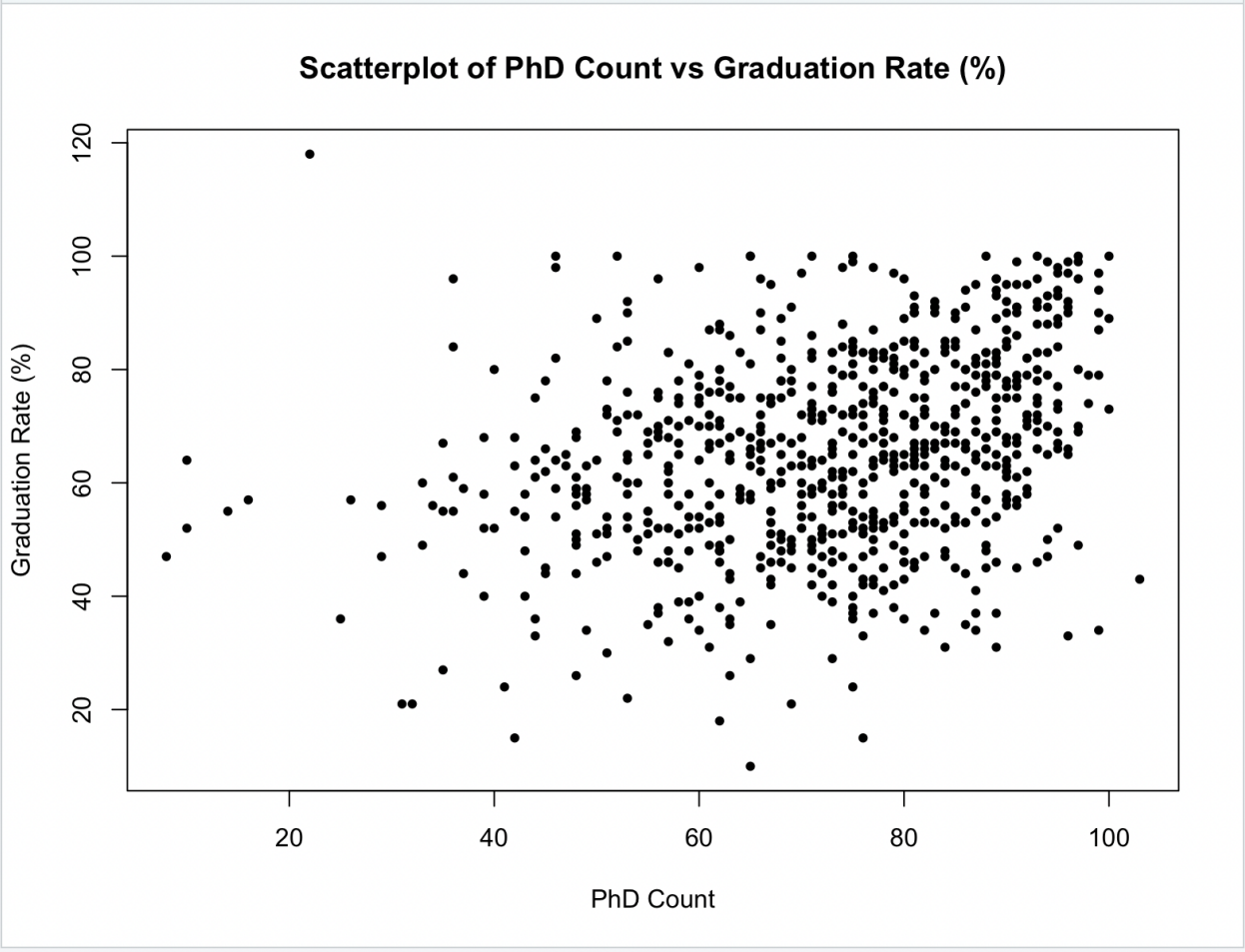
Justin Campbell
Samar Siddiqui

Regression Project

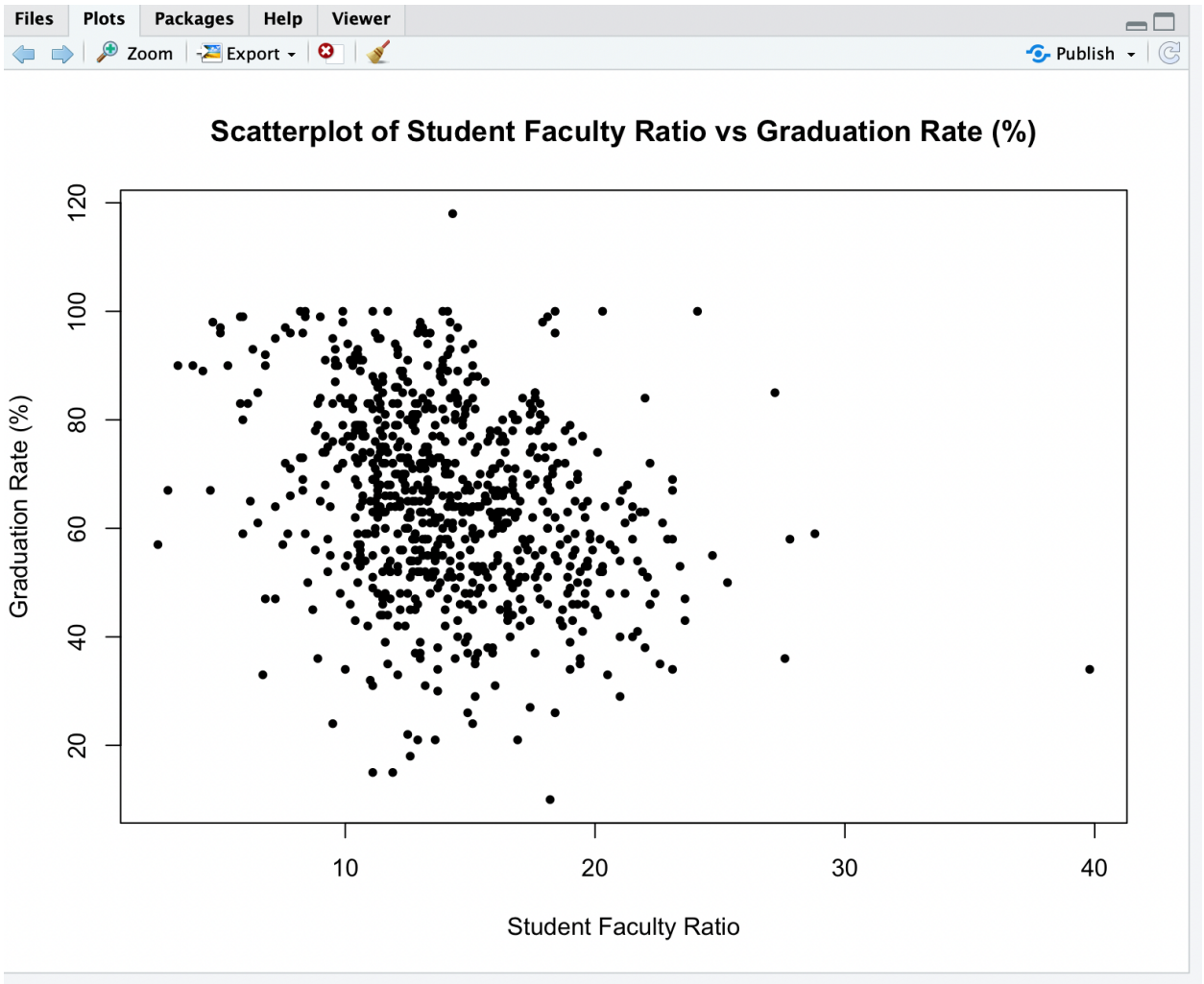
Variables Chosen:

- Outstate
    - This variable was chosen because if out-of-state tuition has a significant impact on the graduation rate, as a consultant I could advise the university to potentially decrease or increase the tuition amount.
- PhD
    - This variable was chosen because if the percentage of faculty with Ph.D. 's significantly impacts the graduation rate, as a consultant I could advise the university to potentially hire more or less faculty with said degrees.
- S.F.Ratio
    - This variable was chosen because if the student/faculty ratio significantly impacts the graduation rate, as a consultant I could advise the university to potentially increase or decrease the ratio, most likely by modifying the amount of faculty, changing the criteria for faculty members to be granted positions, or by changing student acceptance rates.

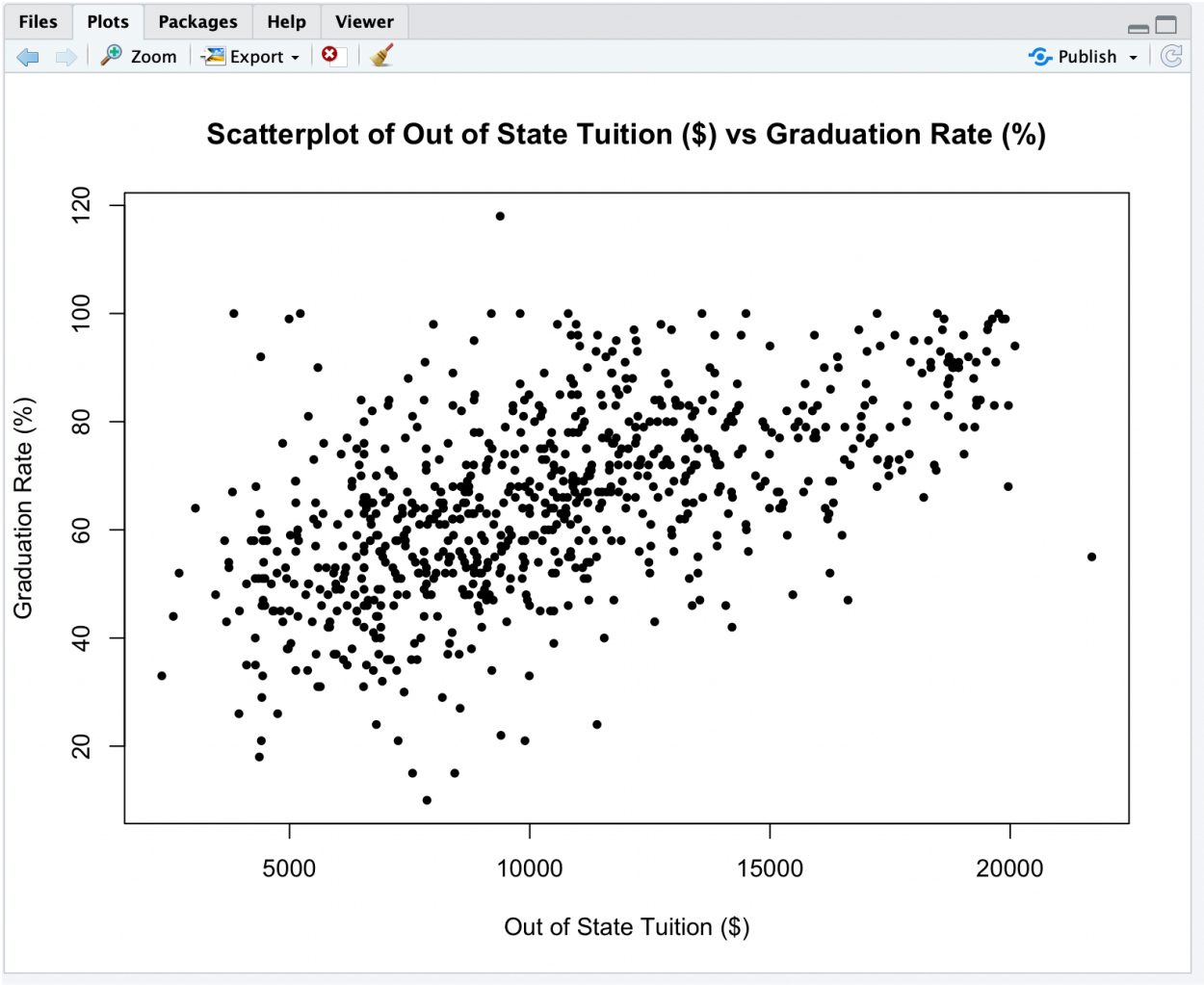
I. Scatterplots of Predictors vs. Response



The scatterplot of PhD Count vs Graduation Rate is moderately straight, moderately strong, and positive association. The straight enough condition is met as there are no significant bends, clumps, or outliers in this scatterplot.



The scatterplot of Student Faculty Ratio vs. Graduation Rate has a negative direction, moderate strength, and has a moderately straight form. In checking the Straight Enough Condition, we can conclude that there are no significant bends, curves, or clumps of data that would suggest other relationships, or lack of independence in the data.



The scatterplot of Out of State Tuition vs Graduation Rate is straight, strong, and has a positive association. The straight enough condition is met as there are no significant bends, clumping, or outliers in this scatterplot.

Looking at all three of the scatterplots, the needed assumptions and conditions are satisfied so we can compute the regression model and find the residuals.

II. Multiple Regression Model:

```
Residuals:
    Min      1Q  Median      3Q     Max
-48.815  -8.846   0.055   8.232  60.316

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.376e+01  3.687e+00   9.157  < 2e-16 ***
Phd_Data                  1.060e-01  3.359e-02   3.157  0.00166 **
S_F_Ratio_Data            1.249e-02  1.539e-01   0.081  0.93530
Out_of_State_Tuition_Data 2.281e-03  1.625e-04  14.041  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.03 on 773 degrees of freedom
Multiple R-squared:  0.3351,    Adjusted R-squared:  0.3325
F-statistic: 129.9 on 3 and 773 DF,  p-value: < 2.2e-16
```

$$\widehat{Grad\ Rate} = 33.7641 + 0.106(PhD) + 0.0125(S.F.Ratio) + 0.0023(Outstate)$$

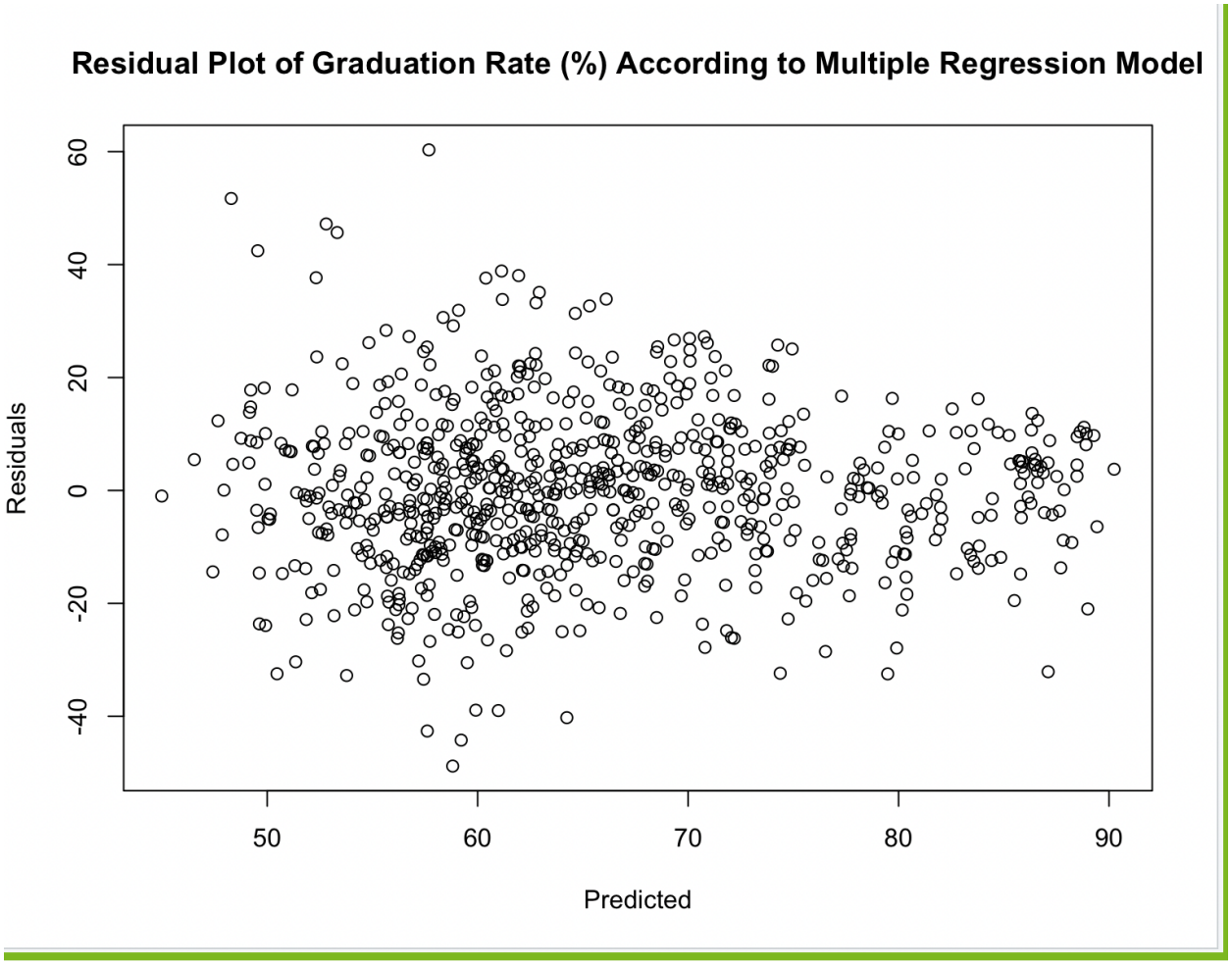
Coefficient Interpretation:

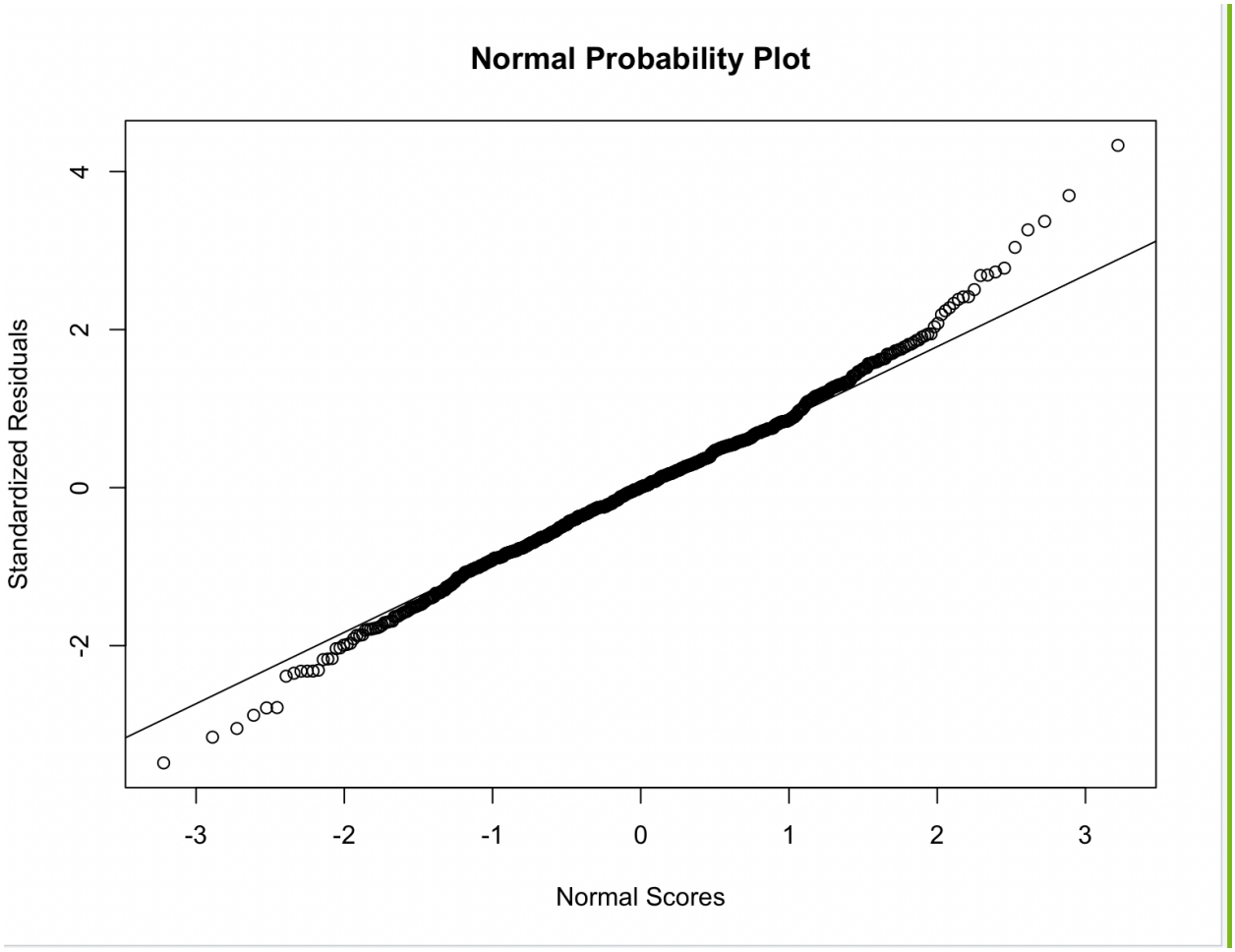
- Phd_Data
  - According to the summary statistic table, graduation rate increases on average about 0.106 percent for each increase in PhD by a count of one, after allowing for the effects of the other variables.
- S_F_Ratio_Data
  - According to the summary statistic table, graduation rate increases on average about 0.0125 percent for each increase in the ratio by a count of one, after allowing for the effects of the other variables. Although the scatterplot displayed a negative association, the coefficient is positive in the model because the other predictor variables have been incorporated.
- Out_of_State_Tuition_Data
  - According to the summary statistic table, graduation rate increases on average about 0.0023 percent for each increase in tuition dollar, after allowing for the effects of the other variables.

III. R^2 and Adjusted R^2 Values

The percent of the variability of the response variable is explained by the R^2. The R^2 in our model is 0.3351 or 33.51%, meaning 33.51% of the variation in graduation rate is accounted for by multiple linear regression model. The adjusted R^2 is less than the original R^2 which tells us the predictor variables improved the model by less than what we expected, meaning they did not add as much value as anticipated.

IV. Residual Plot, Probability Plot, Checks for Assumptions/Conditions





In referencing the scatterplots of the predictors against the response variable developed earlier, we confirmed that the plots are approximately straight and thus satisfy the Straight Enough Condition and Linearity Assumption.

In checking the Independence Assumption, although we cannot conclude that the data are completely independent, nor that the data were sampled randomly from a fixed population, it is reasonable to assume that the values for each of the predictor variables are independent, namely, that the characteristics across colleges did not influence each other. Thus, for the purposes of this regression model, the Independence Assumption holds true.

The scatterplot of residuals against predicted values shows moderately no structure, however there does appear to be a change in spread as the spread tapers from left to right. This could be a potential violation to the Equal Variance Assumption, however it isn't extremely significant so we will proceed with caution.

Looking at the normal probability plot of the residuals, it appears to be fairly straight, therefore the Normality Assumption is met.

As all the assumptions and conditions are met, a full multiple regression analysis is appropriate.

V & VI: Testing Hypothesis for Overall Model and for Each Predictor

H_0: $_{S.F.Ratio} = _{PhD} = _{Outstate} = 0$
H_A: at least one of the $_j \neq 0$

If we reject this hypothesis, then we'll test a null hypothesis for each of the coefficients:

H_0: The S.F. Ratio predictor contributes nothing useful after allowing for the effects of the other predictors in the model: $_{S.F.Ratio} = 0$
H_A: The S.F. Ratio predictor makes a useful contribution to the model: $_{S.F.Ratio} \neq 0$

H_0: The PhD predictor contributes nothing useful after allowing for the effects of the other predictors in the model: $_{PhD} = 0$
H_A: The PhD predictor makes a useful contribution to the model: $_{PhD} \neq 0$

H_0: The Outstate predictor contributes nothing useful after allowing for the effects of the other predictors in the model: $_{Outstate} = 0$
H_A: The Outstate predictor makes a useful contribution to the model: $_{Outstate} \neq 0$

Referencing the generated summary statistic, the F-statistic of 129.9 on 3 and 773 degrees of freedom is very large (considerably larger than 1), so we have sufficient evidence to reject the null hypothesis. Therefore the multiple regression model is significant. So we will examine the individual coefficients using multiple t statistics.

Referencing the generated summary statistic, PhD_Data has a t-value of 3.157, S_F_Ratio_Data has a t value of 0.081, and Out_of_State_Tuition_Data has a t value of 14.041. S_F_Ratio_Data has a relatively small t-value, so we can't be sure its underlying value is not 0, so we have insufficient evidence to reject its null hypothesis. Therefore this predictor does not contribute very much to the model after allowing for the effect of the Out_of_State_Tuition_Data and PhD_Data predictors. However, Out_of_State_Tuition_Data and PhD_Data have large t-values and small p-values which allows us to reject their null hypotheses and conclude that these predictors make a useful contribution to the model.

VII: Conclusion of Usefulness of Model

We conclude that our model is moderately useful as the combination of chosen explanatory variables somewhat predict graduation rates well. There are two predictors that make useful contributions but one that does not, hence the model's moderate usefulness. To add on, although two out of the three chosen predictors have large t-values and thus contribute moderately well to the model, the S.F. Ratio predictor has a very small t-value, and thus does not contribute very well to the model. Our recommendation to the university is to make modifications to their out of state tuition, and the number of faculty members with PhD's. In particular, we recommend that the university improve their methods of recruiting professors with

PhD's and/or increase the out-of-state tuition rates. We do not have any recommendations regarding the student/faculty ratio as we failed to discover any useful contributions it made to the model.

Statement of Contributions:

Both of us worked through the entire project and each section together. Justin executed and provided the majority coding done in R Studio and Samar provided assistance in areas of code. Samar typed and provided majority explanations for each section and Justin provided assistance in areas. Both of us worked on formatting. Work was equally divided.