NORTHWESTERN UNIVERSITY

SCHOOL OF PROFESSIONAL STUDIES

Week 9 Assignment - Computational: Poisson and Zero-Inflated Poisson Regression
MSDS 410

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

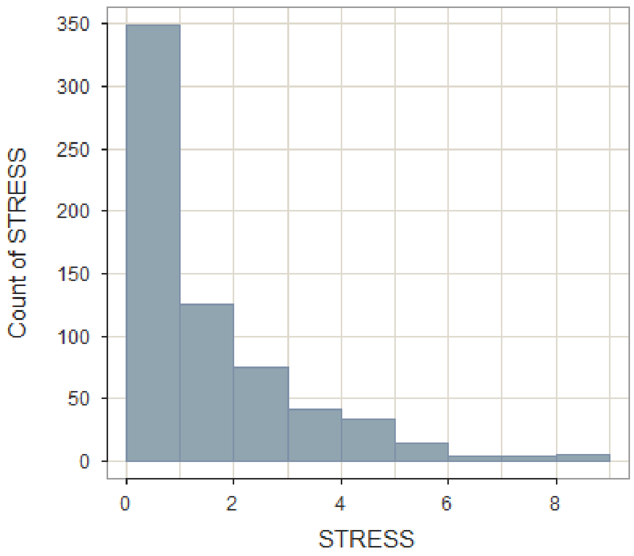
Any computations that involve "the log function", denoted by log(x), ***are always meant to mean the natural log function (which will show as ln() on a calculator).*** The only time that you should ever use a log function other than the natural logarithm is if you are given a specific base.

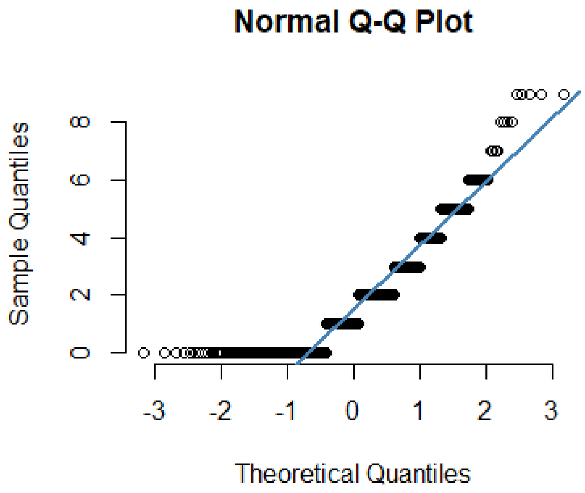
For this assignment, we will be using the STRESS dataset. This includes information from about 650 adolescents in the US who were surveyed about the number of stressful life events they had experienced in the past year (STRESS). STRESS is an integer variable that represents counts of stressful events.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

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

Each problem is worth 5 points.

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





```
--- STRESS ---

        n    miss      mean       sd      min      mdn      max
      651       0      1.73     1.85     0.00     1.00     9.00
```

Stress is not normally distributed. According to the histogram chart and summary of statitics (mean = 1.73 > median 1), it is highly right skewed. QQ plot confirms that the stress variable is not a continous but discrete variable. Since stress response varaible is a count/discrete data and has an excess of zero count, we should use zero-inflated poisson (ZIP) regression. Furthermore, the excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently. Thus, the zip model has two parts, a Poisson count model and the logit model for predicting excess zeros.

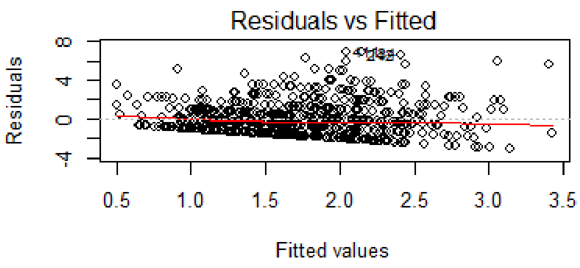
2. Fit an OLS regression model to predict STRESS (Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Obtain the typical diagnostic information and graphs. Discuss how well this model fits. Obtain predicted values (Y_hat) and plot them in a histogram. What issues do you see?

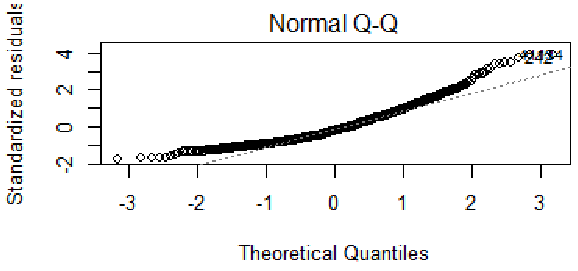
- `Y_hat <- 5.713 - 0.023*COHES - 0.041*ESTEEM - 0.042*GRADES - 0.030*SATTACH`
- Interpretation:
  - Family cohesion reduces the count of stressful events by 0.23 for each level increase.
  - Esteem reduces the count of stressful events by 0.041 for each level increase in self-esteem.
  - Grades reduces the count of stressful events by 0.042 for each increase in previous year grades.
  - School attachment reduces the count of stressful event by 0.030 for each increase in attachment level.
- R-squared is 0.083 and we can see how much of the variation in stressful events is actually explained by those four predictors. The answer is not much. Those fours varaibles explains

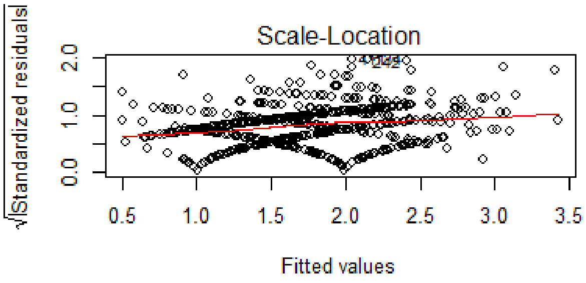
only about 8.3% of the variation in stressful events. That means that 91.7 percent of the stressful events variation for these users is left unexplained.

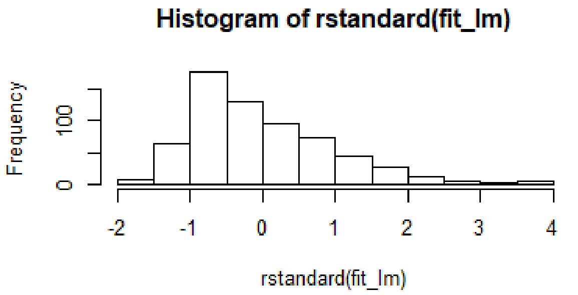
- P-values of esteem, grades, and school attachment predictors are high. With 95% confidence (two-tail test), we do not reject the null hypothesis and conclude that those three variables are not significantly helping us in predicting the stressful events.
- Residuals are not spread equally especially close to zero area. According to the QQ-plot the residuals are not normally distributed in both ends. The residual histogram has long tail in right direction, which confirms non-normality of residual distribution.
- Predicted stress value which is Y_hat is somewhat distributed normally. It did not predict excess zero and distribution does not look the same as original data. It is not predicting more than four stressful events. Histogram range is between 0 and 4.

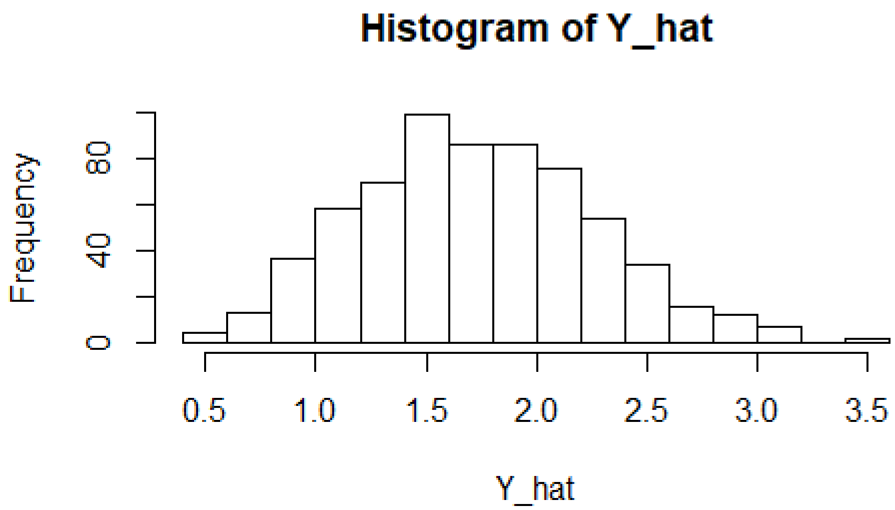
```
Call:
lm(formula = STRESS ~ COHES + ESTEEM + GRADES + SATTACH, data = mydataS)

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
```





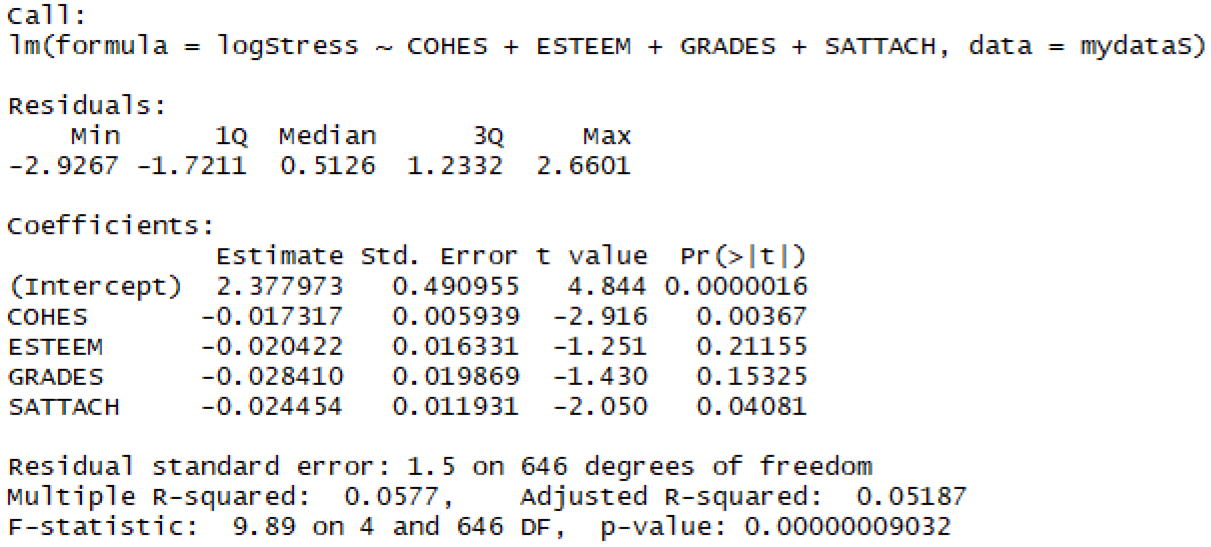




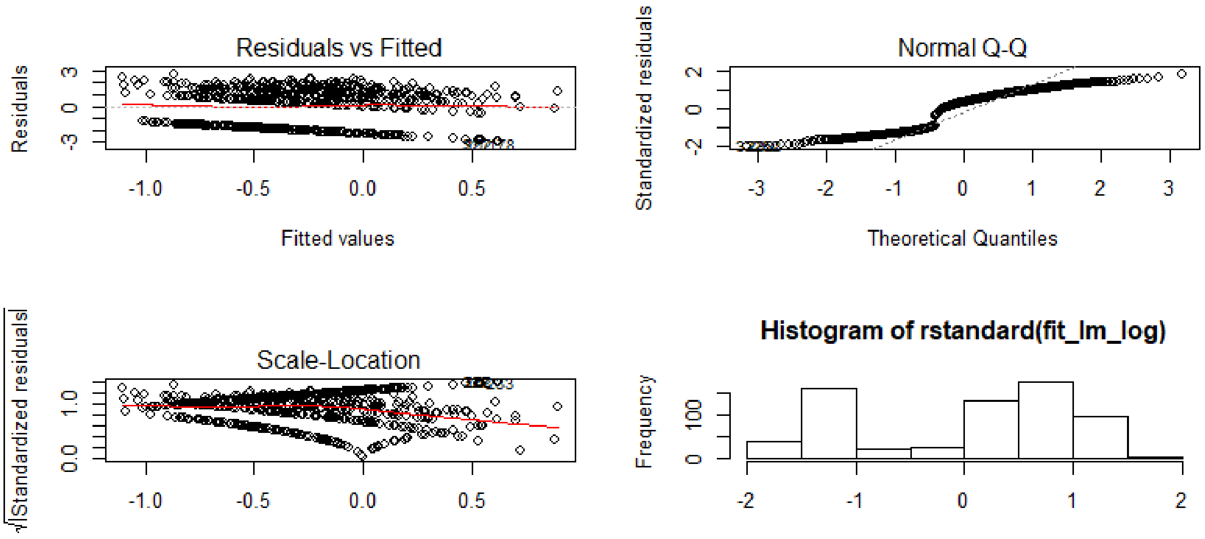


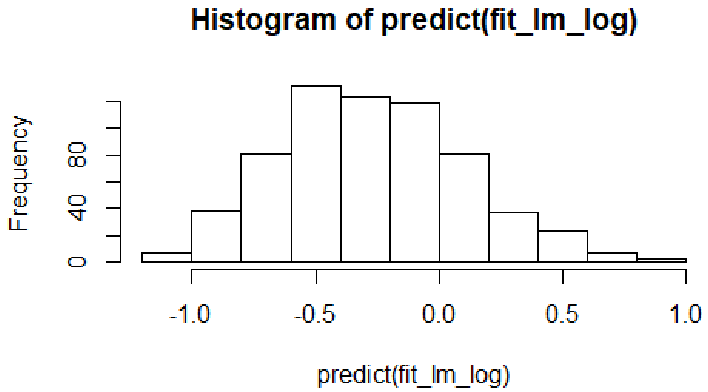
3. Create a transformed variable on Y that is LN(Y). Fit an OLS regression model to predict LN(Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Obtain the typical diagnostic information and graphs. Discuss how well this model fits. Obtain predicted values (LN(Y)_hat) and plot them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

- When we do log (ln) transformation, there is an issue with 0 values. We converted our zero values into 0.1.
- `mydataS$logStress<-log(ifelse(mydataS$STRESS == 0,0.1, mydataS$STRESS))`
- `Y_hat_log <- 2.378 - 0.017*COHES - 0.020*ESTEEM - 0.028*GRADES - 0.025*SATTACH`
- R-squared value worsen, and those three variables did not improve in predicting the count of stressful events. Residuals are not distributed normally. The natural log corresponds to viewing variation through relative or percentage changes rather than through absolute changes. We can also see that predicted histogram ranges from -1 to 1.



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

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9267 -1.7211  0.5126  1.2332  2.6601 

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  2.377973   0.490955   4.844 0.0000016
COHES       -0.017317   0.005939  -2.916   0.00367
ESTEEM      -0.020422   0.016331  -1.251   0.21155
GRADES      -0.028410   0.019869  -1.430   0.15325
SATTACH     -0.024454   0.011931  -2.050   0.04081

Residual standard error: 1.5 on 646 degrees of freedom
Multiple R-squared:  0.0577,	Adjusted R-squared:  0.05187 
F-statistic:  9.89 on 4 and 646 DF,  p-value: 0.00000009032
```

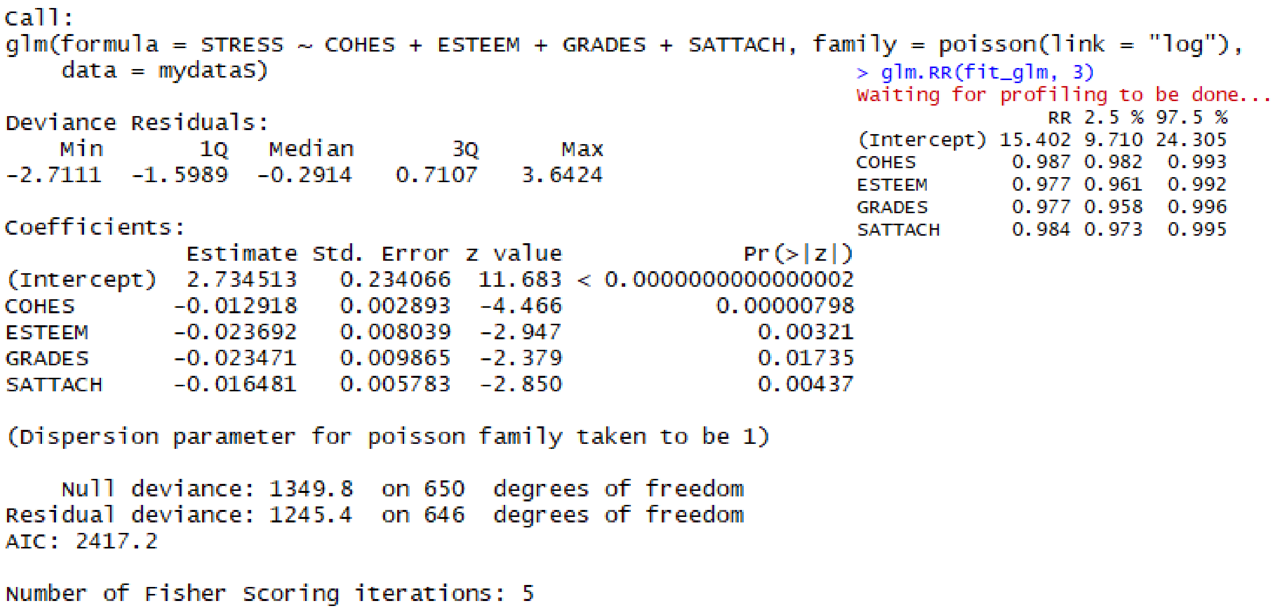




4. Use the glm() function to fit a Poisson Regression for STRESS (Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Interpret the model's coefficients and discuss how this model's results compare to your answer for part 3). Similarly, fit an over-dispersed Poisson regression model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

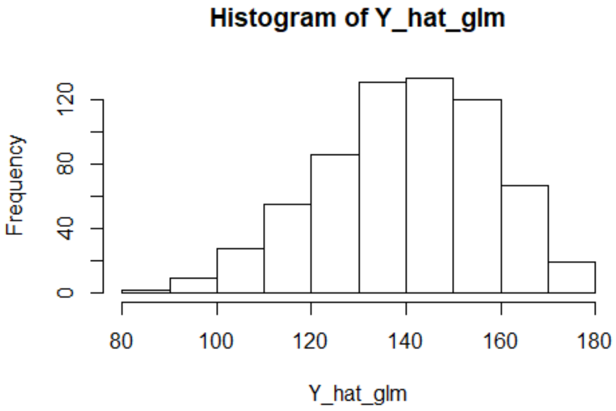
   - Y_hat_glm = 2.735 - 0.013*COHES - 0.024*ESTEEM - 0.023*GRADES - 0.017*SATTACH
   - Goodness Fit:
     - Res. Dev/df = 1245.4/646 = 1.927 => not good fit because of missing predictor, missing data, or overdispersion
     - If the model goodness fit number is close to one, then this model is a good fit. In this case, it is bigger than 1, which is bad.
   - Chi-square test :
     - pvalue <- 1 - pchisq((fit_glm$null.deviance-fit_glm$deviance), (fit_glm$df.null-fit_glm$df.residual))
     - pvalue = 0

    - There is no F test when we deal with count data /categorical data then it's all about chi squared value. Since the p-value is 0, the model is significantly better than the intercept. So overall model is significant.
- Interpretation:
    - estimated event rate when all predictors are 0, it will be exp(2.735) =15.402
    - a one unit increase in COHES reduces STRESS by 0.987% exp(-0.013) = 0.987
    - a one unit increase in ESTEEM reduces STRESS by 0.977% exp(-0.024) = 0.977
    - a one unit increase in GRADES reduces STRESS by 0.977% exp(-0.023) = 0.97)
    - a one unit increase in SATTACH reduces STRESS by 0.984% exp(-0.016) = 0.984



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

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
```

```
> glm.RR(fit_glm, 3)
Waiting for profiling to be done...
                RR 2.5 % 97.5 %
(Intercept) 15.402 9.710 24.305
COHES        0.987 0.982  0.993
ESTEEM       0.977 0.961  0.992
GRADES       0.977 0.958  0.996
SATTACH      0.984 0.973  0.995
```

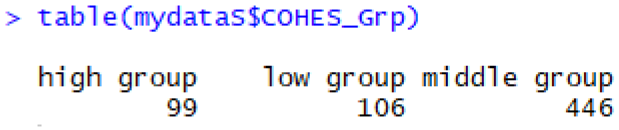


Over-dispersed Poisson regression model:

`Y_hat_OD = 2.579- 0.0134*COHES – 0.0231*ESTEEM – 0.0244*GRADES -0.0168*SATTACH`

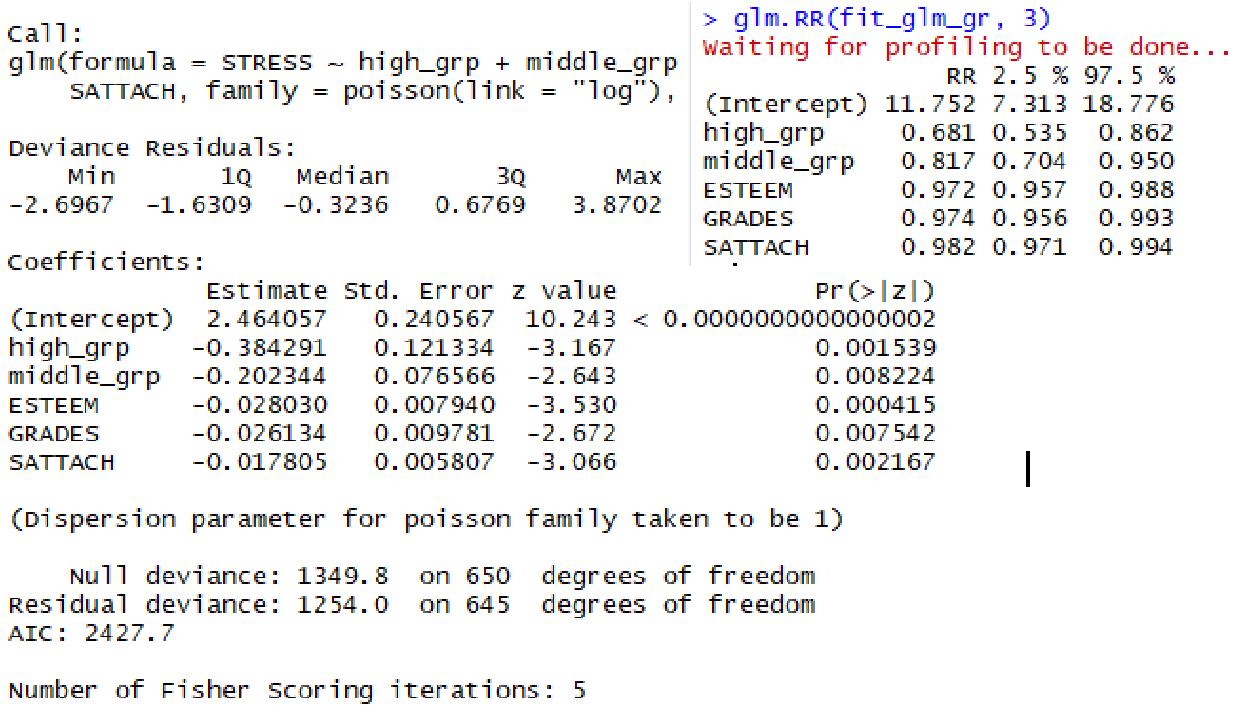
Th over-dispersed Poisson model is very similar to the previous model (`Y_hat_glm`). If we round them, they will be almost the same. However, I have lower AIC score compare to the previous one which it means the model has stronger predictability than the previous one. The difference of AIC: 2417 – 2284 = 133.

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?

- 

```
> table(mydataS$COHES_Grp)

  high group    low group middle group 
          99          106          446 
```

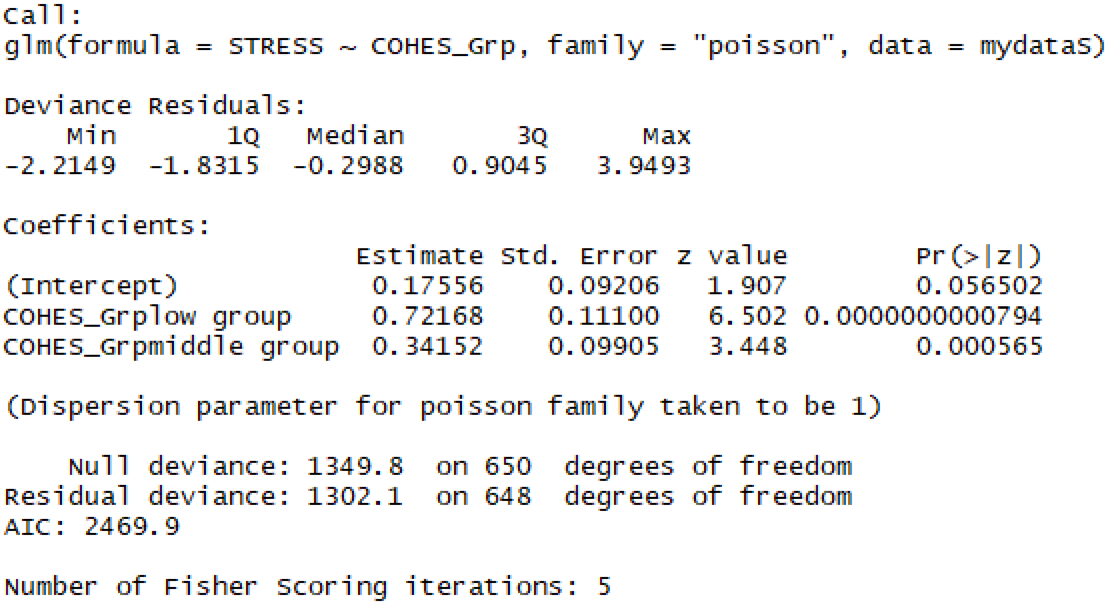
- `Y_hat_glm_grp = 11.752 + 0.681*high_grp + 0.817*middle_grp + 0.972*ESTEEM + 0.974*GRADES + 0.982*SATTACH`
- Goodness Fit: `Res. Dev/df = 1254/645 = 1.944` => not good fit because of missing predictor, missing data, or overdispersion
- Chi-square test :
    - `pvalue <- 1 - pchisq((fit_glm_gr$null.deviance-fit_glm_gr$deviance), (fit_glm_gr$df.null-fit_glm_gr$df.residual))`
    - `pvalue = 0`
- The stressful event ratio RR is 0.681 for high level of family cohesion and 0.817 for middle level of family cohesion. We know that baseline family cohesion is low level because it is coded as a zero. So, after adjusting for the esteem, grades, and s attach variables, the risk ratio is 0.681 for high and 0.817 low levels of family cohesion compare to low level. After adjusting for the those three varaibles, the ratio in high level 0.681 times higher than low level. Another way to say that if I take the exact same 3 varaibles between low level and high level, the stressful event is 0.681 times higher in high level of family cohesion; so that's what I mean by after adjusting for the other variables.



```
Call:
glm(formula = STRESS ~ high_grp + middle_grp
    SATTACH, family = poisson(link = "log"),

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6967  -1.6309  -0.3236   0.6769   3.8702

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept)  2.464057   0.240567  10.243 < 0.0000000000000002
high_grp    -0.384291   0.121334  -3.167             0.001539
middle_grp  -0.202344   0.076566  -2.643             0.008224
ESTEEM      -0.028030   0.007940  -3.530             0.000415
GRADES      -0.026134   0.009781  -2.672             0.007542
SATTACH     -0.017805   0.005807  -3.066             0.002167

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1349.8  on 650  degrees of freedom
Residual deviance: 1254.0  on 645  degrees of freedom
AIC: 2427.7

Number of Fisher Scoring iterations: 5
```

```
> glm.RR(fit_glm_gr, 3)
Waiting for profiling to be done...
                RR 2.5 % 97.5 %
(Intercept) 11.752 7.313 18.776
high_grp     0.681 0.535  0.862
middle_grp   0.817 0.704  0.950
ESTEEM       0.972 0.957  0.988
GRADES       0.974 0.956  0.993
SATTACH      0.982 0.971  0.994
```

If we use family cohesion grouping only to answer the question, high (intercept/baseline) = 0.176 and low = 0.17556 + 0.72168 = 0.897. The percentage difference of high and low is (high – low)/((high + low)/2) = -1.344



```
Call:
glm(formula = STRESS ~ COHES_Grp, family = "poisson", data = mydataS)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2149  -1.8315  -0.2988   0.9045   3.9493

Coefficients:
                     Estimate Std. Error z value        Pr(>|z|)
(Intercept)           0.17556    0.09206   1.907        0.056502
COHES_Grplow group    0.72168    0.11100   6.502 0.0000000000794
COHES_Grpmiddle group 0.34152    0.09905   3.448        0.000565

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1349.8  on 650  degrees of freedom
Residual deviance: 1302.1  on 648  degrees of freedom
AIC: 2469.9

Number of Fisher Scoring iterations: 5
```

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?

   Task # 4 Model 1

   - `AIC(fit_glm) [1] 2417.219`

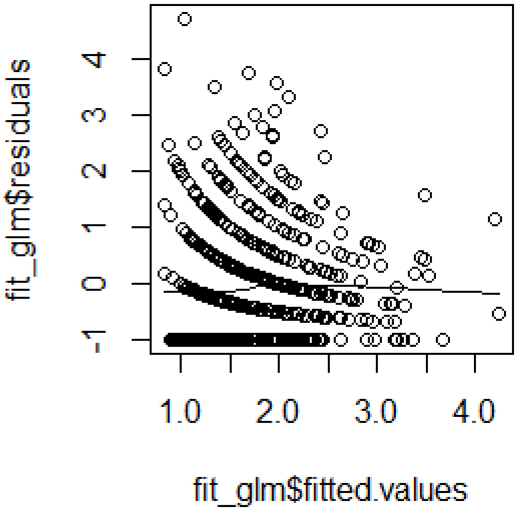
- BIC(fit_glm) [1] 2439.612

Task # 4 Model 2

- AIC(fit_glm_2) [1] 2283.590
- BIC(fit_glm_2) [1] 2310.461

The lowest AIC and BIC values are found in the second model (over-dispersed) of task 4, which means the second model has stronger predictability than the first one.

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.

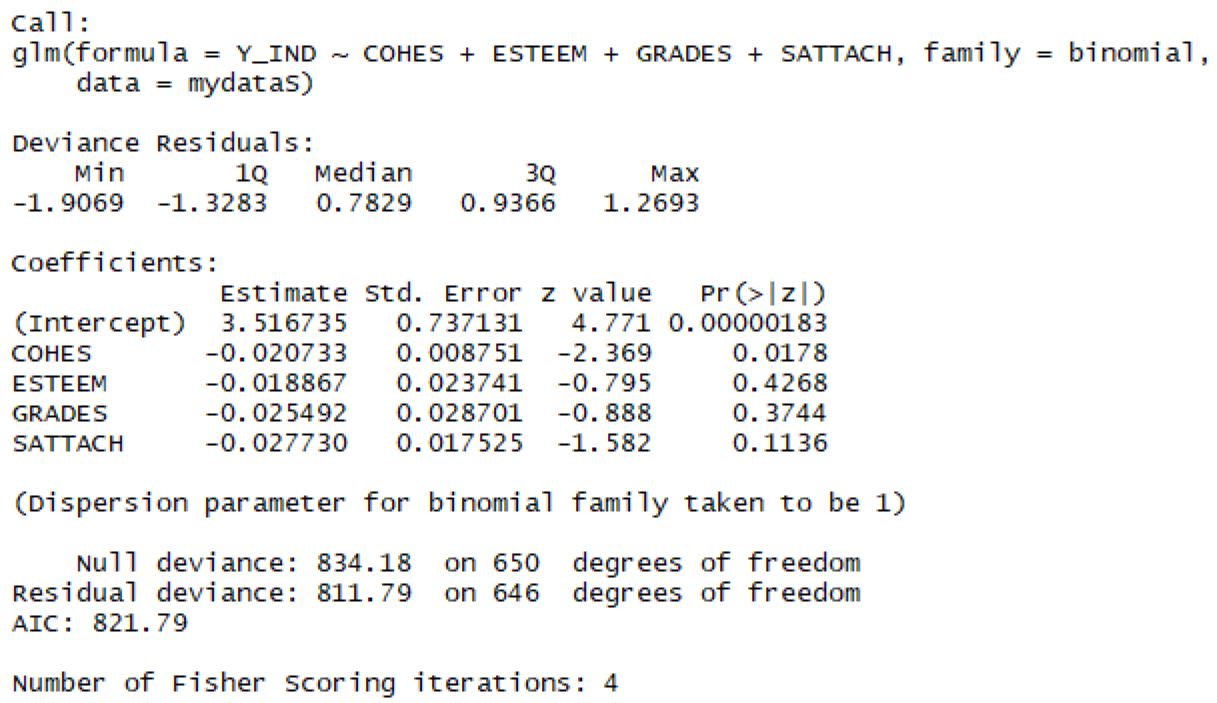


We expect a 'random' pattern in this plot. But it is difficult to see in GLM since the response variable is not continuous. I used smooth line to check whether the line was flat or not. The line is not flat, and it means that we need to consider extra terms in our model such as polynomial terms. Thus, this model is 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?

According to the summary logistic regression function below, only family cohesion p-value is statistically significant.

```
Logit_Y = 3.517 - 0.021*COHES - 0.019*ESTEEM - 0.026*GRADES –
0.028*SATTACH

AIC: 821.786
BIC: 844.178
COHES: exp(-0.020733)-1 = -0.021
ESTEEM: exp(-0.018867)-1 = -0.019
GRADES: exp(-0.025492)-1 = -0.025
SATTACH: exp(-0.027730)-1 = -0.027
```



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

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
```

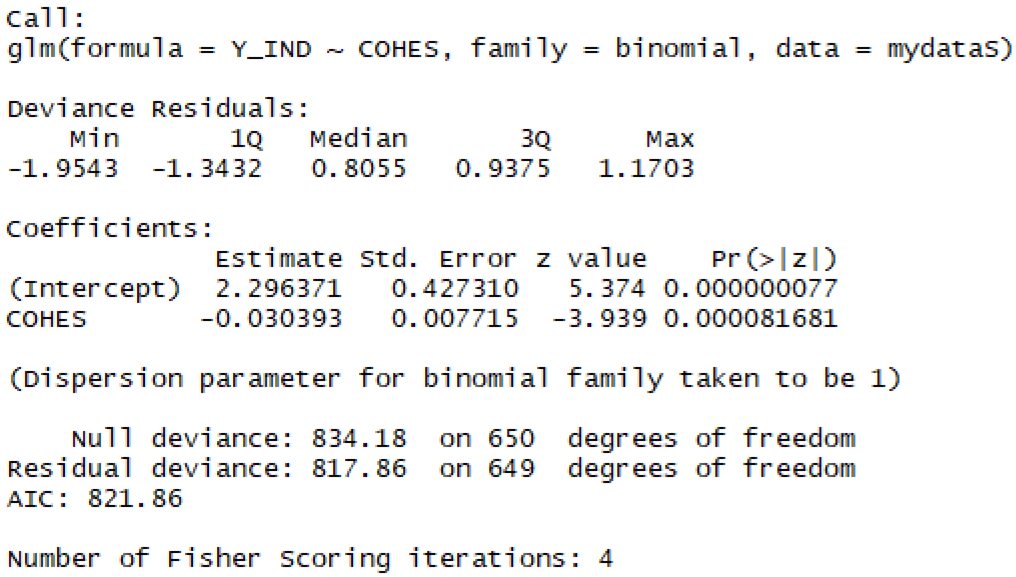
The table shows the coefficient estimates for a logistic regression model that uses family cohesion, self-esteem, grades, and school attachment to predict the probability of having stressful events. The p-value associated with family cohesion predictor is very small, indicating that this variable is associated with the probability of having stressful events.

A one-unit increase in family cohesion, self-esteem, previous year grades, and school attachment predictors are associated with a decrease in the log-odds of having stressful events by 0.021, 0.019, 0.025, and 0.027 units respectively. Or, a one-unit increase in family cohesion, self-esteem, previous year grades, and school attachment levels are associated with a decrease in the odds of having stressful events by 2.1%, 1.9%, 2.5%, and 2.7% respectively. Since all predictors are negative, any additional improvement of those predictors, they help to reduce the stressful events.

Goodness Fit: `Res. Dev/df = 811.79/646 = 1.26` => goodness fit is improved but it is still not a good fit because of missing predictor, missing data, or overdispersion. AIC and BIC improved significantly.

Chi-square test: `1 - pchisq((fit_Y_IND$null.deviance - fit_Y_IND$deviance),(fit_Y_IND$df.null - fit_Y_IND$df.residual)) = 0.` Since the chi-squared test is 0, this model is significantly better than the intercept. So overall model is significant.

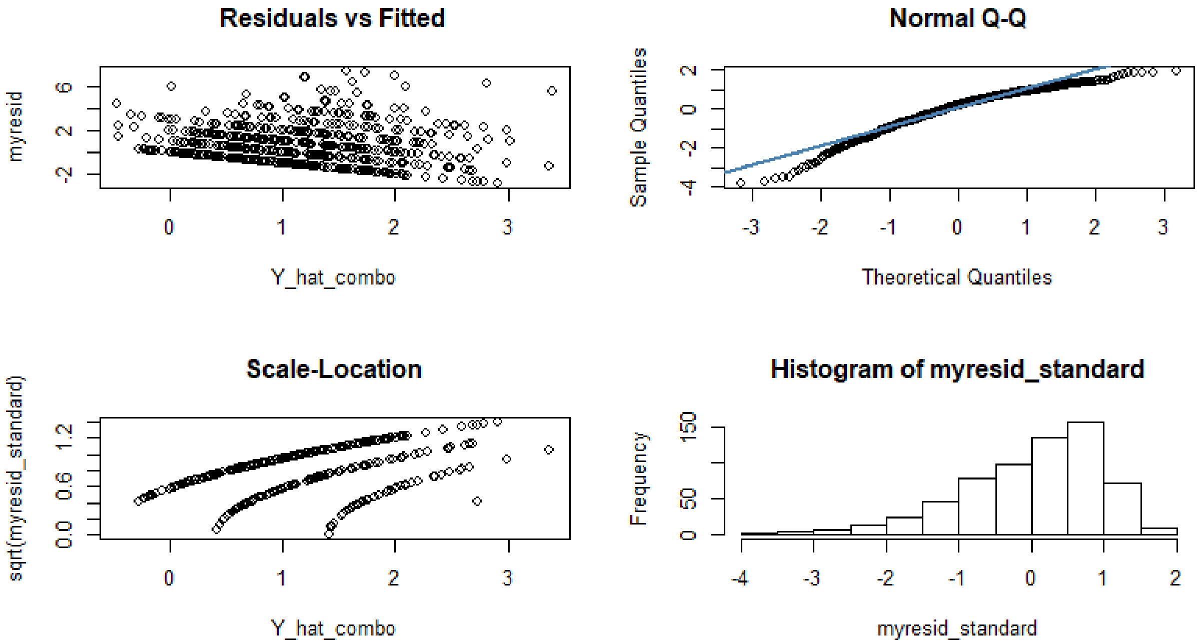
I can rerun it with a family cohesion, significant varaible only. However, it did not improve the AIC, BIC, and goodness fit much.



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

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 0.000000077
COHES       -0.030393   0.007715  -3.939 0.000081681

(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
```

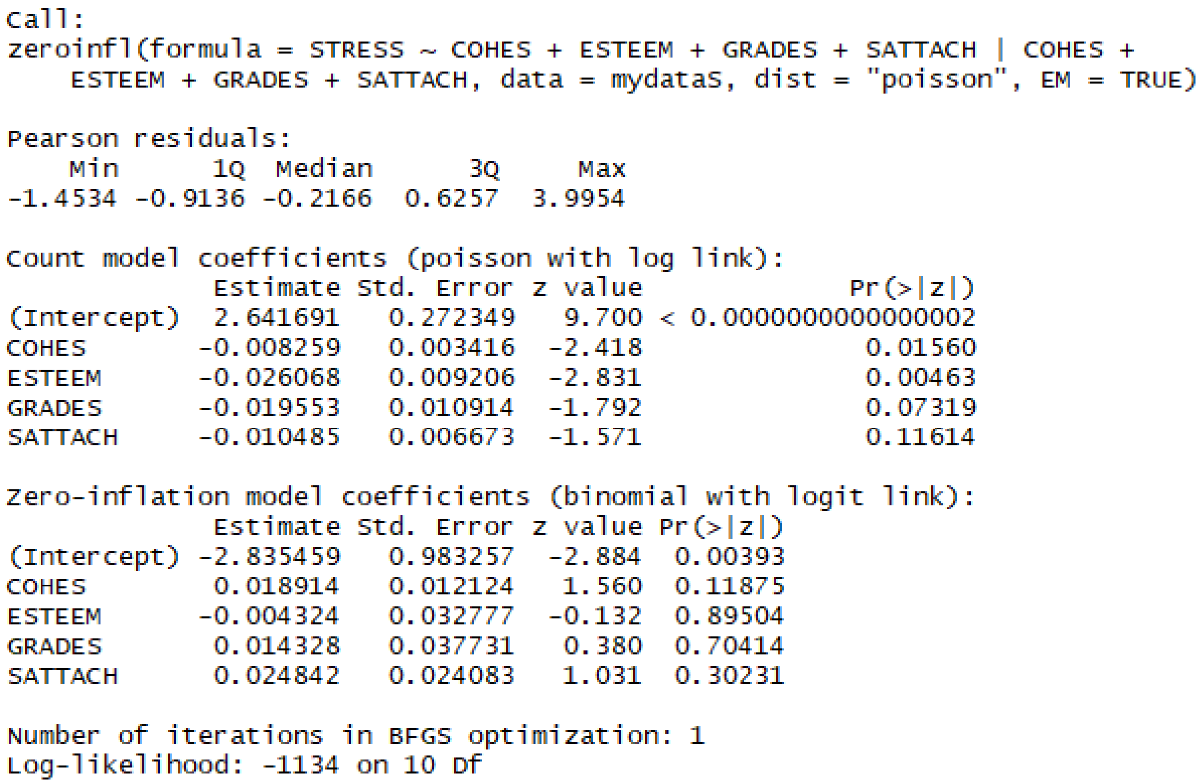
9. It may be that there are two (or more) process at work that are overlapped and generating the distributions of STRESS(Y). What do you think those processes might be? To conduct a ZIP regression model by hand, fit a Logistic Regression model to predict if stress is present (Y_IND), and then use a Poisson 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? Note: It is Logistic Regression First to predict (0=nothing, 1=something where counts are 1 or more). Then Poisson Regression for number of counts. It is not a simple as plug and chug. If you use the counts variable for the Poisson Regression model, there are all the 0's in there that are causing the problem. So, the Poisson Regression part has to be conditional on counts being 1 or more. You will have to select those records (i.e. conditioning) then fit the Poisson Model.

- In task 4, we created the Poisson regression:
  - `2.735 - 0.013*COHES - 0.024*ESTEEM - 0.023*GRADES - 0.017*SATTACH`
- In task 8, we created the logistic regression:
  - `3.517 - 0.021*COHES - 0.019*ESTEEM - 0.026*GRADES - 0.028*SATTACH`
- Combining two regressions:
  - `Y_hat_combo <- 6.252 - 0.034*COHES - 0.043*ESTEEM - 0.049*GRADES - 0.045*SATTACH`
- Obtained predicted values and residuals
  - `myresid <- mydataS$STRESS - Y_hat_combo`



- Residuals are not spread equally, they have pattern. According to the QQ-plot the residuals are not normally distributed in both ends. The residual histogram has long tail in left direction, which confirms non-normality of residual distribution.

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.



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

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)  2.641691   0.272349   9.700 < 0.0000000000000002
COHES       -0.008259   0.003416  -2.418              0.01560
ESTEEM      -0.026068   0.009206  -2.831              0.00463
GRADES      -0.019553   0.010914  -1.792              0.07319
SATTACH     -0.010485   0.006673  -1.571              0.11614

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.835459   0.983257  -2.884  0.00393
COHES        0.018914   0.012124   1.560  0.11875
ESTEEM      -0.004324   0.032777  -0.132  0.89504
GRADES       0.014328   0.037731   0.380  0.70414
SATTACH      0.024842   0.024083   1.031  0.30231

Number of iterations in BFGS optimization: 1
Log-likelihood: -1134 on 10 Df
```