1. Use the data set crabs to fit a Poisson regression model with canonical link for the count variable Satellites using predictor variables Width, Dark, GoodSpine.

Answer:
A poisson regression model of formula

Satellites ~ Width + Dark + GoodSpine

Gives the following result

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.82009    0.57086   -4.94  7.8e-07 ***
Width        0.14920    0.02075    7.19  6.5e-13 ***
Darkyes     -0.26566    0.10497   -2.53    0.011 *
GoodSpineyes -0.00204    0.09799   -0.02    0.983

Residual deviance: 560.96  on 169  degrees of freedom
```

There is overdispersion present since residual deviance is much higher than degrees of freedom

If we compare the deviance with a chi square of 169 degrees of freedom we get a p value of close to zero so we can reject the null hypothesis that the model is adequate.

Including the interaction terms

A poisson regression model of formula

Satellites ~ Width * Dark * GoodSpine

Gives the following result

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.4144     1.0051   -3.40  0.00068 ***
Width                       0.1713     0.0366    4.68  2.8e-06 ***
Darkyes                    -1.0490     1.6561   -0.63  0.52647
GoodSpineyes                2.2686     1.3281    1.71  0.08761 .
Width:Darkyes               0.0299     0.0620    0.48  0.62954
Width:GoodSpineyes         -0.0840     0.0485   -1.73  0.08329 .
Darkyes:GoodSpineyes       -7.4078     3.4831   -2.13  0.03344 *
Width:Darkyes:GoodSpineyes  0.2751     0.1266    2.17  0.02972 *

Residual deviance: 549.49  on 165  degrees of freedom
```

If we compare the deviance with a chi square of 165 degrees of freedom we get a p value of close to zero so we can reject the null hypothesis that the model is adequate.

As we can see width and dark are significant variables in the first model lets fit a model based on only these two

A poisson regression model of formula

Satellites ~ Width * Dark

Gives the following result

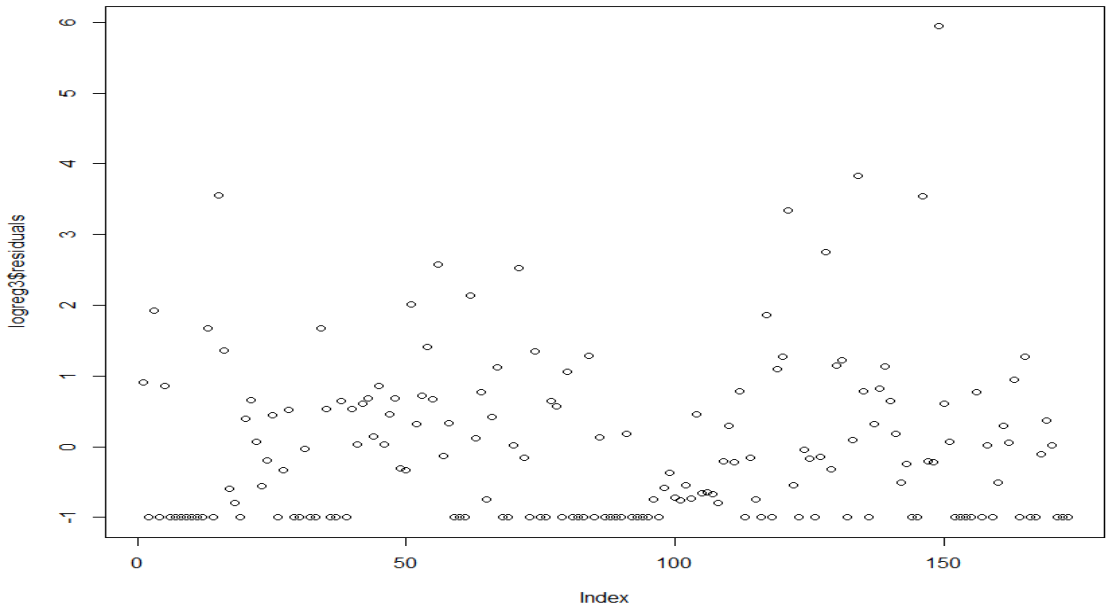
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.0992     0.6511   -3.22   0.0013 **
Width          0.1228     0.0238    5.17  2.4e-07 ***
Darkyes       -3.3930     1.3446   -2.52   0.0116 *
Width:Darkyes  0.1178     0.0503    2.34   0.0191 *

Residual deviance: 555.42  on 169  degrees of freedom
```

which is a upgrade to the first model but still not adequate.

We stick to the second model but with quasiposisson family to handle the overdispersion.

This is a plot of residuals of model



Which does not show any pattern

2. Use the data set heart to fit a logistic regression with canonical link to Death using covariates AgeGroup, Severity, Delay, Region. Note here that the data is in matrix form: death and Total, you want to input the data to the model as death and non-death.

Answer:

We first treat each of the covariates as factors and fit a logistic regression model with binomial family

A model of
(Death,NonDeath) ~ Agegroup+Severity+Delay+Region
Gives the following results

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -4.1040     0.0953  -43.08  < 2e-16 ***
as.factor(AgeGroup)2   1.1479     0.0935   12.28  < 2e-16 ***
as.factor(AgeGroup)3   2.1974     0.1002   21.93  < 2e-16 ***
as.factor(Delay)2      0.0716     0.0790    0.91    0.365
as.factor(Delay)3      0.2566     0.0933    2.75    0.006 **
as.factor(Severity)2   0.8275     0.0828    9.99  < 2e-16 ***
as.factor(Severity)3   2.0762     0.1434   14.48  < 2e-16 ***
as.factor(Region)2     0.0532     0.2049    0.26    0.795
as.factor(Region)3     0.8014     0.1346    5.96  2.6e-09 ***

Residual deviance:  113.11  on 65  degrees of freedom
```

If we compare the deviance with a chi square of 65 degrees of freedom we get a p value of 0.0002044 so we can reject the null hypothesis that the model is adequate.

With the interaction terms included
A model of
(Death,NonDeath) ~ Agegroup*Severity*Delay*Region
shows that Agegroup* delay is the significant interaction term.

So we proceed with the model of
(Death,NonDeath) ~ Agegroup *Delay+Severity+Region

Which gives the statistics

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -3.9340     0.1268  -31.03  < 2e-16 ***
as.factor(AgeGroup)2                0.7648     0.1588    4.81  1.5e-06 ***
as.factor(AgeGroup)3                2.2183     0.1675   13.25  < 2e-16 ***
as.factor(Delay)2                  -0.2546     0.1801   -1.41   0.1573
as.factor(Delay)3                   0.0912     0.2164    0.42   0.6735
as.factor(Severity)2                0.8317     0.0829   10.03  < 2e-16 ***
as.factor(Severity)3                2.0847     0.1437   14.50  < 2e-16 ***
```

```
as.factor(Region)2                        0.0446     0.2050    0.22   0.8277
as.factor(Region)3                        0.8014     0.1347    5.95  2.7e-09 ***
as.factor(AgeGroup)2:as.factor(Delay)2    0.6480     0.2144    3.02   0.0025 **
as.factor(AgeGroup)3:as.factor(Delay)2    0.0341     0.2271    0.15   0.8807
as.factor(AgeGroup)2:as.factor(Delay)3    0.3809     0.2560    1.49   0.1367
as.factor(AgeGroup)3:as.factor(Delay)3   -0.0342     0.2686   -0.13   0.8987

Residual deviance:   97.567 on 61  degrees of freedom
```

So Residual deviance is decreased with the loss of 4 degrees of freedom

If we compare the deviance with a chi square of 61 degrees of freedom we get a p value of 0.002042 so we can reject the null hypothesis that the model is adequate.

Let us fit a model omitting Delay altogether.

A model of
(Death,NonDeath) ~ Agegroup+Severity+Region
Gives Residual deviance: 120.95 on 67 degrees of freedom
So Deviance has increased.

If we include interaction terms
(Death,NonDeath) ~ Agegroup*Severity*Region
The residual deviance is 70.897 on 49 degrees of freedom.

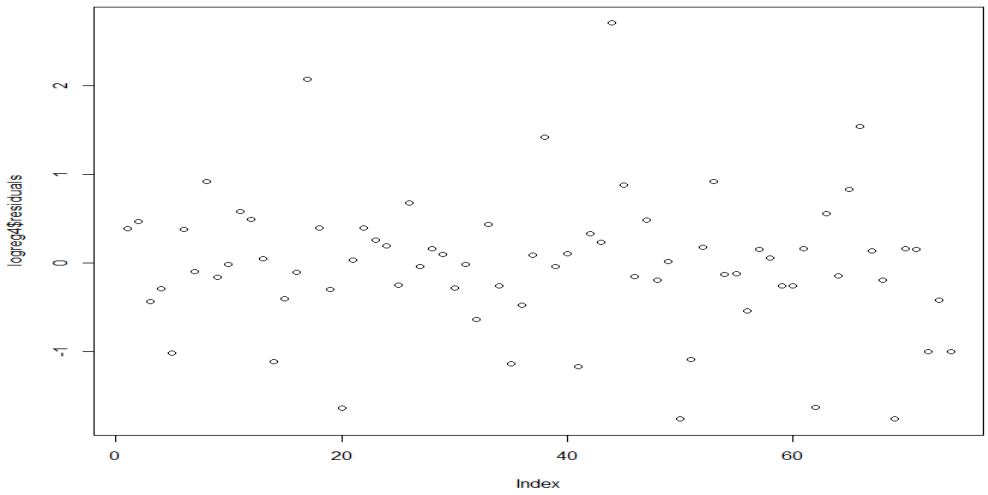
So with the loss of freedom we get a chi square based p value of 0.02203 so this model is not adequate.

Now if we include the Delay term
(Death,NonDeath) ~ Agegroup*Severity*Region + Delay
The residual deviance is 62.885 on 47 degrees of freedom

Now we get a chi square p value of 0.06051 which means we cannot reject the null hypothesis that the model is adequate.

So we stick to the model but with quasibinomial family.
(Death,NonDeath) ~ Agegroup*Severity*Region + Delay

This is a plot of residuals



3. Use the data set heart to fit a logistic regression using a probit link to the variable Death using covariates AgeGroup, Severity, Delay and Region.

Answer:

This is the same problem as 2 but so far we have fitted models with canonical link but now we go ahead and fit with probit link.

A model of
(Death,NonDeath) ~ Agegroup+Severity+Delay+Region
Gives the following results

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.1739     0.0414  -52.46  < 2e-16 ***
as.factor(AgeGroup)2  0.5185     0.0408   12.70  < 2e-16 ***
as.factor(AgeGroup)3  1.0739     0.0471   22.81  < 2e-16 ***
as.factor(Delay)2     0.0311     0.0382    0.82   0.4150
as.factor(Delay)3     0.1228     0.0461    2.66   0.0077 **
as.factor(Severity)2  0.4357     0.0433   10.05  < 2e-16 ***
as.factor(Severity)3  1.1471     0.0839   13.66  < 2e-16 ***
as.factor(Region)2    0.0117     0.1008    0.12   0.9075
as.factor(Region)3    0.3918     0.0696    5.63  1.8e-08 ***

Residual deviance:   98.546  on 65  degrees of freedom
```

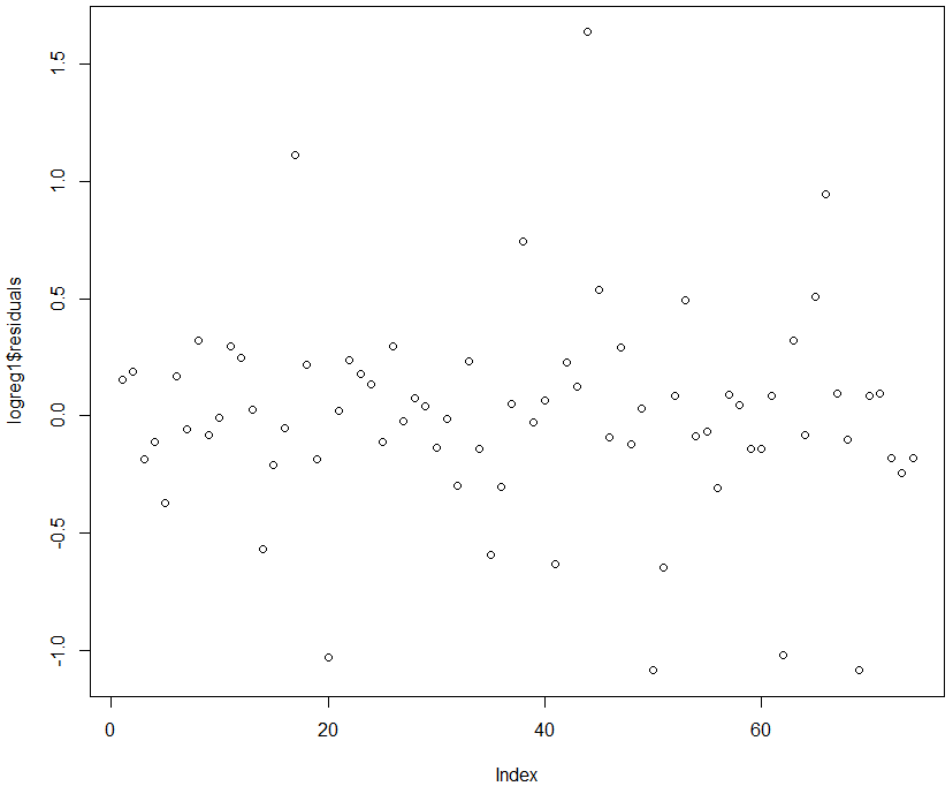
This amounts to a p value of 0.00459 so we reject that the model is adequate.

If we include all the interaction terms possible we see the interaction agegroup*delay*region is significant.

A model of
(Death,NonDeath) ~ Agegroup*Delay*Region +Severity
Gives the residual deviance 63.036 on 45 degrees of freedom which amounts to a p value of 0.03903 meaning we still reject the model.

A model of
(Death,NonDeath) ~ Agegroup*Severity*Region + Delay
Gives the residual deviance of 63.29 on 47 degrees of freedom which amounts to a p value of 0.0565 meaning we now are in a position where we cannot reject the model.

This is a plot of residuals



We stick to this model but with quasibinomial family since over dispersion is present.