NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Computational Assignment #6: Poisson and Zero-Inflated Poisson Regression
MSDS 410

In this assignment we will be fitting models and calculating the various summative statistics that are associated with Poisson and Zero-Inflated Poisson Regression. In addition, we will be fitting logistic regression models and interpreting the results. Students are expected to show all work in their computations. A good practice is to write down the generic formula for any computation and then fill in the values need for the computation from the problem statement. Throughout this assignment keep all decimals to three places, i.e. X.xxx. Students are expected to use correct notation and terminology, and to be clear, complete and concise with all interpretations of results. This computational assignment is worth 50 points. The points associated with each problem are given with the specific question.

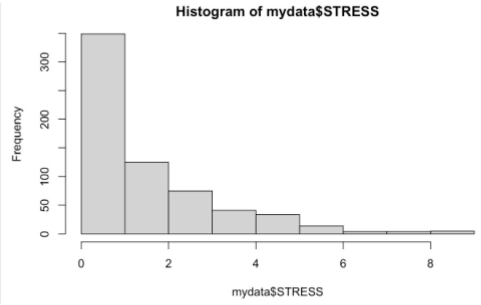
Any computations that involve “the log function”, denoted by log(x), ***are always meant to mean the natural log function (which will show as ln() on a calculator).*** The only time that you should ever use a log function other than the natural logarithm is if you are given a specific base.

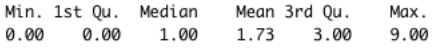
For this assignment, we will be using the STRESS dataset. This includes information from about 650 adolescents in the US who were surveyed about the number of stressful life events they had experienced in the past year (STRESS). STRESS is an integer variable that represents counts of stressful events. The dataset also includes school and family related variables, which are assumed to be continuously distributed. These variables are:

COHES = measure of how well the adolescent gets along with their family (coded low to high)
ESTEEM = measure of self-esteem (coded low to high)
GRADES = past year’s school grades (coded low to high)
SATTACH = measure of how well the adolescent likes and is attached to their school (coded low to high)

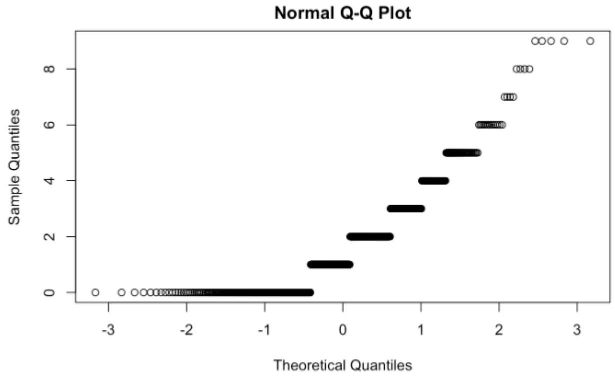
Each problem is worth 5 points.

1. For the STRESS variable, make a histogram and obtain summary statistics. Obtain a normal probability (Q-Q) plot for the STRESS variable. Is STRESS a normally distributed variable? What do you think is its most likely probability distribution for STRESS? Give a justification for the distribution you selected.



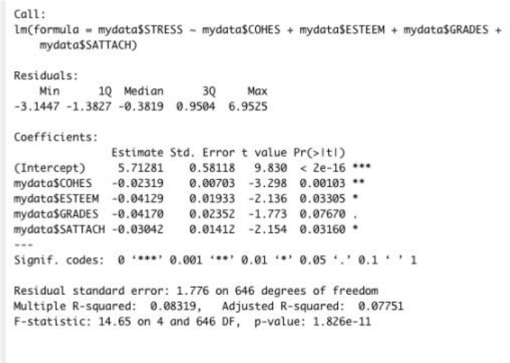


```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00    0.00    1.00    1.73    3.00    9.00
```

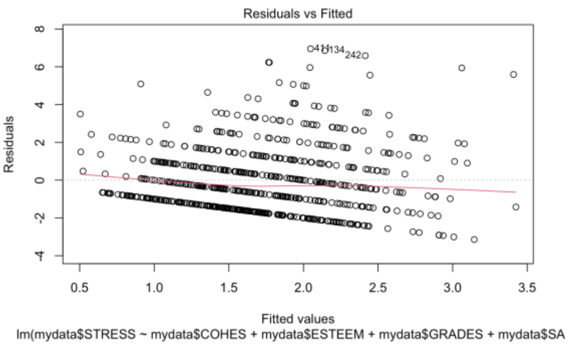


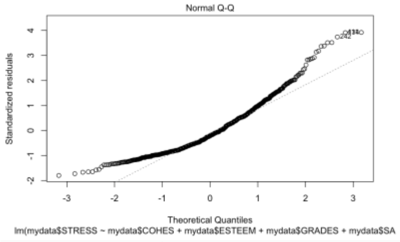
The STRESS variable is not normally distributed, as it is skewed to the right a lot, due to such high amounts of individuals have little to no stress. This appears to have a Poisson distribution, as the events do not affect each other and the probability that an event occurs does not change overtime.

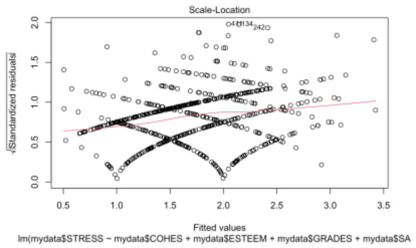
2. Fit an OLS regression model to predict STRESS (Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Obtain the typical diagnostic information and graphs. Discuss how well this model fits. Obtain predicted values (Y_hat) and plot them in a histogram. What issues do you see?

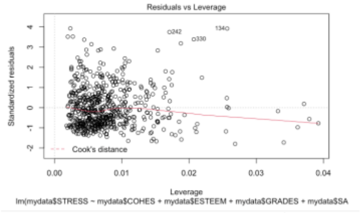


```
Call:
lm(formula = mydata$STRESS ~ mydata$COHES + mydata$ESTEEM + mydata$GRADES +
    mydata$SATTACH)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1447 -1.3827 -0.3819  0.9504  6.9525

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.71281    0.58118   9.830  < 2e-16 ***
mydata$COHES   -0.02319    0.00703  -3.298  0.00103 **
mydata$ESTEEM  -0.04129    0.01933  -2.136  0.03305 *
mydata$GRADES  -0.04170    0.02352  -1.773  0.07670 .
mydata$SATTACH -0.03042    0.01412  -2.154  0.03160 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.776 on 646 degrees of freedom
Multiple R-squared:  0.08319,	Adjusted R-squared:  0.07751
F-statistic: 14.65 on 4 and 646 DF,  p-value: 1.826e-11
```

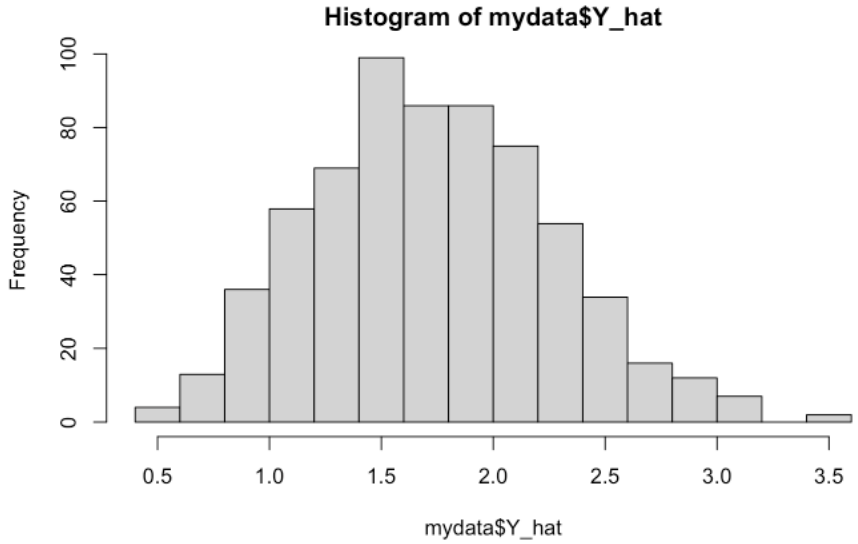




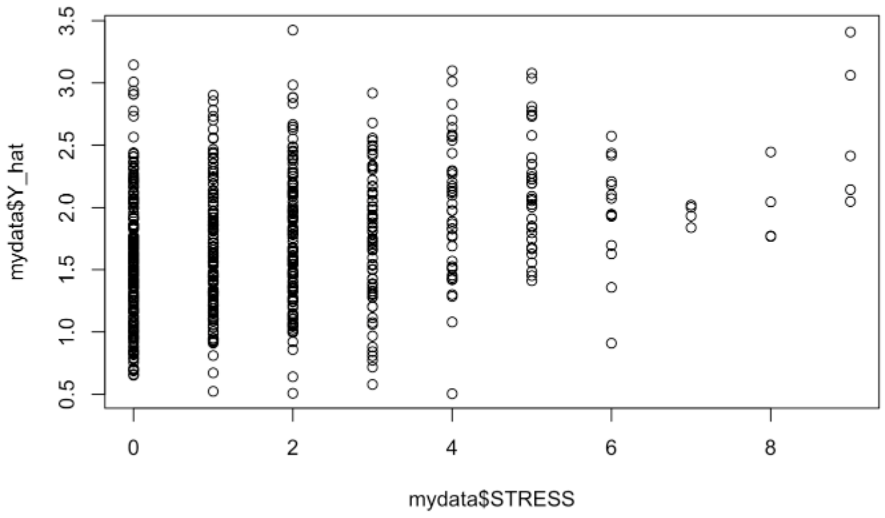




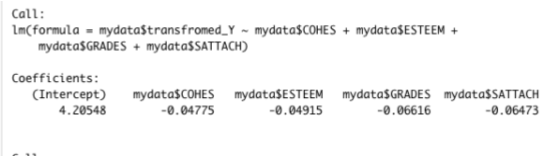
The residuals vs. fitted values plot does not seem to be truly random, and the Q-Q plot also has an almost linear trend with the theoretical quantities and the standardised resudals. This should not occur in a well fitting model. The scale-location plot also is not truly random, indicating this model does not fit well. This is further emphasized by the low R-value of the model.



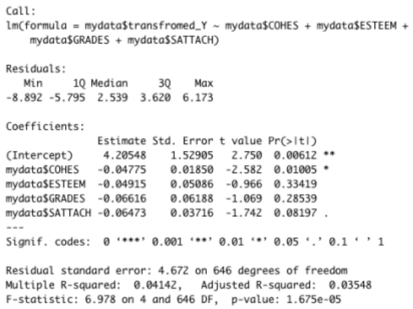
This histogram seems to have a normal distribution, but we can see from the STRESS vs. Y_hat plot how the linear model is not the proper way to go in this case:



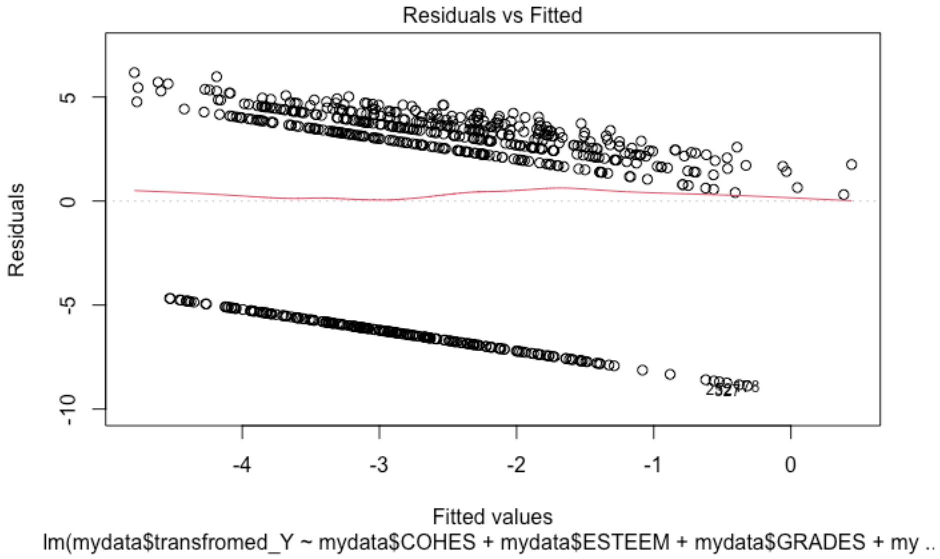
3. Create a transformed variable on Y that is LN(Y). Fit an OLS regression model to predict LN(Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Obtain the typical diagnostic information and graphs. Discuss how well this model fits. Obtain predicted values (LN(Y)_hat) and plot them in a histogram. What issues do you see? Does this correct the issue?



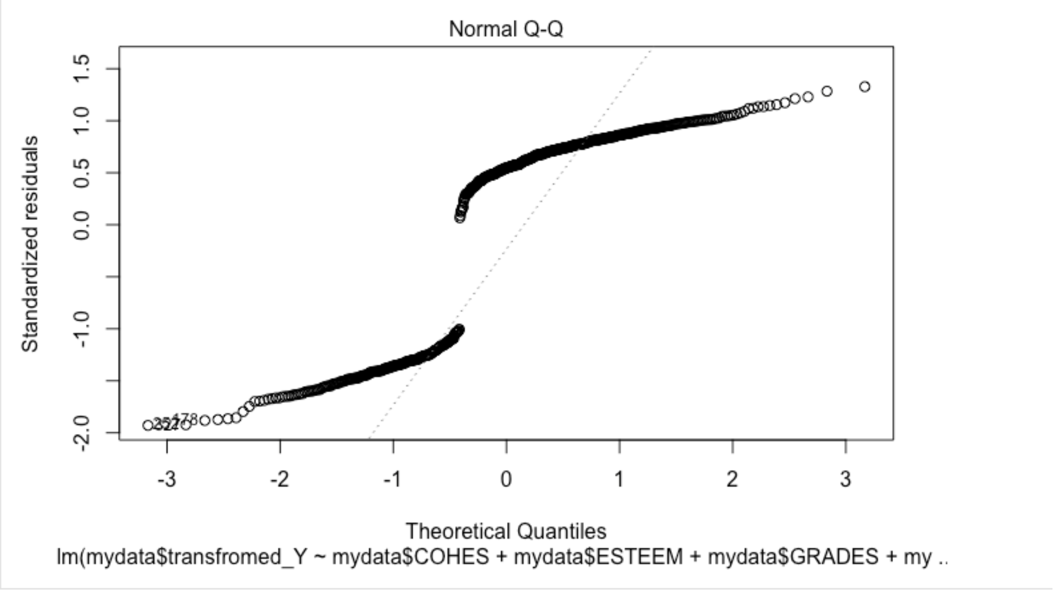
```
Call:
lm(formula = mydata$transfromed_Y ~ mydata$COHES + mydata$ESTEEM +
    mydata$GRADES + mydata$SATTACH)

Coefficients:
   (Intercept)    mydata$COHES   mydata$ESTEEM   mydata$GRADES  mydata$SATTACH
       4.20548        -0.04775        -0.04915        -0.06616        -0.06473
```



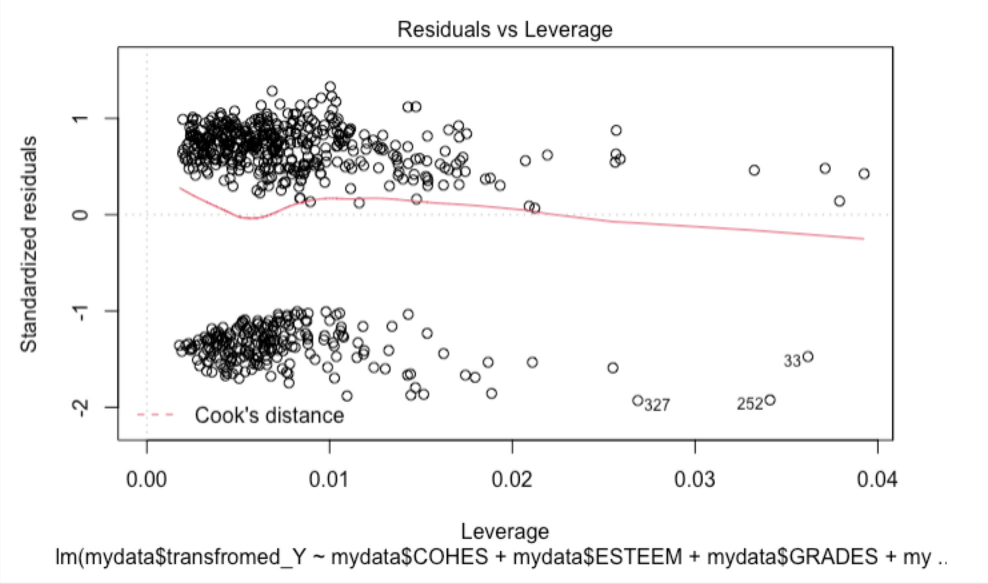
```
Call:
lm(formula = mydata$transfromed_Y ~ mydata$COHES + mydata$ESTEEM +
    mydata$GRADES + mydata$SATTACH)

Residuals:
   Min     1Q Median     3Q    Max
-8.892 -5.795  2.539  3.620  6.173

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.20548    1.52905   2.750  0.00612 **
mydata$COHES   -0.04775    0.01850  -2.582  0.01005 *
mydata$ESTEEM  -0.04915    0.05086  -0.966  0.33419
mydata$GRADES  -0.06616    0.06188  -1.069  0.28539
mydata$SATTACH -0.06473    0.03716  -1.742  0.08197 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.672 on 646 degrees of freedom
Multiple R-squared:  0.04142,   Adjusted R-squared:  0.03548
F-statistic: 6.978 on 4 and 646 DF,  p-value: 1.675e-05
```



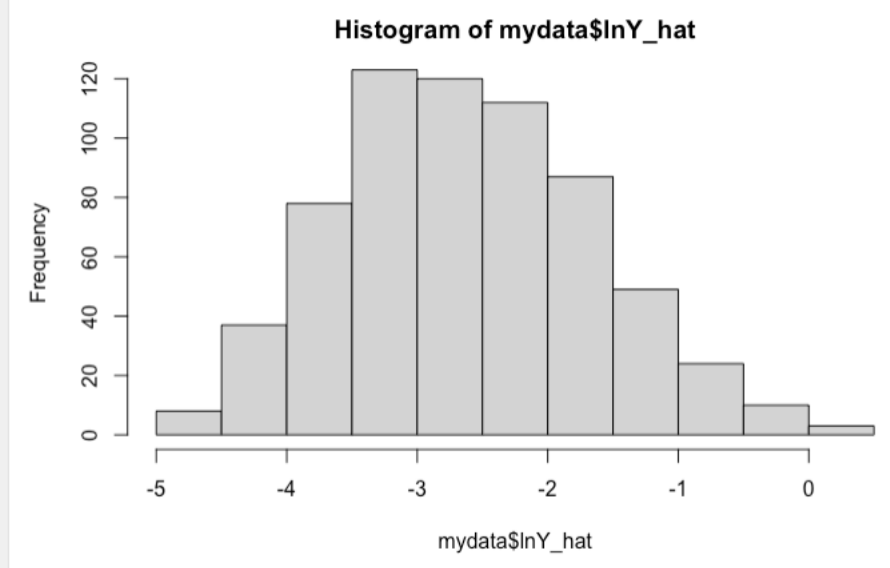
The residuals vs fitted plot still does not seem to show a random distribution of the relationship between the two, and there is still a trend in the plot.



The Q-Q plot for the theoretical quantities and the standardized residuals have a large gap in between two halves of the data, as is the case with the residuals vs. leverage plot.

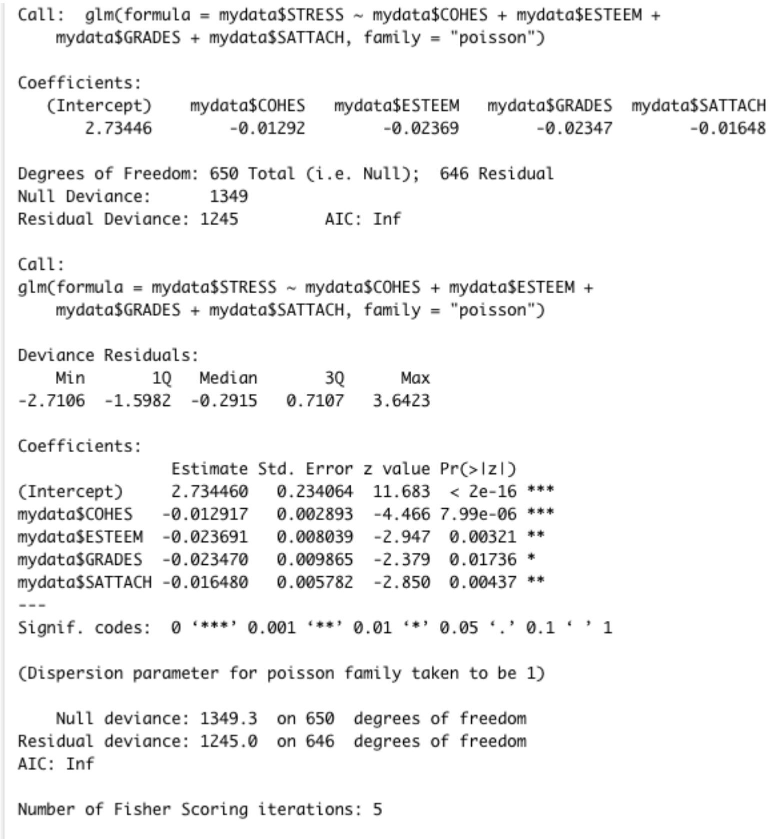


The Cook's line in splitting the data in half when finding outliers, indicating this is not a well-fitting model still.



The histogram is normally fitting, however. Still this model has an even lower R^2 value, and is not a good fit for this data.

4. Use the glm() function to fit a Poisson Regression for STRESS (Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Interpret the model's coefficients and discuss how this model's results compare to your answer for part 3). Similarly, fit an over-dispersed Poisson regression model using the same set of variables. How do these models compare?



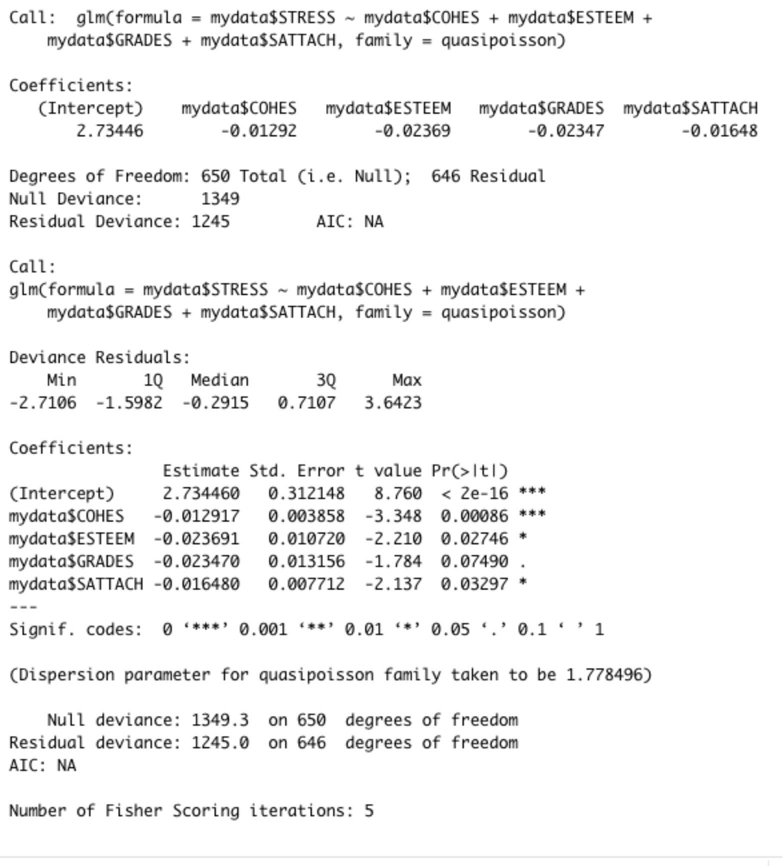
```
Call:  glm(formula = mydata$STRESS ~ mydata$COHES + mydata$ESTEEM +
    mydata$GRADES + mydata$SATTACH, family = "poisson")

Coefficients:
   (Intercept)    mydata$COHES   mydata$ESTEEM   mydata$GRADES  mydata$SATTACH
       2.73446        -0.01292        -0.02369        -0.02347        -0.01648

Degrees of Freedom: 650 Total (i.e. Null);  646 Residual
Null Deviance:      1349
Residual Deviance: 1245        AIC: Inf

Call:
glm(formula = mydata$STRESS ~ mydata$COHES + mydata$ESTEEM +
    mydata$GRADES + mydata$SATTACH, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7106  -1.5982  -0.2915   0.7107   3.6423

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.734460   0.234064  11.683  < 2e-16 ***
mydata$COHES   -0.012917   0.002893  -4.466 7.99e-06 ***
mydata$ESTEEM  -0.023691   0.008039  -2.947  0.00321 **
mydata$GRADES  -0.023470   0.009865  -2.379  0.01736 *
mydata$SATTACH -0.016480   0.005782  -2.850  0.00437 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1349.3  on 650  degrees of freedom
Residual deviance: 1245.0  on 646  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 5
```

Y= 2.73446 - 0.01292*B1 - 0.02369*B2 - 0.02347*B3 - 0.01648*B4

2.73446 is the intercept, - 0.01292 is the coefficient for COHES, - 0.02369 is the coefficient for ESTEEM,

-0.02347 is the coefficient for GRADES, and -0.01648 is the coefficient for SATTACH.

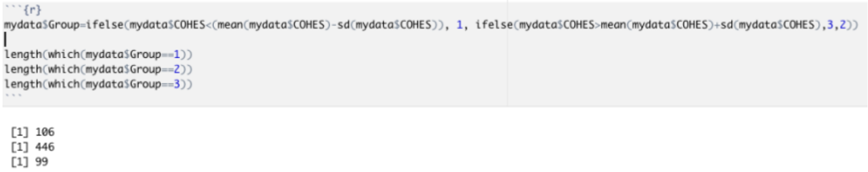
To find the effect of each coefficient on the variable, we must use e^(coefficient) to find the true effect in a Poisson distribution. So the coefficients have an effect of 0.987, 0.9765, 0.9768, and 0.98366 respectively on the variable in the above model, which appears to make sense in this case.



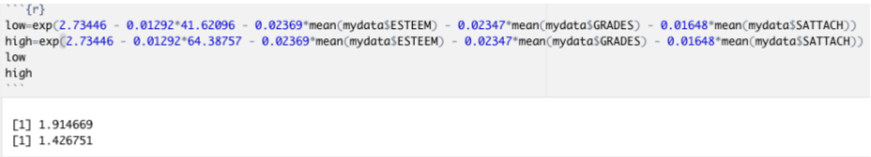
```
Call:  glm(formula = mydata$STRESS ~ mydata$COHES + mydata$ESTEEM + 
    mydata$GRADES + mydata$SATTACH, family = quasipoisson)

Coefficients:
   (Intercept)    mydata$COHES   mydata$ESTEEM   mydata$GRADES  mydata$SATTACH  
       2.73446        -0.01292        -0.02369        -0.02347        -0.01648  

Degrees of Freedom: 650 Total (i.e. Null);  646 Residual
Null Deviance:      1349 
Residual Deviance: 1245         AIC: NA

Call:
glm(formula = mydata$STRESS ~ mydata$COHES + mydata$ESTEEM + 
    mydata$GRADES + mydata$SATTACH, family = quasipoisson)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7106  -1.5982  -0.2915   0.7107   3.6423  

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.734460   0.312148   8.760  < 2e-16 ***
mydata$COHES   -0.012917   0.003858  -3.348  0.00086 ***
mydata$ESTEEM  -0.023691   0.010720  -2.210  0.02746 *  
mydata$GRADES  -0.023470   0.013156  -1.784  0.07490 .  
mydata$SATTACH -0.016480   0.007712  -2.137  0.03297 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.778496)

    Null deviance: 1349.3  on 650  degrees of freedom
Residual deviance: 1245.0  on 646  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

The over-dispersed Poisson regression model seems to be very similar to the original Poisson model.

5. Based on the Poisson model in part 4), compute the predicted count of STRESS for those whose levels of family cohesion are less than one standard deviation below the mean (call this the low group), between one standard deviation below and one standard deviation above the mean (call this the middle group), and more than one standard deviation above the mean (high).   What is the expected percent difference in the number of stressful events for those at high and low levels of family cohesion?



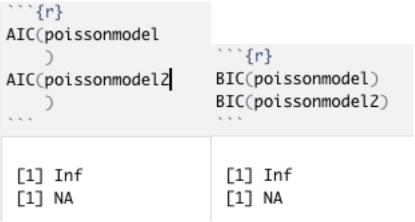
~~~
```{r}
mydata$Group=ifelse(mydata$COHES<(mean(mydata$COHES)-sd(mydata$COHES)), 1, ifelse(mydata$COHES>mean(mydata$COHES)+sd(mydata$COHES),3,2))

length(which(mydata$Group==1))
length(which(mydata$Group==2))
length(which(mydata$Group==3))
```

 [1] 106
 [1] 446
 [1] 99
~~~



~~~
```{r}
low=exp(2.73446 - 0.01292*41.62096 - 0.02369*mean(mydata$ESTEEM) - 0.02347*mean(mydata$GRADES) - 0.01648*mean(mydata$SATTACH))
high=exp(2.73446 - 0.01292*64.38757 - 0.02369*mean(mydata$ESTEEM) - 0.02347*mean(mydata$GRADES) - 0.01648*mean(mydata$SATTACH))
low
high
```

 [1] 1.914669
 [1] 1.426751
~~~

There is about a 25.485% increase in low vs. high stress counts.

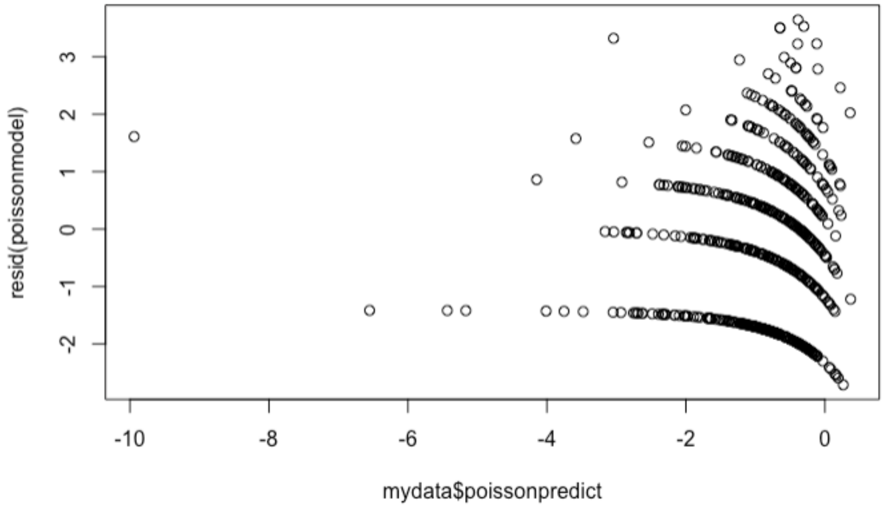
6. Compute the AICs and BICs from the Poisson Regression and the over-dispersed Poisson regression models from part 4). Is one better than the other?



~~~
```{r}
AIC(poissonmodel
   )
AIC(poissonmodel2
   )
```

[1] Inf
[1] NA
~~~

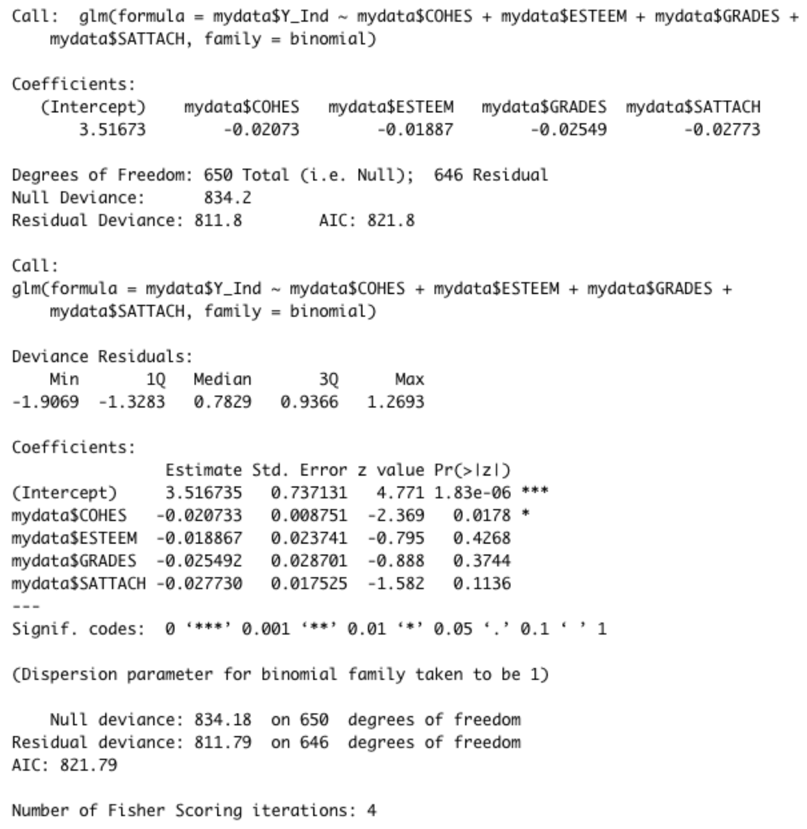
~~~
```{r}
BIC(poissonmodel)
BIC(poissonmodel2)
```

[1] Inf
[1] NA
~~~

7. Using the Poisson regression model from part 4), plot the deviance residuals by the predicted values. Discuss what this plot indicates about the regression model.



This plot still does not indicate a good fitting model, due to the excessive residuals not lining up with the predictions.

8. Create a new indicator variable (Y_IND) of STRESS that takes on a value of 0 if STRESS=0 and 1 if STRESS>0. This variable essentially measures is stress present, yes or no. Fit a logistic regression model to predict Y_IND using the variables using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Report the model, interpret the coefficients, obtain statistical information on goodness of fit, and discuss how well this model fits. Should you rerun the logistic regression analysis? If so, what should you do next?



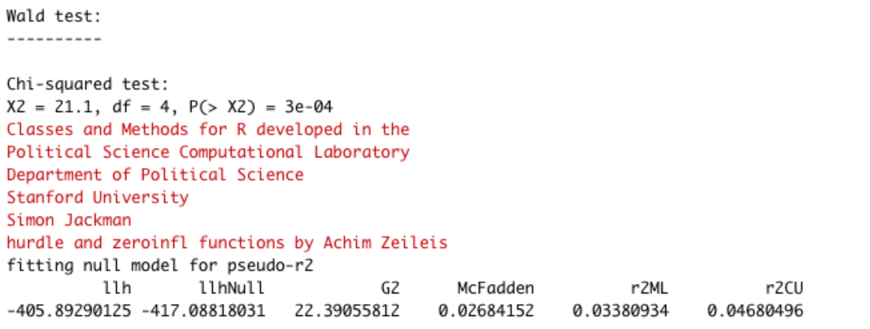
```
Call:  glm(formula = mydata$Y_Ind ~ mydata$COHES + mydata$ESTEEM + mydata$GRADES +
    mydata$SATTACH, family = binomial)

Coefficients:
   (Intercept)    mydata$COHES   mydata$ESTEEM   mydata$GRADES  mydata$SATTACH
       3.51673        -0.02073        -0.01887        -0.02549        -0.02773

Degrees of Freedom: 650 Total (i.e. Null);  646 Residual
Null Deviance:      834.2
Residual Deviance: 811.8        AIC: 821.8

Call:
glm(formula = mydata$Y_Ind ~ mydata$COHES + mydata$ESTEEM + mydata$GRADES +
    mydata$SATTACH, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9069  -1.3283   0.7829   0.9366   1.2693

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.516735   0.737131   4.771 1.83e-06 ***
mydata$COHES   -0.020733   0.008751  -2.369   0.0178 *
mydata$ESTEEM  -0.018867   0.023741  -0.795   0.4268
mydata$GRADES  -0.025492   0.028701  -0.888   0.3744
mydata$SATTACH -0.027730   0.017525  -1.582   0.1136
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 834.18  on 650  degrees of freedom
Residual deviance: 811.79  on 646  degrees of freedom
AIC: 821.79

Number of Fisher Scoring iterations: 4
```

Y= 3.51673-0.02073*B1-0.01887*B2-0.02549*B3-0.02773*B4

3.51673 is the intercept, - 0.02073 is the coefficient for COHES, -0.01887 is the coefficient for ESTEEM,

-0.02549 is the coefficient for GRADES, and -0.02773 is the coefficient for SATTACH.

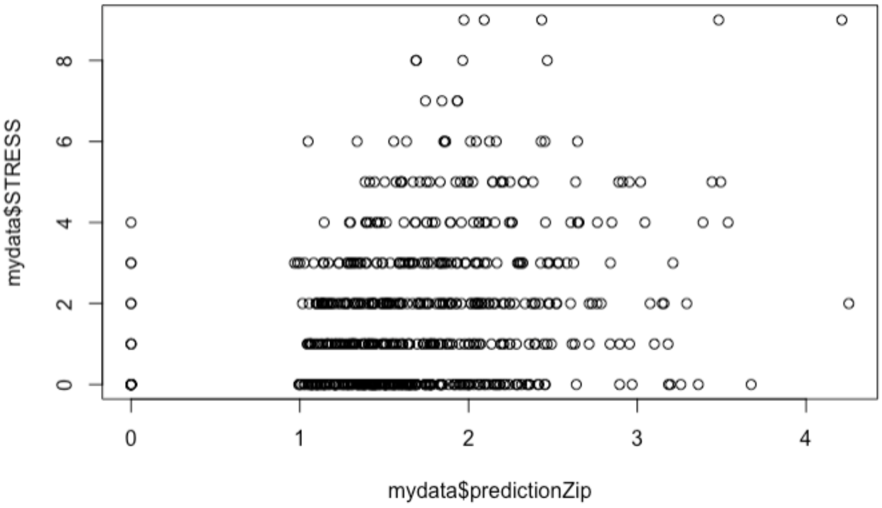
To find the effect of each coefficient on the variable, we must use e^(coefficient) to find the true effect in a Poisson distribution.

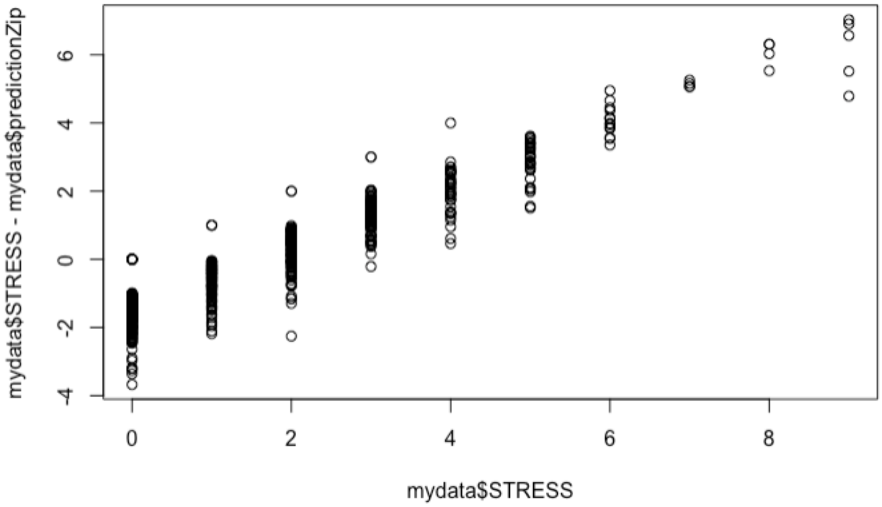


```
Wald test:
----------

Chi-squared test:
X2 = 21.1, df = 4, P(> X2) = 3e-04
Classes and Methods for R developed in the
Political Science Computational Laboratory
Department of Political Science
Stanford University
Simon Jackman
hurdle and zeroinfl functions by Achim Zeileis
fitting null model for pseudo-r2
          llh       llhNull            G2      McFadden          r2ML          r2CU
-405.89290125 -417.08818031   22.39055812    0.02684152    0.03380934    0.04680496
```

The pseudo r^2 still indicates poor fitment, and the p-value is extremely small in the Wald test. The AIC does seem relatively low however compared to previously.

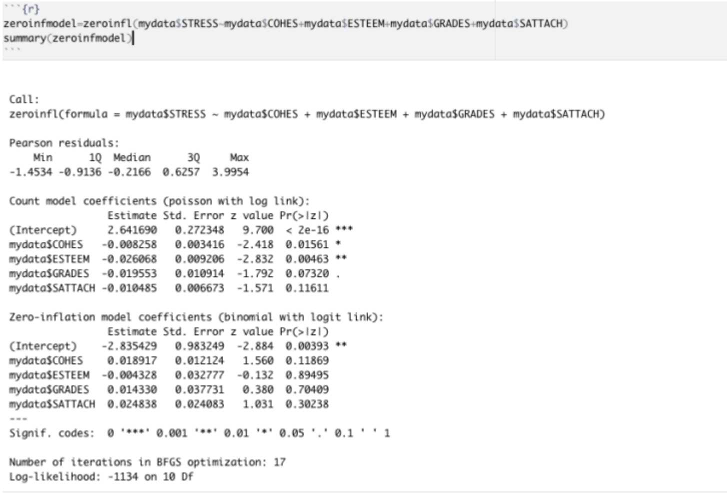
9. It may be that there are two (or more) process at work that are overlapped and generating the distributions of STRESS(Y). What do you think those processes might be? To conduct a ZIP regression model by hand, fit a Logistic Regression model to predict if stress is present (Y_IND), and then use a Poisson Regression model to predict the number of stressful events (STRESS) conditioning on stress being present. Is it reasonable to use such a model? Combine the two fitted model to predict STRESS (Y). Obtained predicted values and residuals. How well does this model fit? HINT: You have to be thoughtful about this. It is not as straight forward as plug and chug!





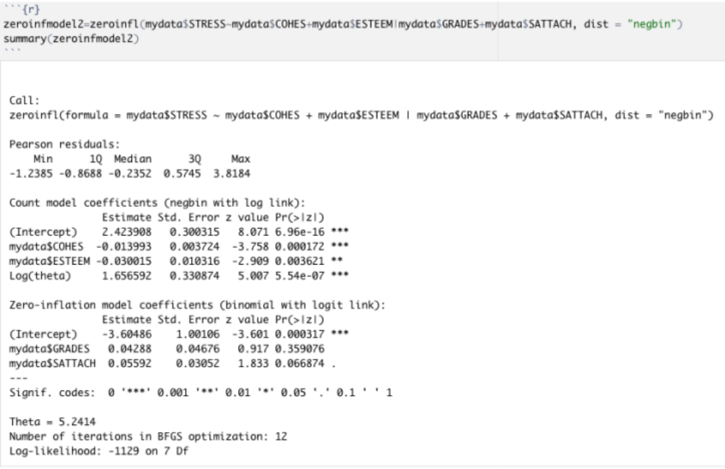
The model does seem to have improved based on this plot, as a relationship now seems to be visible. The residuals are higher early on.

10. Use the pscl package and the zeroinfl() function to Fit a ZIP model to predict STRESS(Y). You should do this twice, first using the same predictor variable for both parts of the ZIP model. Second, finding the best fitting model. Report the results and goodness of fit measures. Synthesize your findings across all of these models, to reflect on what you think would be a good modeling approach for this data.



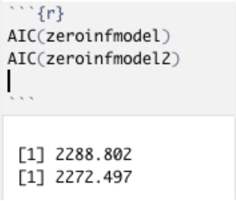
~~~
```{r}
zeroinfmodel=zeroinfl(mydata$STRESS~mydata$COHES+mydata$ESTEEM+mydata$GRADES+mydata$SATTACH)
summary(zeroinfmodel)
```
~~~

```
Call:
zeroinfl(formula = mydata$STRESS ~ mydata$COHES + mydata$ESTEEM + mydata$GRADES + mydata$SATTACH)

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.4534 -0.9136 -0.2166  0.6257  3.9954 

Count model coefficients (poisson with log link):
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)     2.641690   0.272348   9.700  < 2e-16 ***
mydata$COHES   -0.008258   0.003416  -2.418  0.01561 *  
mydata$ESTEEM  -0.026068   0.009206  -2.832  0.00463 ** 
mydata$GRADES  -0.019553   0.010914  -1.792  0.07320 .  
mydata$SATTACH -0.010485   0.006673  -1.571  0.11611    

Zero-inflation model coefficients (binomial with logit link):
                Estimate Std. Error z value Pr(>|z|)   
(Intercept)    -2.835429   0.983249  -2.884  0.00393 **
mydata$COHES    0.018917   0.012124   1.560  0.11869   
mydata$ESTEEM  -0.004328   0.032777  -0.132  0.89495   
mydata$GRADES   0.014330   0.037731   0.380  0.70409   
mydata$SATTACH  0.024838   0.024083   1.031  0.30238   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Number of iterations in BFGS optimization: 17 
Log-likelihood: -1134 on 10 Df
```



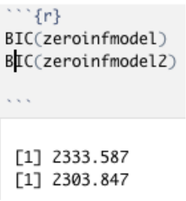
~~~
```{r}
zeroinfmodel2=zeroinfl(mydata$STRESS~mydata$COHES+mydata$ESTEEM|mydata$GRADES+mydata$SATTACH, dist = "negbin")
summary(zeroinfmodel2)
```
~~~

```
Call:
zeroinfl(formula = mydata$STRESS ~ mydata$COHES + mydata$ESTEEM | mydata$GRADES + mydata$SATTACH, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max 
-1.2385 -0.8688 -0.2352  0.5745  3.8184 

Count model coefficients (negbin with log link):
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)    2.423908   0.300315   8.071 6.96e-16 ***
mydata$COHES  -0.013993   0.003724  -3.758 0.000172 ***
mydata$ESTEEM -0.030015   0.010316  -2.909 0.003621 ** 
Log(theta)     1.656592   0.330874   5.007 5.54e-07 ***

Zero-inflation model coefficients (binomial with logit link):
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -3.60486    1.00106  -3.601 0.000317 ***
mydata$GRADES   0.04288    0.04676   0.917 0.359076    
mydata$SATTACH  0.05592    0.03052   1.833 0.066874 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Theta = 5.2414 
Number of iterations in BFGS optimization: 12 
Log-likelihood: -1129 on 7 Df
```



~~~
```{r}
AIC(zeroinfmodel)
AIC(zeroinfmodel2)

```
~~~

```
[1] 2288.802
[1] 2272.497
```



~~~
```{r}
BIC(zeroinfmodel)
BIC(zeroinfmodel2)

```
~~~

```
[1] 2333.587
[1] 2303.847
```

The better fitting model is run using the “negbin” distribution and having the COHES and ESTEEM variables run as count model coefficients and the GRADES and SATTACH variables run as zero-inflation model coefficients. This improves the AIC value, lowering it to 2272.497 from 2288.802. The BIC also lowers from 2333.587 to 2303.847.

11. Conclusion.

This assignment has added more ways to conduct models to what we have learned, including the Poisson distribution and zero-inf. This helps add more ways to conduct better fitting models depending on the data on hand, especially if it is not normally distributed initially.