Assignment 3

2022-11-23

Question 1

a)

We have

$$
\begin{aligned}
\log f(y; r, \lambda_i) &= \log y^{r-1} + \log e^{-y/\lambda_i} - \log \lambda_i^r - \log \Gamma(x) \\
&= (r-1)\log y - \frac{y}{\lambda_i} - r\log\lambda_i - \log\Gamma(x) \\
&= (r-1)\log y - yr\theta_i - r\log\lambda_i - \log\Gamma(x) \\
&= (r-1)\log y - yr\theta_i - r\log\left((r\theta_i)^{-1}\right) - \log\Gamma(x) \\
&= (r-1)\log y - yr\theta_i + r\log(r\theta_i) - \log\Gamma(x) \\
&= (r-1)\log y - yr\theta_i + r\log(r) + r\log(\theta_i) - \log\Gamma(x) \\
&= -yr\theta_i + r\log\theta_i + r\log(r) + (r-1)\log y - \log\Gamma(x) \\
&= -r(y\theta_i - \log\theta_i) + r\log(r) + (r-1)\log y - \log\Gamma(x) \\
&= \frac{y\theta_i - \log\theta_i}{-1/r} + r\log(r) + (r-1)\log y - \log\Gamma(x) \\
&= \frac{y\theta_i - \log\theta_i}{-1/r} + \left(r\log r + (r-1)\log y - \log\Gamma(r)\right).
\end{aligned}
$$

Thus, it follows that $f(y; r, \lambda_i)$ belongs to the exponential family with $b(\theta_i) = \log\theta_i$, $\phi/A_i = -1/r$ with $\phi = 1/r$ and $A_i = -1$.

b)

We have

$$b(\theta_i) = \log\theta_i$$

so $\mathrm{E}\, Y_i = b'(\theta_i) = \frac{1}{\theta_i} = r\lambda_i = \mu_i$. Similarly,

$$\mathrm{Var}(Y_i) = b''(\theta_i)\phi/A_i = \frac{-1}{\theta_i^2}\frac{-1}{r} = \frac{1}{\theta_i^2 r} = \frac{r^2\lambda_i^2}{r} = r\lambda_i^2.$$

Furthermore, the CLF $g(\mu_i) = (b')^{-1}(\mu_i) = \frac{1}{\mu_i}$.

c)

We have the Pearson's statistic

$$\begin{aligned}
P &= \phi \sum_{i=1}^{n} \frac{(Y_i - \mathrm{E}_{\hat{\beta}} Y_i)^2}{\mathrm{Var}_{\hat{\beta}}(Y_i)} \\
&= \frac{1}{r} \sum_{i=1}^{n} \frac{\left(Y_i - \frac{1}{\hat{\theta}_i}\right)^2}{1/(\hat{\theta}_i^2 r)} \\
&= \frac{1}{r} \sum_{i=1}^{n} \frac{\left(\frac{Y_i \hat{\theta}_i}{\hat{\theta}_i} - \frac{1}{\hat{\theta}_i}\right)^2}{1/(\hat{\theta}_i^2 r)} \\
&= \frac{1}{r} \sum_{i=1}^{n} \frac{\left(Y_i \hat{\theta}_i - 1\right)^2}{\hat{\theta}_i^2/(\hat{\theta}_i^2 r)} \\
&= \frac{1}{r} \sum_{i=1}^{n} \frac{(Y_i \hat{\theta}_i)^2 - 2 Y_i \hat{\theta}_i + 1}{\hat{\theta}_i^2/(\hat{\theta}_i^2 r)} \\
&= \sum_{i=1}^{n} \left((Y_i \hat{\theta}_i)^2 - 2 Y_i \hat{\theta}_i + 1\right).
\end{aligned}$$

Furthermore, $\tilde{\theta}_i = (b')^{-1}(Y_i) = 1/Y_i$ then

$$\begin{aligned}
D &= 2 \sum_{i=1}^{n} A_i (Y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)) \\
&= -2 \sum_{i=1}^{n} \left(Y_i \left(\frac{1}{Y_i} - \hat{\theta}_i \right) - \log \frac{1}{Y_i} + \log \hat{\theta}_i \right) \\
&= -2 \sum_{i=1}^{n} \left(Y_i \left(\frac{1}{Y_i} - \hat{\theta}_i \right) + \log Y_i + \log \hat{\theta}_i \right) \\
&= -2 \sum_{i=1}^{n} (1 - Y_i \hat{\theta}_i + \log Y_i + \log \hat{\theta}_i).
\end{aligned}$$

Finally, for the working matrix $\hat{W}$,

$$\begin{aligned}
\hat{w}_{ii} &= \frac{A_i}{[g'(\mu_i)]^2 b''(\theta_i)} \\
&= \frac{1}{\frac{1}{\hat{\mu}_i^4} \frac{-1}{\hat{\theta}^2}} \\
&= \frac{1}{\hat{\theta}_i^4 \frac{-1}{\hat{\theta}^2}} \\
&= \frac{-1}{\hat{\theta}_i^2}.
\end{aligned}$$

Question 2

a)

There are 18 observations. The estimated parameters $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (-0.021, 0.017, 0.01)$. Similarly, $\tilde{\phi} = \frac{D_\Omega}{n-p-1} = \frac{0.3}{18-2-1} = 0.02$. To test for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, we compute $t_1 = \frac{\hat{\beta}_1}{\sqrt{\left(\hat{I}_F^{-1}\right)_{11}}} =$

$\frac{0.017}{0.001} = 17$ where $\left(\hat{I}_F^{-1}\right)_{11}$ is the square of the standard error of the first variable. Since the test statistic $t_1 = 17$ is larger than $t_{15;1-0.05/2} = 2.13$, we reject H_0. The 96% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{15;1-0.04/2}\sqrt{\left(\hat{I}_F^{-1}\right)_{11}} = 0.017 \pm 2.25 \times 0.01 = [-0.0055, 0.0395].$$

b)

To investigate the relevance of the variables $X1$ and $X2$, it suffices to look at the corresponding p-values in the output table of the command $summary(model)$. This shows that the p-value of $X2$ is smaller than 0.05 so we can not dismiss the relevance of $X2$. Similarly, we test for the relevance of $X1$ in a) and rejected the null hypothesis that $\beta_1 = 0$ so $X1$ is relevant.

c)

If not given, the standard errors in the first table could have been recovered as the square root of the variance of the sample. If the estimates are given, then the t-values could be (approximately) recovered by the formula

$$t_i = \frac{\hat{\beta}_i}{\sqrt{\left(\hat{I}_F^{-1}\right)_{ii}}}$$

where $\sqrt{\left(\hat{I}_F^{-1}\right)_{11}}$ is the standard error of the i predictor. It follows that the p-values could be recovered from the test statistic and the degree of freedom. Similarly, if the t-values are given, the estimates could be recovered.

For the second table, the degree of freedom can be recovered from the number of observations and the number of predictors. If the residuals deviance are given then we could calculate the deviances and the p-values since the test statistic will also be known. On the other hand, we are given the residual deviances for the variables but not the NULL one, then we could not calculate the deviance of the first variable since it depends on the residual deviance of the NULL one. Furthermore, if the deviances are known then we could not compute the residual one because the system of equation will be underdetermined. Also, by the result of Exercise 1, the residual deviance D of the models can be determined by the responses Y_i and the fitted canonical parameters $\hat{\theta}_i$ but since $\hat{\theta}_i = x_i^T\hat{\beta}$, we could recover $\hat{\theta}_i$ and therefore, the deviances of the models. And if the deviances are known, everything in the second table can be recovered.

d)

We use that option because the dispersion parameter ϕ is not known beforehand It is not possible to derive what we would have obtained if we used the Chi-square test option because the Chisquare test assumes that ϕ is known and the test statistics use ϕ instead of the estimated one in the F test. ### e) Since we have the dispersion parameter $\hat{\phi}$, the Pearson statistic P can be derived as

$$P = \hat{\phi}(n - p - 1) = 0.001958 * (18 - 2 - 1) = 0.0029.$$

It is not possible to derive $\hat{W}$ because as we have shown in exercise 1, the element $\hat{w}_{ij}$ of the working matrix $\hat{W}$ depends on the estimated $\hat{\theta}_i$ which in turn depends on the input x_i. Since we do not know x_i (not given in the question), we can not derive $\hat{w}_{ij}$.

Question 3

a)

```r
data3=read.table("psi.txt",header=TRUE)
attach(data3)
```

The following table gives us some insight about the dataset. Note that the students that have not been instructed with the PSI method are, a priori, more expected to fail. This observation is supported by the difference in the passing rate of both groups. However, we still need more information to draw any conclusions.

```r
n00 = length(data3$gpa[data3$passed == 0 & data3$psi == 0])
n01 = length(data3$gpa[data3$passed == 0 & data3$psi == 1])
n10 = length(data3$gpa[data3$passed == 1 & data3$psi == 0])
n11 = length(data3$gpa[data3$passed == 1 & data3$psi == 1])

percent0 = n10 / (n00 + n10); percent0
```

```
## [1] 0.1666667
```

```r
percent1 = n11 / (n01 + n11); percent1
```

```
## [1] 0.5714286
```

```r
table_data = matrix(c(n00,n01,n10,n11), ncol=2, byrow=TRUE)
colnames(table_data) = c('No PSI','PSI')
rownames(table_data) = c('Failed','Passed')
table=as.table(table_data)
table
```

```
##        No PSI PSI
## Failed     15   6
## Passed      3   8
```

The table below shows the mean of the students for each of the combinations of the binary variables 'passed' and 'psi'. The only significant difference is that the mean of the students that passed and did not have PSI is higher than the average grade of the students that passed and did have PSI. However, this could be due to the fact that only 3 students passed. Probably, only the outstanding students managed.

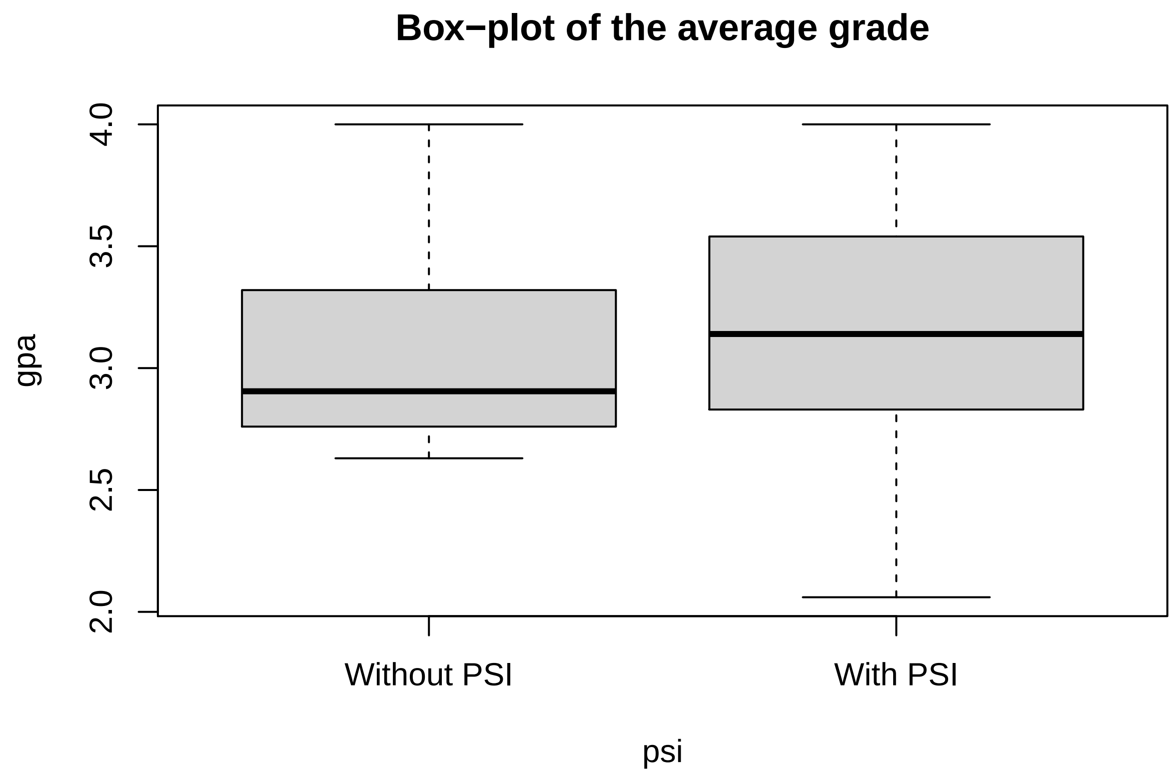
```r
m00 = mean(data3$gpa[data3$passed == 0 & data3$psi == 0]);
m01 = mean(data3$gpa[data3$passed == 0 & data3$psi == 1]);
m10 = mean(data3$gpa[data3$passed == 1 & data3$psi == 0]);
m11 = mean(data3$gpa[data3$passed == 1 & data3$psi == 1]);

table_data2 = matrix(c(m00,m01,m10,m11), ncol=2, byrow=TRUE);
colnames(table_data2) = c('No PSI','PSI');
rownames(table_data2) = c('Failed','Passed');
table2=as.table(table_data2);
table2
```

```
##          No PSI      PSI
## Failed 2.976000 2.891667
## Passed 3.726667 3.322500
```

The following will give us more insight about the data. Note how the students that did have PSI have a higher average and also more variance. The plot suggests a better overall performance of the students with PSI.

```r
boxplot(gpa~psi,data=data3,names=c("Without PSI","With PSI"), main='Box-plot of the average grade')
```



Now, we can try different models and compare how they fit, after converting the variable 'psi' into a factor:

```
model3=glm(passed~gpa + psi,data=data3,family=binomial)
drop1(model3, test = 'Chisq')
```

```
## Single term deletions
##
## Model:
## passed ~ gpa + psi
##        Df Deviance    AIC    LRT Pr(>Chi)
## <none>      26.253 32.253
## gpa     1   35.342 39.342 9.0885 0.002572 **
## psi     1   32.418 36.418 6.1647 0.013033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(glm(passed~psi * gpa,data=data3,family=binomial), test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: passed
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
```

```
## NULL                       31     41.183
## psi      1   5.8418        30     35.342 0.015650 *
## gpa      1   9.0885        29     26.253 0.002572 **
## psi:gpa  1   1.8725        28     24.381 0.171189
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The default logistic model includes all the explanatory variables with their interaction terms. We then check if either of these variables is significant or not, including the interaction term. Since the p-value of the interaction term is (> 0.05) non-significant,we can say that there is no interaction between the variables 'psi' and 'passed'. On the other hand, the p-values of both explanatory variables are smaller than 0.05 which mean that we can not dismiss them. Thus, a good model would be one that contains both explanatory variables.

b)

```
summary(model3)
```

```
##
## Call:
## glm(formula = passed ~ gpa + psi, family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8396  -0.6282  -0.3045   0.5629   2.0378
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -11.602      4.213  -2.754  0.00589 **
## gpa            3.063      1.223   2.505  0.01224 *
## psi            2.338      1.041   2.246  0.02470 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 41.183  on 31  degrees of freedom
## Residual deviance: 26.253  on 29  degrees of freedom
## AIC: 32.253
##
## Number of Fisher Scoring iterations: 5
```

It can be seen that the estimated coefficients for the model are $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (-11.6, 2.34, 3.06)$. It follows that the probability that a student with gpa equals to 3 who received psi passes the assignment is

$$\frac{\exp(-11.6 + 1 \times 2.34 + 3 \times 3.06)}{1 + \exp(-11.6 + 1 \times 2.34 + 3 \times 3.06)} = 0.48.$$

For the student who does not receive psi:

$$\frac{\exp(-11.6 + 0 \times 2.34 + 3 \times 3.06)}{1 + \exp(-11.6 + 0 \times 2.34 + 3 \times 3.06)} = 0.08.$$

Thus, it can be seen that having psi positively affect the passing probability, which is evident from the positive coefficient of psi. ### c) The odd is

$$o = \frac{P(passes)}{P(fails)} = \exp(-11.6 + 2.34 + gpa \times 3.06).$$

This number means that the probability of passing is o times as big as the probability of failing. So the bigger the odd, the higher the probability of passing. When *gpa* increases in one unit, the odd is multiplied by $e^{3.06}$, which is larger than 1 so this means that the higher the *gpa*, the more likely the probability of passing. Thus, this number is dependent on *gpa*. For an estimate of the odd, we calculate the mean of the *gpa*, which is 3.12 then the odd is $\exp\left(-11.6 + 2.34 + 3.12 \times 3.06\right)$

Question 4

a)

We first build the model without taking *math* into account using Poisson regression.

```
model4=glm(num_awards~prog,data=data4,family=poisson)
anova(model4,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: num_awards
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                   199     228.83
## prog  2   12.733       197     216.10 0.001718 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results for the ANOVA with the Chi Squared test yield a significant (< 0.05) p-value for the factor *prog*.

```
summary(model4);
```

```
##
## Call:
## glm(formula = num_awards ~ prog, family = poisson, data = data4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5306  -1.0750  -0.1625   0.5027   3.1536
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.5486     0.1961  -2.797  0.00515 **
## prog2         0.7068     0.2158   3.275  0.00106 **
## prog3         0.4432     0.2463   1.799  0.07199 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 228.83  on 199  degrees of freedom
## Residual deviance: 216.10  on 197  degrees of freedom
## AIC: 512.42
##
## Number of Fisher Scoring iterations: 5
```

```r
o1 = predict(model4, newdata=data.frame(prog=factor(1)), se.fit=TRUE);
exp(o1$fit)
```

```
##         1
## 0.5777778
```

```r
o2 = predict(model4, newdata=data.frame(prog=factor(2)), se.fit=TRUE);
exp(o2$fit)
```

```
##        1
## 1.171429
```

```r
o3 = predict(model4, newdata=data.frame(prog=factor(3)), se.fit=TRUE);
exp(o3$fit)
```

```
##   1
## 0.9
```

It can be seen that the second program has the most award $1.17 > 0.9 > 0.57$. Thus, it is the best for this model. ### b)

Now we are taking into account the variable *math* for our model. When performing the first test, we notice that the interaction term between the factor 'prog' and the variable 'math' does not turn out to be relevant. Checking the AIC values for the models we. note that indeed the model that excludes interaction term seems to work better. Therefore, we decide to drop that term.

```r
model4_1=glm(num_awards~prog*math,data=data4,family=poisson)
anova(model4_1, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: num_awards
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        199     228.83
## prog       2  12.7334       197     216.10  0.001718 **
## math       1  18.0527       196     198.05 2.149e-05 ***
## prog:math  2   3.6911       194     194.35  0.157935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
drop1(model4_1)
```

```
## Single term deletions
##
## Model:
## num_awards ~ prog * math
##           Df Deviance    AIC
## <none>         194.35 496.67
## prog:math  2   198.05 496.36
```

Now we need to check the significance of our new model. It can be seen that both of the p-values of *prog* and *math* are smaller than 0.05 so those two variables are important to the model.

```r
model4_2=glm(num_awards~prog+math,data=data4,family=poisson)
drop1(model4_2, test = 'Chisq')
```

```
## Single term deletions
##
## Model:
## num_awards ~ prog + math
##        Df Deviance    AIC     LRT  Pr(>Chi)
## <none>      198.05 496.36
## prog    2   204.23 498.55  6.1837   0.04542 *
## math    1   216.10 512.42 18.0527 2.149e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To compute 95%-confidence intervals, we first need to obtain the Covariance matrix to compute our confidence intervals. By default, this is done. in R by taking the Normal quantiles. However, we opt to be more conservative and not make any extra assumptions by computing it with. t-quantiles. The. resulting interval for each of the coefficients of the model is provided below.

```r
confint(model4_2, level=0.95)
```

```
## Waiting for profiling to be done...

##                   2.5 %     97.5 %
## (Intercept) -3.31216355 -1.4475568
## prog2        0.02817683  0.9126913
## prog3        0.08459774  1.0595014
## math         0.01936098  0.0520766
```

c)

```r
summary(model4_2)
```

```
##
## Call:
## glm(formula = num_awards ~ prog + math, family = poisson, data = data4)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.96335  -1.14818  -0.01392   0.35710   2.52541
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.372577   0.475525  -4.989 6.06e-07 ***
## prog2        0.452621   0.224746   2.014   0.0440 *
## prog3        0.561720   0.247482   2.270   0.0232 *
## math         0.035779   0.008344   4.288 1.80e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 228.83  on 199  degrees of freedom
## Residual deviance: 198.05  on 196  degrees of freedom
## AIC: 496.36
```

```
##
## Number of Fisher Scoring iterations: 5
```

Since the link function is *log*, which is an increasing monotonic function so an increase in an input will result in an increase in the output. Also, the estimated coefficients are $\hat{\beta} = (-2.37, 0.45, 0.56, 0.036) = (program1, program2, program3, math)$. Since the coefficient of $program3$ is the highest, it follows that the number of awards from $program3$ is highest as well.

```
out = predict(model4_2, newdata=data.frame(prog=factor(1), math=55), type="response", se.fit=TRUE)
exp(out$fit)
```

```
##         1
## 0.6671683
```

```
out = predict(model4_2, newdata=data.frame(prog=factor(2), math=55), type="response", se.fit=TRUE)
exp(out$fit)
```

```
##        1
## 1.049074
```

```
out = predict(model4_2, newdata=data.frame(prog=factor(3), math=55), type="response", se.fit=TRUE)
exp(out$fit)
```

```
##        1
## 1.170004
```

The number of awards for the vocational program and math score 55 is 1.05. To confirm the previous result, the number of awards for the academic program is 1.17, highest among all programs.

d)

We can see that the Pearson chi-squared statistic $P = 179.07$ and the deviance $D = 198.04$. It can be seen that the two values differ quite a bit which can indicate that the model does not fit the data well.

```
P=sum(residuals(model4_2,type="pearson")^2)
D=deviance(model4_2)
P
```

```
## [1] 179.0771
```

```
D
```

```
## [1] 198.0461
```

e)

To check for overdispersion, we check both

$$\tilde{\phi} = \frac{D}{n-p-1}$$

and

$$\hat{\phi} = \frac{P}{n-p-1}$$

to see if they are smaller than 1 or not.

```
D/df.residual(model4_2);
```

```
## [1] 1.01044
```

```
P/df.residual(model4_2);
```

```
## [1] 0.9136588
```

Since $\tilde{\phi} = 1.01 > 1$, we could technically say that there is an overdispersion problem. However, the value for $\hat{\phi}$ is satisfactory and $\tilde{\phi}$ is sufficiently close to the desired value 1 for us to disregard it. This indicates that we are unlikely to have an overdispersion problem in this scenario.