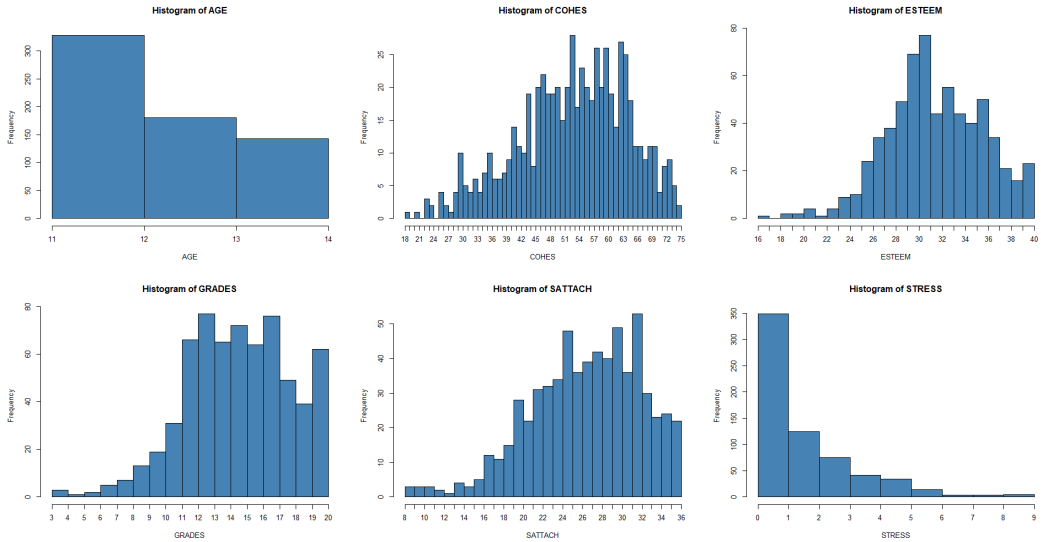
Dataset

The dataset STRESS has 651 rows and 7 columns. None of the values are missing.
Names of the columns are: "AGE" "COHES" "ESTEEM" "GRADES" "SATTACH" "STRESS" "NEWID" .

```
> summary(df)
      AGE            COHES           ESTEEM          GRADES          SATTACH          STRESS           NEWID       
 Min.   :11.00   Min.   :18.00   Min.   :16.19   Min.   : 3.656   Min.   : 8.228   Min.   :0.00   Min.   :   5.0  
 1st Qu.:12.00   1st Qu.:46.00   1st Qu.:29.00   1st Qu.:13.000   1st Qu.:23.000   1st Qu.:0.00   1st Qu.: 945.5  
 Median :12.00   Median :54.00   Median :31.32   Median :15.000   Median :27.000   Median :1.00   Median :1790.0  
 Mean   :12.51   Mean   :53.00   Mean   :31.86   Mean   :14.933   Mean   :26.812   Mean   :1.73   Mean   :1758.6  
 3rd Qu.:13.00   3rd Qu.:61.66   3rd Qu.:35.00   3rd Qu.:17.000   3rd Qu.:31.000   3rd Qu.:3.00   3rd Qu.:2607.5  
 Max.   :14.00   Max.   :75.00   Max.   :40.00   Max.   :20.000   Max.   :36.000   Max.   :9.00   Max.   :3356.0  
```

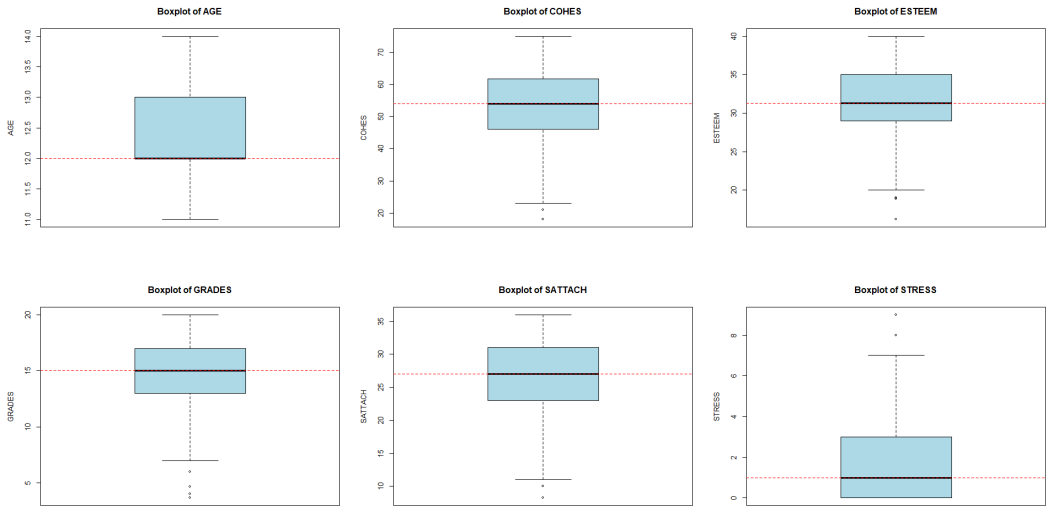
- The column NEWID seems to be some kind of identification number and will not be used in the modelling exercise. All the other columns seem to have numeric values.
- AGE ranges from 11 to 14, indicating a young cohort.
- ESTEEM scores vary widely, suggesting differing self-esteem levels among individuals.
- STRESS levels have a minimum at 0, suggesting some individuals reported no stress.

Plotting Histograms:



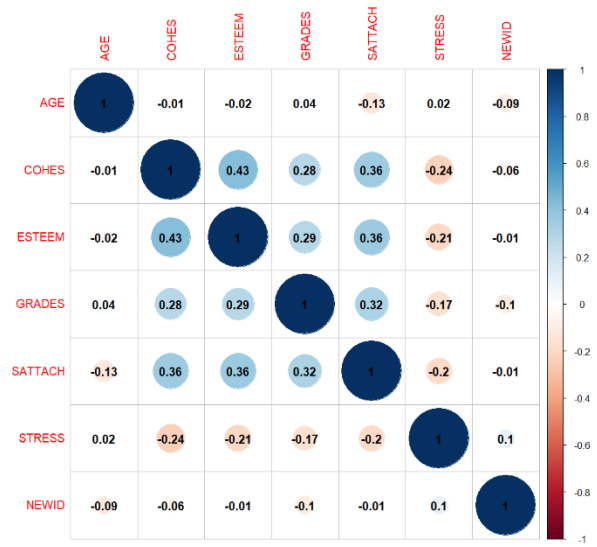
- AGE distribution is left-skewed, most individuals are 12 years old.
- ESTEEM is right-skewed, with the median around 31.
- STRESS has a significant number of zeros.

Boxplots:



- AGE and STRESS boxplots show relatively small interquartile ranges, indicating less variability.
- Except AGE all variables have outliers, suggesting some values that are significantly different from the rest.

Corelation:



- COHES has a moderate positive correlation with ESTEEM (0.43), GRADES (0.28), and SATACH (0.36).
- ESTEEM has a moderate positive correlation with GRADES (0.29) and SATACH (0.36).
- GRADES and SATACH have a moderate positive correlation (0.32).
- Most of the other variables show little to no correlation with each other as indicated by circles closer to white with values around 0.
- STRESS has moderate negative correlation with COHES (-0.24), ESTEEM (-0.21), and GRADES (-0.17).

Task 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.

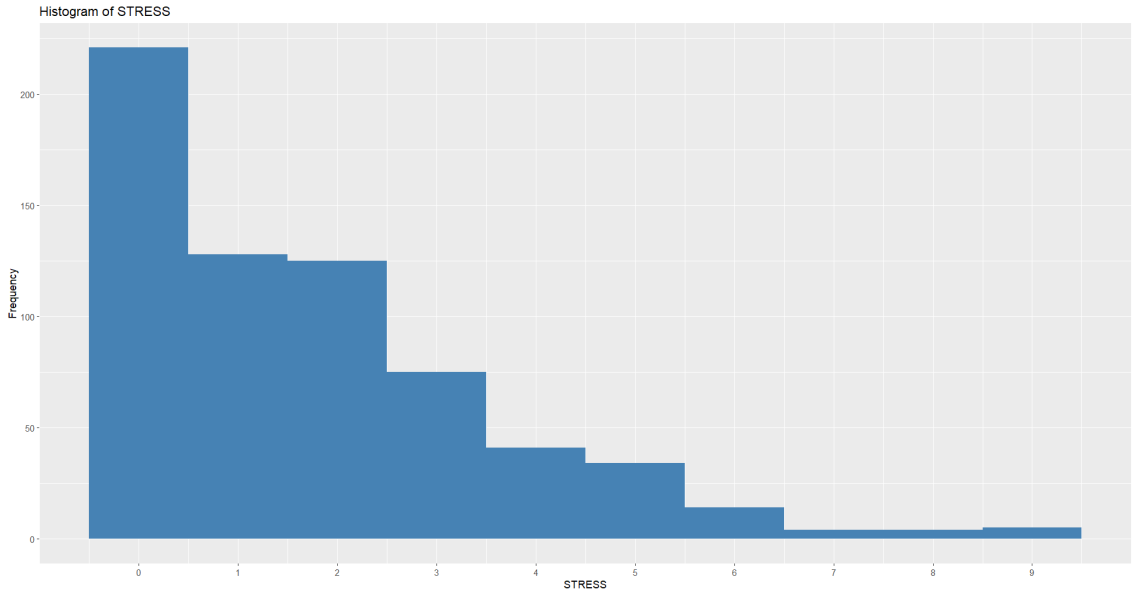
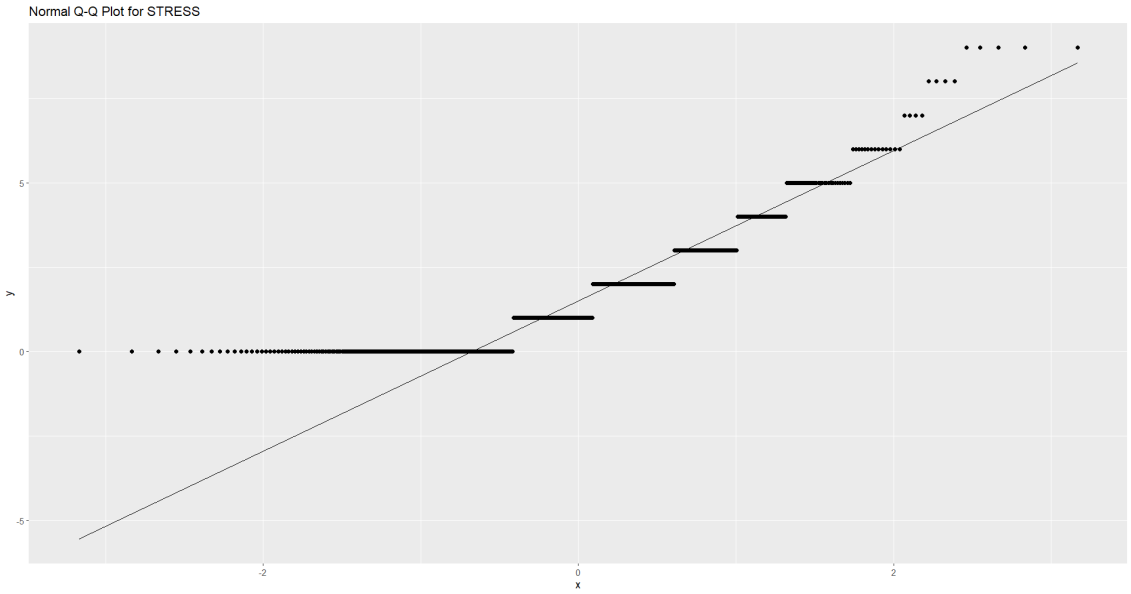


Table: Summary Statistics of STRESS

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Standard Deviation
0	0	1	1.729647	3	9	1.849082

The histogram indicates that the STRESS variable has a range from 0 to 9. The distribution is skewed to the left, with a higher frequency of lower STRESS values and fewer occurrences as the STRESS level increases. Majority of the data points fall within the 0 to 2 range, suggesting that lower stress levels are more common in this dataset.

- The minimum value of STRESS is 0, indicating that there are individuals or entries with no stress reported.
- The first quartile is also 0, which means that 25% of the data are zeros, reinforcing the skew towards lower stress levels.
- The median (or second quartile) is 1, meaning that half of the data points have a STRESS level of 1 or less.
- The mean is approximately 1.73, which is higher than the median, further confirming the left skew of the data since mean can be affected by high values.
- The third quartile is 3, indicating that 75% of the data have a STRESS level of 3 or less.
- The maximum value is 9, showing that there are some instances of very high stress levels, although these are outliers compared to the rest of the data.
- The standard deviation is about 1.85, which suggests there is some variability in the STRESS levels, but given the range of 0 to 9, this is relatively moderate.
- Overall, the data shows that while stress levels vary, the distribution is concentrated towards the lower end of the scale.



- Straight Line: The reference line indicates where the points would lie if STRESS were normally distributed. In a perfect normal distribution, all points would fall exactly on this line.
- Deviations from the Line: The points in the Q-Q plot deviate from the reference line, indicating that the distribution of STRESS deviates from normality.
- Left Side (Lower Tail): The points on the left side of the plot lie below the line, indicating that the lower end of the STRESS distribution has fewer data points than would be expected in a normal distribution (a shorter left tail).
- Center of the Plot: The points in the center of the plot adhere relatively close to the line, suggesting that the middle part of the STRESS distribution is similar to that of a normal distribution.
- Right Side (Upper Tail): The points on the right side of the plot (the upper tail) are above the line, indicating that the higher end of the STRESS distribution has more data points than would be expected in a normal distribution (a longer right tail).

Overall, the pattern suggests that the distribution of STRESS is not normal. It has a left-skewed distribution, as indicated by the concentration of points below the line at the lower end, and a longer right tail, as indicated by the points above the line at the higher end. This interpretation is consistent with the earlier histogram which showed a higher frequency of lower STRESS values.

The dataset does not appear to follow a Poisson distribution for several reasons, which suggest that a negative binomial distribution might be a better fit:

1. Overdispersion: The Poisson distribution assumes that the mean and variance of the data are equal. In this dataset, the variance (standard deviation squared, or 1.85^2) is significantly larger than the mean (1.73), which indicates overdispersion. The negative binomial distribution can account for overdispersion by incorporating an additional parameter that allows the variance to be greater than the mean.
2. High Frequency of Zeros: Although Poisson can handle a fair number of zero counts, when the number of zeros is excessive, it may suggest that another process is generating these zeros, which is not captured by the Poisson distribution. Zero-inflated models or hurdle models are typically used in such cases, but the negative binomial distribution, with its flexibility, might still provide a better fit than the Poisson for overdispersed count data with many zeros.
3. Shape of the Distribution: The histogram shows a long tail of counts, which is more characteristic of a negative binomial distribution. In contrast, a Poisson distribution would typically show a peak at or near the mean and then quickly decline, not showing the long tail that is evident in your data.
4. Q-Q Plot Deviation: The Q-Q plot shows that the data does not follow a normal distribution, which is not a direct indicator of whether a Poisson or negative binomial distribution is more appropriate. However, the pattern of deviation — with more extreme values in the tails than would be expected under a normal distribution — is consistent with the long-tailed distributions like the negative binomial.

In summary, the presence of overdispersion and a high number of zeros in the dataset are the primary indicators that the negative binomial distribution may be more appropriate than the Poisson distribution. The negative binomial distribution's flexibility in modeling count data with variance greater than the mean often makes it a suitable model for real-world count datasets that do not conform to the strict assumptions of the Poisson distribution.

Task 2

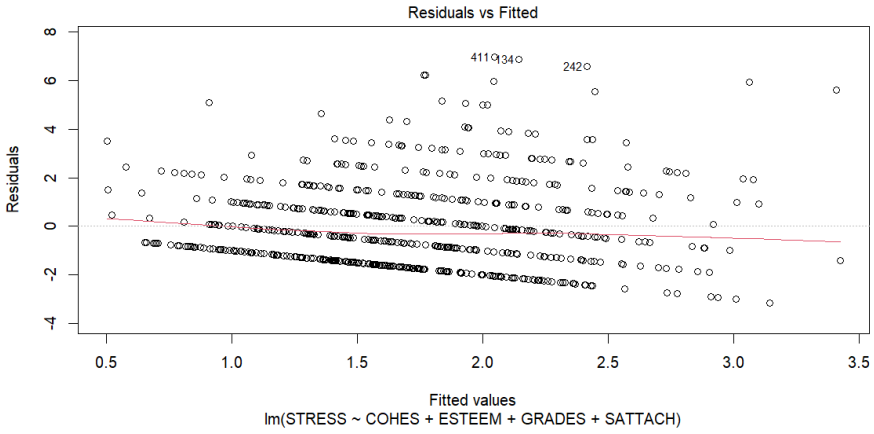
Fit an OLS regression model to predict STRESS (Y) using COHES, ESTEEM, GRADES, SATTACH as explanatory variables (X). Obtain the typical diagnostic information and graphs. Discuss how well this model fits. Obtain predicted values (Y_hat) and plot them in a histogram. What issues do you see?

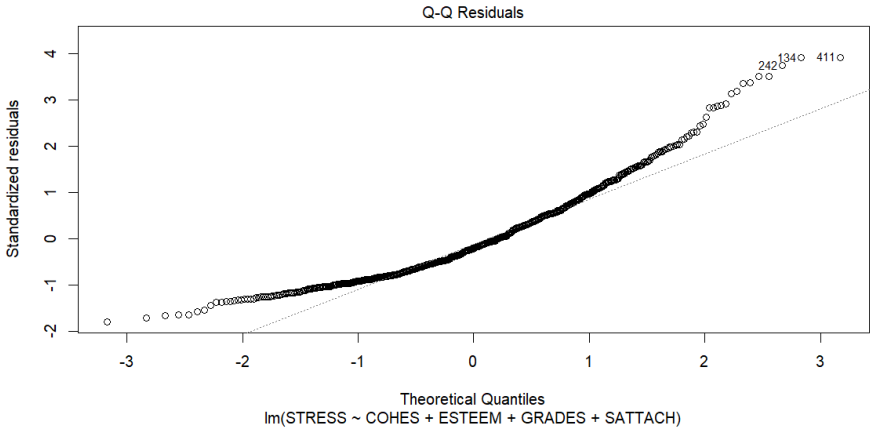
```
Call:
lm(formula = STRESS ~ COHES + ESTEEM + GRADES + SATTACH, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1447 -1.3827 -0.3819  0.9504  6.9525

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.71281    0.58118   9.830  < 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
```

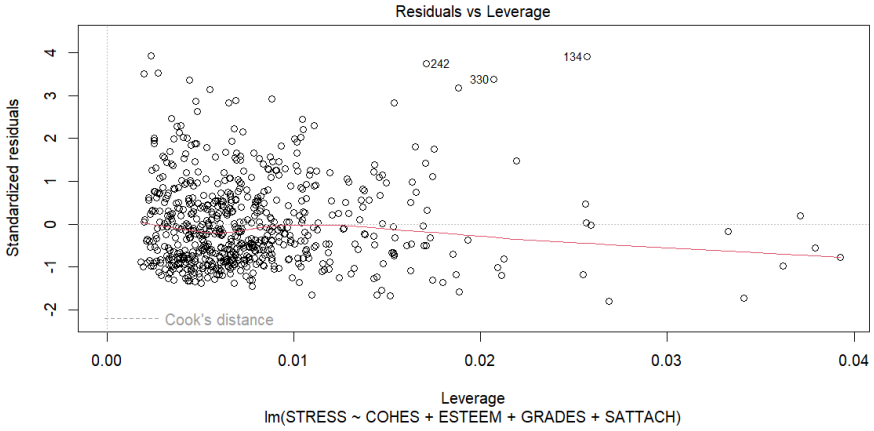
- Residuals:
  - Min and Max suggest there are some large outliers since they are quite far from zero.
  - The 1Q (first quartile) and 3Q (third quartile) indicate that the middle 50% of the residuals fall between -1.3827 and 0.9504, which does not suggest a symmetric spread around the median.
  - The Median is close to zero (-0.3819), which is good as it suggests there's no systematic bias in the predictions — they are not consistently over- or under-predicting.
- Coefficients:
  - The Intercept (5.71281) represents the estimated value of STRESS when all explanatory variables are 0.
  - COHES, ESTEEM, GRADES, and SATTACH have negative coefficients, indicating that increases in these predictors are associated with a decrease in the predicted value of STRESS.

    - COHES has a coefficient of -0.02319, meaning that for each one-unit increase in COHES, the STRESS score is predicted to decrease by 0.02319 units, holding all else constant.
    - ESTEEM and SATTACH have similar interpretations with their respective coefficients.
    - The p-values (Pr(>|t|)) for COHES, ESTEEM, and SATTACH are less than 0.05, indicating that these variables have statistically significant relationships with STRESS at the 5% significance level.
    - GRADES has a p-value above 0.05, indicating a no relationship with STRESS that may be statistically significant at the 5% level.
- Residual Standard Error (RSE): The RSE of 1.776 measures the standard deviation of the residuals. It indicates the average distance that the observed values fall from the regression line.
- R^2 Values:
    - The Multiple R-squared of 0.08319 suggests that approximately 8.32% of the variability in STRESS can be explained by the model.
    - The Adjusted R-squared of 0.07751 is a modified version of R^2 that has been adjusted for the number of predictors in the model. It's slightly lower than the R^2, which is normal, especially when adding variables that do not contribute much to the model.
- F-Statistic: The F-statistic of 14.65 and its associated p-value (1.826e-11) test the null hypothesis that all of the regression coefficients are equal to zero. The very small p-value suggests that we can reject the null hypothesis, meaning the model as a whole is statistically significant.

In summary, the model finds significant relationships between STRESS and COHES, ESTEEM, & SATTACH, but the overall fit of the model explains a relatively small portion of the variability in STRESS. There is also evidence of outliers or extreme values, and potentially influential points, which should be further investigated with diagnostic plots and influence measures.





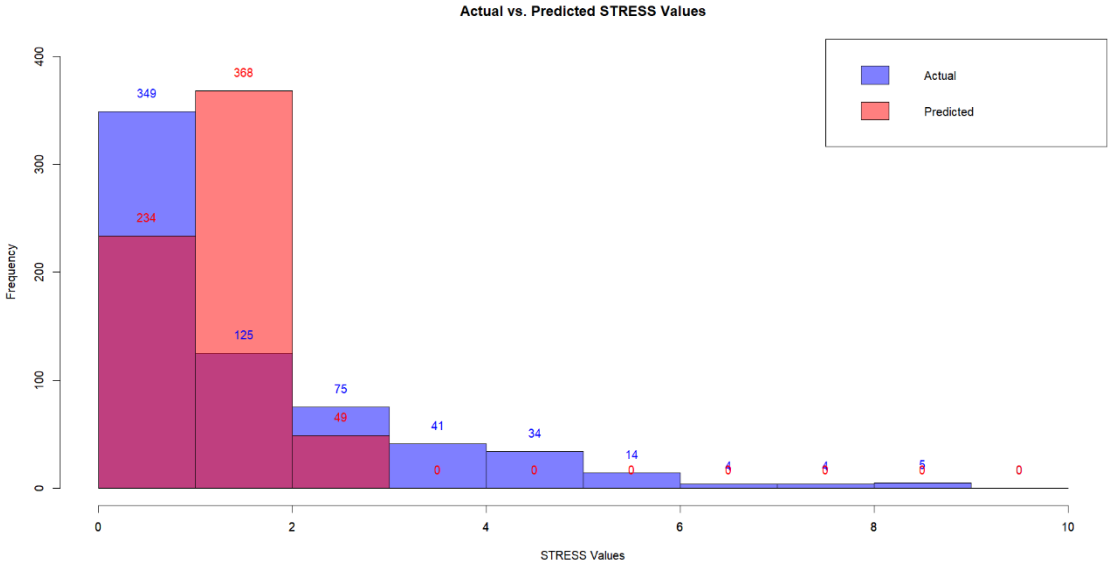


1. Residuals vs Fitted Plot: The residuals do not appear to fan out or in, which is a good sign for equal variance (homoscedasticity). However, there is a pattern that suggests non-linearity since the residuals are not randomly dispersed around the horizontal line at 0. The presence of several outliers is also noted.

2. Q-Q Plot of Standardized Residuals: The plot shows that while many residuals fall along the line, there is a clear deviation at both tails, indicating that the residuals have heavier tails than a normal distribution. This suggests that the normality assumption may be violated.

3. Residuals vs Leverage Plot: The plot indicates a few points with higher leverage, but most do not appear to have large residuals. The points labeled are those that stand out for their leverage and/or residual values and should be investigated further.

In summary, the diagnostic plots suggest that there are some concerns about the validity of the model assumptions. There are indications of non-linearity and potential violations of the normality assumption of the residuals.



- The actual number of cases with stress level 0 is 234, against the predicted value of 349.
- The actual number of cases with stress level 1 is 125, against the predicted value of 368.
- The actual number of cases with stress level 2 is 75, against the predicted value of 49.
- The model has predicted 0 cases for all the stress levels.

Task 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?

Given that we have 349 records with 0 as STRESS, we add 1 to STRESS, before taking log.

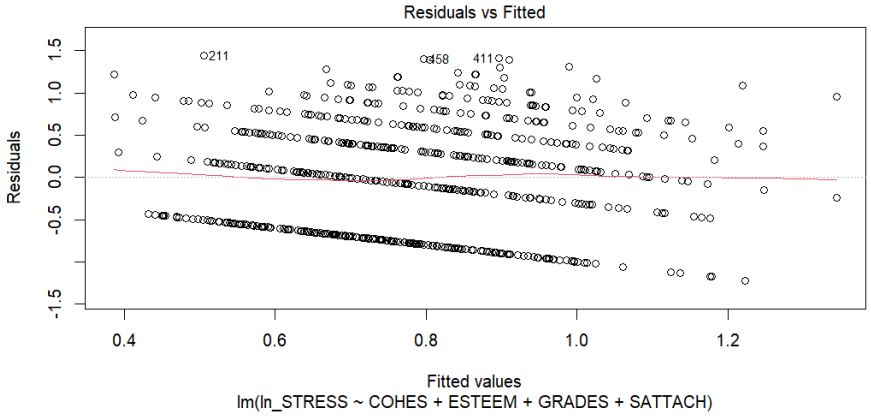
```
# Add 1 to STRESS values before taking the natural logarithm
df$ln_STRESS <- log(df$STRESS + 1)
```

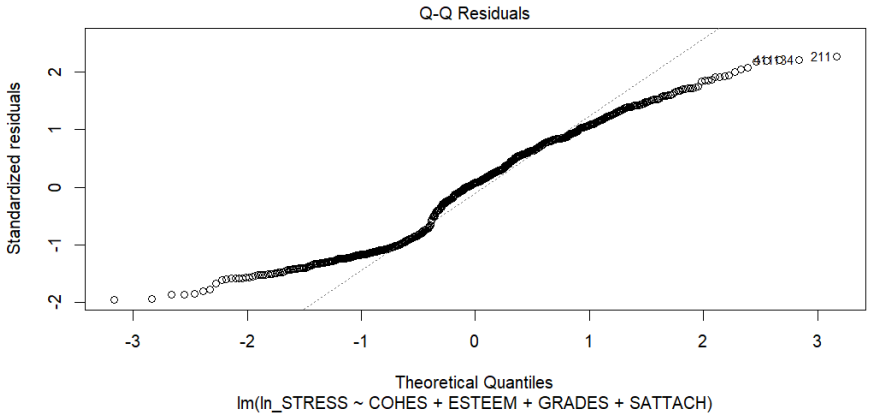
```
> # Fit the OLS regression model for ln(STRESS) after handling missing/infinite values
> model <- lm(ln_STRESS ~ COHES + ESTEEM + GRADES + SATTACH, data = df)
>
> # Summarize the model
> summary(model)

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

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
```

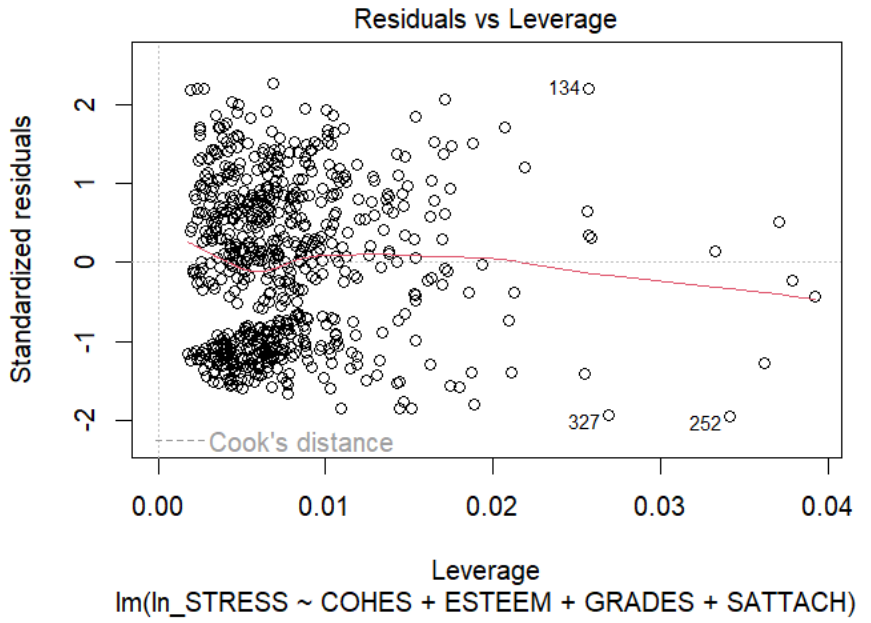
- Residuals:
  - Min and Max suggest there are no large outliers since they are close to zero.
  - The 1Q (first quartile) and 3Q (third quartile) indicate that the middle 50% of the residuals fall between –0.63438 and 0.51763, which does not suggest a symmetric spread around the median.
  - The Median is not close to zero (0.4982), which is not good as it suggests there seems to be a systematic bias in the predictions — they are consistently over- or under-predicting.
- Coefficients:
  - The Intercept (2.073322) represents the estimated value of STRESS when all explanatory variables are 0.
  - COHES, ESTEEM, GRADES, and SATTACH have negative coefficients, indicating that increases in these predictors are associated with a decrease in the predicted value of STRESS.
  - COHES has a coefficient of -0.007947, meaning that for each one-unit increase in COHES, the STRESS score is predicted to decrease by 0.007947units, holding all else constant.
  - The p-values (Pr(>|t|)) for COHES, and SATTACH are less than 0.05, indicating that these variables have statistically significant relationships with STRESS at the 5% significance level.
  - ESTEEM and GRADES have a p-value above 0.05, indicating a no relationship with STRESS that may be statistically significant at the 5% level.
- Residual Standard Error (RSE): The RSE of 0.639 measures the standard deviation of the residuals. It indicates the average distance that the observed values fall from the regression line.
- R^2 Values:
  - The Multiple R-squared of 0.07154 suggests that approximately 7.15% of the variability in STRESS can be explained by the model.
  - The Adjusted R-squared of 0.06579 is a modified version of R^2 that has been adjusted for the number of predictors in the model. It's slightly lower than the R^2, which is normal, especially when adding variables that do not contribute much to the model.
- F-Statistic: The F-statistic of 12.44 and its associated p-value (9.333e-10) test the null hypothesis that all of the regression coefficients are equal to zero. The very small p-value suggests that we can reject the null hypothesis, meaning the model as a whole is statistically significant.

In summary, the model finds significant relationships between STRESS and COHES & SATTACH, but the overall fit of the model explains a relatively small portion of the variability in STRESS. There is also evidence of outliers or extreme values, and potentially influential points, which should be further investigated with diagnostic plots and influence measures.

The R-squared value has reduced from 8.32% to 7.15%, after we took $\log_e$ of STRESS. As expected the variability of Residuals did decrease on taking the $\log_e$ of STRESS.

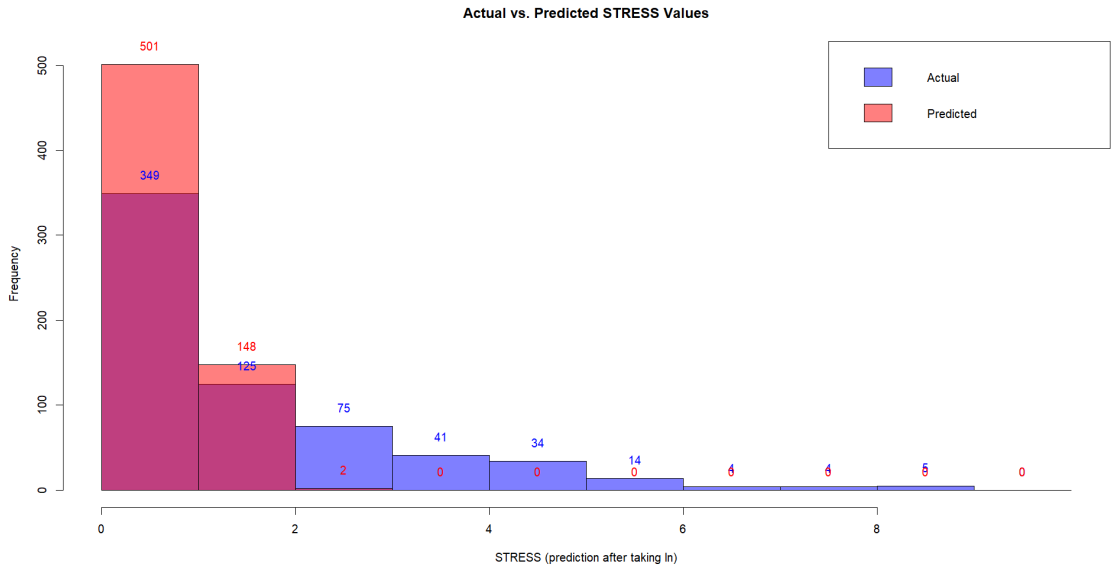






4. Residuals vs Fitted Plot: The residuals do not appear to fan out or in, which is a good sign for equal variance (homoscedasticity). However, there is a pattern that suggests non-linearity since the residuals are not randomly dispersed around the horizontal line at 0. The presence of several outliers is also noted.
5. Q-Q Plot of Standardized Residuals: The plot shows that while many residuals fall along the line, there is a clear deviation at both tails – greater than what was observed when the $\log_e$ of STRESS was not taken, indicating that the residuals have heavier tails than a normal distribution. This suggests that the normality assumption may be violated.
6. Residuals vs Leverage Plot: The plot indicates a few points with higher leverage, but most do not appear to have large residuals. The points labeled are those that stand out for their leverage and/or residual values and should be investigated further.

In summary, the diagnostic plots suggest that there are some concerns about the validity of the model assumptions. There are indications of non-linearity and potential violations of the normality assumption of the residuals.



The number of zeros 501 predicted by this model is much higher than 234 that were predicted without taking the $\log_e$ of STRESS. Also, this model does not predict any numbers beyond 3.

Conclusion: Taking into account the fact that the R-squared value decreased, and that much higher number of zeros predicted by taking $\log_e$ of STRESS, we can conclude that taking $\log_e$ has not helped.

Task 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?

```
> # Fit a Poisson regression model
> model <- glm(STRESS ~ COHES + ESTEEM + GRADES + SATTACH, family = poisson, data = df)
>
> # Summary of the model
> summary(model)

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

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

Number of Fisher Scoring iterations: 5
```

- Coefficients:
    - Intercept (2.734513): This is the expected log count of STRESS when all the explanatory variables are zero.
    - COHES (-0.012918): This coefficient suggests that for a one-unit increase in COHES, the log count of STRESS decreases by approximately 0.012918 units, holding other variables constant. This effect is statistically significant ($p < 0.001$).
    - ESTEEM (-0.023692): For each unit increase in ESTEEM, the log count of STRESS decreases by about 0.023692 units. This relationship is statistically significant ($p = 0.00321$).
    - GRADES (-0.023471): Each unit increase in GRADES is associated with a decrease of about 0.023471 units in the log count of STRESS, which is significant at the 0.05 level ($p = 0.01735$).
    - SATTACH (-0.016481): An increase in SATTACH by one unit is associated with a decrease of 0.016481 units in the log count of STRESS, and this relationship is statistically significant ($p = 0.00437$).
- Deviance:
    - Null Deviance (1349.8): This represents the goodness of fit of a model that includes only the intercept (no predictors). It serves as a baseline to compare against the model including predictors.
    - Residual Deviance (1245.4): This is lower than the null deviance, indicating that the model with predictors fits better than the null model.

AIC (2417.2): The Akaike Information Criterion, a measure for model comparison, with lower values indicating a better fit.

Overall, the model indicates significant relationships between STRESS and all four explanatory variables, with all variables showing a negative association with STRESS. This means that increases in these variables are associated with decreases in the level of stress, according to the model.

In the Poisson model, all predictors are statistically significant, while in the logged OLS model in Task 3, ESTEEM and GRADES are not significant at the traditional 0.05 level.

```
> # Fit an over-dispersed Poisson regression model (Negative Binomial model)
> model_nb <- glm.nb(STRESS ~ COHES + ESTEEM + GRADES + SATTACH, data = df)
>
> # Summary of the model
> summary(model_nb)

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

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.759032   0.341531   8.078 6.56e-16 ***
COHES       -0.013391   0.004136  -3.238  0.00121 **
ESTEEM      -0.023058   0.011477  -2.009  0.04453 *
GRADES      -0.024360   0.013969  -1.744  0.08118 .
SATTACH     -0.016750   0.008296  -2.019  0.04349 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

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

Number of Fisher Scoring iterations: 1


              Theta:  1.865
          Std. Err.:  0.257

 2 x log-likelihood:  -2271.590
```

- Coefficients:
    - Intercept (2.759032): This is the expected log count of STRESS when all explanatory variables are zero.
    - COHES (-0.013391): For a one-unit increase in COHES, the log count of STRESS is expected to decrease by approximately 0.013391 units, holding other variables constant. This effect is statistically significant ($p = 0.00121$).
    - ESTEEM (-0.023058): Each unit increase in ESTEEM is associated with a decrease of about 0.023058 units in the log count of STRESS, significant at the 0.05 level ($p = 0.04453$).
    - GRADES (-0.024360): A unit increase in GRADES is associated with a decrease of 0.024360 units in the log count of STRESS, though this relationship is not significant ($p = 0.08118$).

    - SATTACH (-0.016750): An increase in SATTACH by one unit is associated with a decrease of 0.016750 units in the log count of STRESS, which is statistically significant (p = 0.04349).
- Dispersion Parameter: The dispersion parameter for the Negative Binomial family is 1.8653, which indicates the level of over-dispersion in the data. This parameter being significantly greater than 1 suggests that the Negative Binomial model is appropriate for the data.
- Deviance:
    - Null Deviance (792.47): This represents the goodness of fit of a model with only the intercept. It serves as a baseline to compare against the model with predictors.
    - Residual Deviance (738.53): This is lower than the null deviance, indicating that the model with predictors fits better than the null model.
- AIC (2283.6): The Akaike Information Criterion is used for model comparison, with lower values indicating a better fit.
- Theta (1.865 with Std. Err. 0.257): Theta is the parameter that models the over-dispersion in the Negative Binomial distribution. The standard error of theta gives an idea of the uncertainty around the over-dispersion estimate.
- 2 x log-likelihood (-2271.590): This is a measure of the model's overall fit, used in computing the AIC.

Overall, the model suggests significant relationships between STRESS and the explanatory variables, with all predictors showing a negative association with STRESS. This means that increases in these variables are associated with decreases in the level of stress. The use of the Negative Binomial model addresses the over-dispersion issue, which might not be adequately handled by a standard Poisson model.

Comparing the Negative Binomial model and the Poisson model for the variable `STRESS` with `COHES`, `ESTEEM`, `GRADES`, and `SATTACH` as explanatory variables can provide insight into which model better fits the data, especially in the context of over-dispersion.

1. Model Assumptions and Appropriateness:
    - The Poisson model assumes that the mean and variance of the dependent variable are equal, which is often not the case with count data.
    - Negative Binomial model, on the other hand, accounts for over-dispersion (where the variance exceeds the mean) by introducing an additional parameter (theta). The theta value in the Negative Binomial model (1.865) indicates significant over-dispersion, suggesting that this model might be more appropriate for the data.
2. Coefficients and Their Significance:
    - The direction and significance of the coefficients are fairly consistent across both models for `COHES`, `ESTEEM`, and `SATTACH`. However, the magnitude of the coefficients differs slightly, which is expected due to the different handling of dispersion in the two models.
    - `GRADES` shows no significance in the Negative Binomial model but is significant in the Poisson model, highlighting how model choice can affect inference.
3. Goodness of Fit:
    - The AIC (Akaike Information Criterion) values for both models can be compared to assess model fit. The Negative Binomial model has a lower AIC (2283.6) compared to the Poisson model (2417.2), suggesting a better fit.
    - The residual deviance for the Negative Binomial model is lower (738.53) compared to the Poisson model (1245.4), which also indicates a better fit, especially when considering over-dispersion.
4. Model Complexity:
    - The Negative Binomial model is more complex due to the additional dispersion parameter (theta). This complexity is justified if it significantly improves the model fit, as indicated by the AIC and residual deviance.
5. Interpretation of the Dispersion Parameter:
    - The dispersion parameter in the Negative Binomial model provides an estimate of over-dispersion. In contrast, the Poisson model always assumes this parameter to be 1, which may not be suitable for over-dispersed data.
6. Number of Iterations for Convergence:

- The Negative Binomial model converged in fewer iterations (1) compared to the Poisson model (5), which might indicate a more straightforward fitting process for the data at hand.

In conclusion, while both models show that the explanatory variables have a significant relationship with `STRESS`, the Negative Binomial model seems to provide a better fit for the data, as indicated by lower AIC and residual deviance values. This model is more appropriate when dealing with count data that exhibits over-dispersion, as it appears to be the case here.

Task 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?

```
> # Calculate mean and standard deviation of COHES
> mean_cohes <- mean(df$COHES, na.rm = TRUE)
> sd_cohes <- sd(df$COHES, na.rm = TRUE)
>
> # Define the Cohesion Groups
> low_cohes <- mean_cohes - sd_cohes
> high_cohes <- mean_cohes + sd_cohes
>
> # Extract coefficients from the model
> intercept <- coef(model)["(Intercept)"]
> beta_cohes <- coef(model)["COHES"]
>
> # Calculate predicted counts for low, middle, and high groups
> predicted_low <- exp(intercept + beta_cohes * low_cohes)
> predicted_middle <- exp(intercept + beta_cohes * mean_cohes)
> predicted_high <- exp(intercept + beta_cohes * high_cohes)
>
> # Calculate expected percent difference between high and low groups
> percent_difference <- ((predicted_high - predicted_low) / predicted_low) * 100
>
> # Output the results
> list(
+   predicted_low = predicted_low,
+   predicted_middle = predicted_middle,
+   predicted_high = predicted_high,
+   percent_difference = percent_difference
+ )
$predicted_low
(Intercept)
   8.996837

$predicted_middle
(Intercept)
   7.766572

$predicted_high
(Intercept)
   6.704539

$percent_difference
(Intercept)
  -25.47893

> # Print the predicted counts and percent difference, rounded to two decimal places
> print(paste("Predicted count of STRESS for low COHES group:", round(predicted_low, 2)))
[1] "Predicted count of STRESS for low COHES group: 9"
> print(paste("Predicted count of STRESS for middle COHES group:", round(predicted_middle, 2)))
[1] "Predicted count of STRESS for middle COHES group: 7.77"
> print(paste("Predicted count of STRESS for high COHES group:", round(predicted_high, 2)))
[1] "Predicted count of STRESS for high COHES group: 6.7"
> print(paste("Expected percent difference between high and low COHES groups:", round(percent_difference, 2)))
[1] "Expected percent difference between high and low COHES groups: -25.48"
```

1. Predicted Count of STRESS for Different Levels of COHES:
   - Low COHES Group: The predicted count of stressful events for individuals with a level of family cohesion (COHES) less than one standard deviation below the mean is approximately 9. This suggests that lower family cohesion is associated with a higher count of stressful events.
   - Middle COHES Group: For individuals with family cohesion levels between one standard deviation below and one standard deviation above the mean, the predicted count of stressful events is approximately 7.77. This is a moderate level compared to the low and high groups.
   - High COHES Group: For individuals with a level of family cohesion more than one standard deviation above the mean, the predicted count of stressful events is approximately 6.7. This indicates that higher levels of family cohesion are associated with a lower count of stressful events.
2. Expected Percent Difference Between High and Low COHES Groups:
   - The expected percent difference in the number of stressful events between individuals with high and low levels of family cohesion is approximately -25.48%. This negative value indicates that individuals with high

family cohesion experience about 25.48% fewer stressful events compared to those with low family cohesion.

Overall, these results suggest a negative relationship between family cohesion (COHES) and the number of stressful events (STRESS). Higher family cohesion is associated with fewer stressful events, highlighting the potential protective role of family cohesion against stress.

Task 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?

```
> # Print AIC and BIC for the Poisson model, rounded to three decimal places
> print(paste("AIC for Poisson model:", round(aic_poisson, 3)))
[1] "AIC for Poisson model: 2417.219"
> print(paste("BIC for Poisson model:", round(bic_poisson, 3)))
[1] "BIC for Poisson model: 2439.612"
>
> # Print AIC and BIC for the Negative Binomial model, rounded to three decimal places
> print(paste("AIC for Negative Binomial model:", round(aic_nb, 3)))
[1] "AIC for Negative Binomial model: 2283.59"
> print(paste("BIC for Negative Binomial model:", round(bic_nb, 3)))
[1] "BIC for Negative Binomial model: 2310.461"
```

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values for the Poisson and Negative Binomial models provide information about the relative quality of each model for the given data, considering both the goodness of fit and the complexity of the model.

1. AIC Values:
   - Poisson Model: The AIC for the Poisson model is 2417.219.
   - Negative Binomial Model: The AIC for the Negative Binomial model is 2283.59.
   - Lower AIC values indicate a better fit of the model to the data when accounting for the number of parameters used. In this case, the Negative Binomial model has a lower AIC, suggesting it fits the data better than the Poisson model.
2. BIC Values:
   - Poisson Model: The BIC for the Poisson model is 2439.612.
   - Negative Binomial Model: The BIC for the Negative Binomial model is 2310.461.
   - Similar to the AIC, a lower BIC value indicates a better model. BIC also penalizes the number of parameters more strongly than AIC. Again, the Negative Binomial model shows a lower BIC, suggesting it is the more preferable model in terms of both fit and complexity.

Overall, both AIC and BIC suggest that the Negative Binomial model is a better choice for this data compared to the Poisson model. This likely reflects the fact that the Negative Binomial model can account for over-dispersion in the data, which is a common issue in count data and often not adequately addressed by a Poisson model.

Task 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?

```
> # Fit a logistic regression model
> logistic_model <- glm(Y_IND ~ COHES + ESTEEM + GRADES + SATTACH, family = binomial, data = df)
>
> # Summary of the logistic regression model
> summary(logistic_model)

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

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.516735   0.737131   4.771 1.83e-06 ***
COHES       -0.020733   0.008751  -2.369   0.0178 *
ESTEEM      -0.018867   0.023741  -0.795   0.4268
GRADES      -0.025492   0.028701  -0.888   0.3744
SATTACH     -0.027730   0.017525  -1.582   0.1136
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 834.18  on 650  degrees of freedom
Residual deviance: 811.79  on 646  degrees of freedom
AIC: 821.79

Number of Fisher Scoring iterations: 4
```

1. Coefficients:
   - Intercept (3.516735): This is the log odds of having stress (`Y_IND = 1`) when all explanatory variables are zero.
   - COHES (-0.020733): A one-unit increase in `COHES` is associated with a decrease in the log odds of having stress by 0.020733 units. This effect is statistically significant (p = 0.0178).
   - ESTEEM (-0.018867): Each unit increase in `ESTEEM` is associated with a decrease in the log odds of having stress by 0.018867 units, but this relationship is not statistically significant (p = 0.4268).
   - GRADES (-0.025492): A one-unit increase in `GRADES` corresponds to a decrease in the log odds of having stress by 0.025492 units, though this effect is not statistically significant (p = 0.3744).
   - SATTACH (-0.027730): An increase in `SATTACH` by one unit is associated with a decrease in the log odds of having stress by 0.027730 units, but this effect is not statistically significant (p = 0.1136).
2. Model Fit:
   - Null Deviance (834.18): This represents the goodness of fit of a model with only the intercept. It serves as a baseline for comparison.
   - Residual Deviance (811.79): This is lower than the null deviance, indicating that the model with predictors fits better than the null model. However, the reduction is not substantial, suggesting that other variables or interactions might need to be considered for a better fit.
   - AIC (821.79): Akaike Information Criterion is a measure for model comparison, where lower values indicate a better fit.
3. Interpretation and Next Steps:
   - The model suggests that higher family cohesion (`COHES`) is significantly associated with a lower likelihood of experiencing stress.
   - Since not all variables are significant, and the reduction in deviance is modest, one may consider rerunning the analysis with different variables or interactions. It might also be beneficial to check for multicollinearity or non-linear relationships.

In conclusion, while this logistic regression model shows some relationship between the variables and the presence of stress, further model refinement and exploration of additional variables or relationships may be necessary to improve its predictive power and interpretability.

In light of the logistic regression model's summary, revisiting the analysis with a combined approach encompassing both logistic and Poisson regression models could be highly beneficial. This integrated strategy involves several steps and considerations:

1. Addressing Multicollinearity and Model Fit: Investigate and address any high collinearity among predictors, which can impact the reliability of model estimates. This step is crucial, especially since the current model shows only a modest improvement over the null model, suggesting it might not be fully capturing the data's variability.
2. Incorporating Interaction Terms and Non-linear Relationships: If theory or empirical evidence suggests interactions between predictors, or if relationships between predictors and stress odds appear non-linear, consider adding interaction terms and applying transformations like polynomials or splines.

3. Refining Model Selection with Additional Variables: Employ model selection techniques, such as stepwise regression, to identify a more effective set of predictors. Introduce other relevant variables to better delineate the relationship between predictors and stress presence.
4. Dual-Model Strategy and Sequential Modeling: Integrate insights from logistic regression (analyzing stress occurrence) and Poisson regression (focusing on stress count). First, use logistic regression to predict the likelihood of experiencing any stress (Y_IND). Then, for those predicted to experience stress, use the Poisson model to estimate the count of stress events. This combined approach allows for a more nuanced understanding — logistic regression informs about the factors influencing the probability of experiencing stress, while Poisson regression sheds light on the intensity or frequency of stress.
5. Model Validation, Comparison, and Theoretical Implications: Validate and compare both models independently and in conjunction. Assess them not only based on statistical metrics like AIC and BIC but also considering their explanatory relevance and practical applications. Leverage insights from both models to strengthen theoretical frameworks and guide practical interventions.
6. Diagnostic Checks: Perform thorough residual checks and goodness of fit assessments to further evaluate model fit and assumption adherence.

In summary, merging logistic regression analysis with insights from Poisson regression, while addressing model fit issues, adding interaction terms, exploring non-linear relationships, and including additional variables, could provide a more comprehensive understanding of stress dynamics. This integrated approach, underpinned by both statistical analysis and theoretical knowledge, can potentially offer a richer and more actionable understanding of the factors influencing stress.

Task 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!

Taking into account the histogram and summary statistics for `STRESS`, it appears that a Zero-Inflated Poisson (ZIP) model could indeed be a suitable choice. The histogram displayed a substantial number of zeros, indicative of many individuals not experiencing any stress events. Moreover, the summary statistics revealed the mean to be greater than the median and left skew, suggesting over-dispersion—a condition where the variance exceeds the mean, which is common in count data.

The two processes possibly at play here could be:

1. Zero Generation Process: This determines whether or not an individual experiences stress at all.
2. Count Process: Once an individual is susceptible to stress, this process governs the frequency of stress events.

Creating a ZIP Model:

1. Logistic Regression for Stress Presence: Fit a logistic regression model to predict `Y_IND`, which indicates the presence or absence of stress. This model addresses the zero-generation process.
2. Poisson Regression for Stress Count: Conditioned on stress being present, fit a Poisson regression model to predict the count of stressful events. This model caters to the count process among those already experiencing stress.

Reasonableness of the ZIP Model: Given the excess zeros and the likelihood of over-dispersion, a ZIP model is reasonable. It accounts for the excess zeros by modeling them through a separate logistic process, and it models the count of events with a Poisson process, which is appropriate for non-negative integer data.

The two models can be combined as follows:

1. Use the logistic regression to estimate the probability that `STRESS` is greater than zero.
2. Use the Poisson regression to estimate the expected count of `STRESS` given that it is greater than zero.

For prediction, we would first predict whether stress is present (`Y_IND = 1`). If so, we would then predict the count of `STRESS` using the Poisson model. For those predicted to have no stress (`Y_IND = 0`), the predicted count would simply be zero.

In summary, the ZIP model allows for a differentiated approach to modeling the distribution of `STRESS`, accommodating the excess zeros and providing a mechanism to handle over-dispersion. It is a well-suited approach that can yield more accurate predictions and insights into the data generating processes.

```
> # If you need to print the summaries
> print(logistic_summary)

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

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.516735   0.737131   4.771 1.83e-06 ***
COHES       -0.020733   0.008751  -2.369   0.0178 *
ESTEEM      -0.018867   0.023741  -0.795   0.4268
GRADES      -0.025492   0.028701  -0.888   0.3744
SATTACH     -0.027730   0.017525  -1.582   0.1136
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 834.18  on 650  degrees of freedom
Residual deviance: 811.79  on 646  degrees of freedom
AIC: 821.79

Number of Fisher Scoring iterations: 4


> print(poisson_summary)

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

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.311700   0.239989   9.633   <2e-16 ***
COHES       -0.006254   0.002945  -2.124   0.0337 *
ESTEEM      -0.019521   0.008107  -2.408   0.0160 *
GRADES      -0.014661   0.009689  -1.513   0.1302
SATTACH     -0.008002   0.005836  -1.371   0.1703
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 415.81  on 429  degrees of freedom
Residual deviance: 380.26  on 425  degrees of freedom
AIC: 1552.1

Number of Fisher Scoring iterations: 5
```

The summaries of the logistic regression and Poisson regression models provide the following insights:

Logistic Regression Summary:

- Intercept (3.516735): The positive estimate for the intercept suggests that when all explanatory variables are at their reference levels (typically zero), the log odds of experiencing stress (`Y_IND = 1`) are high.
- COHES (-0.020733): Higher family cohesion (COHES) is associated with a decrease in the log odds of experiencing stress, significant at the 0.05 level.
- ESTEEM (-0.018867) and GRADES (-0.025492): Although the estimates are negative, suggesting a potential decrease in the log odds of experiencing stress with higher self-esteem or grades, these coefficients are not statistically significant.
- SATTACH (-0.027730): Higher social attachment (SATTACH) is also associated with lower log odds of experiencing stress, but this relationship is not statistically significant.
- The model's AIC (821.79) allows for comparison with other models; lower AIC values are typically preferred.

Poisson Regression Summary:

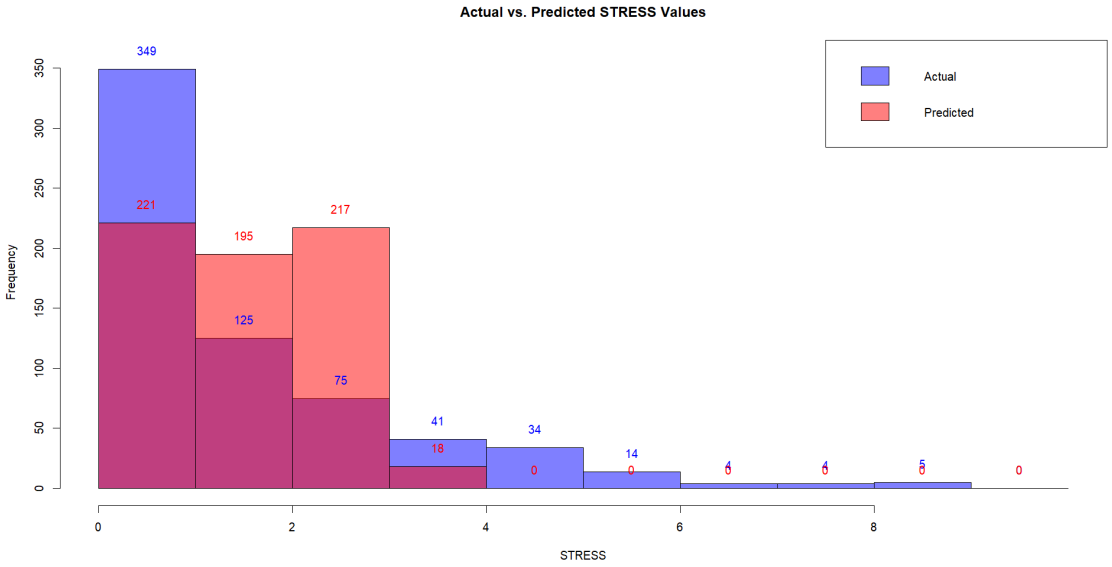
- Intercept (2.311700): There's a positive log count of stress events when all predictors are at their reference levels, indicating a positive number of stress events as the baseline.
- COHES (-0.006254) and ESTEEM (-0.019521): These variables are significant at the 0.05 level and negatively associated with the count of stress events, suggesting that higher family cohesion and self-esteem are related to fewer stress events.
- GRADES (-0.014661) and SATTACH (-0.008002): These coefficients are negative but not statistically significant, suggesting they may not have a strong influence on the count of stress events.
- The AIC (1552.1) here is useful for model comparison, with lower values indicating a more parsimonious fit.

Interpretation and Next Steps:

- The logistic regression model indicates that family cohesion is a significant predictor for the occurrence of stress. In contrast, self-esteem, grades, and social attachment do not show a statistically significant effect.
- The Poisson regression model, applied only to individuals who have experienced stress, shows that both family cohesion and self-esteem significantly impact the number of stress events experienced.
- Since self-esteem is significant in the Poisson model but not in the logistic model, this might suggest that while self-esteem does not influence the likelihood of experiencing stress, it does affect the frequency of stress events for those who experience stress.

Given the significance of variables in both models, the use of a ZIP model is validated, which would combine these insights to predict the presence and count of stress events.

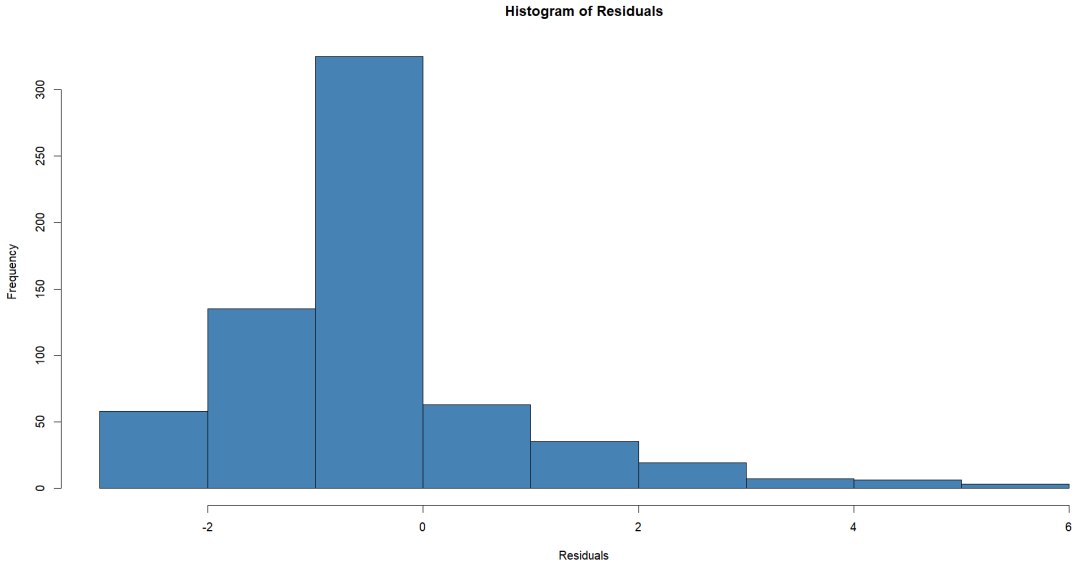
- The relatively high AIC values suggest there may be room for model improvement. This could involve exploring additional predictors, interaction terms, or alternative model specifications.
- To predict STRESS (Y), one would use the logistic regression model to determine the probability of zero events and the Poisson model to predict the number of events when the logistic model predicts the presence of stress. This combined prediction would cater to both processes suggested by the ZIP model.



Residual Summary:

Statistic	Value
Mean	0.0199693
Median	0.0000000
Standard Deviation	1.3130941
Min	-3.0000000
Max	6.0000000

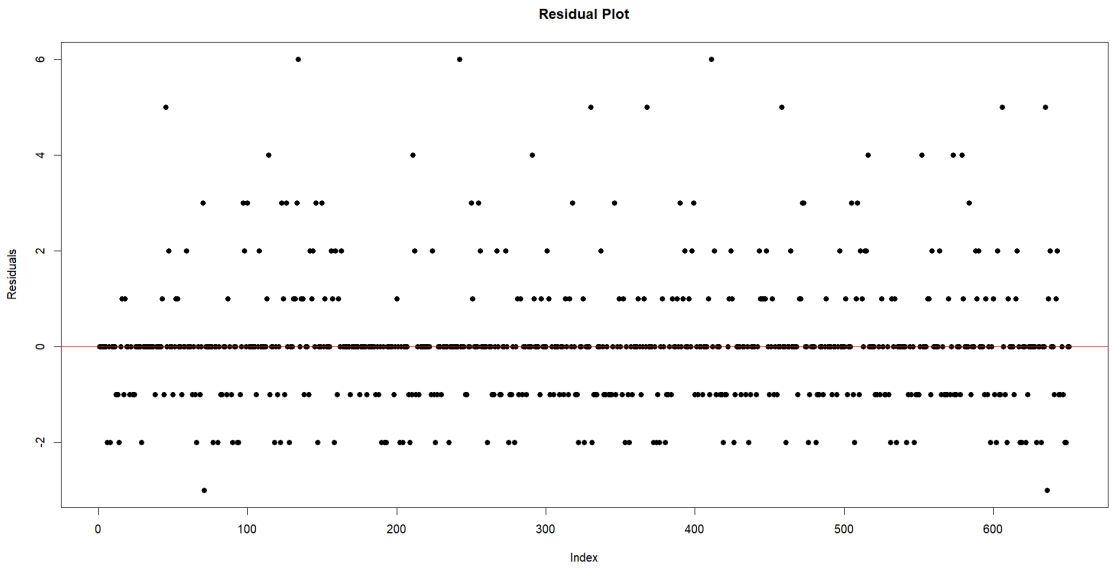
1. Statistical Summary:
   - The mean of the residuals is very close to zero, which suggests that there is no significant bias in the predictions — the model is not systematically over or under-predicting.
   - The standard deviation of the residuals is 1.3130941, which gives us an idea of how spread out the residuals are. In combination with the RMSE, this suggests that there is some variability in the prediction errors.
   - The minimum and maximum residual values are -3 and 6, respectively, indicating the range of the model's prediction errors. The presence of residuals as high as 6 could be a concern, depending on the scale of your dependent variable.



2. Histogram of Residuals: The histogram of residuals can give us an indication of the normality of the errors, which is an assumption for many linear models. Ideally, we would want the residuals to be normally distributed around zero. In this histogram, there appears to be a slight right skew, as there are residuals that extend further to the right, indicating that there might be some outliers, or the model is not capturing all the variability in the data.

```
Mean Absolute Error (MAE): 0.797235
>
> # Calculate RMSE (Root Mean Squared Error)
> rmse <- sqrt(mean((df$STRESS - df$Combined_Predictions)^2))
> cat("Root Mean Squared Error (RMSE):", rmse, "\n")
Root Mean Squared Error (RMSE): 1.312237
```

3. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE): These metrics provide a measure of how well the model's predictions match the actual data. A lower value indicates a better fit.
   - MAE of 0.797235: This suggests that, on average, the model's predictions are about 0.797 units away from the actual values. Whether this is acceptable depends on the context and the scale of the data.
   - RMSE of 1.312237: Since RMSE penalizes larger errors more heavily, this higher RMSE (compared to the MAE) suggests the presence of some larger errors (likely outliers). Again, the acceptability of this value depends on the context of the problem and the variance inherent in the data.



4. Residual Plot: In an ideal situation, the residuals would be randomly scattered around the horizontal axis (which represents a residual value of zero). This plot does not appear to have a clear pattern, which is good, but it does seem to show that the residuals are not equally variable across all observations.

In summary, the model seems to have a decent performance with no significant bias, but there might be some issues with outliers or model assumptions (such as normality and homoscedasticity of residuals) that could be investigated further. It would be beneficial to look into any data points with large residuals to understand why the model is not predicting those correctly. Depending on the context and purpose of the model, different thresholds for MAE and RMSE could be considered acceptable.

Task 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.

```
> # Extract count model coefficients
> count_coef <- coef(zip_model, "count")
>
> # Extract zero-inflation model coefficients
> zero_infl_coef <- coef(zip_model, "zero")
>
> # Print the coefficients
> print(count_coef)
 (Intercept)          AGE        COHES       ESTEEM       GRADES      SATTACH
 2.241921461  0.028593335 -0.008416582 -0.025398862 -0.018839800 -0.009720928
> print(zero_infl_coef)
 (Intercept)          AGE        COHES       ESTEEM       GRADES      SATTACH
-4.567758532  0.129397540  0.018001919 -0.002519984  0.014594871  0.028592663
```

Count Model Coefficients (Poisson Part):

- `(Intercept)`: The baseline log count of `STRESS` when all other variables are 0 is approximately 2.242.
- `AGE`: The log count of `STRESS` increases by approximately 0.029 units for each additional year of age, though this was not statistically significant in the model summary.
- `COHES`: The log count of `STRESS` decreases by approximately 0.008 units for each unit increase in `COHES`, indicating that higher cohesion is associated with lower stress.
- `ESTEEM`: The log count of `STRESS` decreases by approximately 0.025 units for each unit increase in `ESTEEM`, suggesting that higher self-esteem is associated with lower stress.
- `GRADES`: The log count of `STRESS` decreases by approximately 0.019 units for each unit increase in `GRADES`, indicating that better grades might be associated with lower stress, although this was not statistically significant in the model summary.

- `SATTACH`: The log count of `STRESS` decreases by approximately 0.010 units for each unit increase in `SATTACH`, which suggests that higher school attachment might be associated with lower stress, but again, this was not statistically significant in the model summary.

Zero-Inflation Model Coefficients (Binomial Part):

- `(Intercept)`: The baseline log-odds of having a zero count of `STRESS` (as opposed to a count predicted by the Poisson part) when all other variables are 0 is approximately -4.568.
- `AGE`: The log-odds of a zero count of `STRESS` increases by approximately 0.129 units for each additional year of age, though this was not statistically significant in the model summary.
- `COHES`: The log-odds of a zero count of `STRESS` increases by approximately 0.018 units for each unit increase in `COHES`, which was not statistically significant in the model summary.
- `ESTEEM`: The log-odds of a zero count of `STRESS` decreases by approximately 0.003 units for each unit increase in `ESTEEM`, suggesting no significant impact on the likelihood of zero counts of `STRESS`.
- `GRADES`: The log-odds of a zero count of `STRESS` increases by approximately 0.015 units for each unit increase in `GRADES`, but this was not significant.
- `SATTACH`: The log-odds of a zero count of `STRESS` increases by approximately 0.029 units for each unit increase in `SATTACH`, indicating that higher school attachment is associated with a higher probability of a zero count of `STRESS`, but this relationship was not statistically significant.

In both models, the coefficients represent the log transformation of the respective dependent variables (count of `STRESS` for the Poisson part and the odds of a zero count for the binomial part). The significance of these coefficients should be considered in the context of the p-values presented in the model summary output, as the coefficients alone do not convey whether the relationships are statistically significant.

```
> # Display the summary of the model
> summary(zip_model)

Call:
zeroinfl(formula = STRESS ~ AGE + COHES + ESTEEM + GRADES + SATTACH, data = df, dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.4905 -0.9106 -0.2290  0.6319  4.0480

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.241921   0.520343   4.309 1.64e-05 ***
AGE          0.028593   0.032246   0.887  0.37523
COHES       -0.008417   0.003424  -2.458  0.01396 *
ESTEEM      -0.025399   0.009227  -2.753  0.00591 **
GRADES      -0.018840   0.010927  -1.724  0.08469 .
SATTACH     -0.009721   0.006767  -1.437  0.15084

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.56776    1.75039  -2.610  0.00907 **
AGE          0.12940    0.10916   1.185  0.23586
COHES        0.01800    0.01217   1.479  0.13908
ESTEEM      -0.00252    0.03263  -0.077  0.93844
GRADES       0.01460    0.03767   0.387  0.69845
SATTACH      0.02859    0.02464   1.160  0.24587
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 21
Log-likelihood: -1134 on 12 Df
```

The summary of the Zero-Inflated Poisson (ZIP) model gives us several pieces of information regarding the relationship between `STRESS` and the predictors `AGE`, `COHES`, `ESTEEM`, `GRADES`, and `SATTACH`. Here is an interpretation of the output:

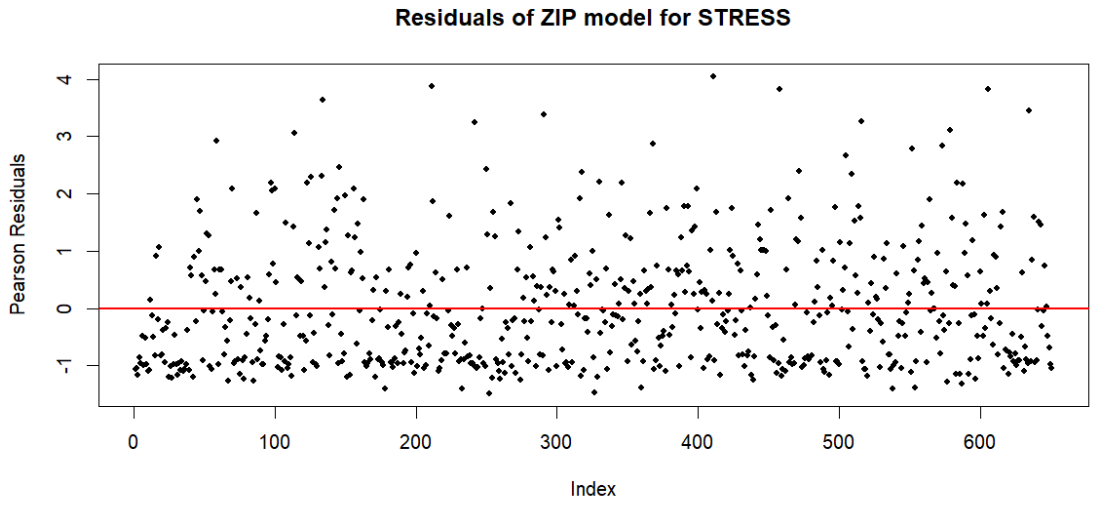
1. Pearson Residuals: The residuals from the model (differences between observed and predicted values) have a minimum value of -1.4905 and a maximum value of 4.0480. The median is close to zero (-0.2290), which is a good sign, but the range indicates that there might be some outliers or influential points, particularly on the positive side, which could affect the model fit.
2. Count Model Coefficients (Poisson Part):
   - The `Intercept` has a highly significant positive coefficient, meaning there's a base level of `STRESS` when all predictors are at zero.
   - `COHES` and `ESTEEM` have negative coefficients and are statistically significant, indicating that higher levels of cohesion and self-esteem are associated with a reduction in the count of `STRESS`.

- `AGE`, `GRADES`, and `SATTACH` are not statistically significant at the conventional 0.05 level, although `GRADES` is marginally significant (p = 0.08469, indicated by `.`, which means significant at the 0.1 level), suggesting a potential trend where better grades might be associated with lower stress.

3. Zero-Inflation Model Coefficients (Binomial Part):
   - The significant negative `Intercept` suggests that there is a tendency for zero counts of `STRESS` that is not captured by the Poisson part of the model.
   - None of the predictor variables are statistically significant, meaning none of them have a significant impact on the likelihood of observing excess zeros in `STRESS`.
4. Model Diagnostics:
   - The BFGS optimization algorithm used to fit the model converged in 21 iterations, which is an indication of the model's computational stability.
   - The log-likelihood value of -1134 on 12 degrees of freedom is a measure of the model's fit to the data, used for comparison with other models or a null model (not provided here).

In summary, the ZIP model indicates that `COHES` and `ESTEEM` are significant predictors for the count of `STRESS`, with higher values being associated with a lower count. The zero-inflation part of the model suggests that the presence of excess zeros is significant overall but is not significantly influenced by the predictors included in the model.

```
> # Print AIC and BIC values for the original model
> print(paste("AIC for original ZIP model:", aic_value_original))
[1] "AIC for original ZIP model: 2291.04583899369"
> print(paste("BIC for original ZIP model:", bic_value_original))
[1] "BIC for original ZIP model: 2344.78795470019"
```



1. AIC and BIC Values: The Akaike Information Criterion (AIC) for the model is approximately 2291.05, and the Bayesian Information Criterion (BIC) is approximately 2344.79.

2. Residual Plot: The residual plot shows the Pearson residuals plotted against the index of observations. The residuals should ideally be randomly scattered around zero, indicating that the model's predictions are unbiased. The plot does show a random scatter of residuals around the horizontal line at zero, which is good. However, there are some residuals that stand out at the top, indicating potential outliers or instances where the model may not be capturing the data's structure as well.

Creating the Zip Model with Significant Variables:

```
> zip_model_sig <- zeroinfl(STRESS ~ COHES + ESTEEM | 1, data = df, dist = "poisson")
>
> # Display the summary of the model
> summary(zip_model_sig)

Call:
zeroinfl(formula = STRESS ~ COHES + ESTEEM | 1, data = df, dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.3069 -0.9676 -0.2559  0.5909  4.1995

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.519799   0.243479  10.349  < 2e-16 ***
COHES       -0.012671   0.003080  -4.114 3.89e-05 ***
ESTEEM      -0.033103   0.008572  -3.862 0.000113 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1184     0.1212  -9.225   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 9
Log-likelihood: -1143 on 4 Df
>
> # Print the coefficients for the significant model
> count_coef_sig <- coef(zip_model_sig, "count")
> zero_infl_coef_sig <- coef(zip_model_sig, "zero")
>
> print(count_coef_sig)
(Intercept)       COHES      ESTEEM
 2.51979893 -0.01267140 -0.03310316
> print(zero_infl_coef_sig)
(Intercept)
  -1.118406
```

The summary of the revised Zero-Inflated Poisson (ZIP) model, which only includes the significant variables from the previous analysis (`COHES` and `ESTEEM` for the count model and only the intercept for the zero-inflation model), provides the following insights:

1. Pearson Residuals: The range of Pearson residuals is from -1.3069 to 4.1995, with a median of -0.2559. Like the previous model, this range suggests some variability in the model's fit, with a few large positive residuals indicating potential outliers or model misfits.
2. Count Model Coefficients (Poisson Part):
   - `(Intercept)`: The baseline log count of `STRESS` is approximately 2.520, and it's highly significant. This suggests a positive base level of `STRESS` when `COHES` and `ESTEEM` are at their reference levels.
   - `COHES`: The coefficient is -0.01267 and is statistically significant ($p < 0.0001$). This indicates that higher cohesion is associated with a lower count of `STRESS`.
   - `ESTEEM`: The coefficient is -0.03310 and is also statistically significant ($p < 0.0001$). This suggests that higher self-esteem is associated with a lower count of `STRESS`.
3. Zero-Inflation Model Coefficients (Binomial Part):
   - `(Intercept)`: The coefficient is -1.1184 and is highly significant. This indicates a strong baseline propensity for zero counts of `STRESS` that is not explained by the Poisson part of the model.

Model Diagnostics:

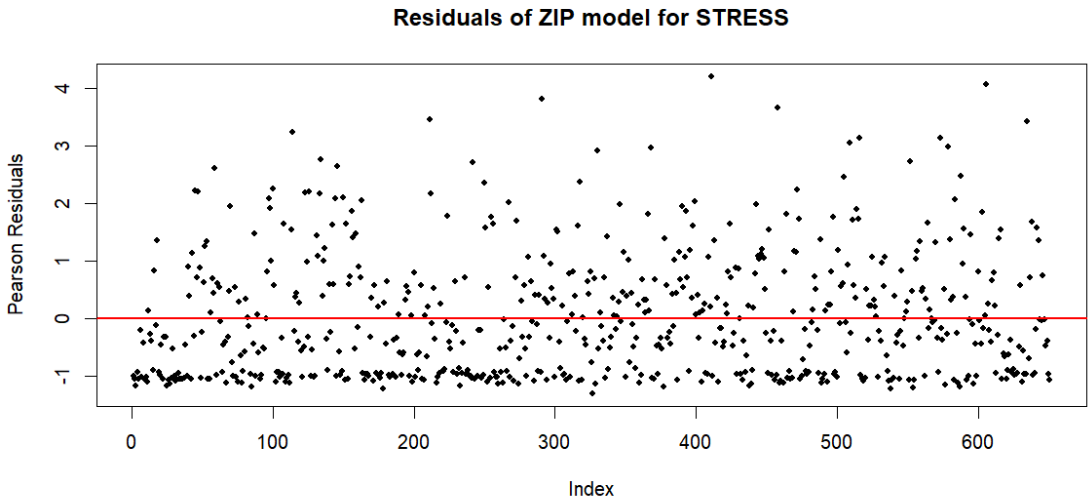
- The model converged in 9 iterations of the BFGS optimization algorithm, which is fewer than the previous model, indicating potentially a more stable model.
- The log-likelihood of the model is -1143 on 4 degrees of freedom, suggesting the overall fit of the model to the data. This value is slightly worse than the previous model (-1134 on 12 Df), indicating that removing the non-significant variables has slightly reduced the model's fit to the data.

4. Interpretation of Coefficients:
   - The negative coefficients for the count model (`count_coef_sig`) confirm the significant negative relationship of `COHES` and `ESTEEM` with the count of `STRESS`.

- The negative coefficient for the zero-inflation model (`zero_infl_coef_sig`) confirms the significant intercept, indicating the propensity for zero counts of `STRESS`.

In summary, the revised model suggests that both `COHES` and `ESTEEM` are significant predictors of `STRESS`, with higher values associated with lower stress levels. The zero-inflation part indicates a significant baseline level of zero `STRESS` counts. The model appears to fit the data reasonably well, although some outliers or misfits are indicated by the range of Pearson residuals.

```
> # AIC and BIC for the model
> aic_value <- AIC(zip_model_sig)
> bic_value <- BIC(zip_model_sig)
>
> # Print AIC and BIC values
> print(paste("AIC:", aic_value))
[1] "AIC: 2293.10532864024"
> print(paste("BIC:", bic_value))
[1] "BIC: 2311.01936720907"
```



1. AIC and BIC Values: The Akaike Information Criterion (AIC) for the model is approximately 2291.05, and the Bayesian Information Criterion (BIC) is approximately 2344.79.

2. Residual Plot: The plot does show a random scatter of residuals around the horizontal line at zero, which is good. However, there are some residuals that stand out at the top, indicating potential outliers or instances where the model may not be capturing the data's structure as well.

When comparing the original `zip_model` with the full set of predictors (AGE, COHES, ESTEEM, GRADES, SATTACH) to the simplified `zip_model_sig` that includes only significant predictors (COHES, ESTEEM), we consider the outputs from the goodness-of-fit measures, likelihood ratio tests, and residual plots.

Goodness-of-Fit Measures:

- The AIC for the original model is slightly lower (2291.05) than for the simplified model (2293.10), indicating that the original model with more predictors may fit the data slightly better when considering the number of parameters.
- The BIC for the original model is higher (2344.79) than for the simplified model (2311.02), suggesting that when penalizing for model complexity, the simplified model is preferred.

Residual Plots: Both residual plots show the Pearson residuals scattered around the zero line, indicating that both models do not have systematic biases. However, the presence of outliers or extreme residuals in both plots suggests that there may be some data points that neither model captures well.

In summary, the original `zip_model` has a slightly better AIC, indicating a marginally better fit to the data considering the number of predictors. The simplified `zip_model_sig` has a better BIC, suggesting that it is a more parsimonious model and may be preferred when penalizing for complexity. The choice between the two models may depend on

the specific context and objectives of the analysis. If the goal is to have a more parsimonious model, the simplified model may be preferred. If the goal is to explain as much variability as possible, the original model with more predictors may be the better choice.

The models reflects a refinement in statistical analysis, moving from simple linear models to more complex count models that account for the distribution of the response variable, `STRESS`.

1. Ordinary Least Squares (OLS) Regression: The first approach was an OLS regression, which assumes a continuous response variable. It provided a baseline model with a certain level of fit as indicated by its residuals, coefficient significance, and R-squared values.
2. Transformation and OLS Regression: To improve the model and possibly meet the assumptions of OLS better, a log transformation of `STRESS` was applied, which can help to normalize the distribution of errors or deal with heteroscedasticity. The residuals improved slightly, but the R-squared remained similar, indicating a marginal improvement in the model's explanation of the variability.
3. Poisson Regression: Recognizing that `STRESS` is count data, a Poisson regression was fitted. This model is more appropriate for count data but assumes that the mean and variance of the response are equal (equidispersion). The Poisson model showed signs of improvement in fit as indicated by the AIC value, but it's also clear that the data exhibited overdispersion (variance greater than the mean), which Poisson models cannot handle well.
4. Negative Binomial Regression: To address overdispersion, a Negative Binomial regression was employed. This model allows the variance to exceed the mean. The AIC for the Negative Binomial model was lower than the Poisson model, indicating a better fit.
5. Zero-Inflated Models: Finally, recognizing that the data might have an excess number of zeros, two Zero-Inflated models were considered: `zip_model` and `zip_model_sig`. These models account for the excess zeros by combining a count model (Poisson or Negative Binomial) with a logit model that predicts the occurrence of zero outcomes.
    - The `zip_model` included all original predictors. It had a good fit as indicated by its log-likelihood, but there was a need to check if all predictors were necessary.
    - The `zip_model_sig` is a more parsimonious version, including only significant predictors. It resulted in a slightly worse log-likelihood but a better BIC value, indicating a more balanced model in terms of fit and complexity.

A good modeling approach for this data?

- Based on AIC and BIC values, the `zip_model_sig` is more parsimonious and balances model fit with complexity better than the full `zip_model`. However, the choice of the 'better' model may also depend on the specific research question or application. If interpretability and model simplicity are valued and the slight loss in fit is acceptable, `zip_model_sig` may be preferable. If the goal is to capture as much variability as possible and every predictor is theoretically justified, the full `zip_model` might be the choice despite its higher complexity.
- The Negative Binomial model has a lower AIC compared to the Poisson model, indicating a better fit due to the overdispersion in the data.
- In comparison to the OLS and transformed OLS models, the count models (Poisson, Negative Binomial, and Zero-Inflated) are more theoretically sound for count data and have shown better goodness-of-fit measures.
- Each model has its merits, and the final selection should consider the purpose of the modeling, the need for parsimony, the theoretical justification of the predictors, and the overall goodness-of-fit measures.