Date: November 30, 2017
From: Briana Schumacher, Kiera Murphy, Abby Koenig, Kanchan Sayers
To: The Pennsylvania State University STAT 470 Class
Re: College Students Case Study

COLLEGE STUDENTS CASE STUDY

ABSTRACT. This project used generalized linear modeling to analyze the relationship between caffeine consumption and the GPA, stress level, gender, and major of college students.

CONTENTS

1. PROJECT DESCRIPTION

This data set was collected from college age students. The purpose of the study was to evaluate how caffeine consumption, and other factors, affect a college student's GPA; we wanted to determine the relative significance of each of the variables.

How the data was collected and reviewed: We created a questionnaire (refer to appendix) with questions regarding GPA, caffeine consumption, stress level, gender, year in college, major, and university attended. This survey was anonymous in order to receive honest, unbiased responses. We aimed to collect a large enough population of responses that would give us information to draw a significant conclusion.

1.1. Research Questions.

The client is targeting the following research questions:

> **Q1:** How does the student's caffeine consumption, stress level, and gender affect their GPA?
> **Q2:** Can we differentiate between STEM and non-STEM students based on their GPA, caffeine consumption, stress level, and gender?
> **Q3:** Is there a relationship between gender and stress level?

1.2. Statistical Questions.
To answer the client's first research question, we investigated the following statistical question:

> **Q1:** Is there a significant relationship between the student's caffeine consumption, stress level, gender, and their GPA?

To answer the client's second research question, we investigated the following statistical question:

> **Q2:** Do STEM and non-STEM students differ significantly based on their GPA, caffeine consumption, stress level, and gender?

To answer the client's third research question, we investigated the following statistical question:

> **Q3:** Is there a significant relationship between gender and stress level?

1.3. Variables of Interest.
There are 7 variables that we collected data for: stress level, gender, GPA, major, academic standing, institution of study, and caffeine consumption. Gender was listed originally as male or female on the survey, but was made binary for the purpose of analysis by logistic regression; females were re-coded as 0's, and males were re-coded as 1's. From the variable for college major, we created a new binary variable (STEM) to differentiate between STEM, 1, and non-STEM, 0, majors, also for the purpose of analysis by logistic regression. For the servings of caffeine we had the options 0, 1, 2, 3, 4, 5-10, and >10 servings; 5-10 was changed to 5 and >10 to 11, in order for R-studio to understand the data. Table 1 provides the name and a brief description of each variable in addition to the associated levels and necessary comments.

Variable	**Description**	**Levels**	**Comments**	**Type**
Stress Level	How stressed the student considers his/herself	1-5	1- the least stressed, 5 - the most stressed	Ordinal (Explanatory)
Gender	Whether the student is a Male or a Female	Male/Female	0 - Female 1 - Male	Binary (Explanatory) (Response)
GPA	College grade point	0.00-4.00	On a 4.00 scale	Continuous

	average			(Explanatory) (Response)
Major	Whether the student is in a STEM major or a non-STEM major	STEM or non-STEM major	STEM (Science, Technology, Engineering, Math) vs. non-STEM major	Binary (Explanatory) (Response)
Caffeine Consumption	Servings of caffeine per day	0, 1, 2, 3, 4, 5-10, >10	0: no caffeine consumed, >10: more than 10 servings of caffeine consumed per day	Ordinal (Explanatory)
Year	Year in college	1,2,3,4,5,6	1-Freshman, 2-Sophomore, 3-Junior, 4- Senior, 5-Super-Senior, 6-Above	Categorical (Explanatory)

Table 1: *The table includes the name, description, level, comments, and type for each variable. Explanatory variables and response variables are noted in the type.*

2. Exploratory Data Analysis (EDA)

The data was reviewed and slightly modified prior to the statistical analysis. We checked for outliers, and excluded a few variables we had asked about in our survey, such as university attended and hours of sleep per night. No data were missing and therefore no further modifications were needed (see appendix for column summary). Table 2 shows the descriptive statistics for the variables GPA and stress. We did not include gender, major, caffeine consumption or year because they are not continuous variables. (*See appendix for the command code*).The variables gender and STEM were recoded as binary numeric variables for ease of analysis.

We looked at the descriptive statistics for a few of our explanatory variables to get an overall sense of the data we had received. Looking at these summaries, we found a few outliers, but did not remove them from our dataset.

Table 2: Descriptive Statistics for Explanatory Variables

Variables	N	Mean	Standard Deviation	Minimum	Maximum

GPA	51	3.39	0.43	2.3	3.96
Stress	51	3.49	1.13	1	5

Table 2: *The table displays each of the explanatory variables. There were a total of 51 observations across the study. Additional information pertaining to the mean, maximum, minimum, and standard deviation are given.*

We looked at the relationship between stress level and a student's gender. In Figure 1, shown below, the spread of the data suggests that, on average, men have a significantly lower stress level. We can see that the mean stress level for males in our dataset was 2.46 and for females it was 4 (*Figure 1*).

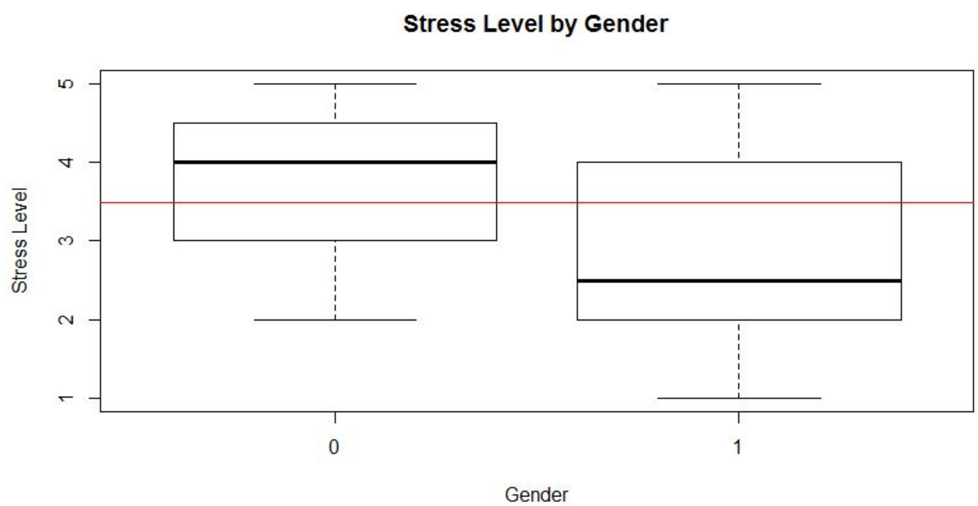


Figure 1. *A boxplot of Gender versus Stress-Level. In this box-plot for Gender, 0 represents Female and 1 represents Male. For the Stress-Level 1 represents the least amount of stress, while 5 is the most amount of stress. The mean for both genders, 3.49, is represented by a red line.*

Finally, we looked at potential interactions between various combinations of the explanatory and response variables but non significantly improved the performance of any of our models.

3. Statistical Analysis

To answer our first statistical question as stated in section 2, we first looked at the relationship between the student's GPA and their caffeine consumption, stress level, and gender. To do this, we first tested a logistic regression model that used all of our explanatory variables to predict the student's GPA. We first performed a chi-squared test to determine each variable's significance to the model (refer to the Appendix). From the output, we can see that the p-value for major is 6.088e-06, which is significant. Because major was significant, the regression model was used to determine which majors were most significant. We found that only business management, mathematics, mechanical engineering, and science were the majors with a p-value below 0.05, and therefore the only ones with a significant effect on GPA.

To answer our second statistical question, we tested to see if the GPA, caffeine consumption, stress level, and gender of STEM students is significantly different than that of non-STEM students. To do this we created a generalized linear model with STEM as the response variable. We coded STEM as a binary response so we could perform a logistic regression. In our output, we found that our only significant explanatory variable was gender, with the next closest being year. The p-value for gender, 0.004, allowed us to conclude that there is a significant relationship between being a STEM major and gender. The student's year of academic standing had a p-value of 0.068779; year was the next closest to being significant, but with alpha= 0.05 we cannot confidently conclude that there exists a significant relationship. We then created a revised model to exclude non-significant variables. Year was ultimately included in this revised model because its exclusion resulted in a model with a higher AIC and lower significance for gender. In this revised model (found in our appendix), gender was again significant with a p-value of 0.00988.

To answer our third statistical question and test if there is a significant relationship between gender and stress level, we used a t-test. It can be visually observed from the boxplot in figure 1 that the relationship between a student's stress level and their gender differs; on average, men appear to be less stressed than women. We can see that the mean stress level for males in our dataset was 2.46 and for females it was 4, the p-value for our t-test, 0.01102, confirms that this difference is in fact significant; we can therefore conclude that a student's relative stress level is related to their gender.

4. Recommendations.

The first research question was: How does the student's caffeine consumption, stress level, and gender affect their GPA?
We found that major had a significant relationship on a student's GPA. Performing further analysis, we found that STEM majors such as science, mechanical engineering

and mathematics and a non-STEM major, business management were the majors with a significant relationship to GPA.

The second research question was: Can we differentiate between STEM and non-STEM students based on their caffeine consumption, stress level, and gender?
We found that the only significant variable was gender. This means that we can only differentiate STEM vs. non-STEM majors based on gender and not by their caffeine consumption, or stress level.

The third research question was: Is there a relationship between gender and stress level?
We found that there is a significant relationship between gender and stress level. Given the gender (female vs. male), there is a significance when determining the amount of stress the student has. We concluded that men are on average less stressed than female students.

5. Resources.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

http://www.statisticssolutions.com/assumptions-of-linear-regression/

6. Considerations.

There are several additional considerations to help to fully understand this study:

- Although we conducted an anonymous study, some of the students may have responded untruthfully. Some may have lied about their GPA, or there also may have been people who have taken the survey who were not college students.
- Students have different gauges of stress. The average stress level is very subjective and different people will have different interpretations of stress. Stress levels can also vary by day, year, etc.
- By just asking our friends, family, and people in our classes, our sample may not be representative of all Penn State students. This also resulted in a small sample size but with equal quantities of observations across the study.
- Different caffeinated drinks contain varying amounts of caffeine. It also may be hard to measure the amount of servings you drink per day. Sometimes you may

consume a beverage with caffeine and may not be aware that the drink has caffeine.

- All of our measurements of data were in whole numbers. Numbers of sleep, servings of caffeine, and year were all asked as whole numbers, which affects our precision. The only continuous variable is GPA.
- The way in which we asked the survey questions, and how we posed our response options can also have a large impact on how people respond to our survey and can be a potential source of bias.

7. Acknowledgment of Work

It is a pleasure to thank our mentor, Matthew Beckman, and our Teaching Assistant, Christian Schmid, for providing helpful suggestions and information related to our data set during the process of finishing this project.

APPENDIX

Survey Monkey We Made and Used to Gather Data from Students:

Stat 470 Caffeine Related Survey

Stat 470 Survey

All of the data you submit will remain anonymous.

1. What is your cumulative GPA (on a 4.0 scale)?

2. On average how much caffeine do you drink per day? (with 1 serving being 1 caffeinated drink)

- None
- 1 Serving
- 2 Servings
- 3 Servings
- 4 Servings
- 5-10 Servings
- Greater than 10 servings

3. How would you rate your average amount of stress during a day at college?

Not Stressed	Minimally Stressed	Normal	A Little Stressed	Very Stressed
☆	☆	☆	☆	☆

4. Are you a Male or Female?

- Male
- Female

5. On average how many hours do you sleep per night at your university?

6. What year are you in college?

- Freshman
- Sophomore
- Junior
- Senior
- Super-Senior
- More than a Super-Senior

7. What major are you? (Ex: Science, Labor Arts, Art, etc.)

8. What university do you attend?

R-Studio Coding:

```
## Recode 'Gender' and 'STEM' as binary numeric variables for ease of analysis

FinalProject$Gender[FinalProject$Gender == "f"] <- "0"
FinalProject$Gender[FinalProject$Gender == "m"] <- "1"
FinalProject$Gender <- as.numeric(FinalProject$Gender)

FinalProject$STEM[FinalProject$STEM == "N"] <- 0
FinalProject$STEM[FinalProject$STEM == "Y"] <- 1
FinalProject$STEM <- as.numeric(FinalProject$STEM)

FinalProject$Servings[FinalProject$Servings == "5-10"] <- 5
FinalProject$Servings[FinalProject$Servings == ">10"] <- 11
FinalProject$Servings <- as.numeric(FinalProject$Servings)
```

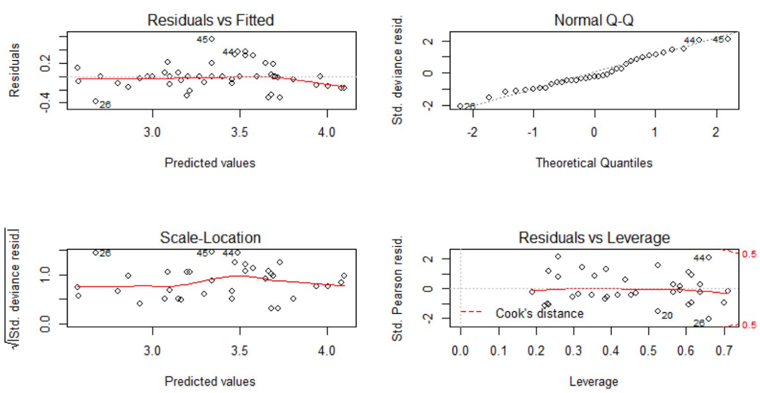
Check for missing values ## Generalized linear model with STEM as response
colSums(is.na(FinalProject))

```
GPA Servings  Stress  Gender  Sleep  Year  Major  STEM  Uni.
 0      0        0       0      0      0      0     0     0
```

Question 1:

Generalized linear model with GPA as the response variable
Model is similar to those with STEM as response but GPA is not
binary so the family is gaussian instead
model2 <- glm(GPA ~ ., family = gaussian(link = "identity"), data = FinalProject)

plot(model2)



anova(model2, test = 'Chisq')

```
Analysis of Deviance Table

Model: gaussian, link: identity

Response: GPA

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        50     9.2477
Servings  1   0.3040        49     8.9436   0.07535 .
Stress    1   0.1065        48     8.8372   0.29268
Gender    1   0.0133        47     8.8239   0.70978
Sleep     1   0.2148        46     8.6091   0.13503
Year      1   0.0802        45     8.5289   0.36105
Major    21   5.9714        24     2.5575 6.088e-06 ***
STEM      0   0.0000        24     2.5575
Uni.      5   0.7308        19     1.8267   0.17961
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

summary(model2)

```
Call:
glm(formula = GPA ~ ., family = gaussian(link = "identity"),
    data = FinalProject)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.37356  -0.10397   0.00000   0.03422   0.56121

Coefficients: (7 not defined because of singularities)
                            Estimate Std. Error t value Pr(>| ***
(Intercept)                  6.24729    1.45120   4.305 0.000 *
Servings                    -0.10489    0.03976  -2.638 0.016
Stress                       0.08315    0.09934   0.837 0.412
Gender                       0.11827    0.18296   0.646 0.525 *
Sleep                       -0.25692    0.09714  -2.645 0.015
Year                        -0.07586    0.10281  -0.738 0.469 **
MajorAnimal Science         -1.13478    0.36177  -3.137 0.005
MajorBusiness                0.43971    0.49932   0.881 0.389 *
MajorBusiness Management    -1.65175    0.73573  -2.245 0.036 ***
MajorCommunications         -1.32358    0.28283  -4.680 0.000
MajorComputer Science       -1.25577    0.62940  -1.995 0.060 .
MajorEducation              -0.35168    0.69680  -0.505 0.619
MajorEnglish                -0.26453    0.35975  -0.735 0.471
MajorGraphic Design         -0.14421    0.29584  -0.487 0.631 *
MajorHospitality            -1.04645    0.38297  -2.732 0.013
MajorInternational Studies  -0.24497    0.63273  -0.387 0.702
MajorKinesiology            -0.59442    0.37344  -1.592 0.127
MajorLiberal Arts           -0.21002    0.38850  -0.541 0.595
MajorMarketing              -0.60239    0.33939  -1.775 0.091 .
MajorMath                   -0.96268    0.30335  -3.173 0.005 **
MajorMechanical Engineering -1.67212    0.63327  -2.640 0.016 *
MajorNursing                -0.71772    0.42844  -1.675 0.110
MajorPsychology             -0.41617    0.38240  -1.088 0.290
MajorPublic Relations       -0.10023    0.37232  -0.269 0.790
MajorRadio TV Film          -0.44370    0.74381  -0.597 0.557
MajorScience                -0.61558    0.23105  -2.664 0.015 *
MajorStatistics             -0.44220    0.26445  -1.672 0.110
STEM                              NA         NA      NA
Uni.Bryant                        NA         NA      NA
Uni.Colgate Uni              0.01408    0.56950   0.025 0.980
Uni.PSU                     -0.40406    0.46515  -0.869 0.395
Uni.Rowan Uni                     NA         NA      NA
Uni.RPI                           NA         NA      NA
Uni.St. Jos.                      NA         NA      NA
Uni.UConn                   -0.80812    0.56480  -1.431 0.168
Uni.Uni. of Buff.            0.14217    0.61283   0.232 0.819
Uni.URI                           NA         NA      NA
Uni.Villanova               -1.10447    0.62479  -1.768 0.093 .
Uni.West. State                   NA         NA      NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

(Dispersion parameter for gaussian family taken to be 0.09614

    Null deviance: 9.2477  on 50  degrees of freedom
Residual deviance: 1.8267  on 19  degrees of freedom
AIC: 40.936

Number of Fisher Scoring iterations: 2
```

Variable reduction for GPA model only increases AIC and makes variables less significant
Some, but not all, majors are statistically significant (i.e. A communications major can expect to
have a GPA lower than average by 1.32358 points

Question 2:

Only significant variable appears to be gender, with year being almost
significant

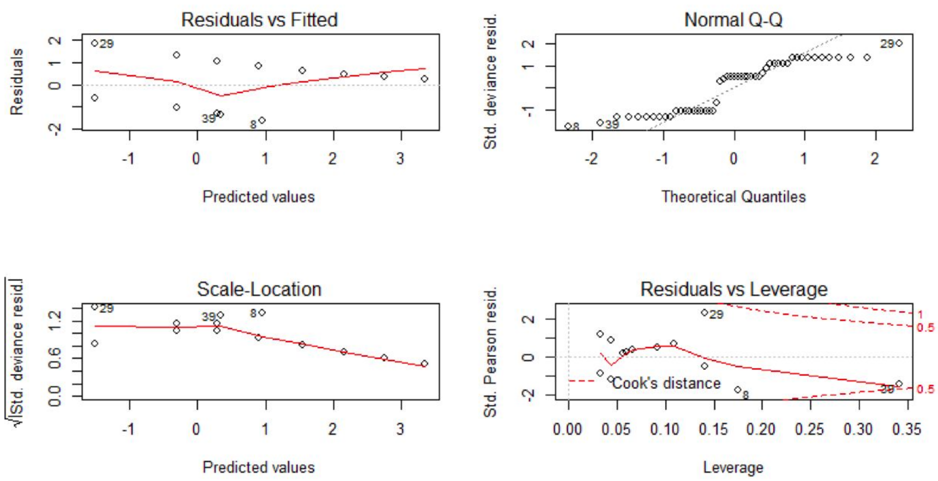
summary(model)

```
Call:
glm(formula = STEM ~ ., family = binomial(link = "logit"), data = FinalProject[,
    -7])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.48999  -0.56869   0.00004   0.65503   1.50555

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -6.3661 10754.0166  -0.001    1.000
GPA                  -1.4026     1.0290  -1.363    0.173
Servings              0.1930     0.3220   0.599    0.549
Stress               -0.9798     0.6269  -1.563    0.118
Gender               19.0261  2832.7338   0.007    0.995
Sleep                -0.4603     0.6789  -0.678    0.498
Year                 -0.1504     0.5412  -0.278    0.781
Uni.Bryant          -21.6375 15470.0347  -0.001    0.999
Uni.Colgate Uni      18.2353 15470.0346   0.001    0.999
Uni.PSU              18.2615 10754.0130   0.002    0.999
Uni.Rowan Uni       -20.7837 15470.0347  -0.001    0.999
Uni.RPI              18.7064 15470.0346   0.001    0.999
Uni.St. Jos.         -2.2928 15208.4710   0.000    1.000
Uni.UConn            18.2722 10754.0131   0.002    0.999
Uni.Uni. of Buff.    38.4737 12308.5261   0.003    0.998
Uni.URI              19.9504 15470.0347   0.001    0.999
Uni.Villanova        -4.0417 15208.4713   0.000    1.000
Uni.West. State      -1.8621 15208.4711   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69.104  on 50  degrees of freedom
Residual deviance: 37.623  on 33  degrees of freedom
AIC: 73.623

Number of Fisher Scoring iterations: 18
```

Model revision to exclude non-significant variables; Year was included
because the model performs worse without it

model.1 <- glm(STEM ~ Year + Gender, family = binomial(link = "logit"), data = FinalProject[, -7])

plot(model.1)



anova(model.1, test = ‘Chisq’) summary(model.1)

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: STEM

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                      50     69.104
Year    1   0.5064        49     68.598 0.476701
Gender  1   9.6208        48     58.977 0.001924 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

summary(model.1)

```
Call:
glm(formula = STEM ~ Year + Gender, family = binomial(link = "logit"),
    data = FinalProject[, -7])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5979  -1.0500   0.4684   1.0549   1.8502

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7178     1.7240  -1.576  0.11493
Year          0.6026     0.3885   1.551  0.12082
Gender        2.4621     0.9543   2.580  0.00988 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69.104  on 50  degrees of freedom
Residual deviance: 58.977  on 48  degrees of freedom
AIC: 64.977

Number of Fisher Scoring iterations: 4
```

Generalized linear model with STEM as the response variable
Family = binomial because the STEM is a binary variable

model <- glm(STEM ~ ., family = binomial(link = "logit"), data = FinalProject[, -7])

anova(model, test = 'Chisq')

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: STEM

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        50     69.104
GPA       1   2.2504        49     66.854 0.133580
Servings  1   1.4117        48     65.442 0.234778
Stress    1   0.6740        47     64.768 0.411661
Gender    1   8.0210        46     56.747 0.004624 **
Sleep     1   0.0216        45     56.726 0.883082
Year      1   3.3119        44     53.414 0.068779 .
Uni.     11  15.7909        33     37.623 0.149071
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test for significant difference in average stress level of males and females
t.test(Stress ~ Gender, data = FinalProject)

```
        Welch Two Sample t-test

data:  Stress by Gender
t = 2.7733, df = 22.228, p-value = 0.01102
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2470335 1.7085220
sample estimates:
mean in group 0 mean in group 1
       3.777778        2.800000
```

Assumptions for Logistic Regression

By looking at the 4-in-1 plot it can be seen that the following assumptions are met for logistic regression.

- **A Linear Relationship:**
  There exists a linear relationship between explanatory variables and the response variables.
- **Multivariate Normality:**
  According to the QQ-plot in each fitted model, we can see that all the points lying on a straight line roughly, which implies that the multivariate normality assumption is met.
- **No Multicollinearity:**
  There is no multicollinearity in the model.
- **Homoscedasticity**
  Can be seen from the residual plots, the residuals points are equally distributed across all values of the independent variables. Thus, homoscedasticity is met.