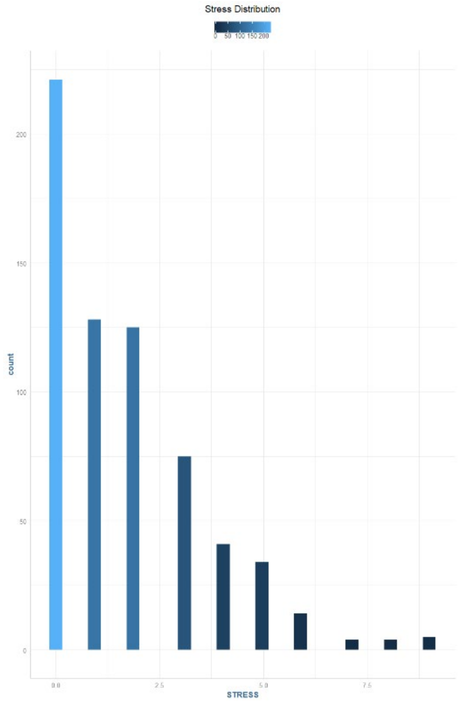
COMPUTATIONAL ASSIGMENT #6
BRANDON MORETZ

COMPUTATION

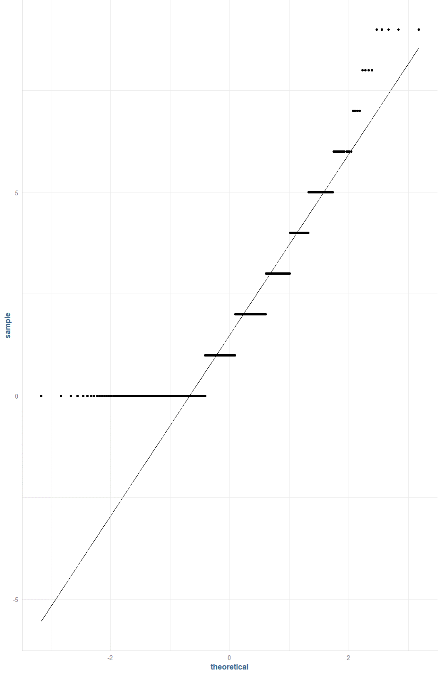
1.)

a. For the STRESS variable, make a histogram and obtain summary statistics.

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



b. Obtain a normal probability (Q-Q) plot for the STRESS variable.

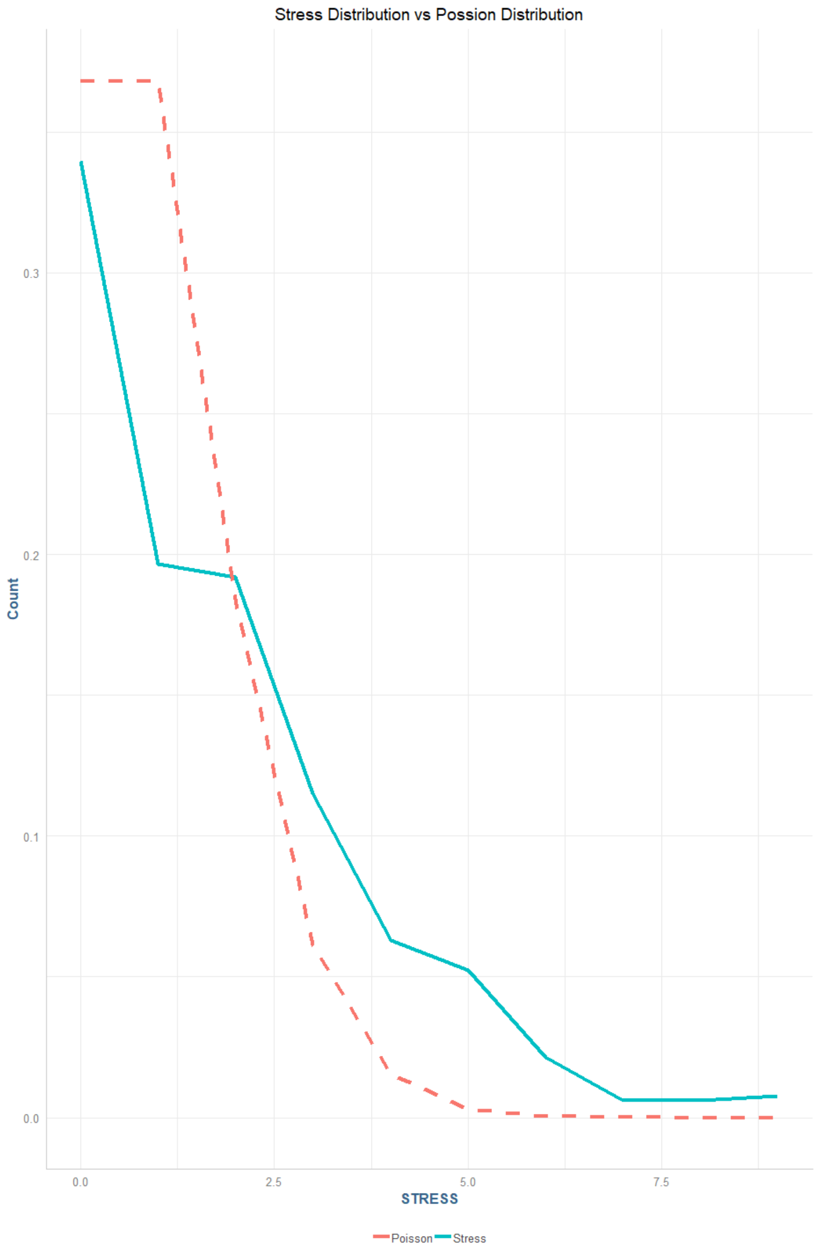


c. *Is STRESS a normally distributed variable?*

Stress is not a normally distributed variable, it has a heavy right skew.

d. *What do you think is its most likely probability distribution for STRESS?*

STRESS looks like it follows a Poisson distribution, which we can see from the below graphical comparison:



e. *Give a justification for the distribution you selected.*

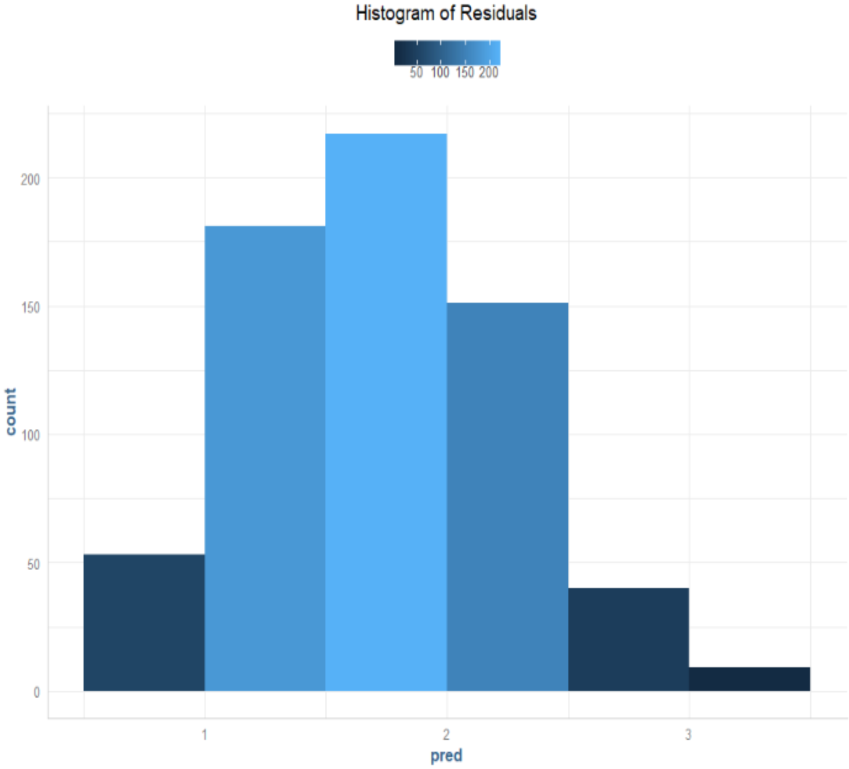
The distribution of the stress variable follows a Poisson distribution with a heavy right tail that slowly tappers off in the end. We can see the comparison of the actual values of stress in blue vs a Poisson distribution in the red dotted line.

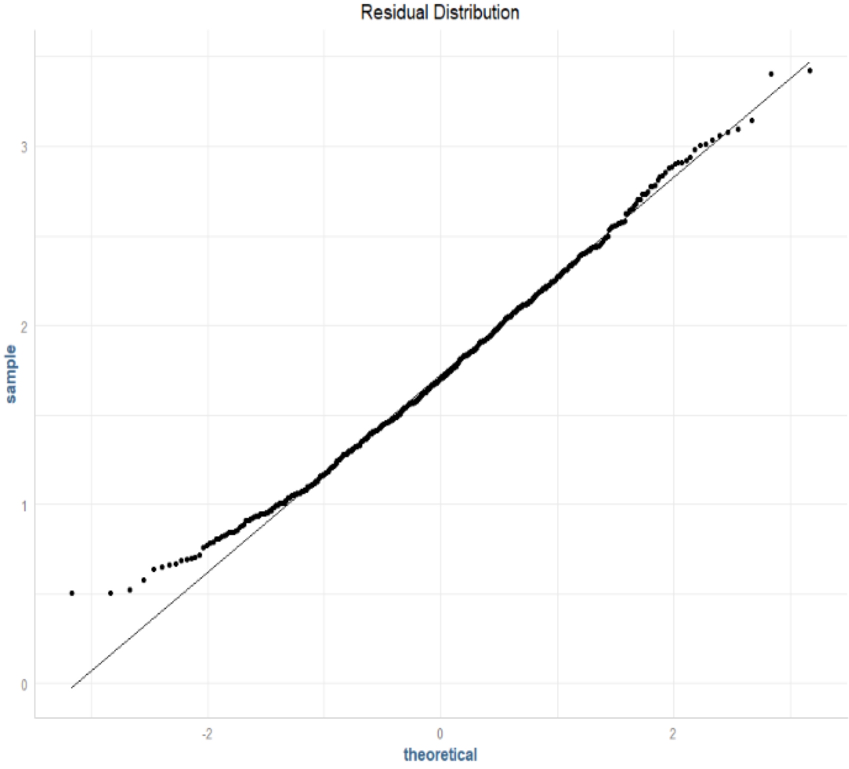
2.)

a. Fit an OLS regression model to predict ***STRESS (Y)*** *using* ***COHES***, ***ESTEEM***, ***GRADES***, ***SATTACH*** *as explanatory variables (X).*

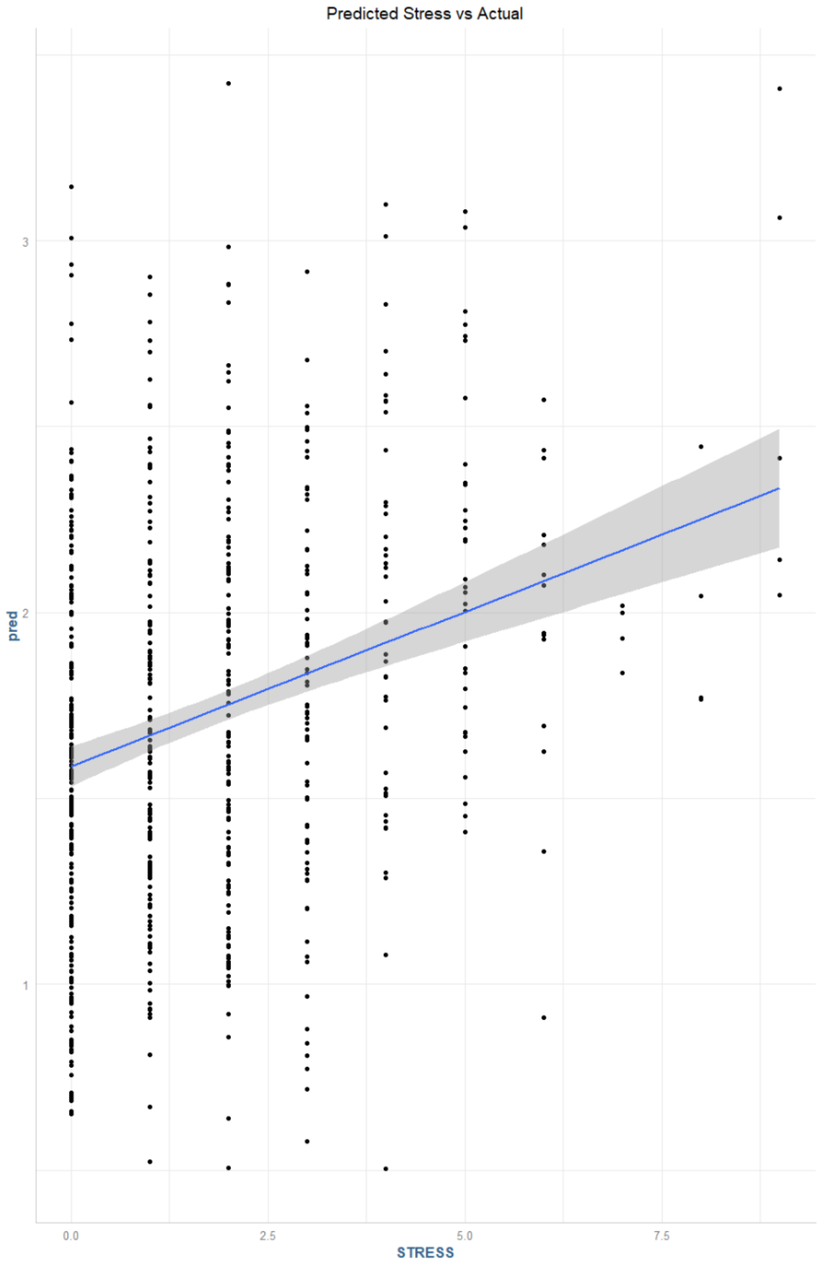
Model 1

$$\hat{y} = 5.713 - 0.023\beta_1 - 0.412\beta_2 - 0.0471\beta_3 - 0.03\beta_4$$





In the model, the intercept term of 5.713 would denote the mean stress level when all the coefficients are zero. Where β_1 is COHES, a measurement of how well an adolescent gets along with their family. A one unit increase here denotes a 0.023 unit decrease in stress. Esteem is represented by β_2, and a one unit increase here represents a 0.412 decrease in an individual's stressful events. Grades is the sum of the grades for the prior year and are represented with the β_3 coefficient, where a one-unit increase represents a 0.0471 decrease to an individual's stress. Finally, we have SATTACH which is a measurement of how well a student is attached to their school, and a one unit increase here would represent a 0.03 unit decrease in an individual's predicted stressful events.



The fitted model has an R² of **.0831**, which denotes it explains approximately **8**% of the variance in the data, which does not denote a good fit for the data. The residuals are approximately normally distributed as we can see in the chart to the left.

*b. 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?*

The distribution of the residuals is approximately normal, which is fine, however, if we look at a plot of the predicted values of stress () vs the actuals and fit a linear model over this data set, we can clearly see there are several issues with using this model to infer an individual's stress.

3.)

*a. Create a transformed variable on Y that is **LN**(Y).*

In order to facilitate this transformation, we must add **0.001** to the dependent variable (**STRESS**) so that the log transformation is not undefined (**LN**(0) = *Undefined*).

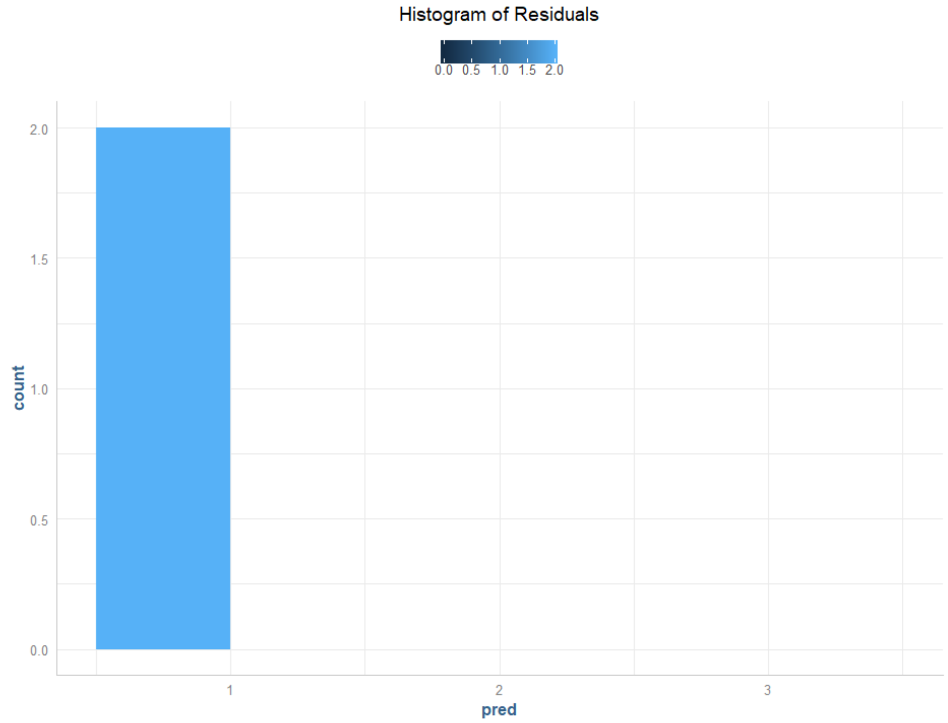
*b. Fit an OLS regression model to predict **LN**(Y) using **COHES**, **ESTEEM**, **GRADES**, **SATTACH** as explanatory variables (X).*

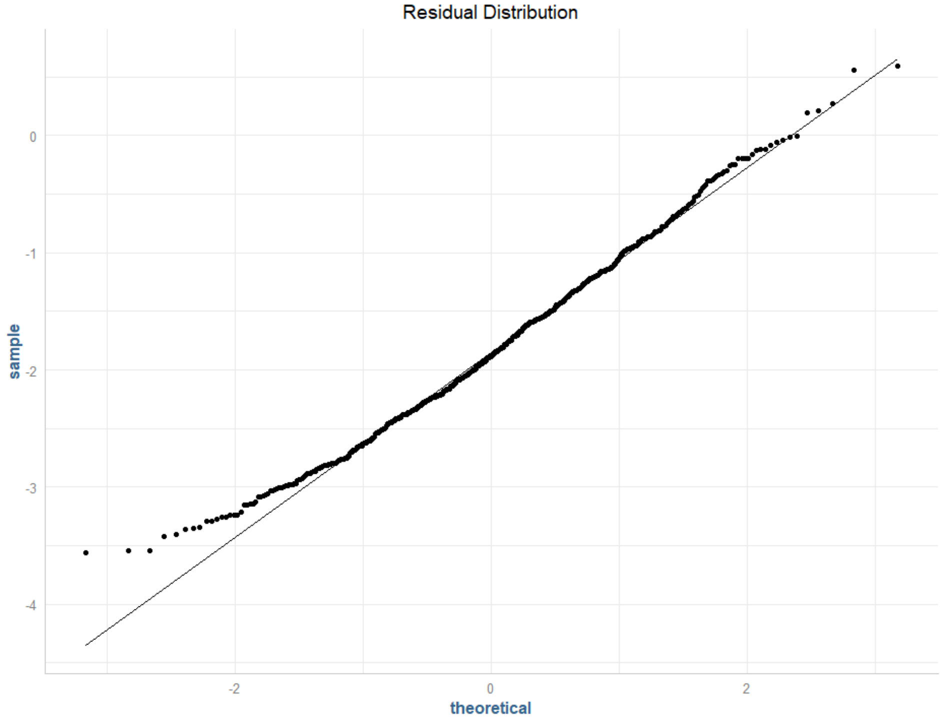
Model 2

$$\hat{y} = 3.597 - 0.0378\beta_1 - 0.04\beta_2 - 0.054\beta_3 - 0.51\beta_4$$

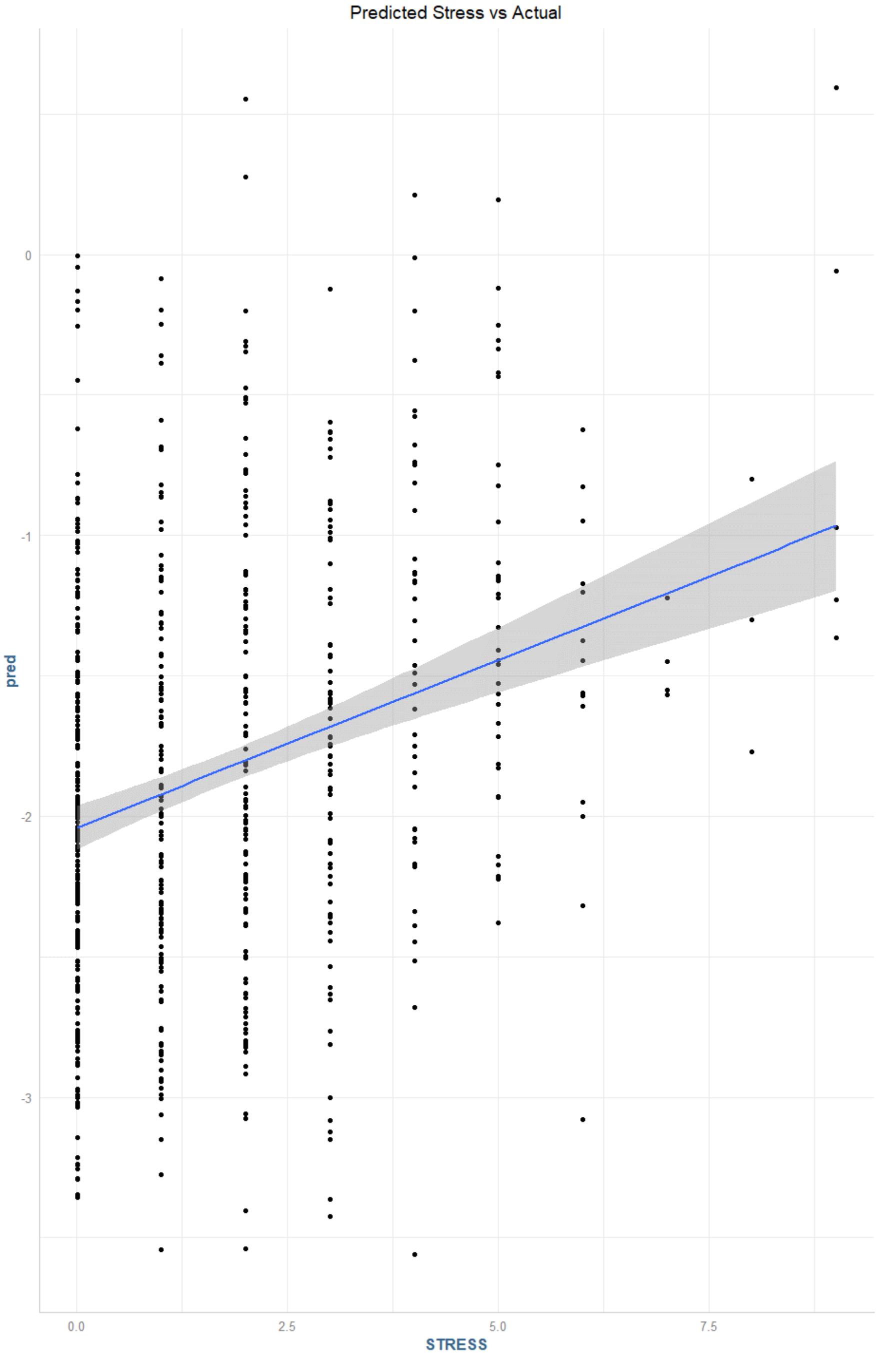
c. *Obtain the typical diagnostic information and graphs. Discuss how well this model fits. Obtain predicted values (**LN**(Y)_hat) and plot them in a histogram.*

The model has an R^2 value of 0.044, indicating that approximately 4.4% of the overall variance in the data is explained by the model, which is essentially no additional information. None of the coefficient terms are statistically significant, including the intercept. By all accounts, the model is useless. The graph below shows a histogram and QQ-Plot of the residuals, which are all distributed into the same bucket.





Below is a chart of the predicted values vs actuals, and we can see that the outliers in this model are the dominate factor, the model would only accurately predict a few edge cases essentially by accident.



d. *What issues do you see? Does this correct the issue?*

An abundance of issues remains with this model. There is almost no explained variance by the model, the residuals are normally distributed due to the logarithmic transformation of the response, however, the core issues of poor fit and wildly inaccurate predictions persist after the transformation of the response variable. This model should be discarded and not used be used for any further analysis.

4.)

a. *Use the glm() function to fit a Poisson Regression for* ***STRESS*** *(Y) using* ***COHES***, ***ESTEEM***, ***GRADES***, ***SATTACH*** *as explanatory variables (X).*

Model 3

$$\hat{Y} = 2.735 - 0.013\beta_1 - 0.024\beta_2 - 0.23\beta_3 - 0.016\beta_4$$

b. *Interpret the model's coefficients and discuss how this model's results compare to your answer for part* ***3****).*

For the coefficients in this model, we can interpret the intercept in this model as a simple placeholder value since:

$$\exp(2.735) = \mathbf{15.4}$$

Which is above the possible value range for stressful events, meaning that without the additional coefficient terms the model will essentially generate garbage values. Where β_1 is COHES, a measurement of how well an adolescent gets along with their family. A one unit increase here denotes an

$$\exp(-0.013) = \mathbf{.987}$$

or a **0.013**% unit decrease in expected rate of stressful events. Esteem is represented by β_2, and a one unit increase here represents a

$$\exp(-0.024) = \mathbf{0.977}$$

0.023% decrease in an individual's expected rate of stressful events. Grades is the sum of the grades for the prior year and are represented with the β_3 coefficient, where a one-unit increase represents an

$$\exp(-0.023) = \mathbf{0.977}$$

0.023% decrease to an individual's expected rate of stressful events. Finally, we have SATTACH which is a measurement of how well a student is attached to their school, and a one unit increase here would represent a

$$\exp(-0.016) = \mathbf{0.984}$$

0.016% unit decrease in an individual's predicted rate of stressful events.

c. *Similarly, fit an over-dispersed Poisson regression model using the same set of variables. How do these models compare?*

Model 4

$$\hat{Y} = 2.759 - 0.013\beta_1 - 0.023\beta_2 - 0.24\beta_3 - 0.017\beta_4$$

The overly dispersed Poisson model has a similar fit to the normal Poisson model, however, looking at the AIC for the two models, 2,483 and 2,283 for model 3 and model 4 respectively, we can infer from the reduction in AIC from model 4 of approximately 133.6, the overly dispersed Poisson model is a better fit to the data.

5.)

a. *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**).*

Cut points:
Low < **41.62** < Medium < **64.38** < High

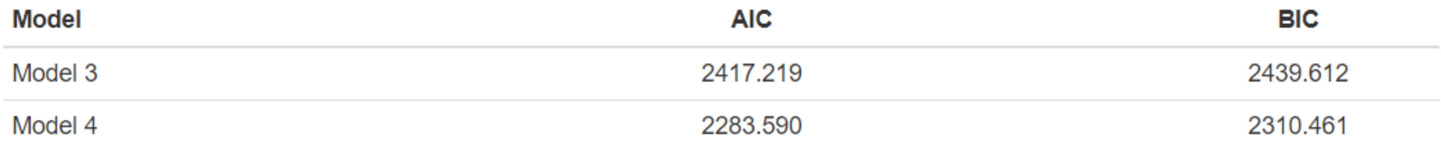
b. *What is the expected percent difference in the number of stressful events for those at high and low levels of family cohesion?*

Using the mean values as placeholder values for the beta 2, 3 and 4 coefficients we can derive the following equations:

$$Low = exp(2.735 - 0.013 * 41.62 – 0.024 * 31.86 – 0.023 * 14.93 – 0.016 * 26.81) = \mathbf{1.915}$$
$$High = exp(2.735 - 0.013 * 64.38 – 0.024 * 31.86 – 0.023 * 14.93 – 0.016 * 26.81) = \mathbf{1.427}$$

Which is equal to roughly a **25%** decrease in expected stressful events for someone with a high level of family cohesion.

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?*



Model	AIC	BIC
Model 3	2417.219	2439.612
Model 4	2283.590	2310.461

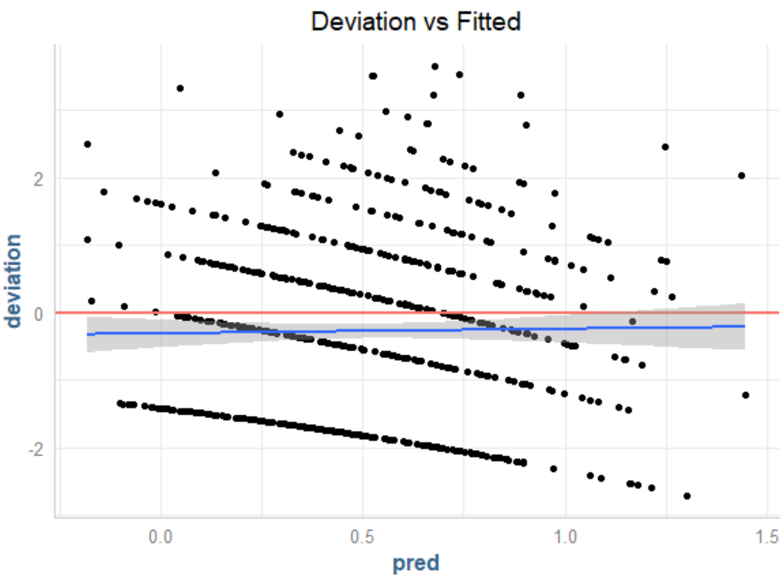
In the above table, Model 3 is the regular Poisson and Model 4 is the overly dispersed Poisson. We can see that the overly dispersed Poisson has a lower AIC and BIC measure, meaning model 4 has a statistically better fit to the data than model 3.

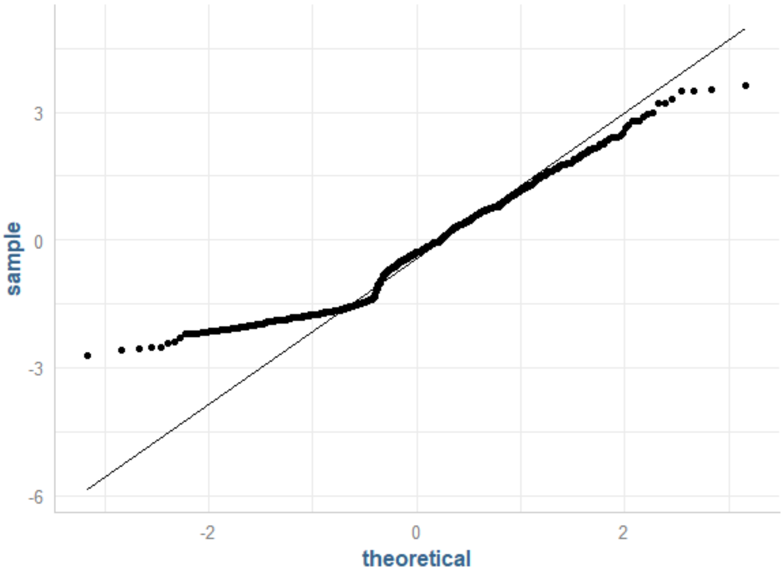
7.) *Using the Poisson regression model from part **4**), plot the deviance residuals by the predicted values. Discuss what this plot indicates about the regression model.*

In the following plot we can see the deviance residuals vs the predicted values for the model as well as a QQ-Plot of the deviance residuals. We see deviations from normality on the QQ-Plot and the deviations vs predicted values from the model shows a great deal of over-dispersion from the model fit to the data. Further, we can look at a Chi-Square “goodness-of-fit” test for this model:

$$X^2 = 1 - pchisq(1245.42, 646) = 0$$

Even though three of our four coefficients show statistical significance, overall the model seems to be a poor fit.





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?*

Model 5

$$\hat{Y} = 3.517 – 0.21\beta_1 – 0.019\beta_2 - 0.025\beta_3 – 0.028\beta_4$$

In this model we have an intercept value is again a simple placeholder given that its interpretation is:

$$\exp(3.517)/(1 + \exp(3.517) = \mathbf{.971}$$

or simply there is roughly a 97.1% chance that someone will experience a stressful event at some point, ignoring the rest of the coefficients. Next, the β_1 coefficient, which is related to the **COHES,** a measurement of how well an adolescent gets along with their family, decreases the chances of having a stressful event by exp(-0.21) = **.997**

or simply a 2.1% chance decrease per unit of **COHES**. The β_2 coefficient is related to **ESTEEM**, and here we can also see a negative correlation like in **COHES**, although generally weaker.

exp(-.019) = **.981**

or simply a 1.9% decrease in the chances of having a stressful event per unit increase in **ESTEEM**. Grades are the cumulation of the prior year and are reflected by the β_3 coefficient and again we see a negative (decreasing) relationship:

exp(-0.025) = **.975**

which shows a 2.52% decrease in the chances of having a stressful event given one unit of increase in grades. The final variable in the model is **SATTACH,** which is a measurement of how well a student is attached to their school, is associated with the β_4 coefficient which we can interpret as:

exp(-0.028) = **.973**

which we can interpret as a 2.73% decrease in the changes of having a stressful event given a one unit increase in student attachment.

The model has only one statistically significant term, COHES, at the 5% level. The other three vary from ~10-40%, however, the model does have a relatively low AIC, 821.79, especially when compared to the other models. However, from a practical sense, the model is mostly meaningless in that practically everyone will have a stressful event at some point.

RESEARCH

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 must be thoughtful about this. It is not as straight forward as plug and chug!*

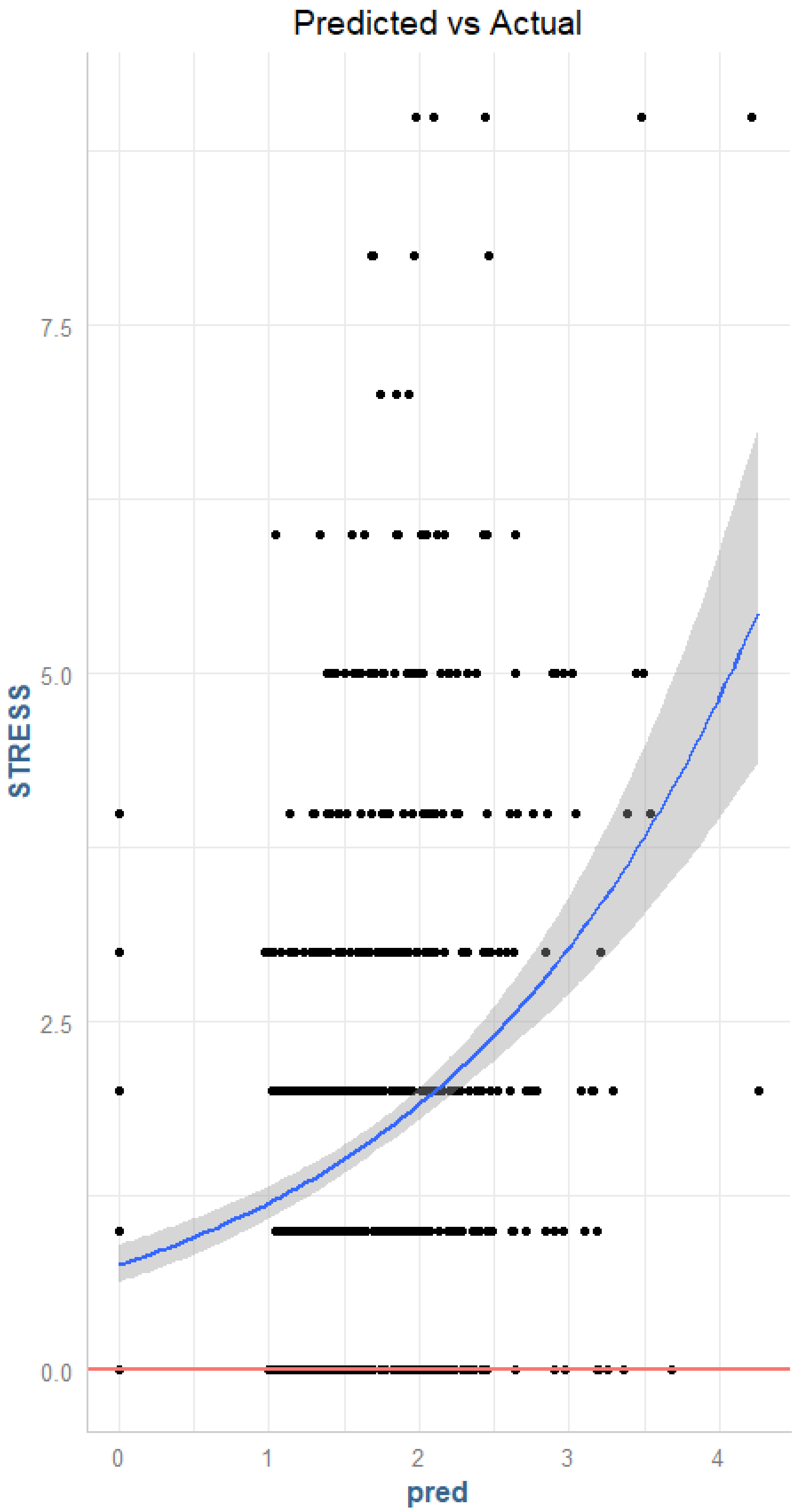
The conceptual basis for this type of modeling is indeed valid, it is used when there are an excess of zero count items in a Poisson. Further, the process that is generating the zeros is theorized to be of a different mathematical form than the one generating the non-zero counts. This type of modeling is known as "ZIP" or zero-inflated Poisson regression. Performing the calculations manually, we get the following logistic model:

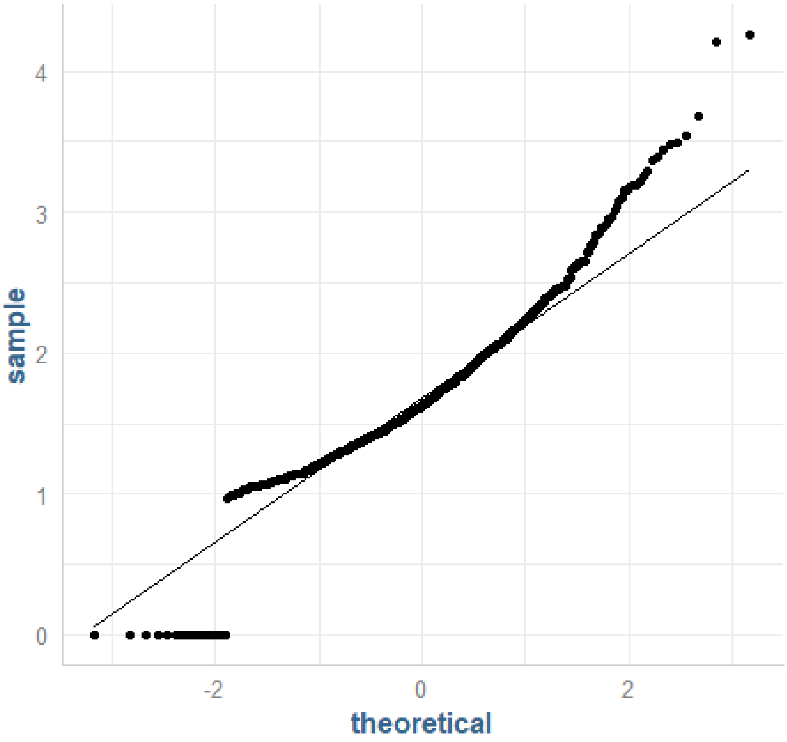
Zip Model 1 (Logistic)

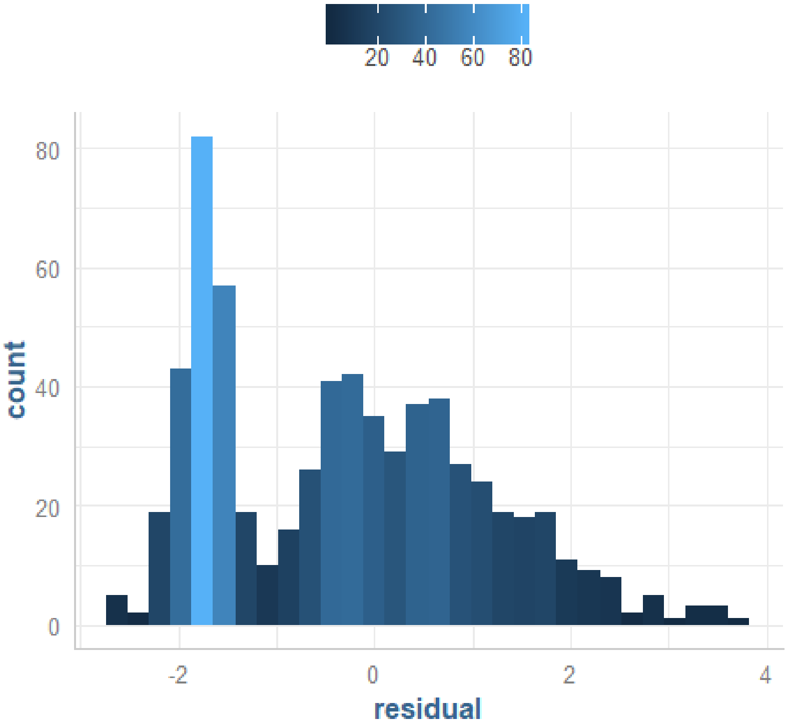
(Intercept)	COHES	ESTEEM	GRADES	SATTACH
3.517	-0.021	-0.019	-0.025	-0.028

Zip Model 2 (Poisson)

(Intercept)	COHES	ESTEEM	GRADES	SATTACH
2.735	-0.013	-0.024	-0.023	-0.016


Predicted vs Actual
STRESS
7.5
5.0
2.5
0.0
0
1
2
3
4
pred





In terms of the model fit, the AIC for the Poisson model here is 2417, which shows no improvement over our prior attempts. The model diagnostics confirm that the zeros are being accounted for, however, the overall distribution of the deviance residuals, non-normality of prediction values, and generally high presence of outliers throughout the data suggest that this is not a terribly useful model although an improvement over the standard Poisson approach.

CONCLUSION

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.*

In our earlier analysis, we saw that the data was better modeled with a negative binomial regression model, so we will fit another model using a zero-inflated negative binomial. Here we will fit one standard Poisson model using the ZIP technique, and then re-fit it using a negative binomial. We will then compare the two ZIP models using a Vuong test, which we can see below:

Vuong Non-Nested Hypothesis Test-Statistic:			
Vuong z-statistic			H_A p-value
Raw	-1.970681	model2 > model1	0.02438
AIC-corrected	-1.970681	model2 > model1	0.02438
BIC-corrected	-1.970681	model2 > model1	0.02438

Here we see that there is less than a 5% chance that the negative binomial ZIP (model 2) model is not the better fitting model.

Overall, this is a difficult data set to forecast accurately given the data at hand. The attempts at Logistic and Standard Linear regression obviously failed due to the nature of the response, however, even fitting with a generalized linear model with a similar distribution of outcomes the models simply do not explain the vast majority of the variance in the data. We can infer basic relationships, such as self-esteem and how well one gets along with their family have a statistically strong relationship to reducing the number of stressful events, by exactly how many though is extremely difficult to quantify with the data at hand.

The introduction of the ZIP technique I thought was interesting, it does make intuitive sense that there are two unrelated mathematical process responsible for generating excessive zeros when modeling these discrete variables. Even with this approach, we are still not satisfied with the results. The distribution of the residuals are all over the place, the outliers in the predicted vs actual are outrageous, and overall the process does not really the theoretical Poisson (the blue line in the top graph) to a degree I would give confidently endorse. The modeling efforts are better than nothing, however, leave much to be desired.