STAT 302: Homework 3

Simple Linear Regression

Jaiden Atterbury

Due: 04-30-23 at 11:59 PM

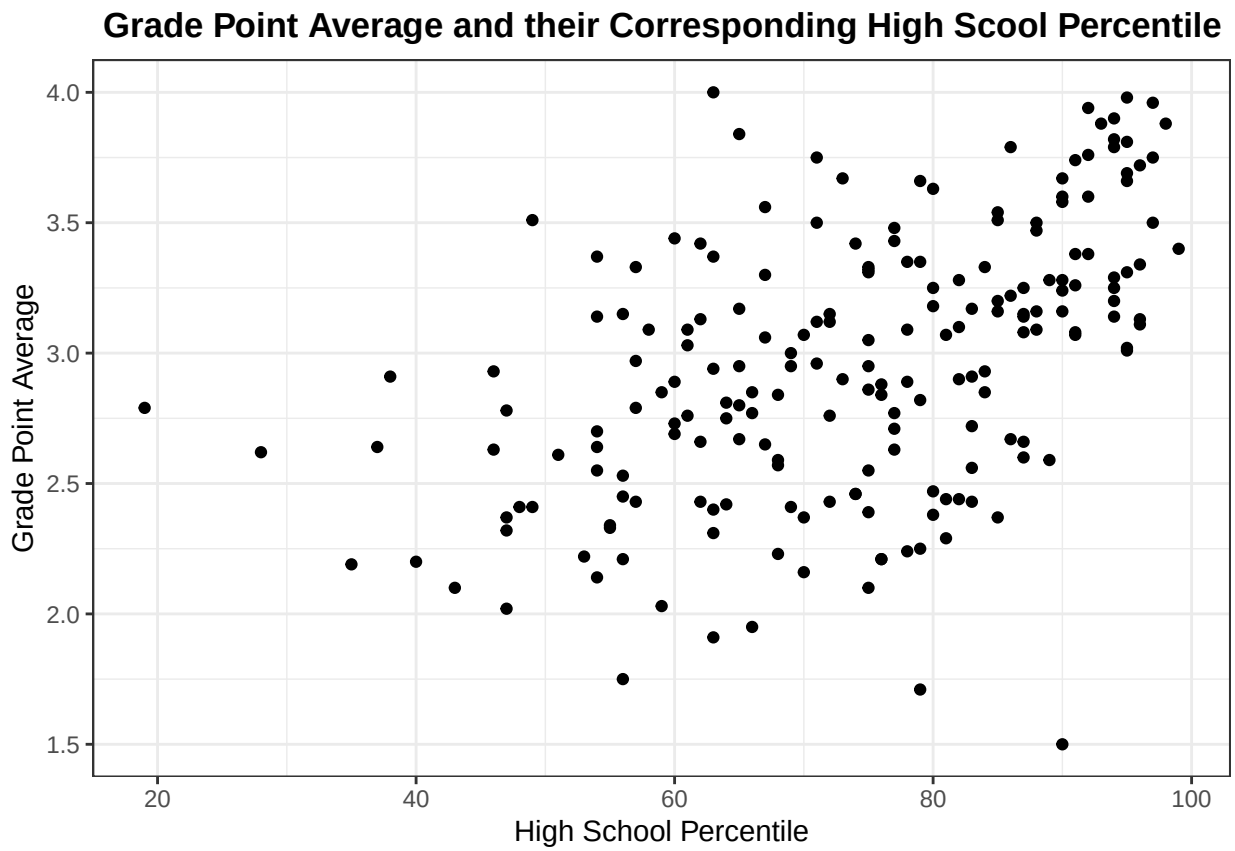
1. The data set for this question is **Focus**. Select any four variables of your choice and use one visualization plot to tell a more complete and compelling story of the data set masking use of all four variables selected. Consider using at least **color, faceting, theme** among many others in ggplot2. Write at least two paragraphs for this. One explaining why you choose those variables and the other what you see from the graph.

The following plots are constructed using the data in the **Focus** data set which contains information on undergraduate students at the University of Wisconsin - Eau Claire (UWEC). In particular, the variables that show up in the below plots are `GPA`, `HSP`, `SEX`, and lastly `CLASS`.

Before I dive into this combined plot though, we will look at the normal scatter-plot with no faceting and color to see how much insight we can gain from our choices of faceting and coloring.

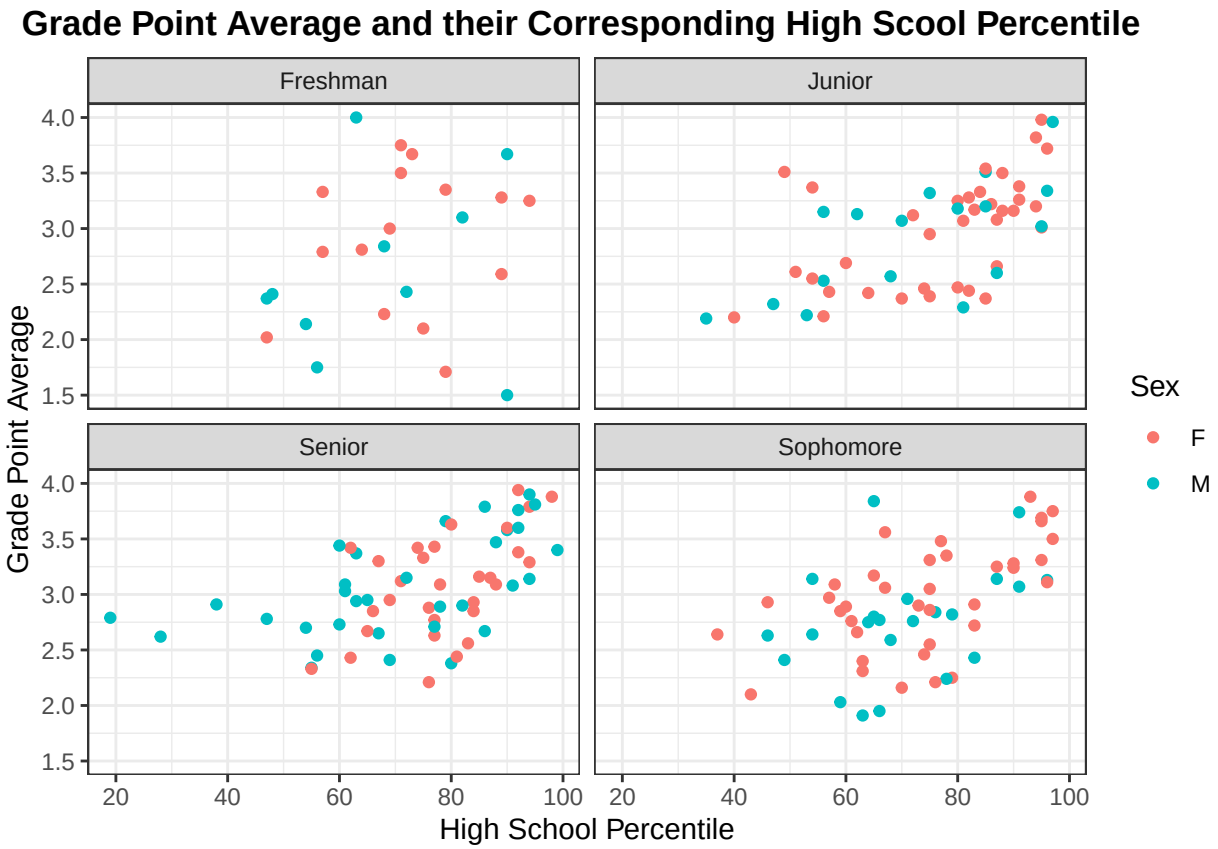
Plot before coloring and faceting:

```
# Plot scatter-plots of grade point average and high school percentile:
ggplot(data = Focus,
       mapping = aes(x = HSP, y = GPA)) +
  geom_point() +
  labs(x = "High School Percentile",
       y = "Grade Point Average",
       title = "Grade Point Average and their Corresponding High Scool Percentile") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.55, face = "bold", size = 13))
```



Plot with 4 variables:

```
# Plot scatter-plots of grade point average and high school percentile split up
# based upon class type and sex:
ggplot(data = Focus,
       mapping = aes(x = HSP, y = GPA, color = SEX)) +
  geom_point() +
  facet_wrap(~CLASS) +
  labs(x = "High School Percentile",
       y = "Grade Point Average",
       title = "Grade Point Average and their Corresponding High Scool Percentile",
       color = "Sex") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.55, face = "bold", size = 13))
```



Why I choose the above plot:
As stated above, the variables I chose to analyze were `GPA` which corresponds to grade point average of a given student in the sample, `HSP` which corresponds to the high school percentile of a given student in the sample, `SEX` which corresponds to the sex of a given student in the sample, and lastly `CLASS` the class type of a given student in the sample. In particular I decided to put `GPA` on the y-axis, `HSP` on the x-axis, faceted on the `CLASS` variable, with the coloring of the points dependent on `SEX`. First off, I decided to make my y-axis `GPA` and my x-axis `HSP`, because I wanted to find out what variables are able to predict `GPA`. After looking at the data, one clear choice for my predictor variable was `HSR`. Hence my main goal of this analysis was to answer "How does the high school percentile of a student predict the GPA of a student in college?" To dig deeper into this question and truly understand the relationship/nuances between the variables, I decided to facet the `GPA` on `CLASS` in order to see how the relationship changes over the different class types, which in essence is how `GPA` changes over time. Lastly, I decided to color each point in the scatter-plots based on `SEX` to see in each class, how does the relationship differ between sexes.

Findings:
As can be seen in the initial scatter-plot, there is a moderate positive linear relationship between a students grade point average and their corresponding high school percentile based on the data from students at UWEC. Once we take a look at the faceted scatter-plots, we see that this trend holds up for the Sophomore, Junior, and Senior class. However, when we take a look at the Freshman class, there seems to be absolutely no association between the two variables. One reason that can explain this phenomenon is that the jump from high school to college can be very difficult for many students independent on how they did in high school, thus even though a student was at the top of their class in high school, this doesn't mean that they will perform above average in college right away. However, once these students get adjusted to the college environment, they tend to perform like their usual selves, hence why we see the positive linear association in the remaining 3 classes. The last important takeaway that can be seen from the graph is that, as other studies have indicated, female students tend to perform better than male students academically. This shows up in both higher average grade point average and high school percentile. This trend holds true for the

Sophomore, Junior, and Senior class, but not so much for the Freshman class for reasons similar to those presented earlier.

2. The data for this question is **Handout 2**.

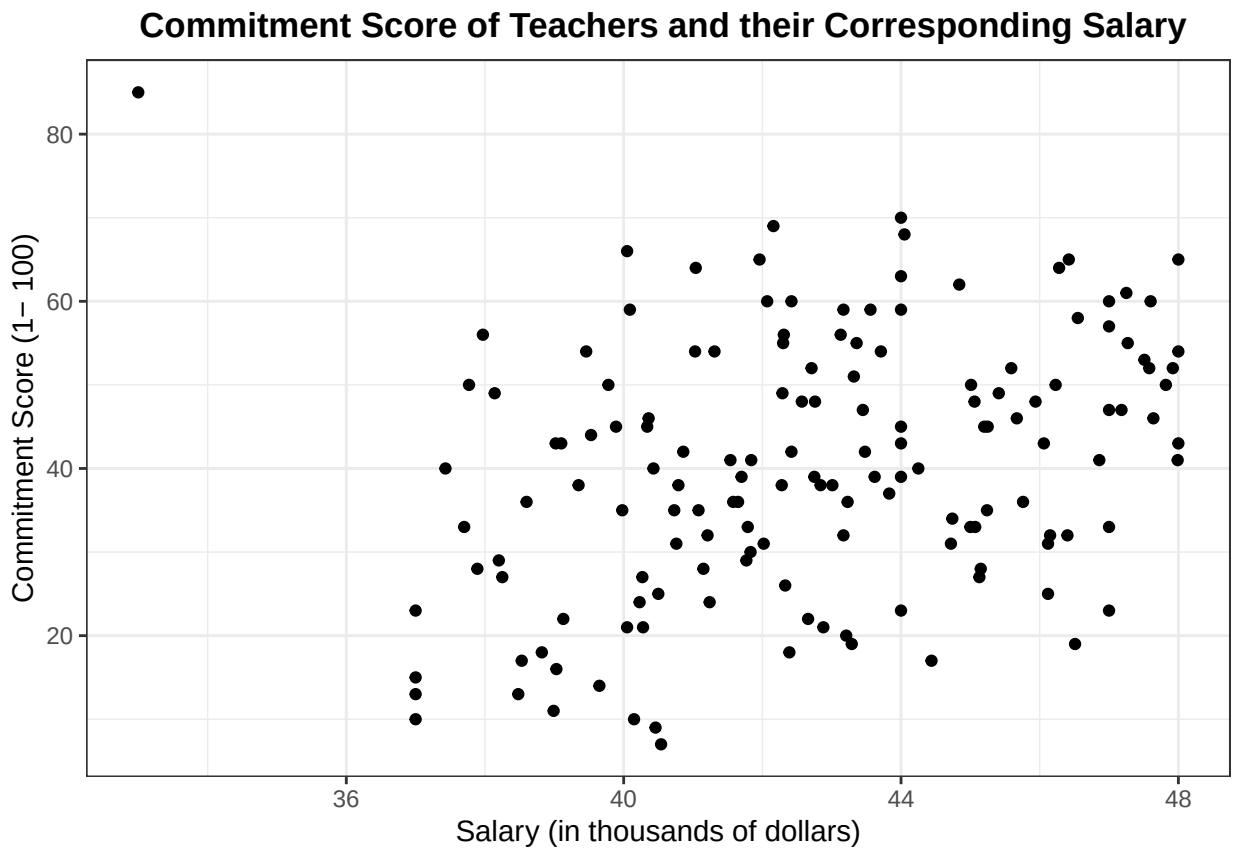
The data in **Handout 2** are from a "hypothetical study" of variables related to a teacher's intention to stay in the profession. The data were collected on 150 teachers who had been teaching for 5 years or less. These teachers responded to a questionnaire about their commitment to teaching and received a score (1 - 100) based on certain aspects of their commitment.

- a) Construct a scatter-plot of COMMIT against SALARY. What can you say about the graph?

Below we will construct a scatter-plot of the variables; `COMMIT` which corresponds to a given teachers commitment score, and `SALARY` which corresponds to a given teachers yearly salary (in thousands of dollars).

Scatter-plot of Commitment versus Salary:

```
# Plot the scatter-plot of commitment versus salary of teachers:
ggplot(data = Handout_2,
       mapping = aes(x = SALARY, y = COMMIT)) +
  geom_point() +
  labs(x = "Salary (in thousands of dollars)",
       y = "Commitment Score (1- 100)",
       title = "Commitment Score of Teachers and their Corresponding Salary") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.55, face = "bold", size = 13))
```



Findings:
As can be seen by the above scatter-plot comparing a teachers commitment score and their corresponding salary, we see that in general, as a teachers salary increases, their commitment score increases as well. For completeness, the range of the `SALARY` variable ranges from around 36,000 to 48,000, while the `COMMIT` variable ranges from around 15 to 65 (with an outlier of 85). This means, that despite the increase in commitment score with each increase in salary, this change isn't that much in terms of the magnitude of the score.

- b) Describe the relationship with respect to linearity, direction and strength of relationship, and the presence of outliers.

In the above scatter-plot on `COMMIT` versus `SALARY`, we see a moderate-to-weak positive linear association between the commitment score of a teacher and their corresponding salary in thousands of dollars. Furthermore, we see one highly influential outlier in the top-left corner, but other than that there doesn't appear to be any other obvious outliers.

- c) Compute the correlation matrix among all variables.

Before we compute the correlation matrix among all variables, we will remove the categorical variables. These variables include: `SEX`, `SCHTYPE`, `SCHLEVEL`. With that said, we will keep all discrete measurements (integer based variables) as they are theoretically on a continuous scale.

Removing categorical variables:

```
# Remove the categorical variables from the Handout_2 data set.
cat_vars_removed <- Handout_2[, -c(3, 4, 5)]
head(cat_vars_removed, 5)
```

```
##   COMMIT AGE SALARY CLASSSIZE RESOURCES AUTONOMY CLIMATE SUPPORT
## 1     38  30  42.28        26         5       12      12      12
## 2     10  30  40.15        21         5       11       9      11
## 3     30  26  41.83        28         7        9      10      12
## 4     10  27  37.00        22         7        5      10       8
## 5     16  24  39.03        25         7       10      11       9
```

Correlation matrix of the remaining variables:

```
cor(cat_vars_removed)
```

```
##                COMMIT         AGE      SALARY  CLASSSIZE    RESOURCES
## COMMIT     1.0000000  0.27316673  0.29565902 -0.2799916  0.274445807
## AGE        0.2731667  1.00000000 -0.04578992  0.1767477  0.060055333
## SALARY     0.2956590 -0.04578992  1.00000000 -0.5540894  0.422335920
## CLASSSIZE -0.2799916  0.17674769 -0.55408940  1.0000000 -0.248026296
## RESOURCES  0.2744458  0.06005533  0.42233592 -0.2480263  1.000000000
## AUTONOMY   0.3406021  0.17125671 -0.03337760  0.2880151  0.005617976
## CLIMATE    0.4470646 -0.02751814  0.18760076 -0.1929396  0.188016361
## SUPPORT    0.2806463  0.12724030  0.25197972 -0.2342537 -0.025971622
##                AUTONOMY     CLIMATE     SUPPORT
## COMMIT     0.340602086  0.44706456  0.28064625
## AGE        0.171256709 -0.02751814  0.12724030
## SALARY    -0.033377601  0.18760076  0.25197972
```

```
## CLASSSIZE  0.288015050 -0.19293962 -0.23425366
## RESOURCES  0.005617976  0.18801636 -0.02597162
## AUTONOMY   1.000000000  0.33387147  0.21051162
## CLIMATE    0.333871465  1.00000000  0.30811075
## SUPPORT    0.210511623  0.30811075  1.00000000
```

- d) Which variable would be the best predictor of COMMIT? Explain.

Based on the above correlation matrix, based purely on correlation coefficients, it appears as if the `CLIMATE` variable, which represents a teachers perception of school climate (0-20), is the best predictor of `COMMIT`. `CLIMATE` has a significantly higher correlation with `COMMIT` than any other variable with a correlation of 0.44706456, and has a coefficient of determination of 0.199866. This means that 19.99% of the variability in `COMMIT` can be explained by `CLIMATE`. `CLIMATE` being the best predictor of `COMMIT` also makes sense intuitively because if a teacher doesn't perceive that the environment of their school is good, they will be less inclined to stay with the school and profession in the long run.

- e) There is one unusual person in this data set. Remove the observation you identified as unusual. Recompute the correlation matrix and report only the correlation between COMMIT and SALARY. How has the relationship changed after removing the observation?

Removing the outlier:

```
# Finding the outlier (highest COMMIT score):
row_index <- sapply(Handout_2, function(x) which.max(x))[1]

# Filter out the outliers:
outlier_removed <- Handout_2[-row_index,]

# Remove the categorical variables:
cat_vars_removed_2 <- outlier_removed[, -c(3, 4, 5)]
```

Recomputing the correlation matrix:

```
cor(cat_vars_removed_2)
```

```
##               COMMIT         AGE      SALARY   CLASSSIZE   RESOURCES   AUTONOMY
## COMMIT     1.0000000  0.18094973  0.38067733 -0.36595143  0.32994994 0.30979764
## AGE        0.1809497  1.00000000  0.10569170  0.04912606  0.17689868 0.08721293
## SALARY     0.3806773  0.10569170  1.00000000 -0.52138855  0.39682904 0.01661215
## CLASSSIZE -0.3659514  0.04912606 -0.52138855  1.00000000 -0.21293023 0.25116551
## RESOURCES  0.3299499  0.17689868  0.39682904 -0.21293023  1.00000000 0.03983696
## AUTONOMY   0.3097976  0.08721293  0.01661215  0.25116551  0.03983696 1.00000000
## CLIMATE    0.4271400 -0.12804799  0.23919658 -0.24608641  0.22124426 0.31358756
## SUPPORT    0.2424249  0.02112233  0.32522504 -0.30817162  0.01128116 0.17733753
##              CLIMATE     SUPPORT
## COMMIT     0.4271400  0.24242491
## AGE       -0.1280480  0.02112233
## SALARY     0.2391966  0.32522504
## CLASSSIZE -0.2460864 -0.30817162
## RESOURCES  0.2212443  0.01128116
## AUTONOMY   0.3135876  0.17733753
## CLIMATE    1.0000000  0.28470265
## SUPPORT    0.2847027  1.00000000
```

After removing the one outlier from the data set, the correlation between COMMIT and SALARY has increased from a correlation coefficient of 0.29565902 to a correlation coefficient of 0.38067733, an increase of around 0.1. This further emphasizes that the outlier is an influential point in the data set.

- f) Perform a regression analysis to predict COMMIT from SALARY (with and without the outlier).

Below we will run a regression analysis to predict COMMIT from SALARY with the outlier.

Regression analysis with the outlier:

```
model_1 <- lm(Handout_2$COMMIT ~ Handout_2$SALARY)
summary(model_1)
```

```
##
## Call:
## lm(formula = Handout_2$COMMIT ~ Handout_2$SALARY)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.093 -10.340   0.192   9.247  59.052
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -22.8285    16.7579  -1.362 0.175173
## Handout_2$SALARY   1.4781     0.3912   3.778 0.000228 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.75 on 149 degrees of freedom
## Multiple R-squared:  0.08741,    Adjusted R-squared:  0.08129
## F-statistic: 14.27 on 1 and 149 DF,  p-value: 0.0002281
```

As computed above, the overall model is statistically significant with an F-statistic p-value of 0.0002281. Furthermore, the SALARY coefficient estimate is 1.4781 and is a significant predictor of COMMIT with a p-value of 0.000228. On the contrary, the intercept estimate is -22.8285 and is not significant with a p-value of 0.175173.

For completeness, here are the 95% confidence intervals for the above estimates:

95% confidence intervals for the slope and intercept:

```
confint(model_1)
```

```
##                        2.5 %    97.5 %
## (Intercept)      -55.9422383 10.285255
## Handout_2$SALARY   0.7049745  2.251197
```

As mentioned above the SALARY coefficient is significant and this is shown by it's interval not containing zero. Also, as mentioned above the intercept estimate is not significant and this is shown by it's interval containing zero.

In total, the regression equation is COMMIT = 1.4781 * SALARY - 22.8285.

Below we will run a regression analysis to predict COMMIT from SALARY without the outlier.

Regression analysis without the outlier:

```
model_2 <- lm(outlier_removed$COMMIT ~ outlier_removed$SALARY)
summary(model_2)
```

```
##
## Call:
## lm(formula = outlier_removed$COMMIT ~ outlier_removed$SALARY)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.7209 -10.4228   0.3067  10.2059  31.2168
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -41.8585    16.3883  -2.554   0.0117 *
## outlier_removed$SALARY   1.9137     0.3821   5.008 1.54e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.92 on 148 degrees of freedom
## Multiple R-squared:  0.1449, Adjusted R-squared:  0.1391
## F-statistic: 25.08 on 1 and 148 DF,  p-value: 1.544e-06
```

As computed above, the overall model is statistically significant with an F-statistic p-value of 1.544e-06. Furthermore, the `SALARY` coefficient estimate is 1.9137 and is a significant predictor of `COMMIT` with a p-value of 1.544e-06. On the same note, the intercept estimate was -41.8585 and was significant with a p-value of -41.8585.

For completeness, here are the 95% confidence intervals for the above estimates:

95% confidence intervals for the slope and intercept:

```
confint(model_2)
```

```
##                             2.5 %   97.5 %
## (Intercept)            -74.24367 -9.47323
## outlier_removed$SALARY   1.15857  2.66873
```

As mentioned above the `SALARY` coefficient is significant and this is shown by it's interval not containing zero. Also, as mentioned above the intercept estimate is also significant and this is shown by it's interval not containing zero.

In total, the regression equation is COMMIT = 1.9137 * SALARY - 41.8585.

Again, as mentioned previously, due to the parameters changing so much with just the inclusion/exclusion of the one outlier, the outlier is thus an influential point.

- g) Interpret the intercept and slope in the regression equation above. What does each of them tell you?

Below we will interpret the intercept and the slope of the regression equation with the outlier included:

Estimate interpretation for the model with outlier:
As computed above, the regression slope was 1.4781 with a 95% confidence interval of [0.7049745, 2.251197]. This means, on average, for every 1 unit increase in the annual salary of a teacher, which equates to a 1000

dollar increase, the mean commitment score of that teacher is estimated to increase by about 1.4781 points over the sampled range of commitment scores.

As computed above, the regression y-intercept was -22.8285 with a 95% confidence interval of [-55.9422383, 10.285255]. This means that the estimated mean commitment score is equal to about -22.8285 when the teacher has an annual salary of zero dollars. Obviously, this interpretation isn't meaningful because you can't have a negative commitment score (the range is 0 - 100).

Below we will interpret the intercept and the slope of the regression equation with the outlier not included:

Estimate interpretation for the model without outliers:
As computed above, the regression slope was 1.9137 with a 95% confidence interval of [1.15857, 2.66873]. This means, on average, for every 1 unit increase in the annual salary of a teacher, which equates to a 1000 dollar increase, the mean commitment score of that teacher is estimated to increase by about 1.9137 points over the sampled range of commitment scores.

As computed above, the regression y-intercept was -41.8585 with a 95% confidence interval of [-74.24367, -9.47323]. This means that the estimated mean commitment score is equal to about -41.8585 when the teacher has an annual salary of zero dollars. Obviously, this interpretation isn't meaningful because you can't have a negative commitment score (the range is 0 - 100).

- h) Check all conditions from both models and comment on your findings.

Below we will check all of the initial and residual conditions for the model with the outlier included.

Checking all of the conditions from the model with the outlier:

The conditions we will be checking are: normality of the response variable, linear relationship between the response and the predictor variable, normality of the residuals, equal variance of the residuals, and no outliers.

Checking for Normality of the Response:

```
shapiro.test(Handout_2$COMMIT)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Handout_2$COMMIT
## W = 0.99047, p-value = 0.4027
```

As can be seen from the above Shapiro-Wilk test, since the p-value is greater than 0.05 we fail to reject the null hypothesis and we assume that the dependent variable is approximately normally distributed.

Checking for Linearity:

```
raintest(model_1)
```

```
##
##  Rainbow test
##
## data:  model_1
## Rain = 0.96721, df1 = 76, df2 = 73, p-value = 0.5576
```

As can be seen from the above Rainbow test, since the p-value is greater than 0.05 we fail to reject the null hypothesis and we assume that the relationship between the variables is linear.

Checking for Normality of the Residuals:

```
resid_1 <- residuals(model_1)

shapiro.test(resid_1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_1
## W = 0.98121, p-value = 0.037
```

As can be seen from the above Shapiro-Wilk test, since the p-value is less than 0.05 we reject the null hypothesis and notice that the normality of the residuals assumption is violated.

Checking for equal variance:

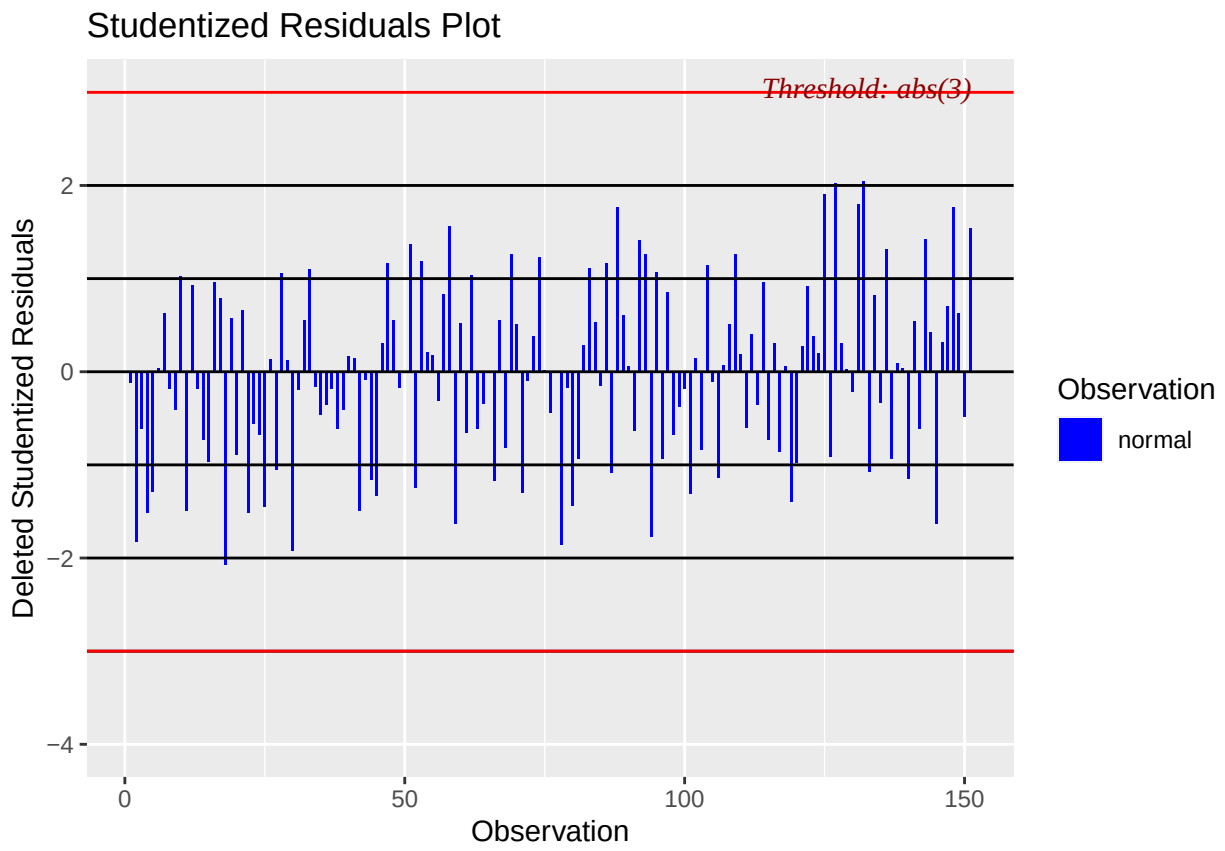
```
ols_test_breusch_pagan(model_1)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                      Data
##  -------------------------------------------
##  Response : Handout_2$COMMIT
##  Variables: fitted values of Handout_2$COMMIT
##
##          Test Summary
##  -----------------------------
##  DF            =    1
##  Chi2          =    17.03746
##  Prob > Chi2   =    3.66497e-05
```

As can be seen from the above Breusch Pagan Test for Heteroskedasticity, since the p-value is less than 0.05 we reject the null hypothesis and we see that the variance of the residuals are not equal/constant

Checking for no outliers:

```
ols_plot_resid_stud(model_1)
```



As can be seen from the above studentized residual plot, there are no outliers. However, we know that there exists one major outlier in our data set.

Based on the fact that the most of the residual assumptions were violated, using this linear model with the outliers to predict `COMMIT` isn't the best idea/is not valid.

Below we will check all of the initial and residual conditions for the model with the outlier not included.

Checking all of the conditions from the model without the outlier:

The conditions we will be checking are: normality of the response variable, linear relationship between the response and the predictor variable, normality of the residuals, equal variance of the residuals, and no outliers.

Checking for Normality of the Response:

```
shapiro.test(outlier_removed$COMMIT)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  outlier_removed$COMMIT
## W = 0.98438, p-value = 0.08778
```

As can be seen from the above Shapiro-Wilk test, since the p-value is greater than 0.05 we fail to reject the null hypothesis and we assume that the dependent variable is approximately normally distributed.

Checking for Linearity:

```
raintest(model_2)
```

```
##
##  Rainbow test
##
## data:  model_2
## Rain = 1.1695, df1 = 75, df2 = 73, p-value = 0.2515
```

As can be seen from the above Rainbow test, since the p-value is greater than 0.05 we fail to reject the null hypothesis and we assume that the relationship between the variables is linear.

Checking for Normality of the Residuals:

```
resid_2 <- residuals(model_2)

shapiro.test(resid_2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_2
## W = 0.9876, p-value = 0.2031
```

As can be seen from the above Shapiro-Wilk test, since the p-value is greater than 0.05 we fail to reject the null hypothesis and assume that the residuals are approximately normally distributed.

Checking for equal variance:

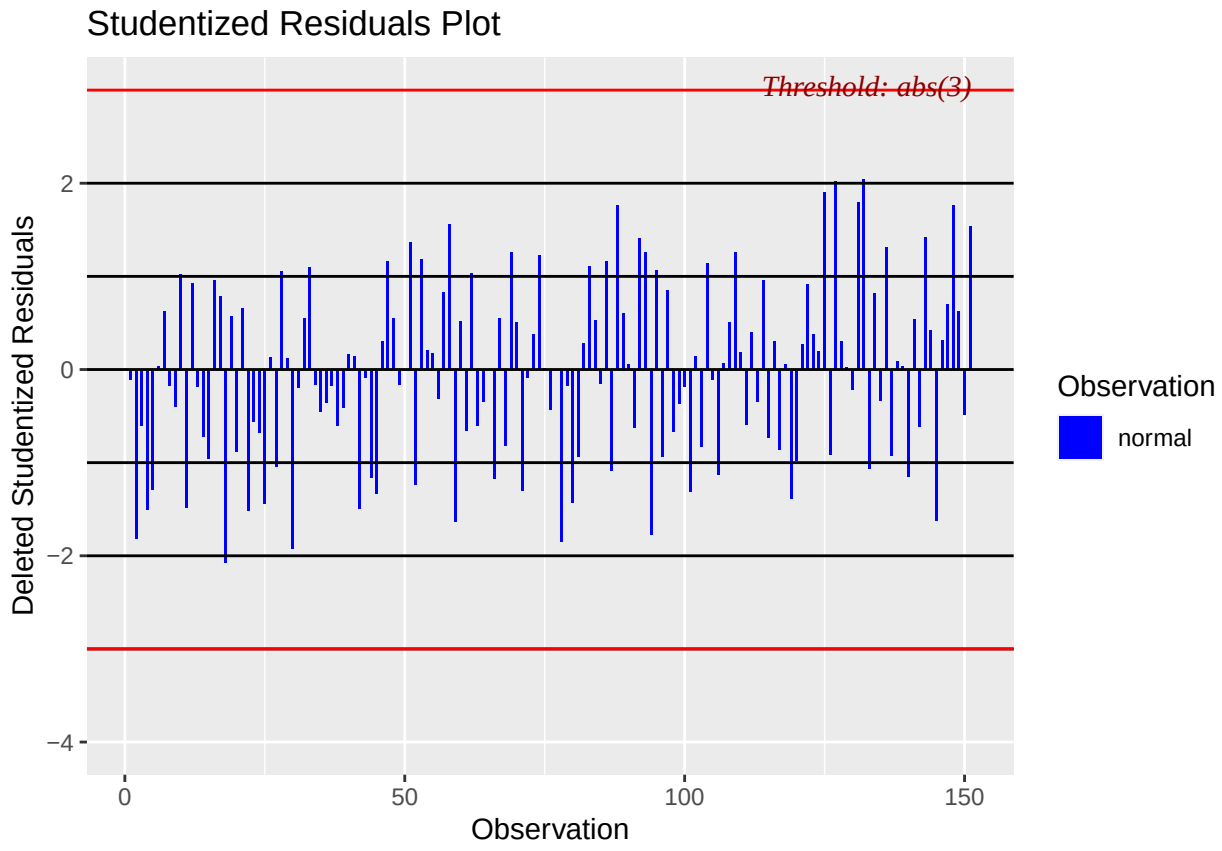
```
ols_test_breusch_pagan(model_2)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                       Data
##  ------------------------------------------------
##  Response : outlier_removed$COMMIT
##  Variables: fitted values of outlier_removed$COMMIT
##
##         Test Summary
##  ---------------------------
##  DF            =    1
##  Chi2          =    1.852941
##  Prob > Chi2   =    0.1734418
```

As can be seen from the above Breusch Pagan Test for Heteroskedasticity, since the p-value is greater than 0.05 we fail to reject the null hypothesis, thus we don't have evidence that the equal/constant variance assumption is violated.

Checking for no outliers:

```
ols_plot_resid_stud(model_1)
```



As can be seen from the above studentized residual plot, there are no outliers, as expected.

Finding the model error rate:

```
# Calculating the model error rate:
error_rate <- sigma(model_2) / mean(outlier_removed$SALARY)
```

Furthermore, since none of our initial, residual, and model assumptions are not violated we will have no problem fitting this model to the data. Due to the moderately high residual standard error of 13.92, with an error rate of 0.32529, the model fits the data moderately well. With that said due to a decently low r^2 value of 0.1449, our model won't have great predicting power for `COMMIT`.

However, for a simple linear model, `SALARY` does a decent job at predicting `COMMIT`, which could be made even better with using multiple linear regression with variables like `CLIAMTE` and `SUPPORT`.

3. Is there a statistical relationship between pH and Mercury? In order to answer this question correctly, argue your choice of dependent and independent variable (if any, provide a literature to support your answer). Using the data **lake.csv.**

Based on my research from researchgate, sciencedirect, and setac onlinelibrary, since decreased pH levels have been shown in numerous studies to decrease the loss of mercury from lake water and increase mercury binding to particulates in the water itself, we will use pH as the independent/predictor variable. and mercury as the dependent/response variable.

- a) First perform correlation analysis. State clearly the null and alternative hypothesis, test statistic, degrees of freedom and p-value, and draw your conclusion.

Below we will perform a correlation analysis on the pH levels and Mercury levels in lake water.

Correlation Analysis of pH and Mercury:

```
cor.test(Lake$pH, Lake$Mercury, method="pearson", conf.level = 0.95)
```

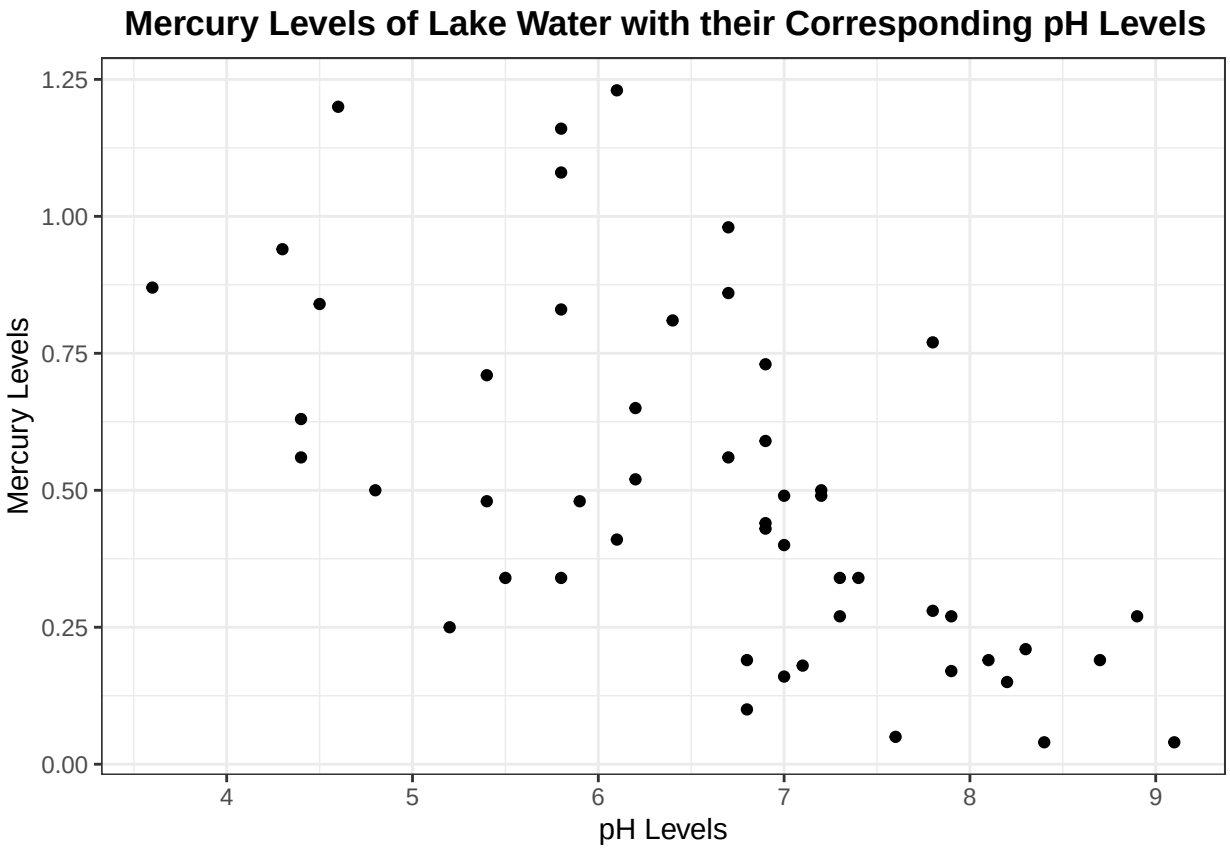
```
##
##  Pearson's product-moment correlation
##
## data:  Lake$pH and Lake$Mercury
## t = -5.3626, df = 49, p-value = 2.207e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7568749 -0.3995094
## sample estimates:
##        cor
## -0.6081381
```

In the above Pearson correlation tests, H_0 : is that the true correlation between pH and Mercury is zero, while H_A : is that the true correlation between pH and Mercury is not zero. The observed t statistic from the sample was -5.3626 on 49 degrees of freedom, where degrees of freedom is the number of data points minus 2, and lastly the p-value was 2.207e-06. Due to the extremely low p-value we reject the null hypothesis that the true correlation between pH and Mercury is zero, and concluded that there is a statistically significant correlation between pH and Mercury. The correlation coefficient between pH and Mercury was -0.6081381, showing a moderately high negative association between the pH and Mercury. Namely, as pH levels increase, Mercury levels decrease, and vice verse. The coefficient of determination is 0.36, which means that roughly 36% of the variability in Mercury can be explained by the pH levels in the water.

Below is a scatterplot to visualize these observation/claims made above.

Scatter-plot of pH and Mercury levels in lake water:

```
# Plot scatter-plots of pH and Mercury levels:
ggplot(data = Lake,
       mapping = aes(x = pH, y = Mercury)) +
  geom_point() +
  labs(x = "pH Levels",
       y = "Mercury Levels",
       title = "Mercury Levels of Lake Water with their Corresponding pH Levels") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.55, face = "bold", size = 13))
```



As mentioned above, there is an obvious moderate-to-high negative linear association between. the two variables.

- b) Second perform a simple linear regression. Report the intercept and slope with its corresponding 95% confidence interval and interpret your results.

We will now run a simple linear regression of Mercury on pH and interpret the results.

Summary of the model of Mercury on pH:

```
model_3 <- lm(Lake$Mercury ~ Lake$pH)
summary(model_3)
```

```
##
## Call:
## lm(formula = Lake$Mercury ~ Lake$pH)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45991 -0.17826 -0.04032  0.09257  0.65472
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.48776    0.18758   7.931 2.43e-10 ***
## Lake$pH     -0.14959    0.02789  -5.363 2.21e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2545 on 49 degrees of freedom
## Multiple R-squared:  0.3698, Adjusted R-squared:  0.357
## F-statistic: 28.76 on 1 and 49 DF,  p-value: 2.207e-06
```

As computed above, the overall model is statistically significant with an F-statistic p-value of 2.207e-06. Furthermore, the `pH` coefficient estimate is -0.14959 and is a highly significant predictor of `Mercury` with a p-value of 2.21e-06. Furthermore, the intercept estimate is 1.48776 and is also highly significant with a p-value of 2.43e-10.

For completeness, here are the 95% confidence intervals for the above estimates:

95% confidence intervals for the slope and intercept:

```
confint(model_3)
```

```
##                   2.5 %      97.5 %
## (Intercept)  1.1108097  1.86470783
## Lake$pH     -0.2056425 -0.09353003
```

As mentioned above the `pH` coefficient is significant and this is shown by it's interval not containing zero. Also, as mentioned above the intercept estimate is also significant and this is shown by it's interval not containing zero.

In total, the regression equation is Mercury = -0.14959 * pH + 1.48776.

Estimate interpretation for the model with outlier:
As computed above, the regression slope was -0.14959 with a 95% confidence interval of [-0.2056425, -0.09353003]. This means, on average, for every 1 unit increase in the pH level of the lake water the mean mercury level of that lake is estimated to decrease by about 0.14959 units over the sampled range of mercury levels.

As computed above, the regression y-intercept was 1.48776 with a 95% confidence interval of [1.1108097, 1.86470783]. This means that the estimated mean mercury level in the lake water is equal to about 1.48776 when the pH level of the lake water is zero.

Overall without running the model assumptions we can't justify the model fit, however due to the significant parameters and low residual standard error we can say that if the model assumptions hold, this linear model fits the data well with pH having decent predicting power for mercury.