Modeling Assignment 9: Poisson and ZIP Regression Models

Assignment Overview

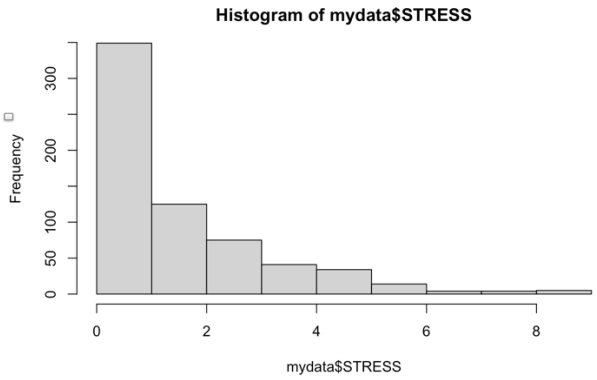
In this assignment we will be fitting models and calculating the various summative statistics that are associated with Poisson and Zero-Inflated Poisson Regression.

The data set for this assignment, STRESS, includes information from about 650 adolescents in the United States who were surveyed about the number of stressful life events they had experienced in the past year (STRESS). **STRESS** is also 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.

The variables in this data set 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).
There is no other information about this data or the variables.

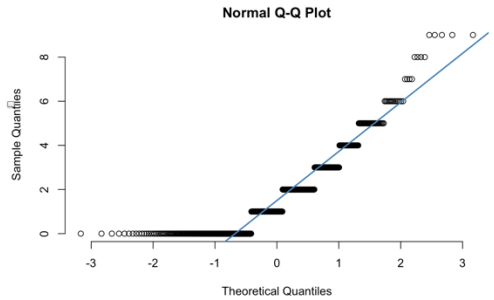
Assignment Tasks

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



Min	0.00
1st quartile	0.00
Median	1.00
Mean	1.73
3rd quartile	3.00
Max	9.00
Sd	1.849082

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.



Stress is not normally distributed. I would classify it as a Poisson distribution because it follows the following assumptions and validity (from wiki):

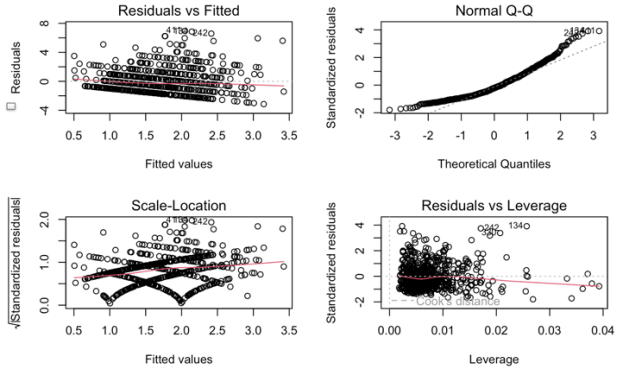
- k is the number of times a stressful life event occurs in an interval and k can take values 0, 1, 2
- The occurrence of one stressful life event does not affect the probability that a second event will occur. That is, events occur independently.
- The average rate at which stressful life events occur is independent of any occurrences. For simplicity, this is usually assumed to be constant, but may in practice vary with time.
- Two stressful life events cannot occur at exactly the same instant; instead, at each very small sub-interval, either exactly one stressful life event occurs, or no stressful life event occurs.

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.

Model Summary

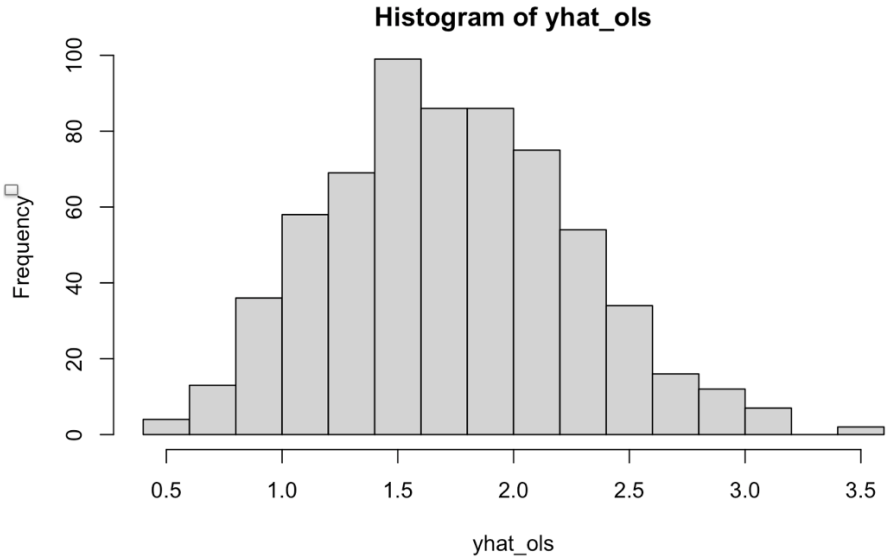
```
Call:
lm(formula = STRESS ~ COHES + ESTEEM + GRADES + SATTACH, data = mydata)

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 ***
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: 1.826e-11
```

Diagnostic Plots



This model fits the data very poorly. There is not a random scatter in the residual vs fitted plot and the scale location plot which indicates there is not a linear relationship between STRESS and the predictors of this model. The poor fit is also corroborated by the very low adjusted R2. Only 7.75% of variation in STRESS can be explained by the predictors in this model.

Obtain predicted values (Y_hat) and plot them in a histogram. What issues do you see?



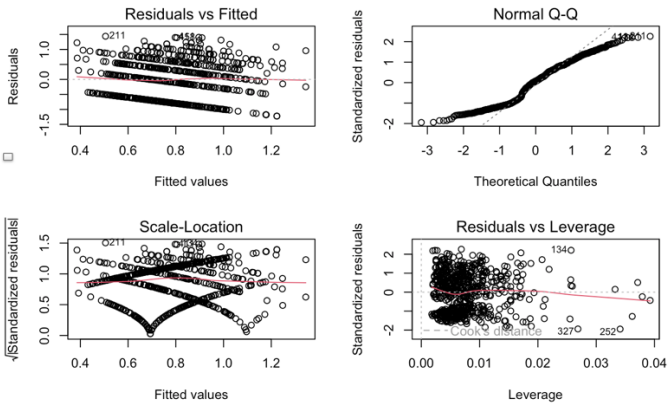
There are 2 problems with the predicted values of STRESS in this model. 1.) The histogram shows a normal distribution whereas the actual distribution is Poisson (skewed right with a median of 1.00). 2.) The max value of Y_hat is 3.5 but the actual distribution contains values ranging from 0 to 9.

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.

Model Summary

```
Call:
lm(formula = log(STRESS + 1) ~ COHES + ESTEEM + GRADES + SATTACH,
    data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.22362 -0.63438  0.04982  0.51763  1.44040

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.073322   0.209097   9.916  < 2e-16 ***
COHES       -0.007947   0.002529  -3.142  0.00175 **
ESTEEM      -0.010915   0.006955  -1.569  0.11706
GRADES      -0.014336   0.008462  -1.694  0.09072 .
SATTACH     -0.011283   0.005081  -2.220  0.02674 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.639 on 646 degrees of freedom
Multiple R-squared:  0.07154,   Adjusted R-squared:  0.06579
F-statistic: 12.44 on 4 and 646 DF,  p-value: 9.333e-10
```

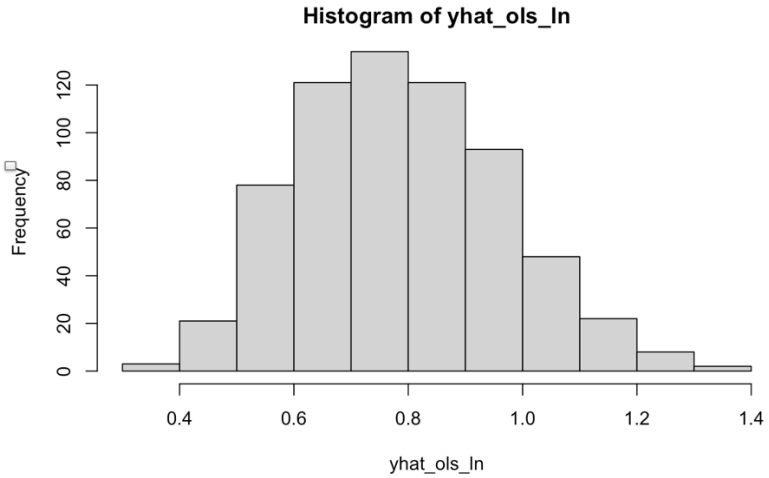
Diagnostic Plots



We see similar concerns in this model as the last model:

This model fits the data very poorly. There is not a random scatter in the residual vs fitted plot and the scale location plot which indicates there is not a linear relationship between STRESS and the predictors of this model. The poor fit is also corroborated by the very low adjusted R2. Only 6.58% of variation in STRESS can be explained by the predictors in this model.

Obtain predicted values (LN(Y)_hat) and plot them in a histogram. What issues do you see? Does this correct the issue?



We see similar concerns in this model as the last model:

There are 2 problems with the predicted values of STRESS in this model. 1.) The histogram shows a normal distribution whereas the actual distribution is Poisson (skewed right with a median of 1.00). 2.) The max value of log(Y_hat+1) is 1.4 but the actual distribution contains values ranging from 0 to ln(9+1) =2.3. Note we add 1 inside the log because ln(0) is undefined.

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).

Model Summary

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

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  < 2e-16 ***
COHES       -0.012918   0.002893  -4.466 7.98e-06 ***
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
```

For each additional COHES, we estimate the odds of a stressful life event occurring in a year *decreases* by exp(-0.012918) – 1 = 1.283492%. For each additional ESTEEM, we estimate the odds of a stressful life event occurring in a year *decreases* by exp(-0.023692) – 1 = 2.341355%. For each additional GRADES, we estimate the odds of a stressful life event occurring in a year *decreases* by exp(-0.023471) – 1 = 2.31977%. For each additional SATTACH, we estimate the odds of a stressful life event occurring in a year *decreases* by exp(-0.016481) – 1 = 1.634593%.

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).
There is no other information about this data or the variables.

From these variables, we may generally conclude that adolescents who get along with their family, have high self esteem, get good grades, and is attached to their school have lower odds of have a stressful event happen during the year.

Relatively this model (AIC = 2417.219) performs worse than the model in part 3 (AIC = 1271.255). This is due to the distribution of STRESS which has a significant number of 0's and the model does not return any predicted stress values of 0. Although the distribution of Y_hat is better, the prediction power is not.

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

The AIC of an overdispersed Poisson regression is 2283.6 which is less than the normal glm Poisson regression. We conclude it is a better fit.

Summary of models thus far:
All models except for OLS ln(Y) perform very similarly in terms of Sum of Squares Error which indicates similar amount of variation in Y from the regression line.

Question	Model	SSE
2	OLS Y	2037.537
3	OLS ln(Y)	2697.992
4	glm Poisson	2037.453
4	glm overdispersion	2038.054

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).

Family Cohesion Level	Definition	Predicted # (Mean) of Stressful Life Events in a Year
Low	< 1 SD below mean	2.524774
Medium	> 1 SD below mean, < 1 SD above mean	1.66493
High	> 1 SD above mean	1.178379

We note that an adolescent with medium family cohesion in our dataset has an average of 1.66 stressful life events in a year. Adolescents with less family cohesion have more stressful life events, and adolescents with more family cohesion have less stressful life events.

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

(2.524774 - 1.178379) / 2.524774
Low cohesion (mean=2.52) has 53.32735% more stressful events than high cohesion (mean=1.17).

(2.524774 - 1.178379) / 1.178379
High cohesion (mean=1.17) has 114.2582% less stressful events than low cohesion (mean=2.52).

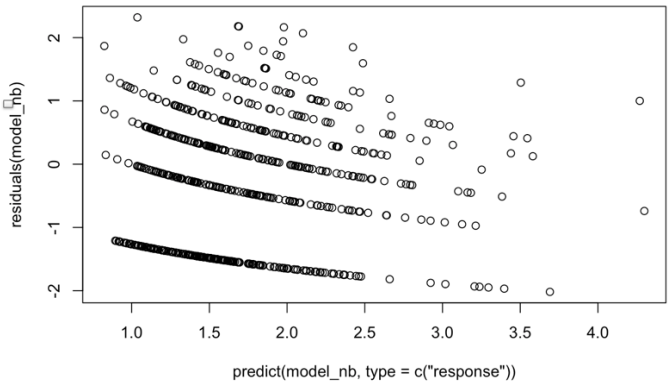
6. **Compute the AICs and BICs** from the Poisson Regression and the over-dispersed Poisson regression models from part 4). Is one better than the other?

Model	AIC	BIC
Poisson Regression	2417.219	2439.612
Over-dispersed Poisson Regression	2283.59	2310.461

The over-dispersed Poisson Regression has lower values of AIC and BIC than the normal Poisson Regression which indicates it is a better model.

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.

Deviance Residual vs Predicted Values Plot



We do note that the points are approximately evenly above and below the residuals = 0 line. However, there does not appear to be a random scatter of points which is not what we want to see from the model.

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 BIC for Y_IND

glm family=poisson	glm negative binomial	glm	glm binomial
# 1241.758	# 1248.245	# 890.9043	# 844.1784
COHES+ESTEEM+GRADES+SATTACH	COHES+ESTEEM+GRADES+SATTACH	COHES+ESTEEM+GRADES+SATTACH	COHES+ESTEEM+GRADES+SATTACH
# 1237.135	# 1243.622	# 890.1174	# 843.4351
ESTEEM+GRADES+SATTACH	ESTEEM+GRADES+SATTACH	ESTEEM+GRADES+SATTACH	ESTEEM+GRADES+SATTACH
# 1235.513	# 1242 COHES+GRADES+SATTACH	# 885.0991	# 838.3319
COHES+GRADES+SATTACH	# 1242.022	COHES+GRADES+SATTACH	COHES+GRADES+SATTACH
# 1235.535	COHES+ESTEEM+SATTACH	# 885.211	# 838.4917
COHES+ESTEEM+SATTACH	# 1242.563 COHES+ESTEEM+GRADES	COHES+ESTEEM+SATTACH	COHES+ESTEEM+SATTACH
# 1236.076 COHES+ESTEEM+GRADES	# 1236.587 COHES+ESTEEM	# 886.9016	# 840.2323
# 1230.1 COHES+ESTEEM	# 1236.543 COHES+GRADES	COHES+ESTEEM+GRADES	COHES+ESTEEM+GRADES
# 1230.056 COHES+GRADES	# 1235.859 COHES+SATTACH	# 881.9305 COHES+ESTEEM	# 835.233 COHES+ESTEEM
# 1229.372 COHES+SATTACH	# 1235.859 ESTEEM+GRADES	# 881.8015 COHES+GRADES	# 835.117 COHES+GRADES
# 1232.09 ESTEEM+GRADES	# 1237.613 ESTEEM+SATTACH	# 879.651 COHES+SATTACH	# 832.8874 COHES+SATTACH
# 1231.126 ESTEEM+SATTACH	# 1238.053 GRADES+SATTACH	# 888.0736 ESTEEM+GRADES	# 841.519 ESTEEM+GRADES
# 1231.566 GRADES+SATTACH	# 1230.798 COHES	# 885.0855 ESTEEM+SATTACH	# 838.4343 ESTEEM+SATTACH
# 1224.312 COHES	# 1233.1 ESTEEM	# 886.3667 GRADES+SATTACH	# 839.6714 GRADES+SATTACH
# 1226.613 ESTEEM	# 1233.971 GRADES	# 877.5257 COHES	**# 830.8144 COHES *****
# 1227.484 GRADES	# 1232.37 SATTACH	# 884.6394 ESTEEM	# 838.1296 ESTEEM
# 1225.884 SATTACH		# 887.2263 GRADES	# 840.7058 GRADES
		# 882.3112 SATTACH	# 835.649 SATTACH

According to BIC (which penalizes harsher on additional variables), the best model for predicting Y_IND, or the yes/no presence of STRESS is the glm binomial model only using COHES as a predictor.

Model Summary

```
Call:
glm(formula = Y_IND ~ COHES, family = binomial, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9543  -1.3432   0.8055   0.9375   1.1703

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.296371   0.427310   5.374 7.70e-08 ***
COHES       -0.030393   0.007715  -3.939 8.17e-05 ***
---
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: 817.86  on 649  degrees of freedom
AIC: 821.86

Number of Fisher Scoring iterations: 4
```

For each additional COHES, we estimate the odds of a stressful life event occurring in a year *decreases* by exp(-0.012918) – 1 = 2.993578%.

```
   FALSE TRUE
0     99  122
1    130  300
```

```
Accuracy: 0.6129032
Sensitivity: 0.4323144
Specificity: 0.7109005
```

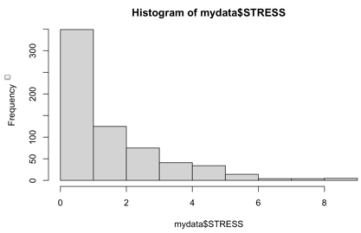
The model accurately predicts the presence or absence of STRESS in the dataset 61.29% of the time.

Although the model performs much better than previous models (SSE = 142.3678), we must note that it can only be used to predict Y_IND which is dichotomous unlike STRESS which is not.

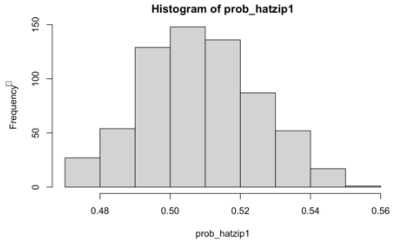
9. It may be that there are two (or more) process at work that are overlapped and generating the distributions of STRESS(Y). What do you think those processes might be? To conduct a ZIP regression model by hand, **fit a Logistic Regression model to predict if stress is present (Y_IND), and then use a 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!

We need 2 models, a logistic regression that first predicts the presence of STRESS (0 or non-zero), and if it is non-zero, we need a poisson regression to predict an integer greater than or equal to 1.

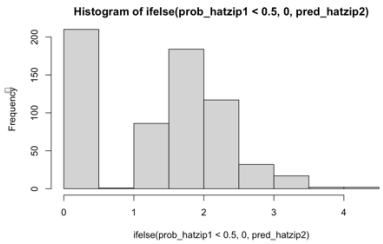
Distribution of Stress



Probability of Y_IND



Combined Model for Y_hat



The left histogram indicates how actual STRESS data is distributed. The model on the right has the predicted values of the combined model. We now see the majority of data is distributed around 0 and also follows a mean value between 1 and 2 like the original data. The middle histogram shows the probability of the presence of a stressful event that has been shifted to match the percent of actual 0 and non-0 stressful life events in the dataset.

Question	Model	SSE (<)	Log Likelihood (>)
2	OLS Y	2037.537	-1295.12 (df=6)
3	OLS ln(Y)	2697.992	-629.6277 (df=6)
4	glm Poisson	2037.453	-1203.61 (df=5)
4	glm overdispersion	2038.054	-1135.795 (df=6)
9	zip by hand	2296.084	N/A

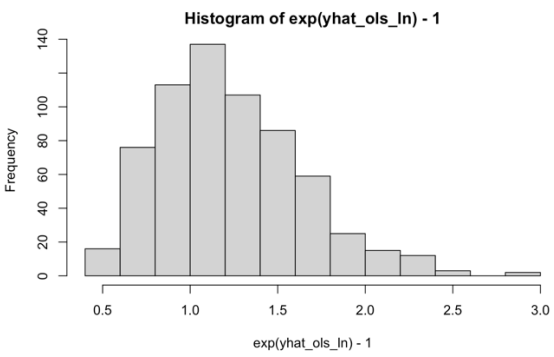
Although the distributions now align better, the performance is still not up to par in comparison to the other models. We will use the pscl package and zeroinfl function to create other zip models and evaluate how they fit.

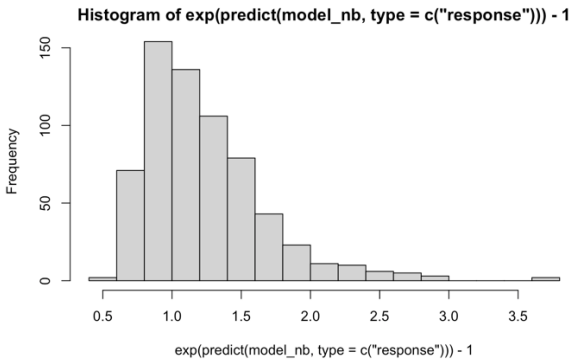
10. Use the pscl package and the zeroinfl() function to Fit a ZIP model to predict STRESS(Y). You should do this twice, first using the same predictor variable for both parts of the ZIP model. Second, finding the best fitting model. **Report the results and goodness of fit measures**. Synthesize your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

Question	Model	SSE (<)	Log Likelihood (>)
2	OLS Y	2037.537	-1295.12 (df=6)
3	OLS ln(Y)	2697.992	-629.6277 (df=6)
4	glm Poisson	2037.453	-1203.61 (df=5)
4	glm overdispersion	2038.054	-1135.795 (df=6)
4	glm Poisson ln(Y)	2696.712	-Inf (df=5)
4	glm overdispersion ln(Y)	2696.712	-684.1042 (df=6)
9	zip by hand	2296.084	N/A
10	zip poisson STRESS ~ COHES+ESTEEM+GRADES+SATTACH | COHES+ESTEEM+GRADES+SATTACH	2035.409	-1134.401 (df=10)
10	zip negative binomial STRESS ~ COHES+ESTEEM+GRADES+SATTACH | COHES+ESTEEM+GRADES+SATTACH	2035.697	-1126.118 (df=11)
10	zip regular STRESS ~ COHES+ESTEEM+GRADES+SATTACH | COHES+ESTEEM+GRADES+SATTACH	2035.409	-1134.401 (df=10)
10	zip poisson STRESS ~ COHES+ESTEEM+GRADES+SATTACH | COHES	2036.261	-1135.139 (df=7)
10	zip negative binomial STRESS ~ COHES+ESTEEM+GRADES+SATTACH | COHES	2035.86	-1126 on 8 Df
10	zip regular STRESS ~ COHES+ESTEEM+GRADES+SATTACH | COHES	2036.501	-1135 on 7 Df
10	zip poisson STRESS ~ COHES | COHES	2104.276	-1147.134 (df=4)
10	zip negative binomial STRESS ~ COHES | COHES	2104.881	-1136.204 (df=5)
10	zip regular STRESS ~ COHES | COHES	2065.82	-1143.841 (df=7)

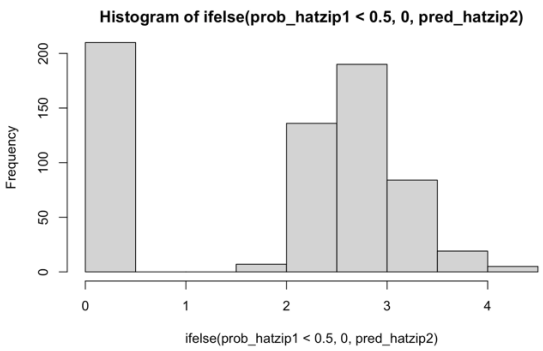
A few notes on the final model:

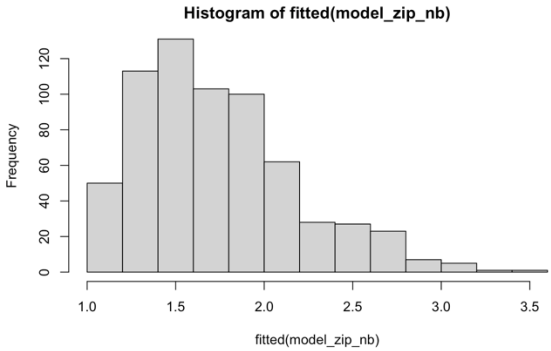
The highest log likelihood (and best performing according to this parameter) models are OLS ln(Y) and glm overdispersion ln(Y). That being said, these models do not do a great job at indicating the presence of 0 event data points, and they have much higher SSE than other models. Sensitivity to 0 event data points is better done in a zip model.





Here we see the zip models performance. The handmade zip model is much more sensitive to identifying 0 event data points. Depending on the context, differentiating between 0 and non-0 events can be crucial in an analysis. In our case, this is not as important, so using our handmade zip model is not necessary. However, the prebuilt zipmodel performs similarly to the OLS and glm overdispersion models in terms of y_hat distribution. All of the prebuilt zip models score much better in terms of lower SSE but moderately in comparison to Log Likelihood.





In terms of methods, zip models are the most logical way to proceed. They are built to take into account both 0 event data points and use a separate model to predict non-0 event datapoints. They perform relatively the best in terms of SSE and moderately in terms of log likelihood (compared to better performing log likelihood stats that have significantly worse SSE values). The model selected was the zip negative binomial model which aligns with our glm overdispersion ln(Y) which performed well in predicting non-0 data points according to Log Likelihood, and only using COHES to predict 0 event data points which aligns with task 8 which indicates the best predictor using BIC. All in all, this model is not perfect, but it does decently well in its goodness of fit evaluators (SSE and Log Likelihood) as well as has the most logical methods to creating it.

Conclusion

In this assignment, we were tasked with working with data that followed a non-normal distribution. More specifically, we looked at a dataset of 650 adolescents in the U.S and how their relationships at home and school affected how many stressful life events they had in a year. As we have typically worked with OLS regressions and generalized linear models, we had to divert from familiar models and explore zip models and generalized linear models that take Poisson and negative binomial distributions. By taking into account sensitivity to 0 vs non-0 based events, these new tool aided greatly in the prediction of Poisson distributed data.

One task from this assignment that was challenging was building a zip model from scratch. By combining both a generalized linear model for a dichotomous outcome with a negative-binomial, generalized linear model for predicting non-zero values, I was able to take control of tweaking sensitivity in the model to how I wanted. It is nice to know exactly what the model is doing, but it takes a lot more time to create a model by hand. I opted to use R packages to create zip models faster and also have access to summary statistics of these models.

Another challenge that I had to overcome in this assignment was evaluating the different models. In the past, we have used R squared for linear regression models or even AUC and2x2 matrices for

dichotomous outcomes to identify type I and type II errors. However, in this case, we were comparing generalized linear models using different distributions as well as zip models which did not all have R2, AUC, or dichotomous outcomes. I first used BIC to aid in predictor selection because it penalizes more on additional variables. On the model comparison, I used SSE to indicate models with less error as well as Log Likelihood which was present in all the different types of models I used.

In the end, there was no model that performed exponentially better than the rest. Ultimately I ended up choosing a zip model which performed greatly in terms of low SSE compared to other models, but performed moderately in terms of Log Likelihood. Any of the other models that performed better in terms of Log Likelihood had significantly higher SSE which indicates that the predicted regression points vary further from their true values. The methodology behind using a zip model for out data distribution was confirmed when creating my own zip model from scratch.

Although this assignment introduced new data distributions, I am proud of the tools I am picking up to overcome new challenges like comparing different model types and building models by hand.