169184 -1.4836587 0.1378995
## avg.time     0.01035264 0.01108367  0.9340446 0.3502809
```

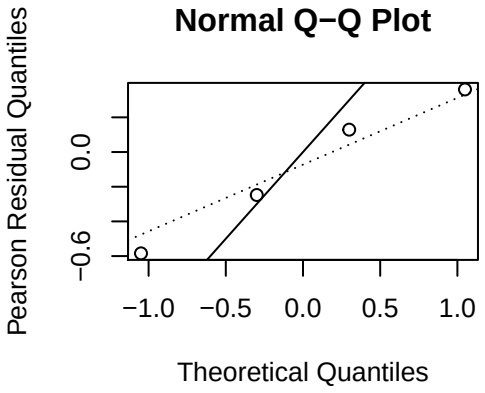
From the output in R, $\beta_1 \approx 0.0104$, and so the odds of a male birth increases by a multiplicative factor of $exp(\beta_1) \approx 1.0104$ for every 1 month increase in average time to conception from the occurrence of the stress event.

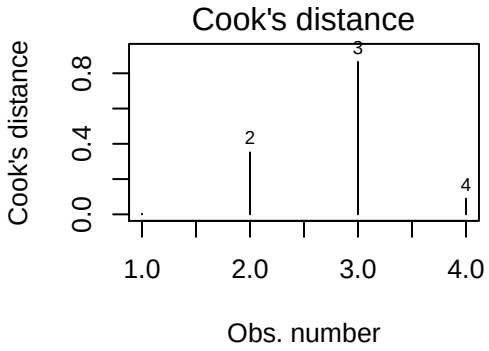
We can confirm the significance of this result by conducting the following hypothesis test for β_1 as follows:

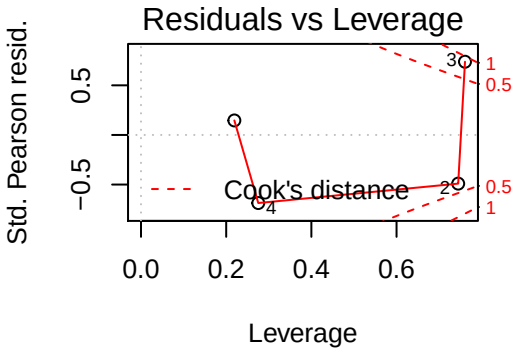
$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

From the above summary output, we see that the corresponding $p - value \approx 0.3503 > \alpha = 0.05$, hence there is insignificant evidence to reject the hull hypothesis and thus we cannot conclude that *avg.time* is a significant predictor of probability of birthing a male.









The Deviance Residuals vs Fitted plot doesn't appear to show too much of an issue with linearity. Since we've standardised the scale, we can see that there aren't any outliers in our model as all deviance residuals lie within ± 2 standard deviations of the mean 0 line.There may be issues with heteroscedasticity however it is important to note that these issues are likely due to there being a very small sample size.

From the Normal Quantile-Quantile plot, all Pearson residuals lie within the (-2,+2) range, as we've seen from the deviance residuals vs linear predictors plot. There doesn't appear to be issues with our assumption of normality as the residuals appear to 'hug' the q-q line.

The Cook's Distance and Residuals vs Leverage plots point out that observation 3 is highly influential as it exceeds our 0.5 Cook's distance level and therefore may be of concern. No observations have significantly high leverage (none exceed the $\frac{2p}{n} = 1$ cut off value).

(c)

We remodel the *avg.time* predictor to now account for the Control Group, and fit the logistic glm using the logit link function as follows:

$$g(\mu_i) = \beta_0 + \beta_1 \times avg.time_i$$

```
modified.male <- ex2117
modified.male$Time <- as.numeric(c(24,14,9,3,0))
avg.time2 <- modified.male$Time
```

```
modified.propn <- modified.male$PctBoys/100
male.glm3 <- glm(modified.propn ~ avg.time2, family = binomial, weights = Number)
summary(male.glm3)$coefficients
```

```
##                 Estimate  Std. Error   z value   Pr(>|z|)
## (Intercept) -0.070034989 0.043934921 -1.594062 0.11092215
## avg.time2    0.004937062 0.001956088  2.523947 0.01160456
```

From the output in R, $\beta_1 \approx 0.0049$, which is about half its value in our model fitted in part (b). It's corresponding z quantile and p-value however are of much greater significance. Whereas in part (b) our *avg.time* covariate was insignificant at the $\alpha = 0.05$ level, after modifying for control group, its $p-value = 0.0116 < 0.05$ and thus we can confirm that average time is a significant predictor for the percentage of male births. Our intercept parameter β_0 increases slightly from -0.0915 to -0.0700 and its corresponding z quantile is still statistically insignificant.

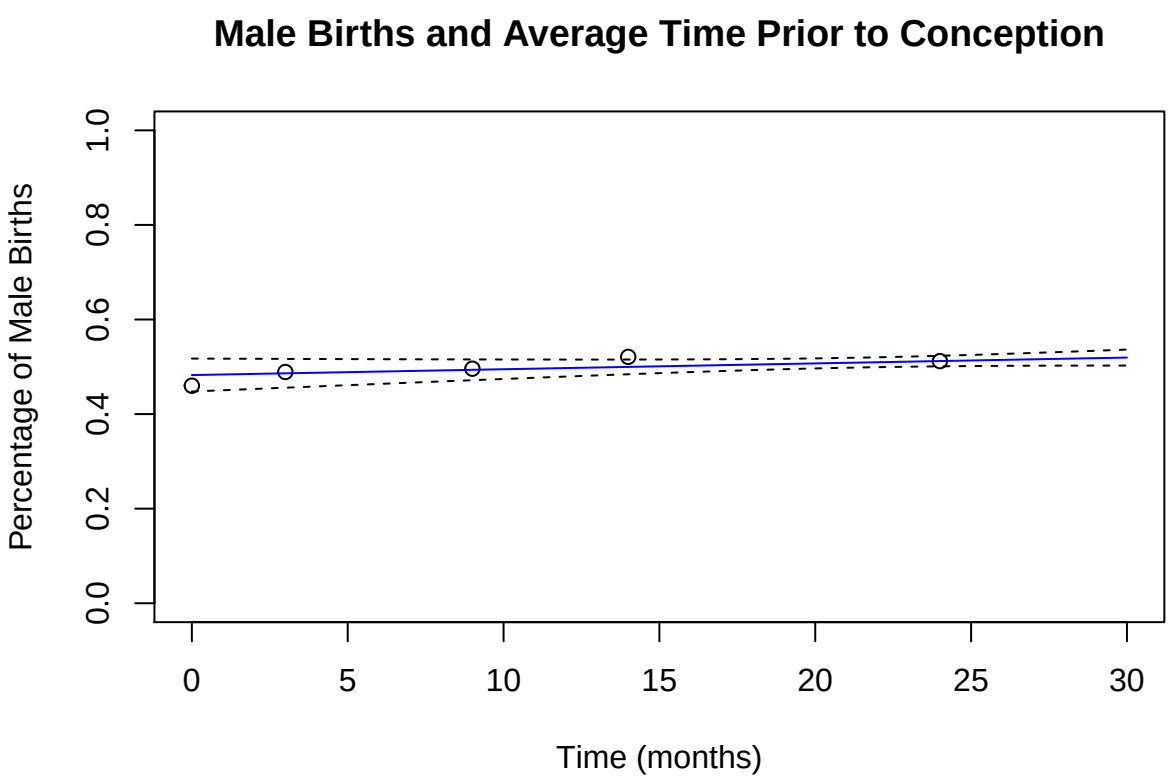
(d)

We would like the plot the following function, as our back-transformed model for probability (percentage) of male births fitted in part (c) against time. This is achieved via the invlogit function:

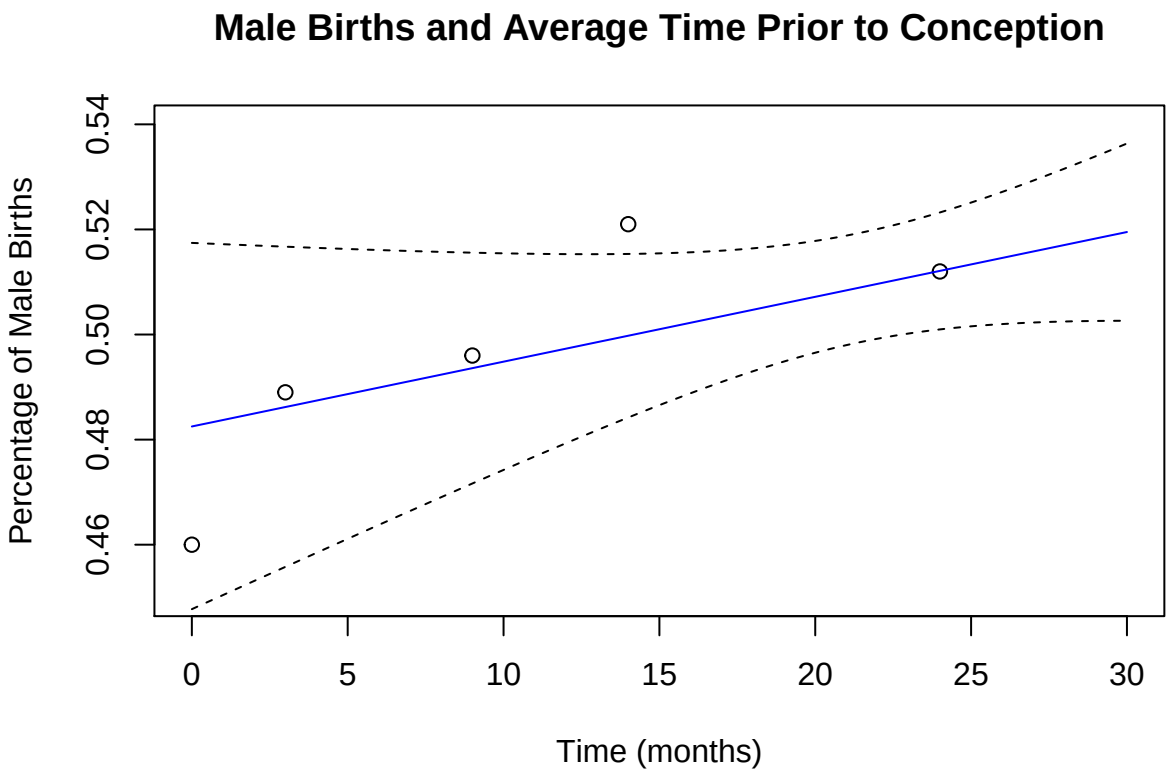
$$\mu_i = g^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

where,

$$\eta_i = \beta_0 + \beta_1 \times avg.time_i$$



From the above plot we can observe that the percentage of male births lies within a relatively narrow range between (0.45,0.55). The 95% confidence interval bands appear relatively narrow although it broadens quite noticeable for lower values of time.



Question 2

Page 161 of the Gelman & Hill text describes data from a study of the effect of integrated pest management on reducing cockroach levels in urban apartments.In this experiment, the treatment and control were applied to 160 and 104 apartments, respectively (though only 158 of the treatment observations appear to have been included in the dataset supplied with the Gelman & Hill text). The outcome measure is y.

(a)

We would like to Model the relationship between the trap rates of roaches and the predictors; *treatment, senior* and *roach1*. We can achieve this by fitting the following log-linear regression model for Poisson rates using the Poisson glm with the corresponding canonical link function as follows:

$$\mu_i \sim Pois(\lambda_i),\ \mu_i = \lambda_i$$

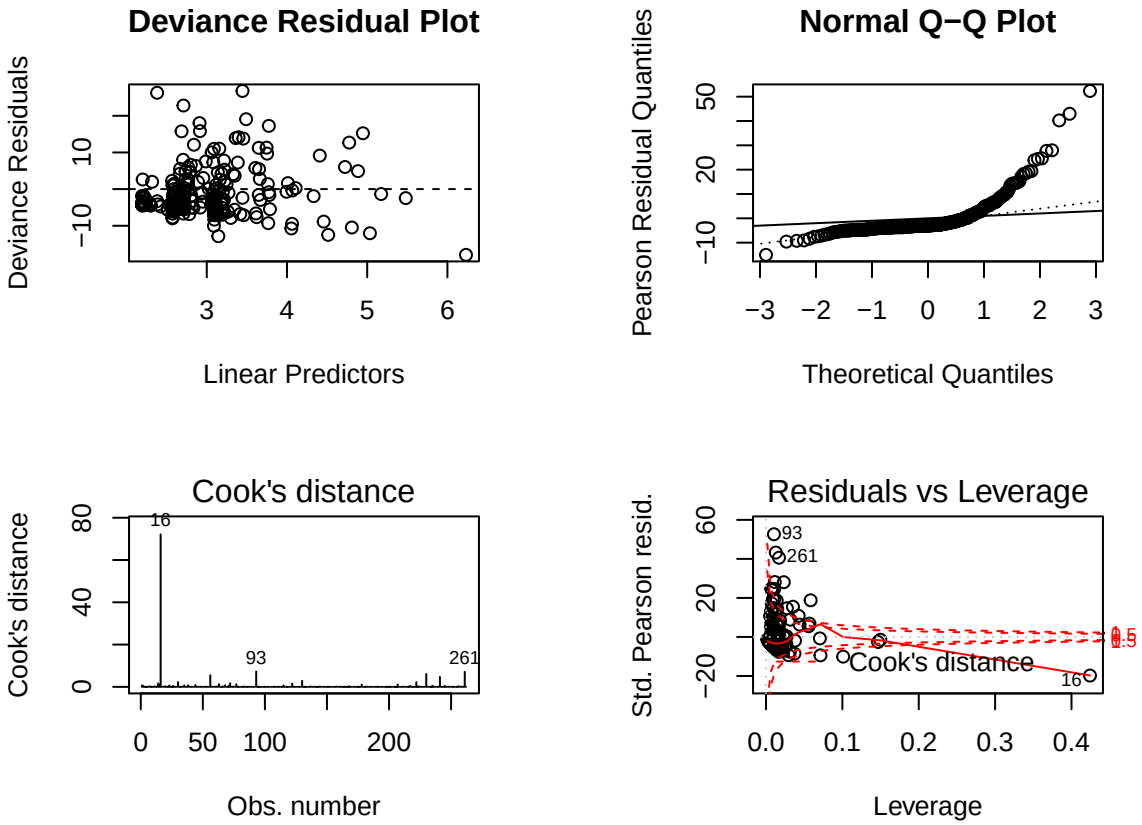
$$\eta_i = log(\mu_i) = \beta_0 + \beta_1 \times treatment_i\ +\ \beta_2 \times senior_i\ +\ \beta_3 \times roach1_i$$

where,

$$treatment_i = \begin{cases} 1 & if\ treatment \\ 0 & if\ control \end{cases}$$

$$senior_i = \begin{cases} 1 & if\ senior \\ 0 & if\ otherwise \end{cases}$$

```
trap.rate <- y/exposure2
roach.glm1 <- glm(trap.rate ~ treatment + senior + roach1,
                  family = poisson, weights = exposure2)
```



The Deviance Residuals vs Linear Predictors plot shows major over-dispersion of the deviance residuals as most points greatly exceed the ± 2 range. We also see a strong indication of issues with homoscedasticity, which violates our constant variance assumption. It is likely that we may need to remove certain observations from our data.

Again from the Normal Quantile-Quantile plot, we can see major issues with normality. A major right-tail in the distribution of the Pearson residuals is also apparent, and thus a major issue for our normality assumption.

The Cook's Distance and Residuals vs Leverage plots identify observations 93, 261 and especially observation 16 to be of extremely high influence. These observations have almost certainly heavily impacted the fit of our model, thereby resulting in the issues with normality and spread. There also appear to be some observations that possess significantly high leverage (exceeding the $\frac{2p}{n} = 4/262$ cut off value), especially observation 16.

(b)

```
round(summary(roach.glm1)$coefficients,4)
```

```
##             Estimate Std. Error  z value Pr(>|z|)
```

```
## (Intercept)   3.0892     0.0212 145.4865        0
## treatment    -0.5167     0.0247 -20.8872        0
## senior       -0.3799     0.0334 -11.3673        0
## roach1        0.0070     0.0001  78.6852        0
```

Despite the major issues with model diagnostics, we can see from the summary output that all the model covariates are statistically significant, so it appears like there is definitely some characteristic structure in the data we can model. We'll assess the overall fit of the model by conducting a 'Goodness-of-fit' Test as follows:

$$H_0 : \phi = 1 \quad H_1 : \phi \neq 1$$

Our observed test statistic $\hat{\phi}$ follows a Chi-Square distribution of n-p=258 degrees of freedom

$$\hat{\phi} = \frac{\Sigma d_i^2}{n-p} \sim \chi^2_{n-p}$$

```
roach.glm1$deviance/summary(roach.glm1)$dispersion
```

```
## [1] 11429.47
```

```
qchisq(p = c(0.025,0.975), df = roach.glm1$df.residual)
```

```
## [1] 215.4017 304.3848
```

As the residual deviance of $\hat{\phi} = 11429.47$ lies outside our middle 95% interval of (215.4017,304.3848), we reject the null hypothesis in favour of the alternative hypothesis and conclude that there is statistically significant over-dispersion in our model. This supports our observations from the residuals plots in part (a), whereby we observed extremely high variability in the spread of the residuals (majority of which exceed the ± 2 range). This is due to there being a strong positive skew in the data and the presence of highly influential points, including both extreme outliers and highly-levered observations which have had a significant influence on the overall fit of the model.

(c)

We can remodel the Poisson glm to account for over-dispersion by assuming our response variables follow a quasipoisson distribution as follows:

```
roach.glm2 <- glm(trap.rate ~ treatment + senior + roach1,
                  family = quasipoisson, weights = exposure2)
round(summary(roach.glm2)$coefficients,4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0892     0.1718 17.9846   0.0000
## treatment    -0.5167     0.2001 -2.5820   0.0104
## senior       -0.3799     0.2703 -1.4052   0.1612
## roach1        0.0070     0.0007  9.7268   0.0000
```

There isn't any apparent change in the coefficients of the modified model, however the *treatment* factor drops in significance to just the $\alpha = 0.05$ level and the *senior* factor drops in statistical significance all together. The quasipoisson glm model takes the dispersion parameter to be $\hat{\phi} = 65.4403$ as opposed to the Poisson glm's estimated dispersion parameter of 1.

Again, we can conduct a 'Goodness-of-fit' Test to assess the fit of our modified model:

$$H_0 : \phi = 1 \quad H_1 : \phi \neq 1$$

```
roach.glm2$deviance/summary(roach.glm2)$dispersion
```

```
## [1] 174.6549
```

```
qchisq(p = c(0.025,0.975), df = roach.glm2$df.residual)
```

```
## [1] 215.4017 304.3848
```

As the residual deviance of $\hat{\phi} = 174.6549$ lies within the middle 95% interval (215.4017,304.3848), we do not reject the null hypothesis in our two-tailed goodness-of-fit test. This suggests that there is no significant over/under-dispersion of the residuals in our new model. We can also conduct a stronger, one-tailed test for over-dispersion:

$$H_0 : \phi = 1 \quad H_1 : \phi > 1$$

```
roach.glm2$deviance/summary(roach.glm2)$dispersion
```

```
## [1] 174.6549
```

```
qchisq(p = 0.95, df = roach.glm2$df.residual)
```

```
## [1] 296.4659
```

Again, our test-statistic does not exceed the 95% quantile, so we do not reject the null hypothesis and thus there isn't any issue with over-dispersion.

(d)

We may experiment with our model fitted in part (c) to account for observation 16, which we've found to be highly influential in previously fitted regression models. We do so by fitting a factor variable for *obs16* into our log-linear model:

$$\eta_i = log(\mu_i) = \beta_0 + \beta_1 \times treatment_i \ + \ \beta_2 \times senior_i \ + \ \beta_3 \times roach1_i \ + \ \beta_4 \times obs16_i$$

where,

$$treatment_i = \begin{cases} 1 & if\ treatment \\ 0 & if\ control \end{cases}$$

$$senior_i = \begin{cases} 1 & if\ senior \\ 0 & if\ otherwise \end{cases}$$

$$obs16_i = \begin{cases} 1 & if\ observation\ 16 \\ 0 & if\ otherwise \end{cases}$$

```
obs16 <- rep(0, length(y))
obs16[16] <- 1
roach.glm3 <- glm(trap.rate ~ treatment + senior + roach1 + obs16,
                  family = quasipoisson, weights = exposure2)
round(summary(roach.glm3)$coefficients,4)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0609     0.1693 18.0844   0.0000
## treatment    -0.6341     0.2011 -3.1530   0.0018
## senior       -0.3454     0.2697 -1.2806   0.2015
## roach1        0.0081     0.0008  9.9098   0.0000
## obs16        -1.8571     0.8468 -2.1931   0.0292
```

From the summary output, there doesn't appear to be any drastic changes in the regression coefficients. The *senior* factor covariate remains insignificant in this model and our newly fitted *obs16* factor is statistically significant. Once again, we conduct a 'Goodness-of-fit' Test to assess the fit of our model:

$$H_0 : \phi = 1 \quad H_1 : \phi \neq 1$$

```
roach.glm3$deviance/summary(roach.glm3)$dispersion
```

```
## [1] 169.3437
```

```
qchisq(p = c(0.025,0.975), df = roach.glm3$df.residual)
```

```
## [1] 214.4881 303.2984
```

As the residual deviance of $\hat{\phi} = 169.3437$ lies within the middle 95% interval (214.4881,303.2984), we do not reject the null hypothesis in our two-tailed goodness-of-fit test. This suggests that there is no significant over/under-dispersion of the residuals in our model. We can also conduct a stronger, one-tailed test for over-dispersion:

$$H_0 : \phi = 1 \quad H_1 : \phi > 1$$

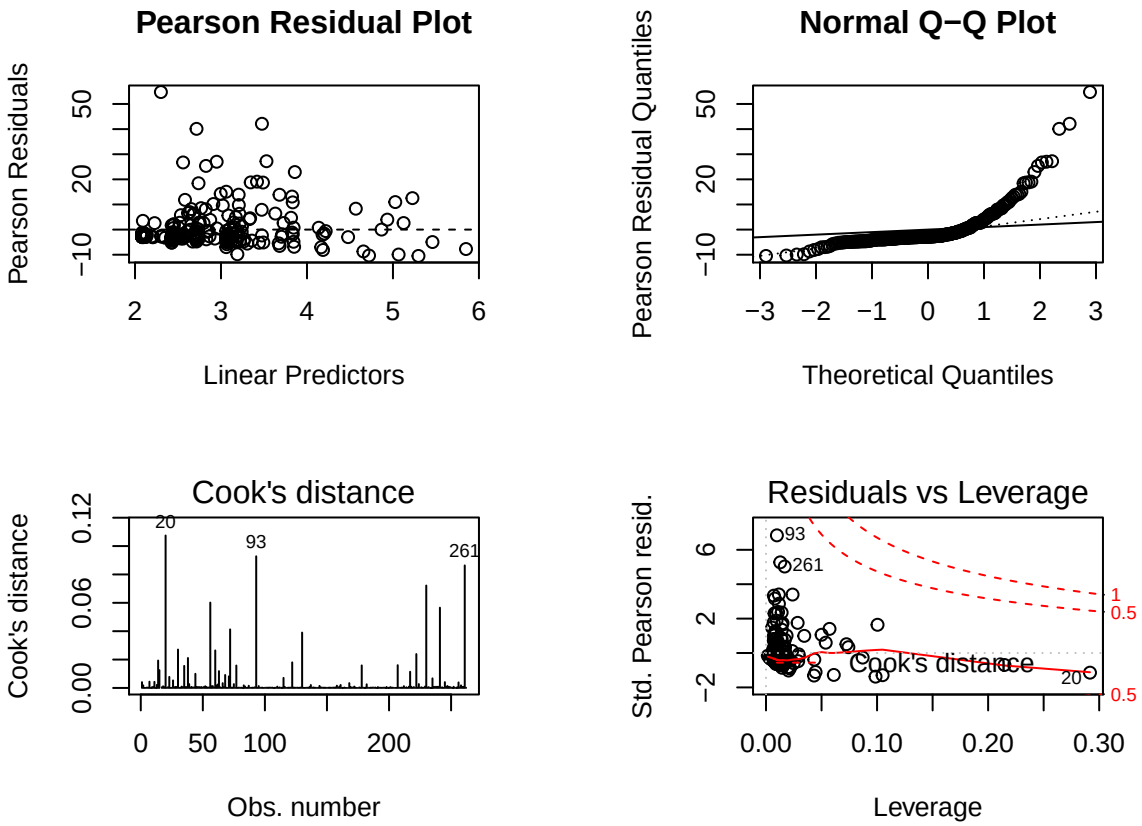
```
roach.glm3$deviance/summary(roach.glm3)$dispersion
```

```
## [1] 169.3437
```

```
qchisq(p = 0.95, df = roach.glm3$df.residual)
```

```
## [1] 295.3934
```

Again, our test-statistic does not exceed the 95% quantile, so we do not reject the null hypothesis and thus there isn't any issue with over-dispersion.



The Pearson Residuals vs Linear Predictors plot still highlights issues with vertical outliers of the Pearson residuals as a large number points greatly exceed the ± 2 range, however this is still an improvement over previous models. We also still see a strong indication of issues with homoscedasticity, although not as severe.

From the Normal Quantile-Quantile plot a major right-tail in the distribution of the Pearson residuals is still apparent, and thus still poses as a major issue for our normality assumption.

The Cook's Distance and Residuals vs Leverage plots are where most of the improvements from our new model are present. None of the observations exceed the 0.5 Cook's distance level, and so we no longer have an issue with highly influential points. There still are some observations that possess significantly high leverage (exceeding the $\frac{2p}{n} = 5/262$ cut off value), however, as no observations are highly influential, it's reasonable to assume that these shouldn't be too much of a concern for our model fit.

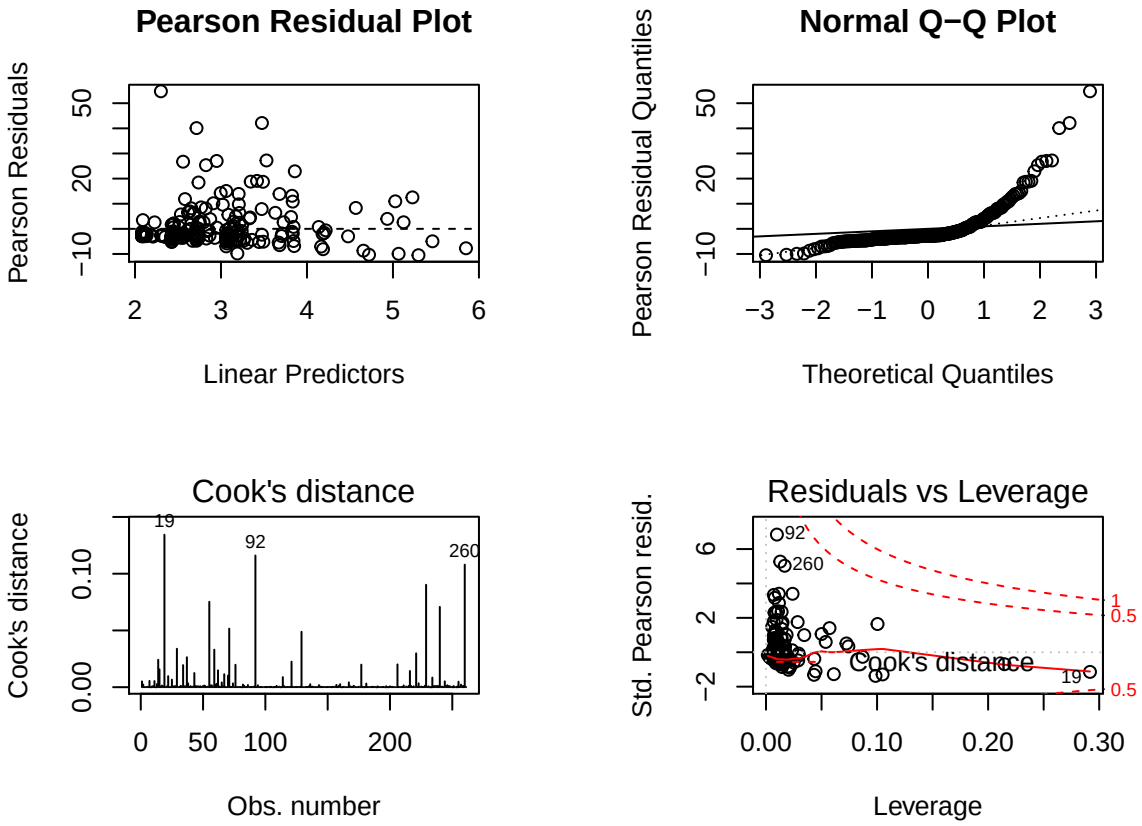
(e)

```
anova(roach.glm3, test = 'Chi')
```

```
## Analysis of Deviance Table
##
## Model: quasipoisson, link: log
##
## Response: trap.rate
##
## Terms added sequentially (first to last)
##
##
```

```
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        261      16954
## treatment  1    571.0       260      16383  0.002966 **
## senior     1    483.1       259      15900  0.006276 **
## roach1     1   4470.1       258      11430 < 2.2e-16 ***
## obs16      1    476.4       257      10953  0.006646 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
roach.glm4 <- glm(trap.rate[-16] ~ treatment[-16] + senior[-16] + roach1[-16],
                  family = quasipoisson, weights = exposure2[-16])
par(mfrow = c(2,2), mar = c(4,4.3,4,4.3))
plot(residuals(roach.glm4, type =  'pearson') ~ roach.glm4$linear.predictors,
     xlab = 'Linear Predictors', ylab = 'Pearson Residuals',
     main = 'Pearson Residual Plot')
abline(h=0, lty = 2)
qqnorm(residuals(roach.glm4, type =  'pearson'),
       ylab = 'Pearson Residual Quantiles', main = 'Normal Q-Q Plot')
qqline(residuals(roach.glm4, type =  'pearson'), lty = 3)
abline(0,1)
plot(roach.glm4, which = c(4,5))
```



Clearly observation 16 is an outlier, so lets fit the model excluding this observation.