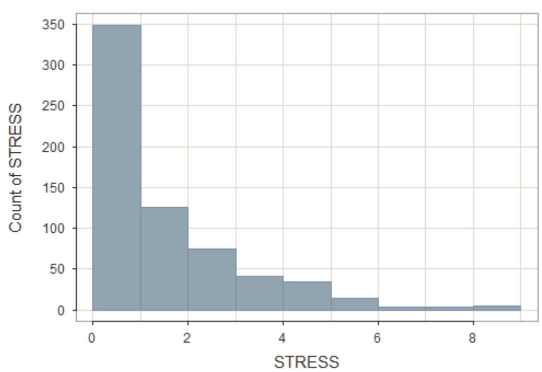
Poisson and Zero-Inflated Poisson Regression

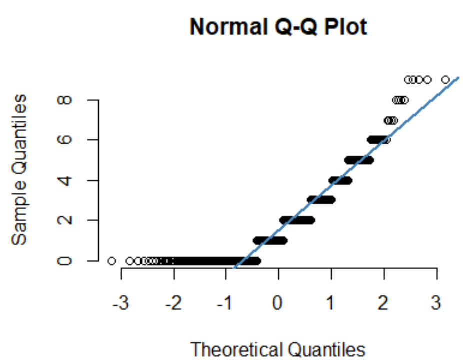
For this assignment, we will be using the STRESS dataset. This includes information from about 650 adolescents in the US who were surveyed about the number of stressful life events they had experienced in the past year (STRESS). STRESS is an integer variable that represents counts of stressful events. 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)

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.

Histogram:





- Based QQplot, STRESS is not normally distributed. Histogram suggest we have a lot of zero associated with STRESS.

Summary statistics:

--- STRESS ---

n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
651	0	1.73	1.85	1.27	1.65	0.00	0.00	1.00	3.00	9.00	3.00

(Box plot) Outliers: 9

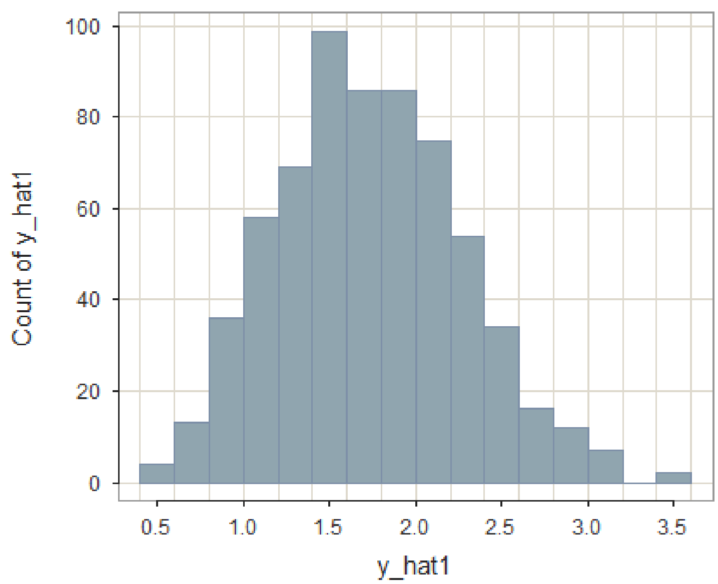
Small	Large
	9.0
	9.0
	9.0
	9.0
	9.0
	8.0
	8.0
	8.0
	8.0

- Poisson distribution: Mean = Variance
- Negative binomial distribution : Mean < Variance
- Since mean is 1.73, variance = sd squared = 1.85 * 1.85= 3.4225. Variance is a bit more than twice the mean. The data is over-dispersed, but of course we haven't considered any covariates yet.
- This is a negative binomial distribution since Mean is less than Variance.

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 = 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 < 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

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
```

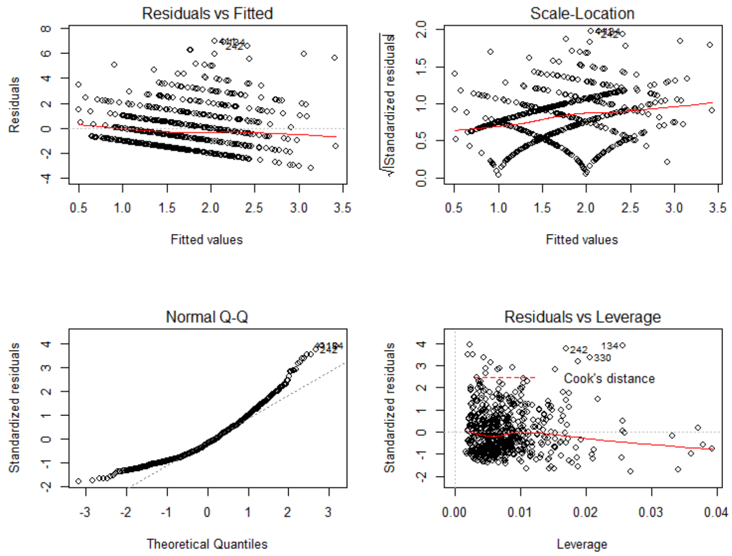
- $Y = 5.713 - 0.023\beta1 - 0.412\beta2 - 0.0471\beta3 - 0.03\beta4$
- Y_hat1 value based on histogram seems to be normal distributed. However, R squared value indicating OLS model can only explain 8.319% of the variance of target variable STRESS.



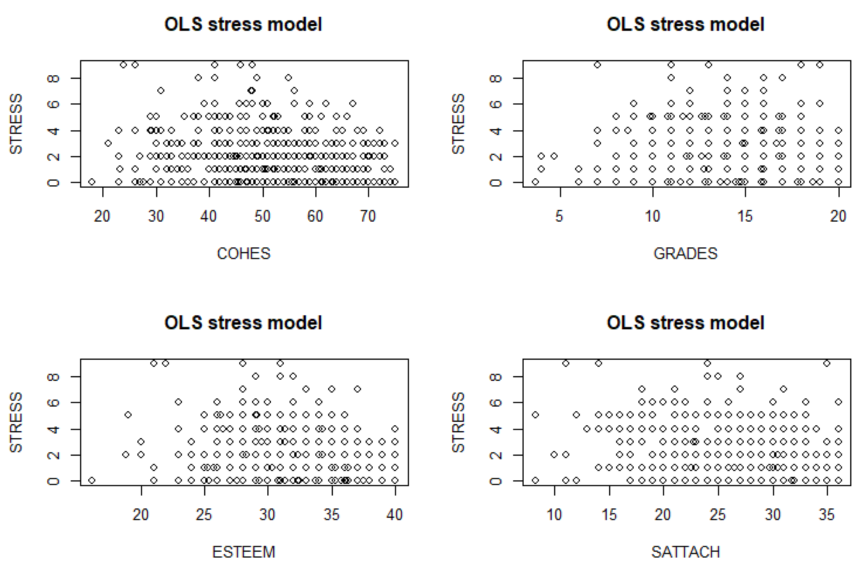
- Closer look at our 2 by 2 diagnostic visualization we noticed a distinct pattern in residuals vs fitted as well as scale-location graph(lattice). This suggests violation of homogeneity of variance. QQ plot shows deviation from 45-degree line which indication of violation of normal distribution. Cook's distance shows we have large numbers of outliers. (242, 1340, 330)

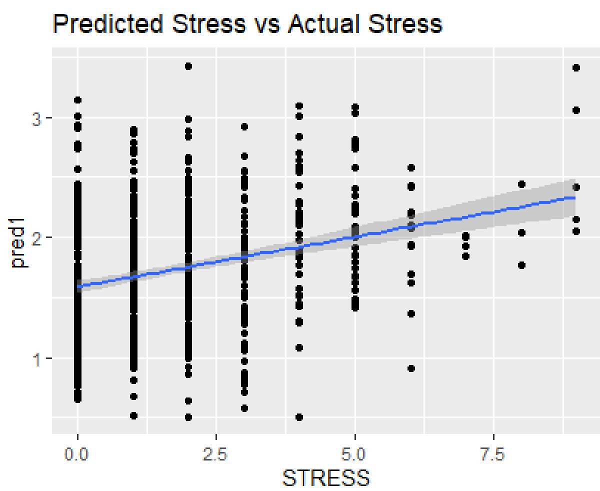
```
No Studentized residuals with Bonferroni p < 0.05
```

```
Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
411 3.963753         0.000082025     0.053398
```



- Plot of individual variables contrasting with STRESS also does not present a distinct relationship.



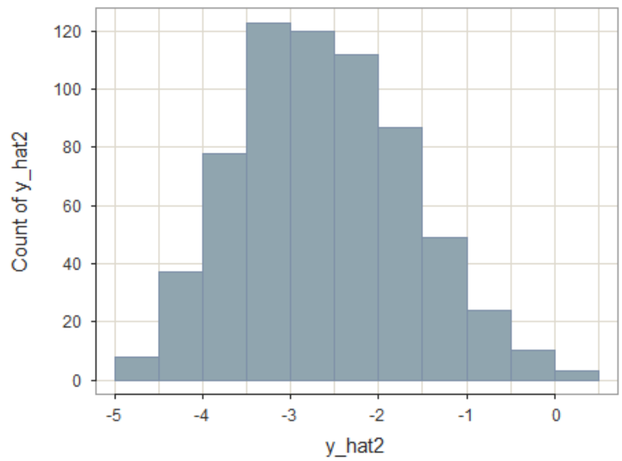


3. Create a transformed variable on Y that is LN(Y). Fit an OLS regression model to predict LN(Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Obtain the typical diagnostic information and graphs. Discuss how well this model fits. Obtain predicted values (LN(Y)_hat) and plot them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

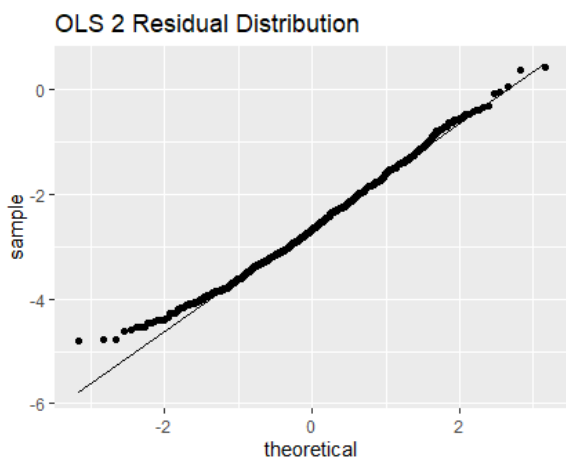
- First was to add 0.0001 to STRESS. Then we can transform the STRESS by avoiding 0. This transformed log Y did not improve our model with R-Squared value. We had 8.319% in the last OLS model, but second model is at 4.142%

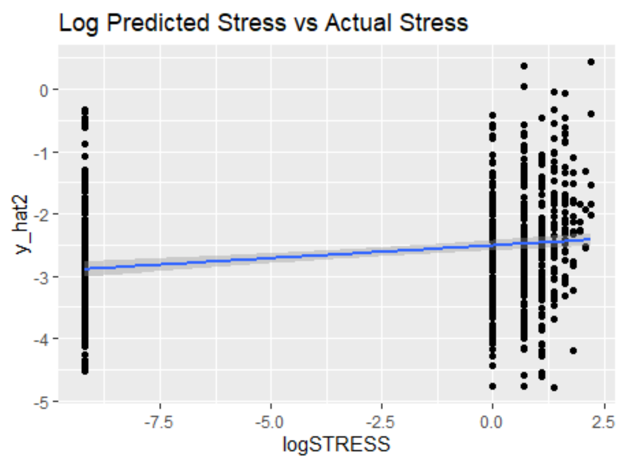
Call:				
lm(formula = logSTRESS ~ COHES + ESTEEM + GRADES + SATTACH, data = mydata)				
Residuals:				
Min	1Q	Median	3Q	Max
-8.893	-5.795	2.539	3.620	6.173
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.20551	1.52906	2.750	0.00612
COHES	-0.04775	0.01850	-2.582	0.01005
ESTEEM	-0.04915	0.05086	-0.966	0.33419
GRADES	-0.06616	0.06188	-1.069	0.28540
SATTACH	-0.06473	0.03716	-1.742	0.08197
Residual standard error: 4.673 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: 0.00001676				

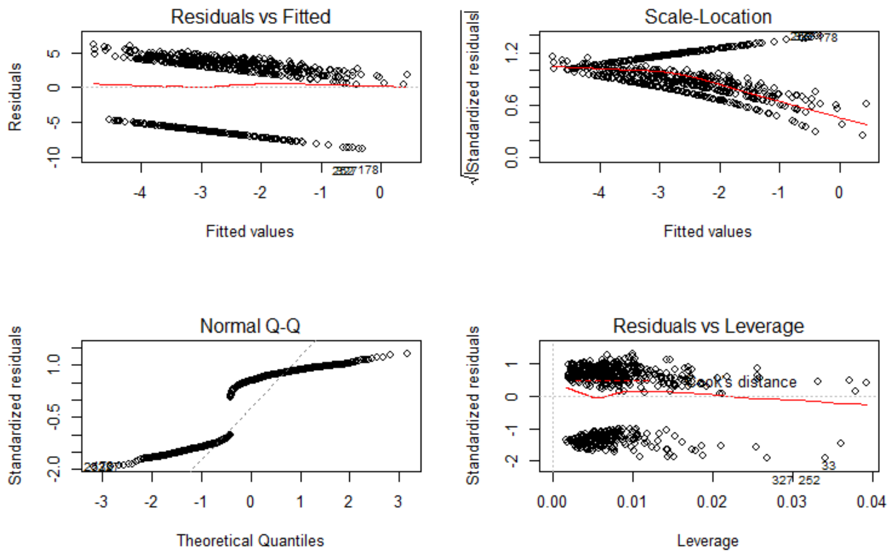
- $Y=4.206 - 0.0478\beta1 - 0.049\beta2 - 0.066\beta3 - 0.065\beta4$



- Based on T-value and P value, intercept and 4 variables are not statistically significant in explaining our target variable logSTRESS.
- Diagnostic graphs do not suggest improvement from last model, and the underlying assumptions are violated across normal distribution, and homogeneity.







- Running our OLS again with transformation but we will run a model by using mydata$level, so we only model against data that indicate stress is not 0.

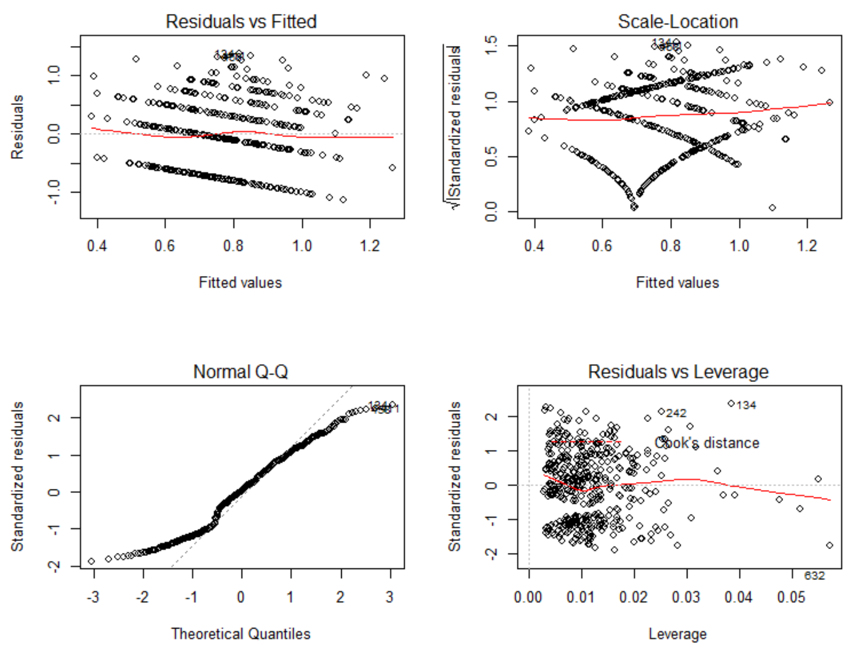
mydata$stressed<-ifelse(mydata$STRESS>0,1,0) #stressed yes or no

mydata$level<-ifelse(mydata$stressed==1, mydata$STRESS,NA) #if not stressed mark it as NA

```
lm(formula = loglevel ~ COHES + ESTEEM + GRADES + SATTACH, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1201 -0.5951  0.0204  0.4655  1.3813

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)  1.960830   0.239249   8.196 0.000000000000000296
COHES       -0.005233   0.002881  -1.816              0.0700
ESTEEM      -0.013694   0.007946  -1.723              0.0855
GRADES      -0.016140   0.009422  -1.713              0.0874
SATTACH     -0.009408   0.005753  -1.635              0.1027

Residual standard error: 0.5955 on 425 degrees of freedom
  (221 observations deleted due to missingness)
Multiple R-squared:  0.0663,    Adjusted R-squared:  0.05752
F-statistic: 7.545 on 4 and 425 DF,  p-value: 0.000007036
```

- Y=1.961 – 0.005β1 – 0.014β2 –0.016β3 – 0.009β4
- Based on T-value and P value, intercept is statistically significant but the 4 variables remained as not statistically significant in explaining our target variable loglevel. R-Squared did slight changed to 6.63% comparing to prior model using logSTRESS.

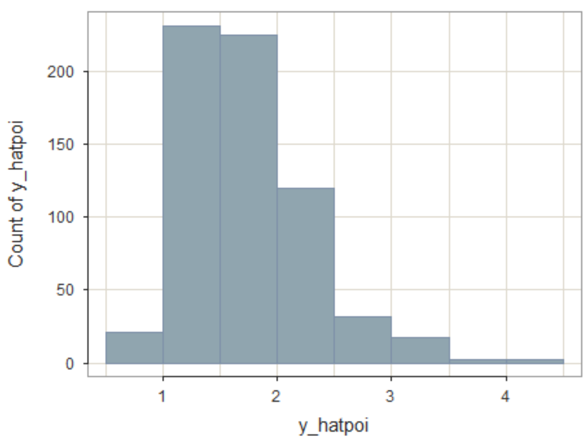
- Diagnostic graphs do not suggest improvement from model 1 or model 2, and the underlying assumptions are violated across normal distribution, and homogeneity. So the overall the goodness of fit is still lacking.

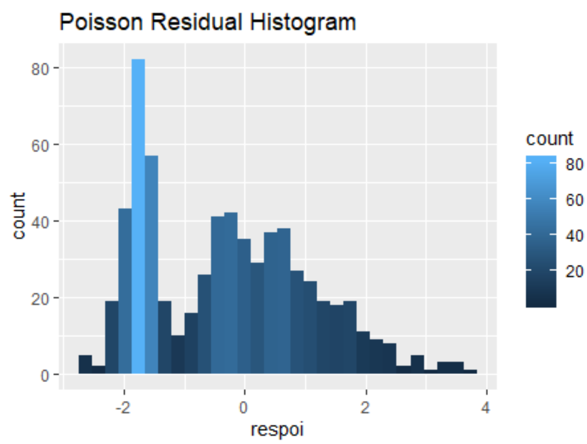


4. Use the glm() function to fit a Poisson Regression for STRESS (Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Interpret the model's coefficients and discuss how this model's results compare to your answer for part 3). Similarly, fit an over-dispersed Poisson regression model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

```
Call:
glm(formula = 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 < 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

(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
```

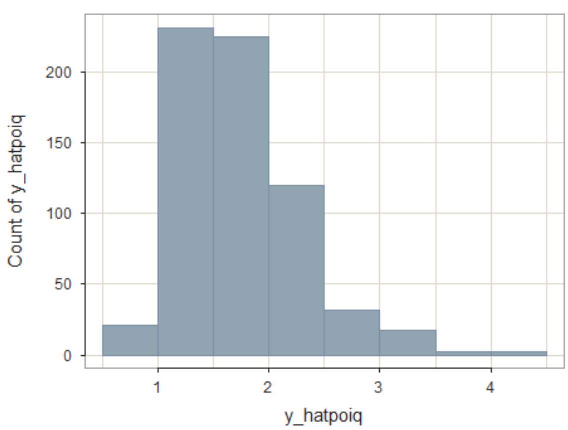
- Y= 2.735 – 0.013β1 – 0.024β2 – 0.23β3 – 0.016β4
- Poisson regression is a log of response Y. so we need to transform the value back for interpretation by using exp() function.
- Intercept exp(2.735)= 15.4. Which is not reasonable considering all variables are 0. Count of stressful events would be 15.4 which is grossly higher than the extreme value in the actual data.
- β1=exp(-0.013) = .987, interpret per unit change of the adolescent's measure of how well is the relationship with the family, this will decrease 0.987 to the count of stressful event encounter.
- B2=exp(-0.024)=0.977, interpret per unit change of the adolescent's measure of self-esteem, this will decrease 0.977 to the count of stressful event encounter.
- B3=exp(-0.023)=0.977, interpret per unit change of the adolescent's past year's school grades, this will decrease 0.977 to the count of stressful event encounter.
- B4=exp(-0.016)=0.984, interpret per unit change of the adolescent's measure of how well adolescent likes or attached to the school, this will decrease 0.984 to the count of stressful event encounter.

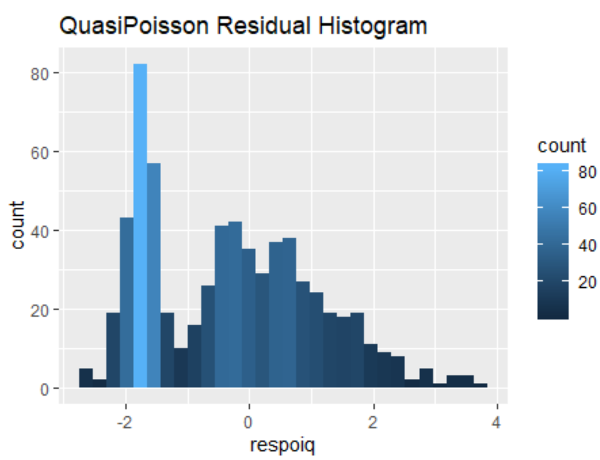




- Histogram of the predicted count of stressful events are clustered around 1-2.5 ranges, which does not represent the actual data.
- Fit a Quasi-Poisson model to see improvement to predicting the count of stressful events.

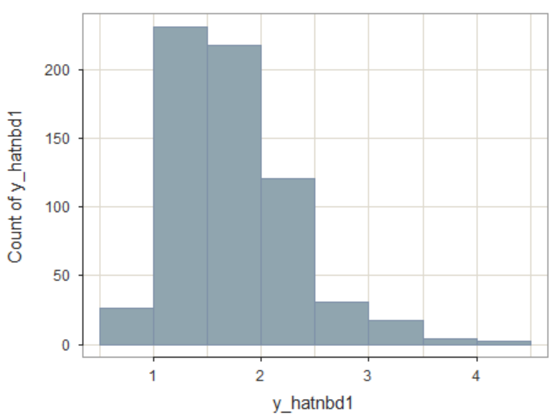
```
glm(formula = STRESS ~ COHES + ESTEEM + GRADES + SATTACH, family = "quasipoisson",
    data = mydata)

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

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  2.734513   0.312160   8.760 < 0.0000000000000002
COHES       -0.012918   0.003858  -3.348              0.00086
ESTEEM      -0.023692   0.010721  -2.210              0.02746
GRADES      -0.023471   0.013157  -1.784              0.07490
SATTACH     -0.016481   0.007712  -2.137              0.03297

(Dispersion parameter for quasipoisson family taken to be 1.778603)

    Null deviance: 1349.8  on 650  degrees of freedom
Residual deviance: 1245.4  on 646  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

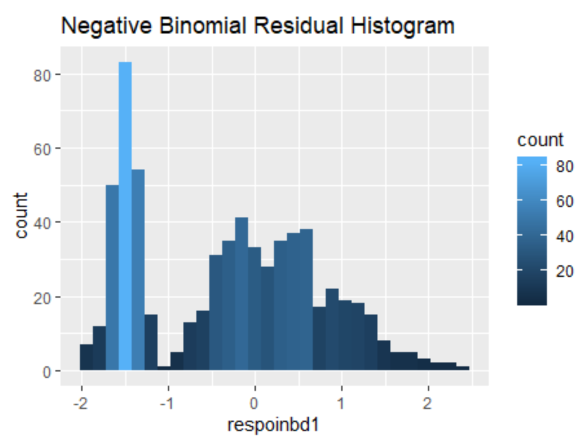




- Histogram and residual graph show not a significant improvement using Quasi Poisson
- Let’s fit to fit negative binomial using the same set of variables since we compared to the mea and variance in our initial evaluation.

```
Call:
glm.nb(formula = STRESS ~ COHES + ESTEEM + GRADES + SATTACH,
    data = mydata, maxit = 100000, 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

(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
```





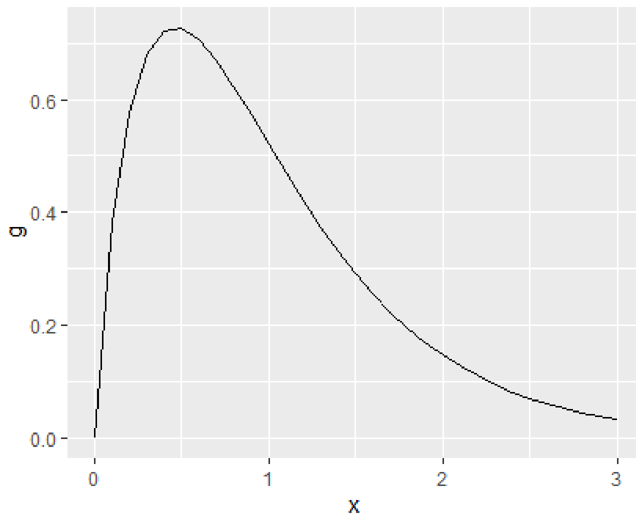
- We see a slight improvement based on the graphs above, we still experience significant residuals between actual versus predicted value. Let's compare these three models closer.

	poisson.coef	quasi.coef	neg.binomial.coef	se.poisson	se.quasi	se.neg.binomial
(Intercept)	2.7345	2.7345	2.7590	0.2341	0.3122	0.3415
COHES	-0.0129	-0.0129	-0.0134	0.0029	0.0039	0.0041
ESTEEM	-0.0237	-0.0237	-0.0231	0.0080	0.0107	0.0115
GRADES	-0.0235	-0.0235	-0.0244	0.0099	0.0132	0.0140
SATTACH	-0.0165	-0.0165	-0.0168	0.0058	0.0077	0.0083

- The negative binomial estimates are not significantly different from those based on the Poisson and QuasiPoisson model, and both sets would lead to the same conclusions.

- Looking at the standard errors across three models, we see that both approaches to over-dispersion lead to very similar estimated standard errors, and that ordinary Poisson regression underestimates the standard errors.

- Unobserved Heterogeneity using dgamma and qgamma with Theta value in negative binomial distribution model.




```
> qgamma((1:3)/4, shape = 1/v, scale = v)
[1] 0.4627460 0.8280858 1.3524631
```

- Adolescents in the US who were surveyed at Q1 of the distribution of unobserved heterogeneity has 54% fewer stressful events than expected from their observed characteristics, while those at the median encountered 17% fewer and those at Q3 encountered 33% more than expected.

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?

- Categorizing the data based on the High Middle and Low, our data can be broken down into these 3 buckets.

```
  High    Low Middle
    99    106    446
```

```
> meanCOHES <- mean(mydata$COHES)
> sdCOHES <- sd(mydata$COHES)
> COHESlow<-meanCOHES-sdCOHES
> COHESlow
```

```
[1] 41.62096

> COHEShigh<-meanCOHES+sdCOHES
> COHEShigh
[1] 64.38757
```

- Calculating the threshold of low and high: 41.621 and 64.388
- Plug them into the Y= 2.735 – 0.013β1 – 0.024β2 – 0.23β3 – 0.016β4

```
> pred_COHES_low<-2.735 - 0.013*COHESlow- 0.024*0 - 0.23*0 - 0.016*0
> pred_COHES_low
[1] 2.193928

> pred_COHES_high<-2.735 - 0.013*COHEShigh- 0.024*0 - 0.23*0 - 0.016*0
> pred_COHES_high
[1] 1.897962
```

- Since Poisson is a log function of response Y, so transform it back using exp()

```
> exp(pred_COHES_low)
[1] 8.970376
> exp(pred_COHES_high)
[1] 6.67228
```

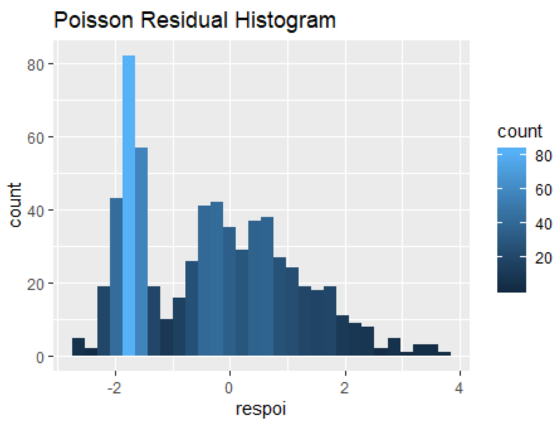
- (exp(pred_COHES_high)-exp(pred_COHES_low))/exp(pred_COHES_low) = **-0.2561872**
- There is 25.619% difference in the number of stressful events for those at high and low levels of family cohesion.

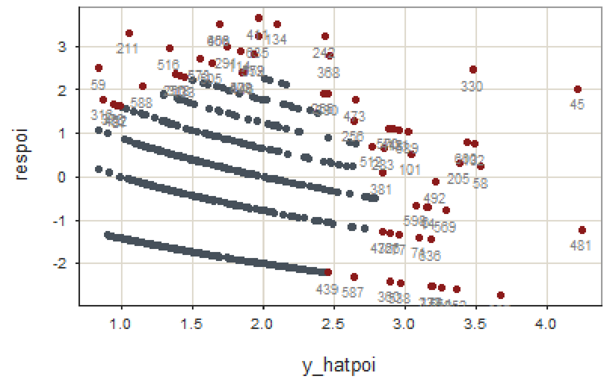
6. Compute the AICs and BICs from the Poisson Regression and the over-dispersed Poisson regression models from part 4). Is one better than the other?

- Negative binomial is better performing than the Poisson regression model based on AIC and BIC values.

```
poisson_AIC   quasi_AIC     nbd_AIC
   2417.219          NA    2283.590
poisson_BIC   quasi_BIC     nbd_BIC
   2439.612          NA    2310.461
```

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.





- Based on the graphs we can see there are lot of outliers, top 10% are highlighted in red on the right side of the graph.
- Poisson regression is also having difficulty predicting zero by displaying large residuals. This leads us to consider modeling the zeros, and the other counts separately.

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?

I previously created a new variable:

mydata$stressed<-ifelse(mydata$STRESS>0,1,0) #stressed yes or no

```
glm(formula = stressed ~ COHES + ESTEEM + GRADES + SATTACH, family = binomial,
    data = mydata)

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

(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
```

- $Y=3.5167 - 0.0207\beta 1 - 0.0189\beta 2 - 0.0255\beta 3 - 0.0277\beta 4$

- β1 is COHES = measure of how well the adolescent gets along with their family

```
➢ odds_dir1 <- round(exp(-0.020733) - 1, digits = 3)
➢ odds_dir1
[1] -0.021
```

- β2 is ESTEEM = measure of self-esteem

```
➢ odds_dir2 <- round(exp(-0.018867) - 1, digits = 3)#ESTEEM
➢ odds_dir2
[1] -0.019
```

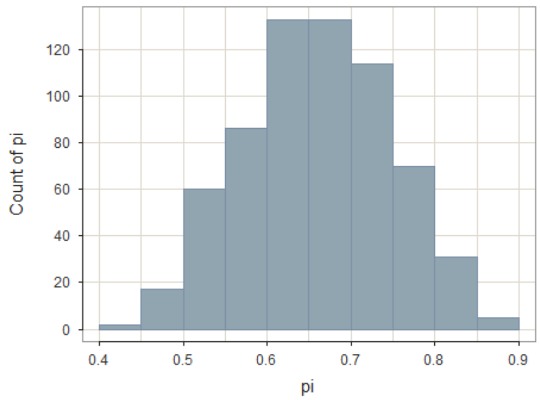
- β3 is GRADES = past year's school grades

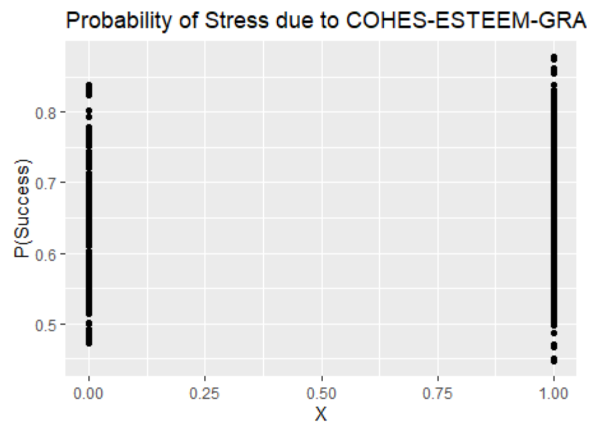
```
➢ odds_dir3 <- round(exp(-0.025492 ) - 1, digits = 3)#GRADES
➢ odds_dir3
[1] -0.025
```

- β4 is SATTACH = measure of how well the adolescent likes and is attached to their school

```
➢ odds_dir4 <- round(exp(-0.027730 ) - 1, digits = 3)#SATTACH
➢ odds_dir4
[1] -0.027
```

- With odds direction created for each variable. It makes the interpretation much easier.
- Interpretation is for change in the adolescent's relationship with the family, as it improves it will decrease by 2.1% of encountering stressful event while everything else is 0.
- For every change in the adolescent's measure of self-esteem, as it improves it will decrease by 1.9% of encountering stressful event while everything else is 0.
- For every change in the adolescent's past year's school grades, as it improves it will decrease by 2.5% of encountering stressful event while everything else is 0.
- For every change in the adolescent's measure of how well adolescent likes or attached to the school, as it improves it will decrease by 2.7% of encountering stressful event while everything else is 0.
- Using probability of success, we can see quite a high probability that's above 50%. But the model is still not predicting 0 well.
- AIC is 821.79 which is the lowest comparing to all previous model. But given the statistically significance of the variables to predict the stress is low. We should consider a combined model in predicting stressed versus not stressed, since we have significant amount of stress occurrence at 0.





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!

- ZIP which is zero inflated poisson regression is a good approach when the data has high counts of zero. We have conducted the first part of the logistic regression previous using “stressed” with 4 explanatory variables. For this round, we will add AGE to the model and see if we can further improve the model by looking at AIC value. Then we will use “level” to review the counts of the adolescents who are stressed.
- Running the logistic model again by adding “AGE”. Unfortunately, we do not see an improvement of the model, and AIC score decreased. Ran through interactive method to by adding one explanatory variable at the time and rank each model as below. Using COHES and ESTEEM we can get to almost identical AIC score rather than using all 4 explanatory variables.

		AIC	RANK
logistic 1	`stressed ~ COHES + ESTEEM + GRADES + SATTACH`	821.79	1
logistic 2	`stressed ~ COHES + ESTEEM + GRADES + SATTACH + AGE,`	822.56	5
logistic 3	`stressed ~ COHES`	821.86	3
logistic 4	`stressed ~ COHES + ESTEEM`	821.80	2
logistic 5	`stressed ~ COHES + ESTEEM + GRADES`	822.32	4

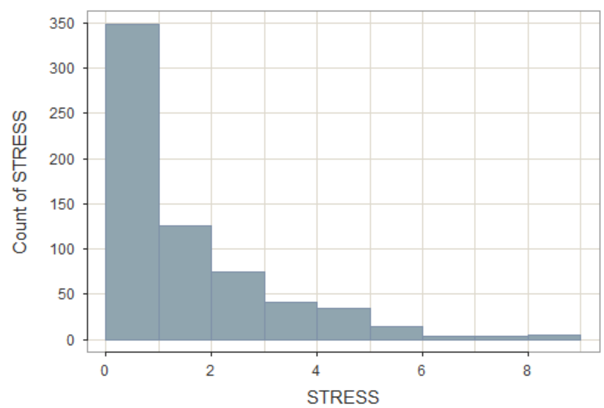
```
Call:
glm(formula = stressed ~ COHES + ESTEEM + GRADES + SATTACH +
    AGE, family = binomial, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9629  -1.3248   0.7718   0.9292   1.2957

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.668426   1.281438   3.643 0.000269
COHES       -0.020527   0.008761  -2.343 0.019125
ESTEEM      -0.018696   0.023771  -0.786 0.431577
GRADES      -0.022959   0.028789  -0.797 0.425164
SATTACH     -0.030743   0.017813  -1.726 0.084361
AGE         -0.089849   0.081274  -1.106 0.268939

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 834.18  on 650  degrees of freedom
Residual deviance: 810.56  on 645  degrees of freedom
AIC: 822.56

Number of Fisher Scoring iterations: 4
```

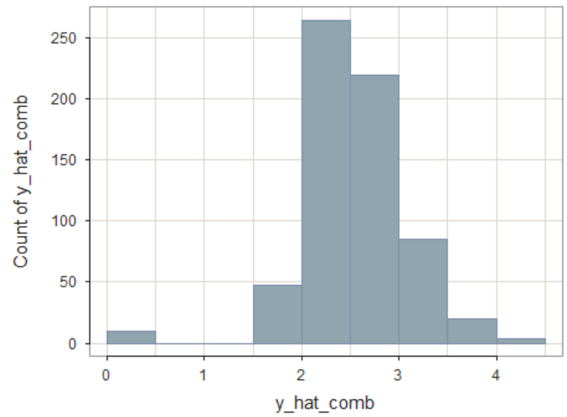
- Let’s fit the Poisson regression against “level” in predicting the count of stress.

Interactive approach by evaluating all variable for Poisson landed us with three variables achieving lowest AIC. When then using negative binomial, we did not see improvement. We will settle with Poisson #5 as the second component of the model.

		AIC	**RANK**
poisson 1	level ~ COHES + ESTEEM + GRADES + SATTACH	1552.10	2
poisson 2	level ~ COHES + ESTEEM + GRADES + SATTACH + AGE,	1553.10	3
poisson 3	level ~ COHES	1560.60	6
poisson 4	level ~ COHES + ESTEEM	1553.30	4
poisson 5	level ~ COHES + ESTEEM + GRADES	1551.90	1
nbd 1	level ~ COHES + ESTEEM + GRADES	1553.90	5

- Using the combined logistic 4 model and poisson 5 model. Our combined model produced the following histogram.





- Reviewing the actual versus the prediction of the model, we can see the model is still grossly missing the ability to predict 0 and failed to predict the stress level/count of encounter of adolescent after 5 completely.

STRESS count

Bin	Midpnt	Count	Prop	Cumul.c	Cumul.p
0 > 1	0.5	349	0.54	349	0.54
1 > 2	1.5	125	0.19	474	0.73
2 > 3	2.5	75	0.12	549	0.84
3 > 4	3.5	41	0.06	590	0.91
4 > 5	4.5	34	0.05	624	0.96
5 > 6	5.5	14	0.02	638	0.98
6 > 7	6.5	4	0.01	642	0.99
7 > 8	7.5	4	0.01	646	0.99
8 > 9	8.5	5	0.01	651	1.00

Y_hat_comb

Bin	Midpnt	Count	Prop	Cumul.c	Cumul.p
0.0 > 0.5	0.25	10	0.02	10	0.02
0.5 > 1.0	0.75	0	0.00	10	0.02
1.0 > 1.5	1.25	0	0.00	10	0.02
1.5 > 2.0	1.75	47	0.07	57	0.09
2.0 > 2.5	2.25	265	0.41	322	0.49
2.5 > 3.0	2.75	220	0.34	542	0.83
3.0 > 3.5	3.25	85	0.13	627	0.96
3.5 > 4.0	3.75	20	0.03	647	0.99
4.0 > 4.5	4.25	4	0.01	651	1.00

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.

- For this exercise I will simply used all 4 variables and review models between using Poisson versus negative binomial distribution for the stress count predictability.

Zip poission

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

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)  2.641693   0.272349   9.700 < 0.0000000000000002
COHES       -0.008258   0.003416  -2.418              0.01561
ESTEEM      -0.026068   0.009206  -2.832              0.00463
GRADES      -0.019553   0.010914  -1.792              0.07320
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) -2.835428   0.983250  -2.884  0.00393
COHES        0.018917   0.012124   1.560  0.11869
ESTEEM      -0.004328   0.032777  -0.132  0.89496
GRADES       0.014330   0.037731   0.380  0.70410
SATTACH      0.024838   0.024083   1.031  0.30238

Number of iterations in BFGS optimization: 16
Log-likelihood: -1134 on 10 Df
```

Zip negative binomial

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

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.2579 -0.8679 -0.2072  0.5817  3.8088

Count model coefficients (negbin with log link):
             Estimate Std. Error z value            Pr(>|z|)
(Intercept)  2.694064   0.343585   7.841 0.00000000000000447
COHES       -0.009199   0.004300  -2.139              0.0324
ESTEEM      -0.026367   0.011510  -2.291              0.0220
GRADES      -0.021654   0.013400  -1.616              0.1061
SATTACH     -0.011805   0.008239  -1.433              0.1519
Log(theta)   1.745944   0.348664   5.008 0.00000055132612184

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.97479    1.34073  -2.219   0.0265
COHES        0.02057    0.01709   1.203   0.2289
ESTEEM      -0.01158    0.04470  -0.259   0.7955
GRADES       0.01013    0.04871   0.208   0.8352
SATTACH      0.02564    0.03321   0.772   0.4402

Theta = 5.7313
Number of iterations in BFGS optimization: 1
Log-likelihood: -1126 on 11 Df
```

- Calculating the ZIP models above in terms of AIC and comparing with previous Poisson and NBD models. We can see the ZIP model using negative binomial distribution produces the best result.

	AIC
Poisson Model	2417.219
NBD Model	2283.590
ZIP_Poisson Model	2288.802
ZIP_NBD Model	2274.236

- Non-nested hypothesis test is another handy way of reviewing the models.

```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A  p-value
Raw                    -1.97073 model2 > model1 0.024377
AIC-corrected          -1.97073 model2 > model1 0.024377
BIC-corrected          -1.97073 model2 > model1 0.024377
```

- Refit the model by using the manual approach from #9. Using AIC as the goodness of fit, that negative binomial distribution ZIP model using all 4 explanatory variables produced slightly better result.

Variables Selection	Model Name	AIC	RANK
COHES + ESTEEM + GRADES + SATTACH	Poisson Model	2417.219	6
COHES + ESTEEM + GRADES + SATTACH	NBD Model	2283.590	3
COHES + ESTEEM + GRADES + SATTACH	ZIP_Poisson Model	2288.802	4
COHES + ESTEEM + GRADES + SATTACH	ZIP_NBD Model	2274.236	1
COHES + ESTEEM | COHES + ESTEEM + GRADES	ZIP_Poisson Model_2	2292.431	5
COHES + ESTEEM | COHES + ESTEEM + GRADES	ZIP_NBD Model_2	2276.416	2

- ZIP_NBD model and ZIP_NBD Model_2 comparison based on non-nested hypothesis testing statistics. If using AIC, ZIP_NBD model is better than ZIP_NBD Model_2. But if looking BIC, ZIP_NBD Model_2 is better than ZIP_NBD.

```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A  p-value
Raw                   1.3574952 model1 > model2 0.087312
AIC-corrected         0.3618523 model1 > model2 0.358731
BIC-corrected        -1.8676459 model2 > model1 0.030906
```

- In conclusion, we explored the modeling thought process and experiment in evaluating a dataset that presents a lot of zero in the response variable. For this dataset STRESS, explanatory variables failed to produce a robust model in predicting the encounter of number of stressful events. As much as the measurement of how well the adolescent gets along with their family is slightly more statistically significant than the other three variables. We still found the final model in predicting stress response of 0, in addition to the count of stressful events taken place was grossly underperforming when leveraging the zero inflated Poisson model. None the less, this exercise laid out the foundational framework on how to handle real world data set if it represents these characteristics.