Scott Jue
MSDS 410
March 5, 2023

Modeling Assignment 9: Poisson and ZIP Regression Models

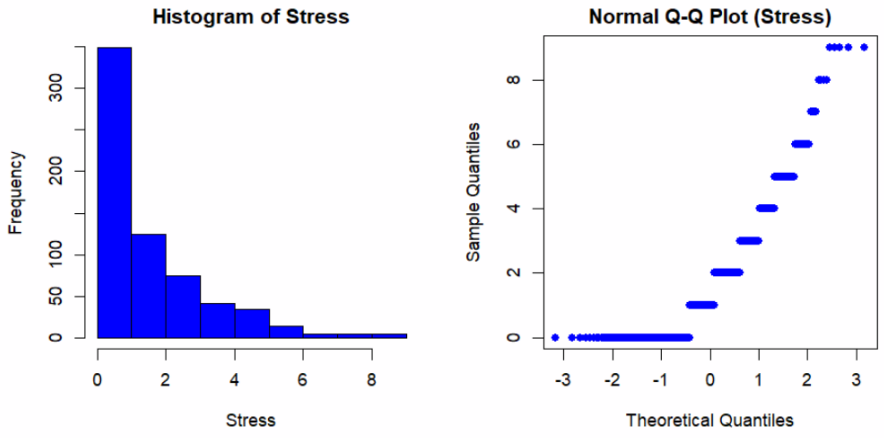
Introduction

In this analysis I have fit several types of regression models using survey data that contains information from 651 adolescents in the United States. The variables provided in the data set describe the the number of stressful life events they had experienced in the past year (STRESS), as well as other school and family related variables that are assumed to be continuously distributed such as measure of how well the adolescent gets along with their family (COHES), measure of self-esteem (ESTEEM), prior year's grades (GRADES), and measure of how the adolescent feels about their school (SATTACH). The STRESS data in this dataset has a distribution that is best suited for Poisson and Zero-Inflated Poisson Regression models. STRESS is an integer variable that represents counts of stressful events. The objective of this analysis is to determine the best fitting model and the most important variables for predicting the number of stressful events (STRESS) in adolescents based on the information from the explanatory variables COHES, ESTEEM, GRADES, and SATTACH. The sample population for this analysis is defined as adolescents living in the United States who attend a school.

Tasks

1. From the below is a table of summary statistics and histogram of STRESS. We can see that STRESS is not normally distributed. There is a sufficiently large amount of zero values. The most likely a probability distribution for STRESS is a Poisson probability distribution or a negative binomial distribution. This is because when we remove the zero values from the STRESS variables the mean and the variance are very close in value

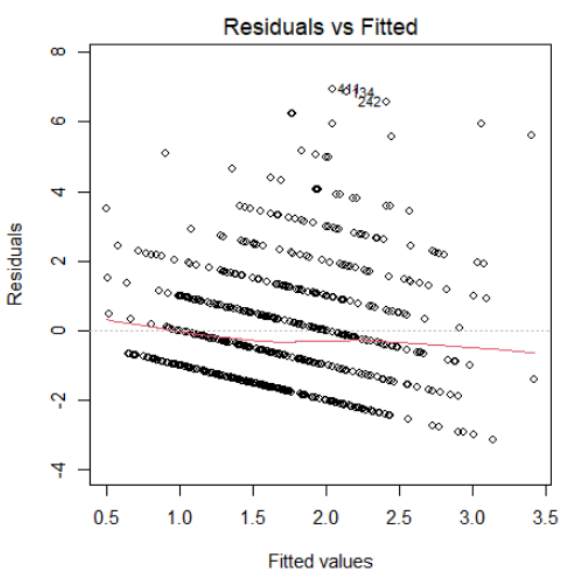
	Mean	Variance
STRESS with zero values	1.73	3.419
STRESS without zero values	2.619	2.847

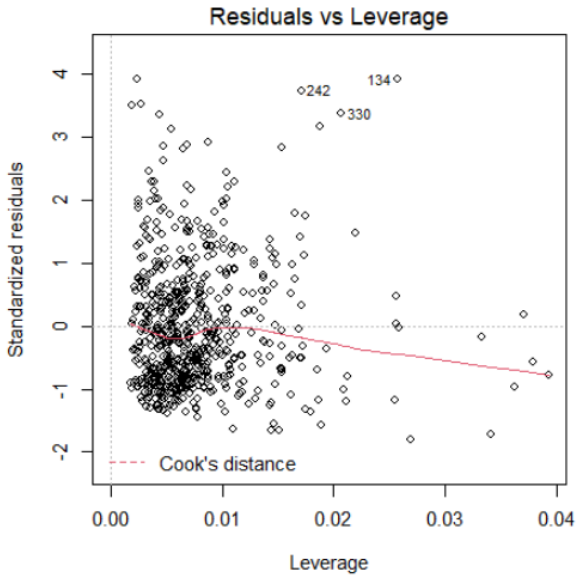


2. First, I fit an OLS regression model to predict STRESS (Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). We can see in the below summary table that the model does a poor job of explaining the variance in STRESS as the R-squared is only 0.08319 or 8.3%. Some possible issues with this model are that STRESS is not normally distributed and the relationships between explanatory variables and STRESS may not be linear. Another issue is that this model does not produce feasible predictions as the values of STRESS are integers and this model has decimal values for the intercept and also the coefficients.

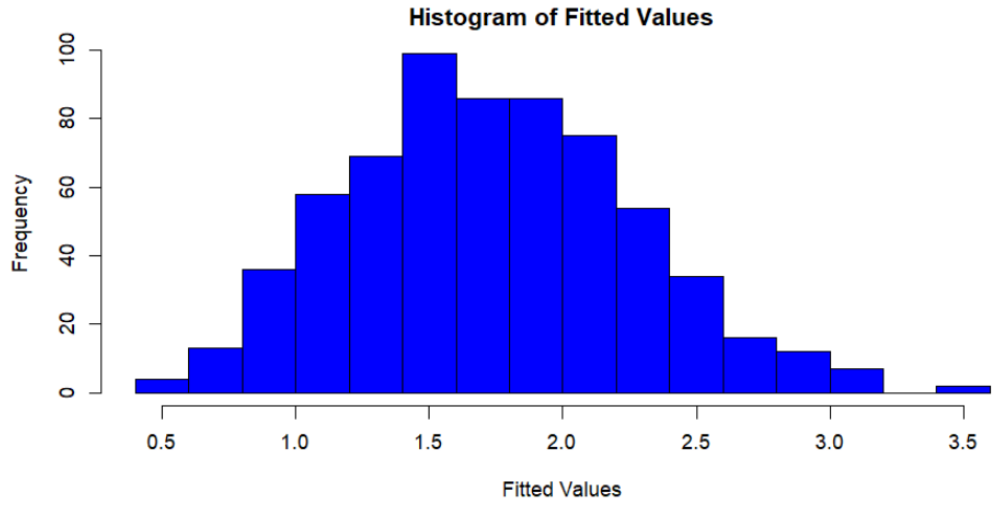
```
Call:
lm(formula = STRESS ~ COHES + ESTEEM + GRADES + SATTACH, data = df)

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 < 0.0000000000000002 ***
COHES       -0.02319    0.00703  -3.298              0.00103 **
ESTEEM      -0.04129    0.01933  -2.136              0.03305 *
GRADES      -0.04170    0.02352  -1.773              0.07670 .
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: 0.00000000001826


 Analysis of Variance Table

 Response: STRESS
           Df  Sum Sq Mean Sq F value          Pr(>F)
 COHES      1  122.93 122.930 38.9749 0.0000000007777 ***
 ESTEEM     1   31.26  31.264  9.9122        0.001718 **
 GRADES     1   16.05  16.052  5.0894        0.024407 *
 SATTACH    1   14.64  14.635  4.6401        0.031602 *
 Residuals 646 2037.54   3.154
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The below scatterplot of the residuals and the fitted values show that there is a heteroscedasticity pattern, so we have violated the assumption of homoscedasticity. Additionally, the below residuals vs leverage plot also shows that there are influential observations in the dataset that have a high leverage. These could be potential outliers in the dataset and would need to be investigated further before determining if they should be removed from the model.





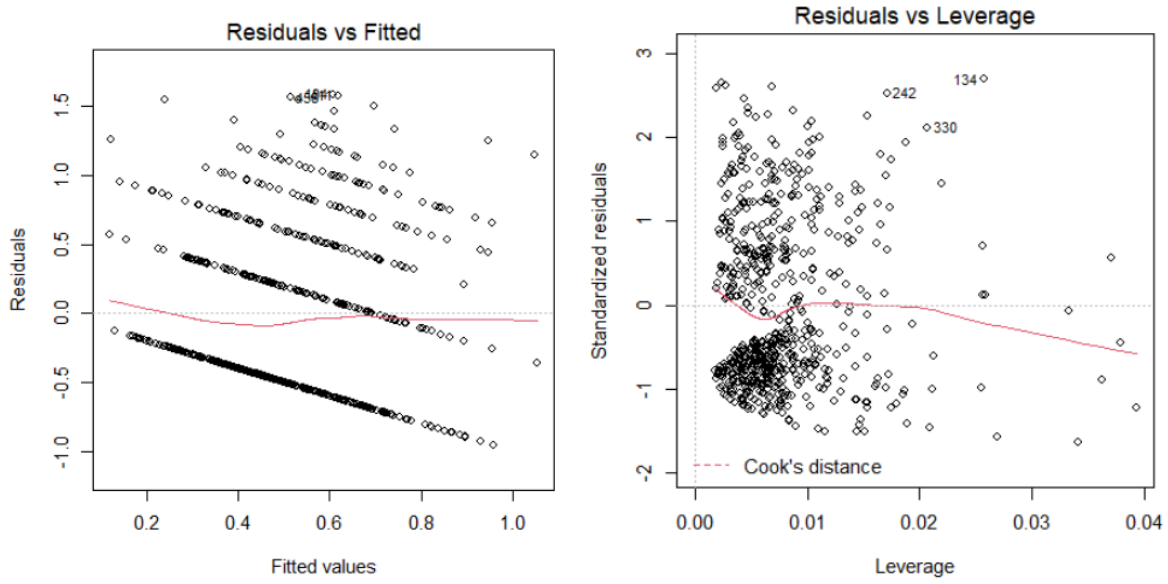
Below is a histogram of the fitted values of the linear regression model. We can see that the distribution is normally distributed is centered around 1.5 – 2. However, the max value is 3.5 while actual values in the go as high as 9. Also, this histogram does not resemble the histogram of actual STRESS counts from part 1, since there are very few values predicted near or at 0. Therefore, we can conclude that this model is not a good fit for this type of data.



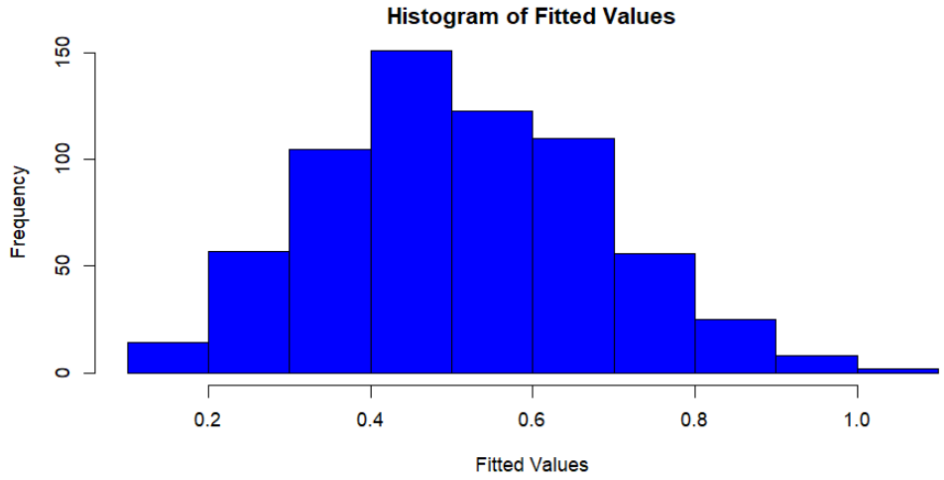
3. Next, I transformed the STRESS variable by performing a log transformation on any STRESS value that was greater than 0, since the value of log (0) is undefined. And then refit a linear regression modeling using this variable as the response variable. Below is the model summary table and ANOVA table.

```
Call:
lm(formula = logSTRESS ~ COHES + ESTEEM + GRADES + SATTACH, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9582 -0.4711 -0.2534  0.4451  1.5908

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  1.768804   0.195173   9.063 < 0.0000000000000002 ***
COHES       -0.007171   0.002361  -3.037              0.00248 **
ESTEEM      -0.010845   0.006492  -1.671              0.09530 .
GRADES      -0.015826   0.007899  -2.004              0.04553 *
SATTACH     -0.011027   0.004743  -2.325              0.02038 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5964 on 646 degrees of freedom
Multiple R-squared:  0.07667,   Adjusted R-squared:  0.07095
F-statistic: 13.41 on 4 and 646 DF,  p-value: 0.0000000001666

Analysis of Variance Table

Response: logSTRESS
           Df  Sum Sq Mean Sq F value         Pr(>F)
COHES       1  12.188 12.1881 34.2644 0.000000007655 ***
ESTEEM      1   2.698  2.6983  7.5856       0.006049 **
GRADES      1   2.271  2.2710  6.3844       0.011750 *
SATTACH     1   1.923  1.9227  5.4053       0.020384 *
Residuals 646 229.787  0.3557
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model performce was similar to the previous model without the log transformation. The R-squared value is relatively the same at 0.077. As we can see, the log transformation of STRESS didn't have any improvement on the model's fit, so this did not correct the issue. Below is a scatterplot of the residuals and the fitted values which show that there is a heteroscedasticity pattern, so we have violated the assumption of homoscedasticity. Additionally, the below residuals Vs leverage plot also shows that there are influential observations in the dataset that have a high leverage. These could be potential outliers in the dataset and would need to be investigated further before determining if they should be removed from the model.



Below is a histogram of the fitted values of the linear regression model. We can see that the distribution is normally distributed. We can see that this histogram does not resemble the histogram of actual STRESS counts from part 1. Also because of the log transformation it is difficult to interpret these fitted values to compare against the actual values for STRESS. Furthermore, because log(0) is undefined this presents another issue with this model as not of the predications would equate to a 0 value for STRESS. Therefore, we can conclude that this model is not a good fit for this type of data.

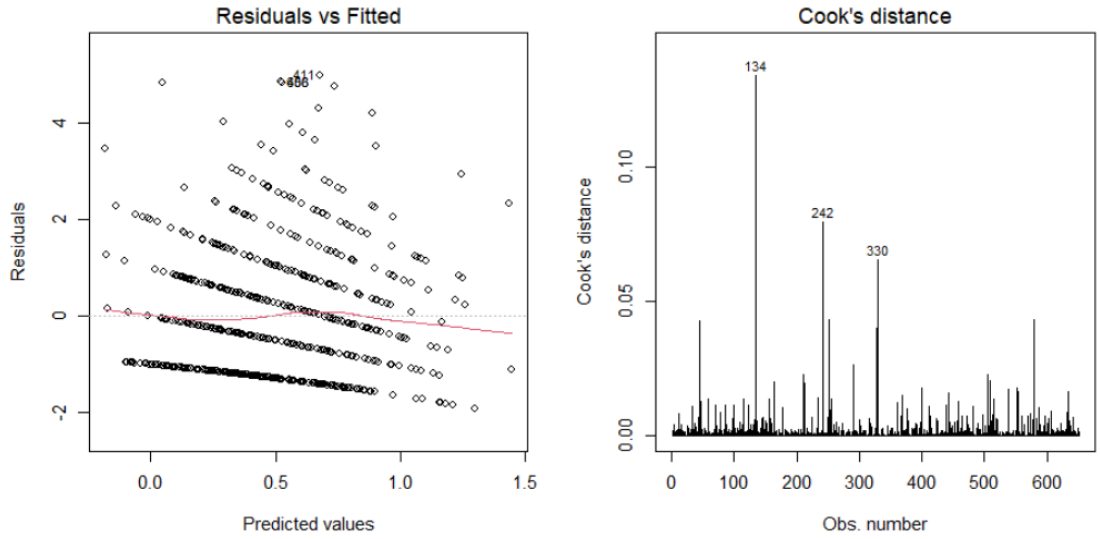


4. Next, using the glm() function in R I fit a Poisson Regression model to predict STRESS using the explanatory variables COHES, ESTEEM, GRADES, and SATTACH. Below is the model summary table for the Poisson model.

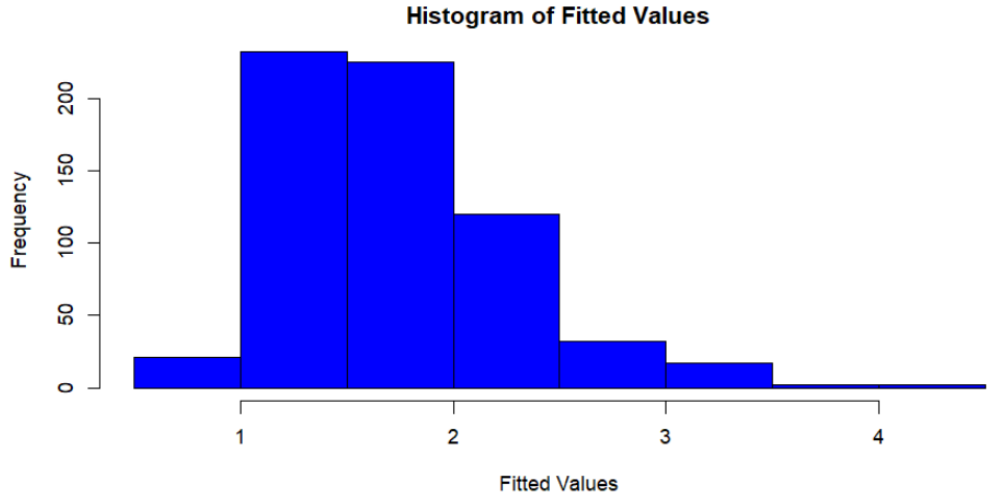
```
Call:
glm(formula = STRESS ~ COHES + ESTEEM + GRADES + SATTACH, family = "poisson",
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7111  -1.5989  -0.2914   0.7107   3.6424

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept)  2.734513   0.234066  11.683 < 0.0000000000000002 ***
COHES       -0.012918   0.002893  -4.466          0.00000798 ***
ESTEEM      -0.023692   0.008039  -2.947              0.00321 **
GRADES      -0.023471   0.009865  -2.379              0.01735 *
SATTACH     -0.016481   0.005783  -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: 1349.8  on 650  degrees of freedom
Residual deviance: 1245.4  on 646  degrees of freedom
AIC: 2417.2

Number of Fisher Scoring iterations: 5
```

For interpretation of the model coefficients, we can see that when all the variables are equal to 0, then the log of STRESS is equal to the intercept value of 2.735. For every 1 unit increase in COHES, the log of STRESS decreases by 0.0129, when all other variables held constant. For every 1

unit increase in ESTEEM, the log of STRESS decreases by 0.0237, when all other variables held constant. For every 1 unit increase in GRADES, the log of STRESS decreases by 0.0235, when all other variables held constant. And lastly, for every 1 unit increase in SATTACH, the log of STRESS decreases by 0.0165, when all other variables held constant. These values are relatively similar to the coefficients from the model in part 3. We can also see in the below plots that there is a heteroscedasticity pattern, so we have violated the assumption of homoscedasticity. Additionally, the below residuals vs leverage plot also shows that there are influential observations in the dataset that have a high leverage. These plots are similar to the plots seen in the log transformation linear regression model from part 3.



We can see in the below histogram that the distribution is right-skewed which is closer to the distribution of the actual STRESS values; however, there still are very few values predicted near 0 even though zero values were the most frequent value in the dataset. Therefore, we can conclude that this model is not a good fit for this type of data.

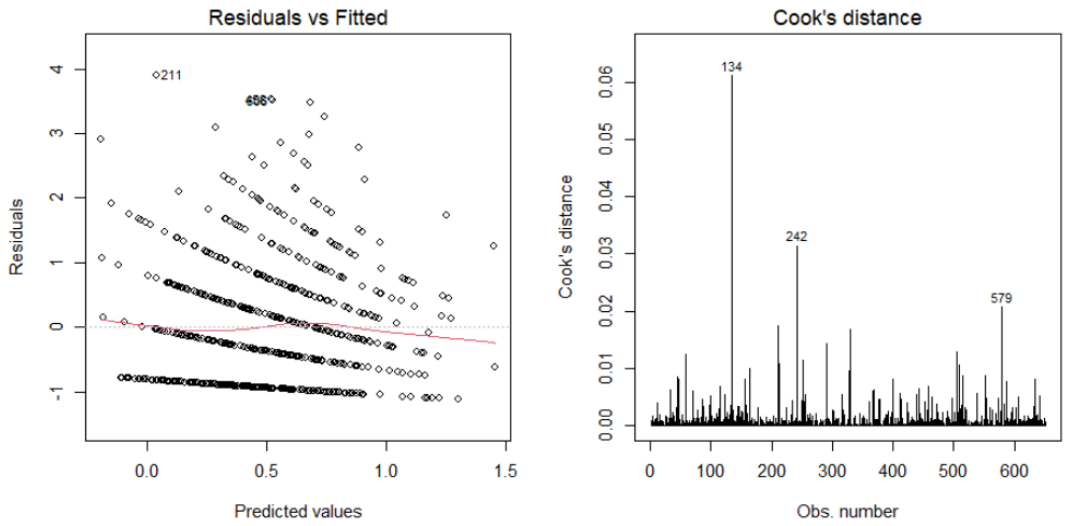


Next, I fit an over-dispersed Poisson regression model using the same set of variables and compared this model to the Poisson model above. Below is the summary of the model coefficients.

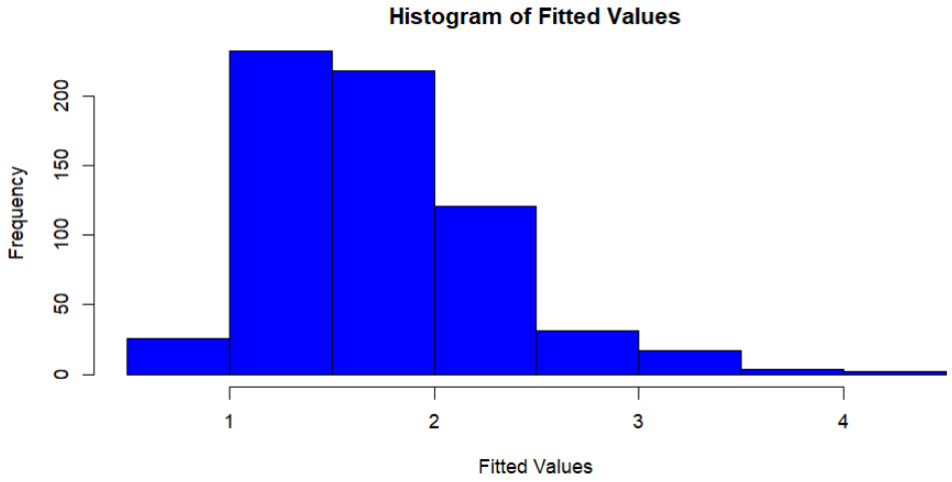
```
Call:
glm.nb(formula = STRESS ~ COHES + ESTEEM + GRADES + SATTACH,
    data = df, init.theta = 1.865329467, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0179  -1.3900  -0.2214   0.4882   2.3199

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept)  2.759032   0.341531   8.078 0.000000000000000656 ***
COHES       -0.013391   0.004136  -3.238              0.00121 **
ESTEEM      -0.023058   0.011477  -2.009              0.04453 *
GRADES      -0.024360   0.013969  -1.744              0.08118 .
SATTACH     -0.016750   0.008296  -2.019              0.04349 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.8653) family taken to be 1)

    Null deviance: 792.47  on 650  degrees of freedom
Residual deviance: 738.53  on 646  degrees of freedom
AIC: 2283.6

Number of Fisher Scoring iterations: 1


              Theta:  1.865
          Std. Err.:  0.257

 2 x log-likelihood:  -2271.590
```

We can see from the above table that the coefficients are similar to the coefficients from the previous Poisson model; however, the over-dispersed model shows a slightly improved AIC of 2,283.6 versus the AIC of 2,417.2 from the previous model. So, the over-dispersed model performs slightly better.

Below we can see that the plot of the residuals versus predicted values and the plot of influential observations is similar to the plots seen in the original Poisson model. The residuals show a heteroscedasticity pattern and there are a few influential observations with high leverage.



The below histogram of the fitted values is nearly identical to the histogram from the previous Poisson model. It is a right-skewed distribution with very little values predicted near 0 even though zero values were the most frequent value in the dataset. Therefore, we can conclude that this model is not a good fit for this type of data.

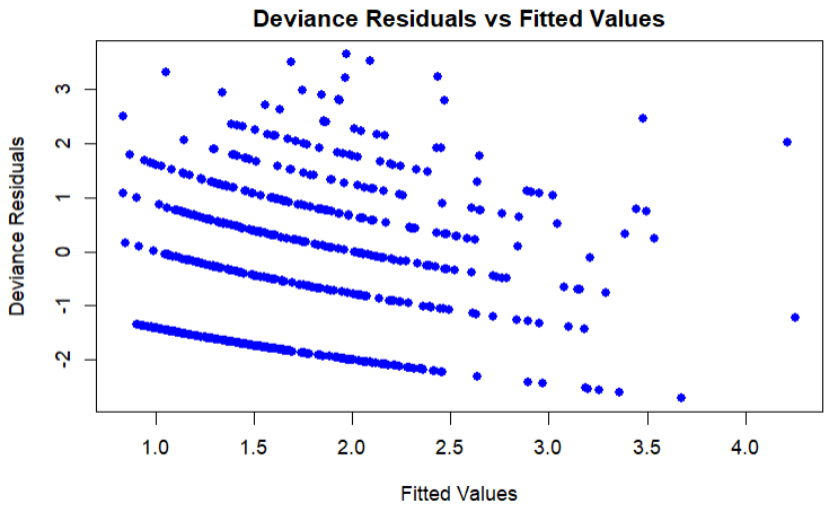


5. Next, I created a categorical variable for COHES called COHES_cat which assigns an observation to a "low" group if the level of family cohesion is less than one standard deviation below the mean, a "middle" group if the level is between one standard deviation below and one standard deviation above the mean, and a "high" group if the level is more than one standard deviation above the mean. Then, I created a table with the mean predicted STRESS counts for each of these groups which can be seen below. From this table we can see that the low family cohesion group has approximately 52% more predicted stressful events than the middle group and that the high family cohesion group has approximately 29% less predicted stressful events compared to the middle group. As such, COHES could be an important determinant in predicting the number of stressful events.

COHES_cat	Mean
Low	2.525
Middle	1.665
High	1.178

6. The AIC from the Poisson Model is 2,417.2 and the AIC from the Negative Binomial Model is 2,283.6. The BIC from the Poisson model is 2,439.6 and the BIC from the Negative Binomial model is 2,310.5. The values are lower for both AIC and BIC for the over-dispersed model. This indicates that the over-dispersed model (negative binomial model) is a better fit when comparing the 2 models.

7. Next, I plotted the deviance residuals by the predicted values which can be seen below. We can see that this plot shows a heteroscedasticity pattern in the residuals. Much of the variance in the residuals occur towards the lower end of the predicted values. Thus, the Poisson model from part 4 violates the assumption of homoscedasticity.



8. Next, I created 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. Then, I fitted a logistic regression model to predict Y_IND using the variables using COHES, ESTEEM, GRADES, SATTACH as explanatory variables. Below is the coefficient summary table for this model.

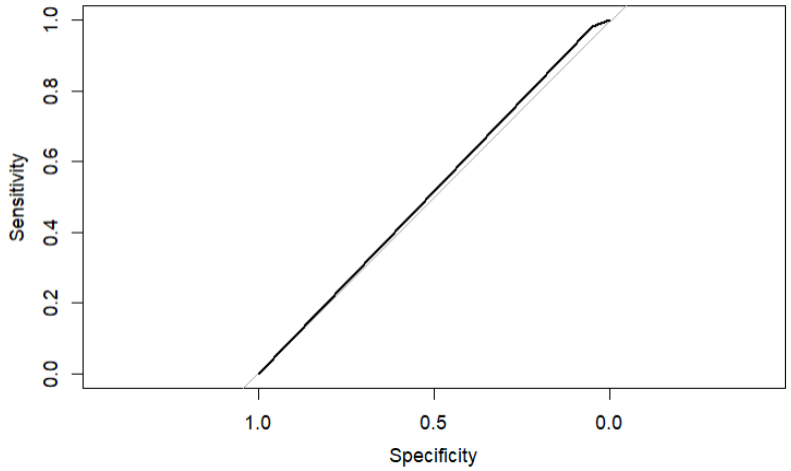
```
Call:
glm(formula = Y_IND ~ COHES + ESTEEM + GRADES + SATTACH, family = "binomial",
    data = df)

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 0.00000183 ***
COHES       -0.020733   0.008751  -2.369     0.0178 *
ESTEEM      -0.018867   0.023741  -0.795     0.4268
GRADES      -0.025492   0.028701  -0.888     0.3744
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
```

Using exponentiation, I converted the coefficients to changes in the odds of having a stressful event for better interpretation. Taking the exponent of the intercept we get an odds ratio equal to 33.67 if all variables are equal to 0. The coefficient for COHES can be interpreted as the odds decrease by 0.979 for every 1 unit increase in CHOES, if everything else is held constant. The coefficient for ESTEEM can be interpreted as the odds decrease by 0.981 for every 1 unit increase in ESTEEM, if everything else is held constant. The coefficient for GRADES can be interpreted as the odds will decrease by 0.975 for every 1 unite increase in GRADES, if everything else is held constant. Lastly, the coefficient for SATTACH can be interpreted as for every 1 unit increase in SATTACH the

odds of having a stressful event decrease by 0.973, if everything else is held constant. Below we can see a confusion matrix for the model.

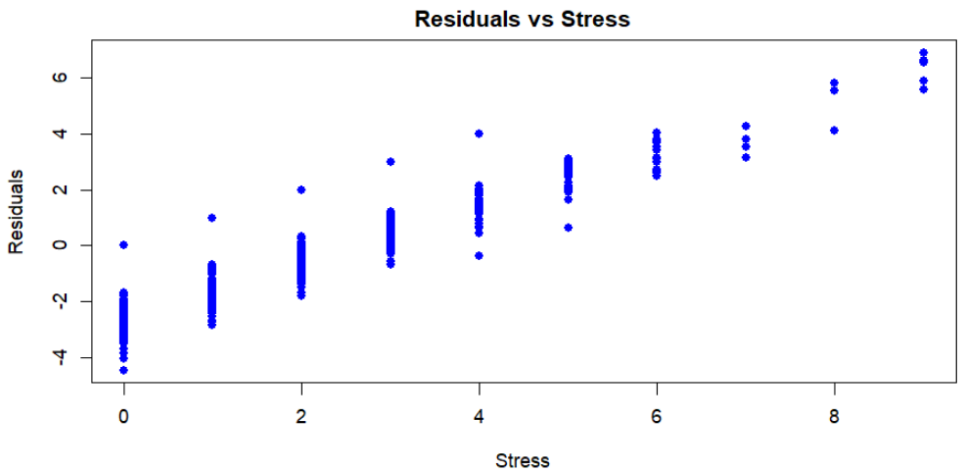
	Predict	
STRESS	0	1
0	11	210
1	8	422

From this table we can see that the model predicted 433 observations correctly out of 651 which equates to approximately 66.5% accuracy. Additionally, this model also has an AUC of 0.516 and has the below ROC curve.

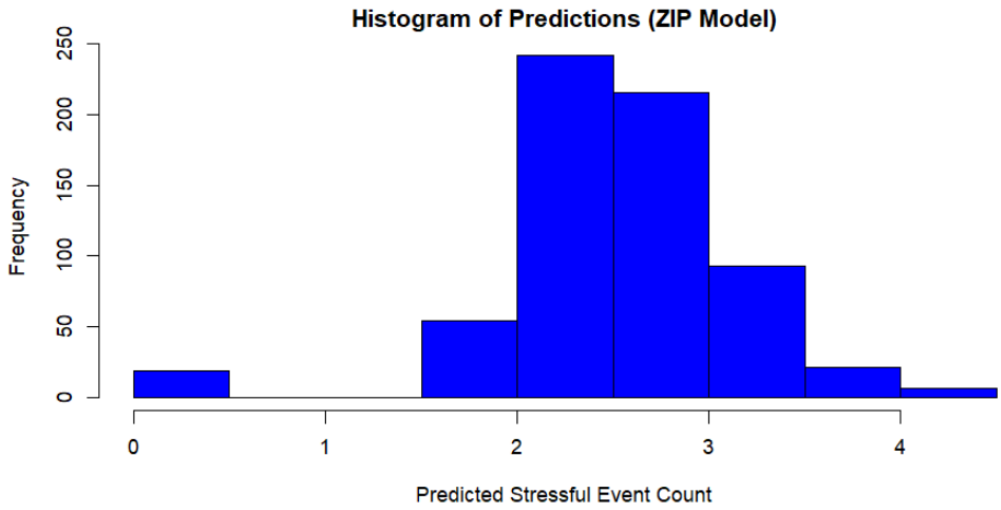


Based on these metrics and the ROC curve, this model is not a good fit. Additionally, this model is only able to predict if the person had a stressful event or not and doesn't predict the count of stressful events which is the objective of this analysis. As such, there is no value in re-running the logistic regression model. Therefore, I will explore another modeling approach that predicts if there is a stressful event or not and if there is can also predict the number of stressful events the adolescent has.

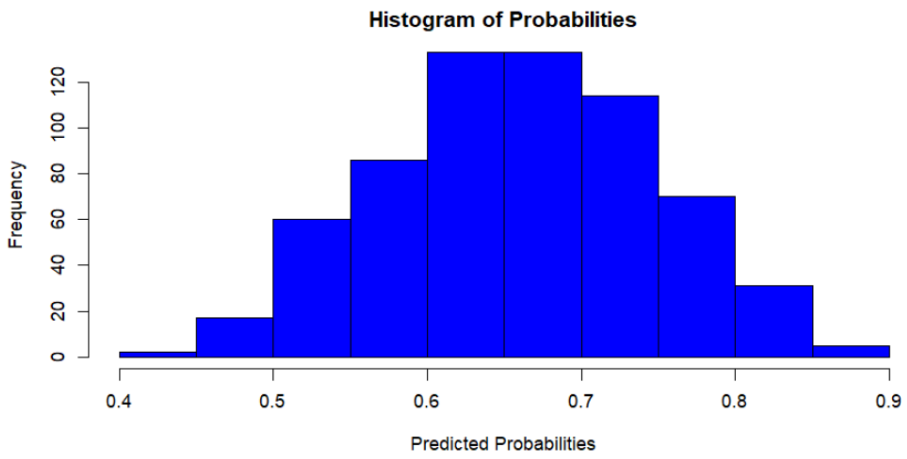
9. For this dataset there appears to be overlapping processes that are generating the distribution of STRESS. First, there is a distribution of the observations that do not have any stressful events and the observations that do have stressful events. Then, there is another distribution that reflects the count of stressful events for the observations have stressful events. As such, I will combine a logistic regression model that predicts if stress is present (Y_IND) with a Poisson regression model that predicts the number of stressful events with a condition on stress being present. With this Zip regression model method, we can then assign a 0 for the observations that are predicted to have no stress using the logistic regression. If the logistic regression predicts stress being present, then we can use the Poisson model to predict the number of stressful events to generate our predictions for the data. Using this approach, I obtained predictions for the number of stressful events which resulted in the below residuals plot.



Compared to previous models, the residuals appear to resemble a homoscedasticity pattern that is also linear, so this assumption has been met by this model. Next, I looked at the histogram of the predictions.



We can see that this distribution is a better approximation to the distribution of the actual stressful events counts as compared to the previous models in this analysis. We can see that for the observations with stress that the distribution is right-skewed which resembles the actual stressful events distribution. Additionally, there are predictions with 0 stressful events which is something that was not present in histograms from the other models. However, the number of zeros predicted is very low compared to what was seen in the original data. This is likely due to the fact the logistic regression model did not perform well with only a 66.5% accuracy. So, while this modeling approach is appropriate, it does a poor job of predicting if stress is present or not. We can see that in the below histogram for the predicted probabilities of the logistic regression model that the majority of the probabilities are above 0.5. So, it appears that logistic model has a difficult time classifying the “no stress” values (0) and classifies that majority of observations as having stress, despite “no stress” (0) being the most frequent value in the original dataset.



10. Next, I will use the pscl package and the zeroinfl() function to fit a ZIP model to predict STRESS(Y). First, I will create a simplified model that uses the same predictor variable, COHES, for both parts of the ZIP model. Below is the coefficient summary table for this model.

```
Call:
zeroinfl(formula = STRESS ~ COHES | COHES, data = df)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.4954 -0.9196 -0.2587  0.5993  4.1250

Count model coefficients (poisson with log link):
             Estimate Std. Error z value             Pr(>|z|)
(Intercept)  1.642574   0.147053  11.170 < 0.0000000000000002 ***
COHES       -0.015427   0.002915  -5.293          0.00000012 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) -2.30569    0.54597  -4.223 0.0000241 ***
COHES        0.02371    0.01005   2.359    0.0183 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 9
Log-likelihood: -1147 on 4 Df
```

Exponentiated Coefficients:

	Count_model	Zero_inflation_model
Intercept	5.168	0.100
COHES	0.985	1.024

By obtaining the exponentiated values of the coefficients in the above table, this zero-inflation portion of the model can be interpreted as the baseline odds of having a stressful event is 0.1. A one unit increase in COHES increases the odds by 1.024. This seems a bit odd as I would have expected COHES to decrease the likelihood of having a stressful event. For the count model, the baseline number of stressful events is 5.168 among those adolescents who have a stressful event. A one unit increase in

COHES decreases this baseline amount by 0.985. Logically, this count model makes sense that increased COHES would lead to less stressful events.

Next, to find the best fitting model I expanded my explanatory variables to use all the variables and removed them one-by-one and compared the models using a chi-square test. For the zero-inflation portion of the model I have chosen to keep all the variables for this part of the model. This is because even with using all the information provided by these variables, the model has a difficult time predicting 0 values. So, removing any variables from zero-inflation model would only decrease the model's performance for predicting if stress is present or not. As such, I only analyzed the removable of variables for the count model portion of the zip regression model.

For reference, the critical chi-square value I will be using for the hypothesis tests of a null hypothesis that $\beta_i = 0$ and an alternative hypothesis that $\beta_i \neq 0$ is 3.841. First, I removed the SATTACH variable due to having a higher p-value. This resulted in a chi-square value of 2.45 meaning that we can fail to reject the null hypothesis can concluded that the SATTACH variable is not contributing any meaningful information to the model any can be removed. Additionally, removing SATTACH only increased the AIC by 0.46 so it is not an important predictor for the model

Next, I removed GRADES from the count model and refit the model. This resulted in a chi-square value of 4.637 meaning that we can reject the null hypothesis can concluded that the GRADES variable is contributing to the model since the chi-square value is greater than the critical value of 3.841. When looking at the change in AIC, the AIC only increases by 2.637. So, it may not be contributing that much information to the model. As such, I will also drop this variable and opt of a simpler model for better interpretation.

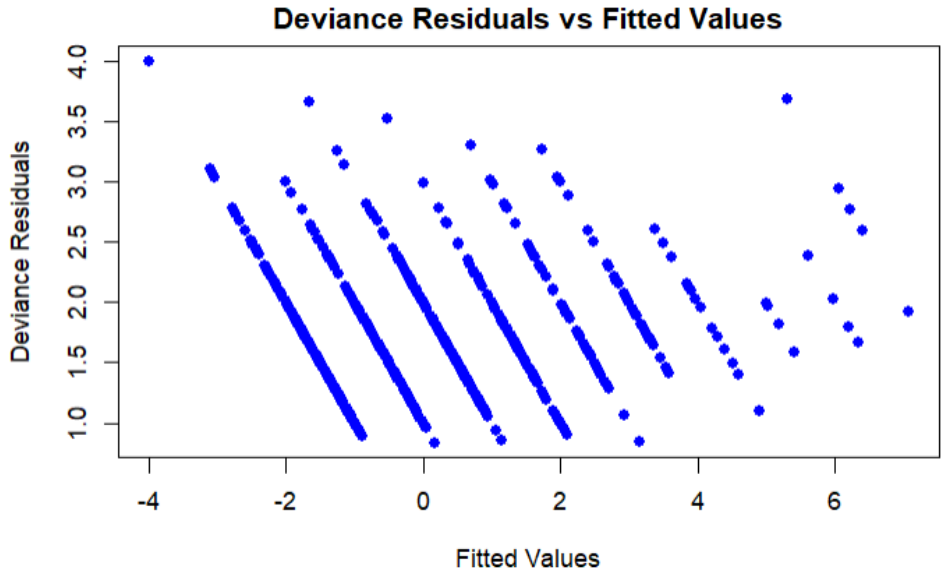
Lastly, I removed the ESTEEM variable and refit the model. This resulted in a chi-square value of 11.779 meaning that we can reject the null hypothesis can concluded that the ESTEEM variable is contributing to the model since the chi-square value is greater than the critical value of 3.841. When looking at the change in AIC, the AIC increases by 9.779. This is a more significant increase in AIC. As such, I will keep ESTEEM in the model. The final model coefficient summary table can be seen below.

```
Call:
zeroinfl(formula = STRESS ~ COHES + ESTEEM | COHES + ESTEEM + GRADES + SATTACH, data = df)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.4611 -0.9134 -0.2473  0.6141  4.1394

Count model coefficients (poisson with log link):
            Estimate Std. Error z value             Pr(>|z|)
(Intercept)  2.395696   0.257200   9.315 < 0.0000000000000002 ***
COHES       -0.011161   0.003205  -3.483             0.000496 ***
ESTEEM      -0.031193   0.009057  -3.444             0.000573 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.12452    0.98594  -3.169  0.00153 **
COHES        0.01469    0.01174   1.251  0.21080
ESTEEM      -0.00600    0.03234  -0.186  0.85283
GRADES       0.03145    0.03684   0.854  0.39332
SATTACH      0.03671    0.02363   1.553  0.12036
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 14
Log-likelihood: -1138 on 8 Df
```

Exponentiated Coefficients:

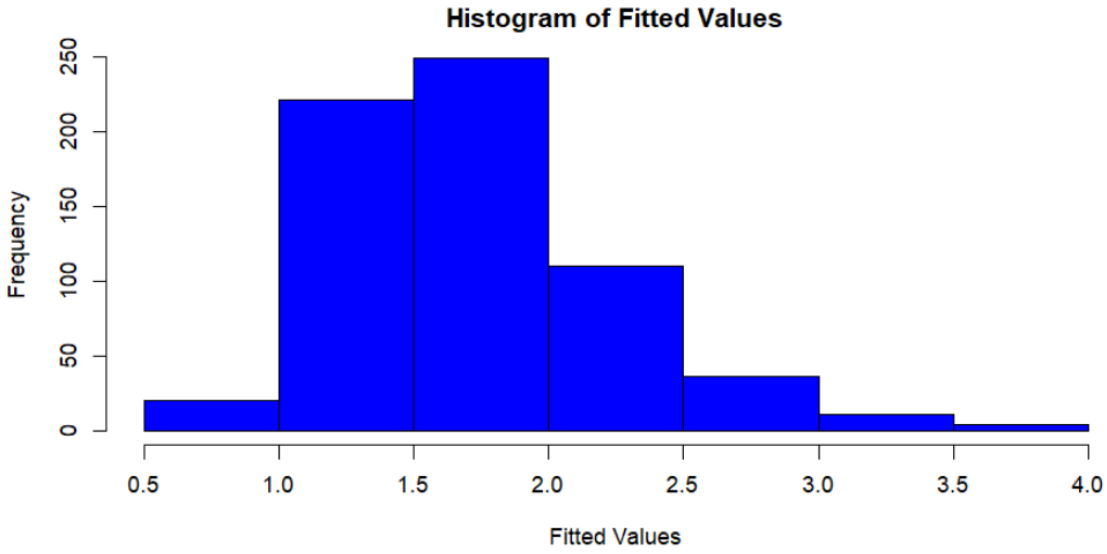
count_(Intercept)	count_COHES	count_ESTEEM	zero_(Intercept)	zero_COHES	zero_ESTEEM	zero_GRADES	zero_SATTACH
10.97583140	0.98890144	0.96928803	0.04395819	1.01480017	0.99401786	1.03194496	1.03739393

By obtaining the exponentiated values of the coefficients in the above table, this zero-inflation portion of the model can be interpreted as the baseline odds of having a stressful event is 0.04 given all other variables are equal to 0. A one unit increase in COHES increases the odds by 1.015, when everything else is held constant. A one unit increase in ESTEEM decreases the odds by 0.994, when everything else is held constant. A one unit increase in GRADES increases the odds by 1.032, when everything else is held constant. And a one unit increase in SATTAH increases the odds by 1.037, when everything else is held constant. For the count model, the baseline number of stressful events is 10.976 among those adolescents who have a stressful event and COHES and ESTEEM are equal to 0. A one unit increase in COHES decreases this baseline amount by 0.989, when everything else is held constant. A one unit increase in ETEEM decreases this baseline amount by 0.969, when everything else is held constant. Logically, this count model makes sense that increases in COHES and ESTEEM would lead to less stressful events.



From the above plot of the deviance residuals vs the fitted values, we can see that there does not appear to be a noticeable heteroscedasticity pattern in the residuals. As such, this model does not violate the assumption of homoscedasticity.

The below histogram shows the distribution for the fitted values of the final model. Except for the 0 value predictions, this model's distribution is similar to that of the actual stressful event counts with the distribution being right skewed and peaking around 1 to 2 stressful events. As previously mentioned these variables do not provide enough information to accurately predict any 0 value stressful event counts. So, it is expected that the model didn't perform well when predicting these values.



Overall, it seems like a zero-inflated model is a good choice for modeling this data and produces predictions that most resemble the distribution of actual stressful events outside of predicting adolescents with no stress. But compared to the other models fitted during this analysis the zero-inflated model is the best approach to use. The OLS regression model was only able to explain about 8% of the variance in the STRESS variable so this approach isn't a good choice for this data. Using a log transformation did not improve the model fit and also was only able to explain about 8% of the variance in STRESS. The Poisson model showed slight improvement in the distribution of predicted values, but as not able to predict if an adolescent did not have any stress. A logistic regression can help with predicting if an adolescent has stress or not, but is not able to predict the number of stressful events. Therefore, combining these approaches by using a ZIP model is the most appropriate modeling approach to use and has shown to produce a distribution of predicted values that most resembles the actual STRESS values out of all the models tested in this analysis.

Conclusion

From this analysis, I was able to learn limitations of OLS regression models, logistic regression models, and Poisson models when there is a large number of 0 values in the dataset. I also learned how to build a ZIP model by hand and using a built-in package. This is my first time building this type of model, so it was definitely a learning process. However, I can now see why this model can be appropriate for certain types of datasets. Although I do wish the model did a better job of predicting 0 values; however, not all models can perform well and it appears that the variables and data provided do not make good predictors to predict if there is stress present in an adolescent. From an overall modeling standpoint, I am getting more comfortable assessing diagnostics and overall goodness of fit for the models. Additionally, I have gotten more proficient at evaluating whether certain variables are providing meaningful information to the model and if they can be removed from the model. I've learned throughout this course that simpler models are often more preferred, so this is an important step in the process to determine if variables can be removed even though they might be statistically significant by

technical standards. As with most processes in modeling, this is also a bit subjective as to what is considered “meaningful”. For example, I removed GRADES because I considered a 2.637 increase in AIC to be not meaningful enough to keep in the model. However, someone else may consider this amount to be meaningful enough to keep on the model. As such, I have learned from this course to communicate and research these details of the modeling process to understand the limitations and/or potential biases in models.