*Advances upon Assignment 3 by using Ridge Regression and LASSO to predict Graduation Rates.*

**Assignment**

**4**

A4

ALY6015 Intermediate Analytics

Assignment 4 – Regularization

**PREPERATION:**

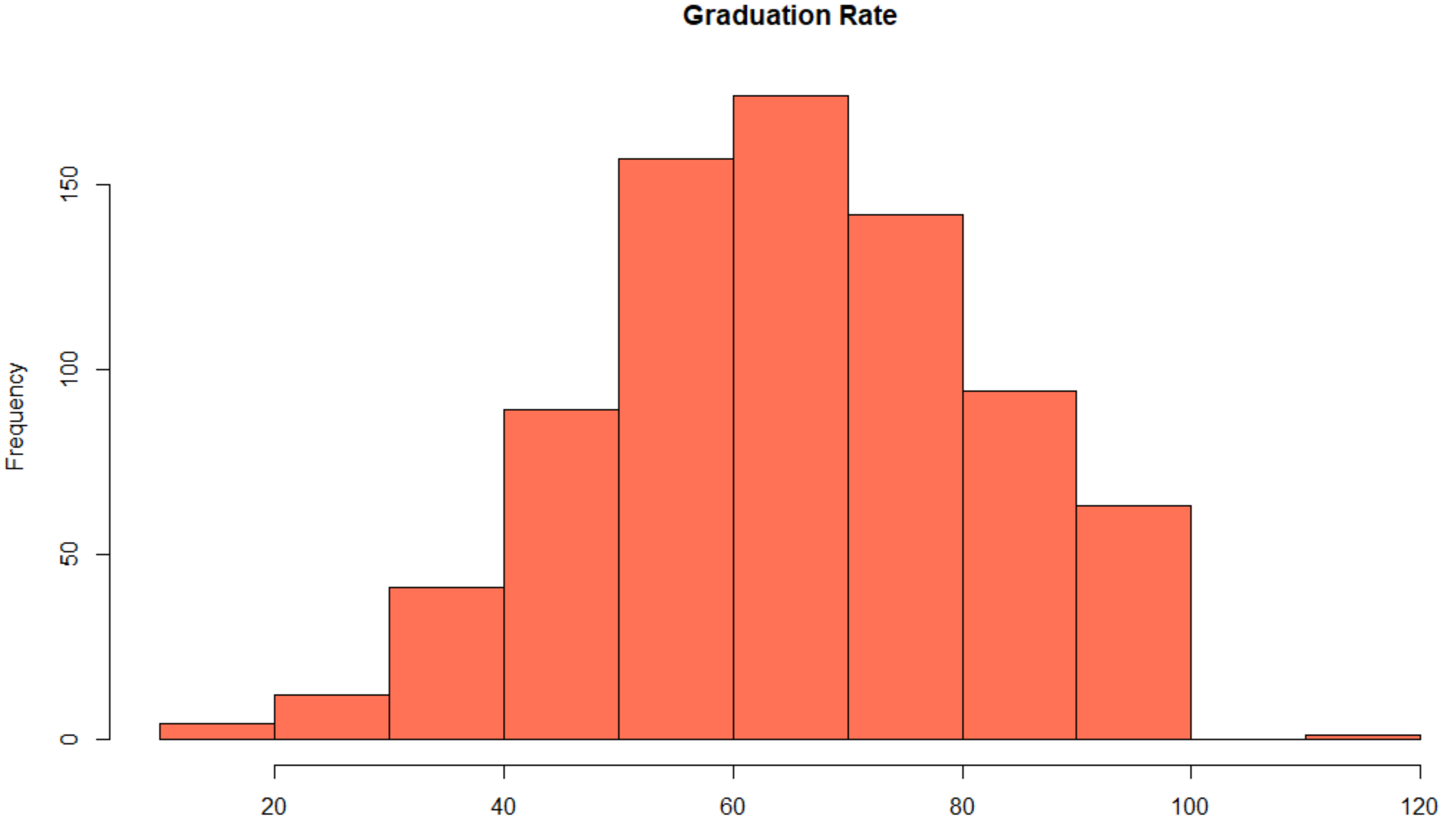
By: John DiSessa

For: Professor Goulding

On: October 26th, 2021

**Introduction**

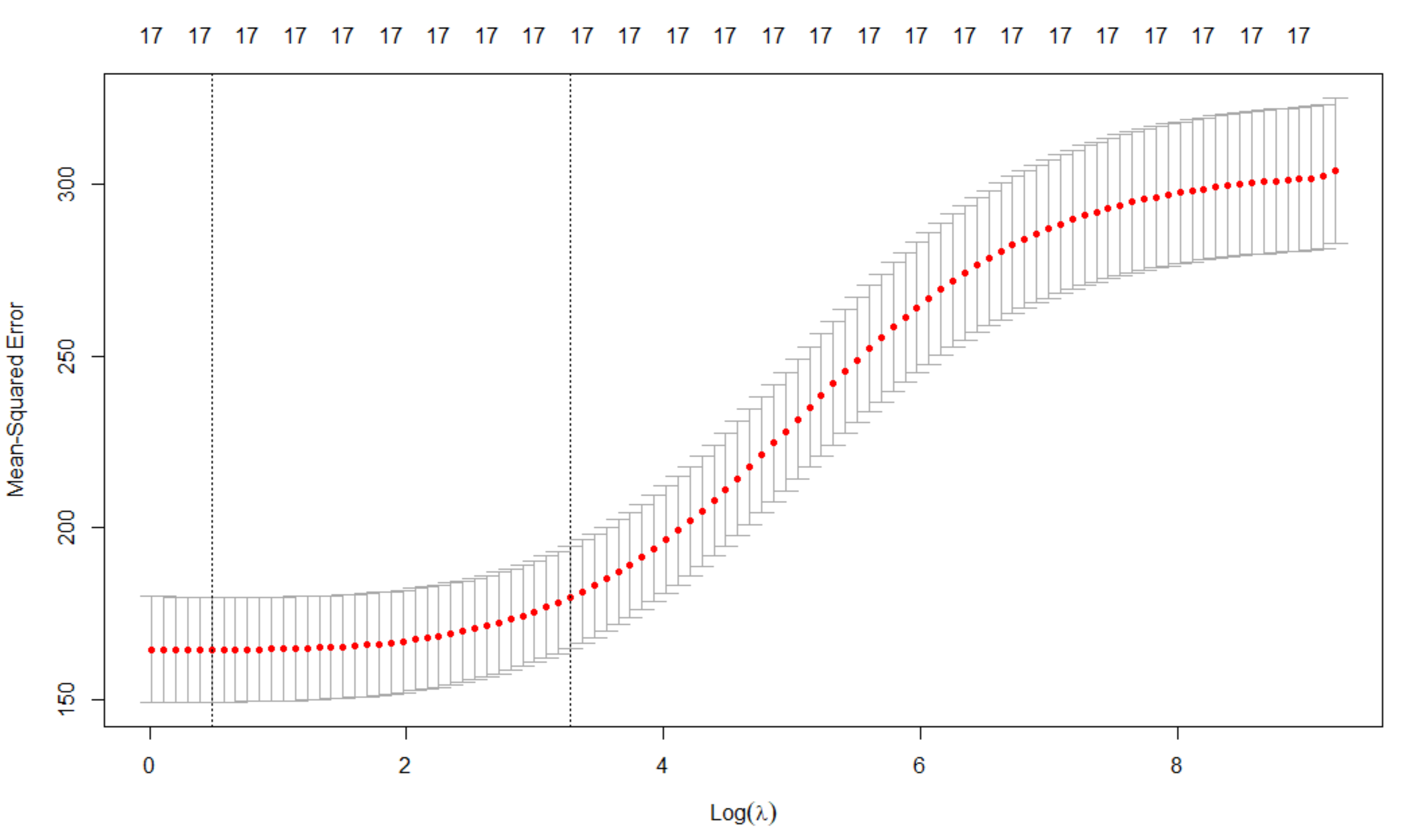
This assignment advances upon our findings in Assignment 3 from the College dataset (ISLR package). We used Ridge Regression and LASSO to conduct regularization on our models in order to best predict graduation rates. Our goal is to find the best model that can reduce the squared error but also reduce the penalty for adding too many variables (λ) and thus, overfitting the model.



As a reminder, our variable Grad.Rate is almost normally distributed but is actually slightly right-skewed with a mean of 65.45% and a median of 65%. We then split our data into training and testing data before starting our regularization. Out of the 777 datapoints, 543 are in the training set (70%) and 234 are in the testing set (30%).

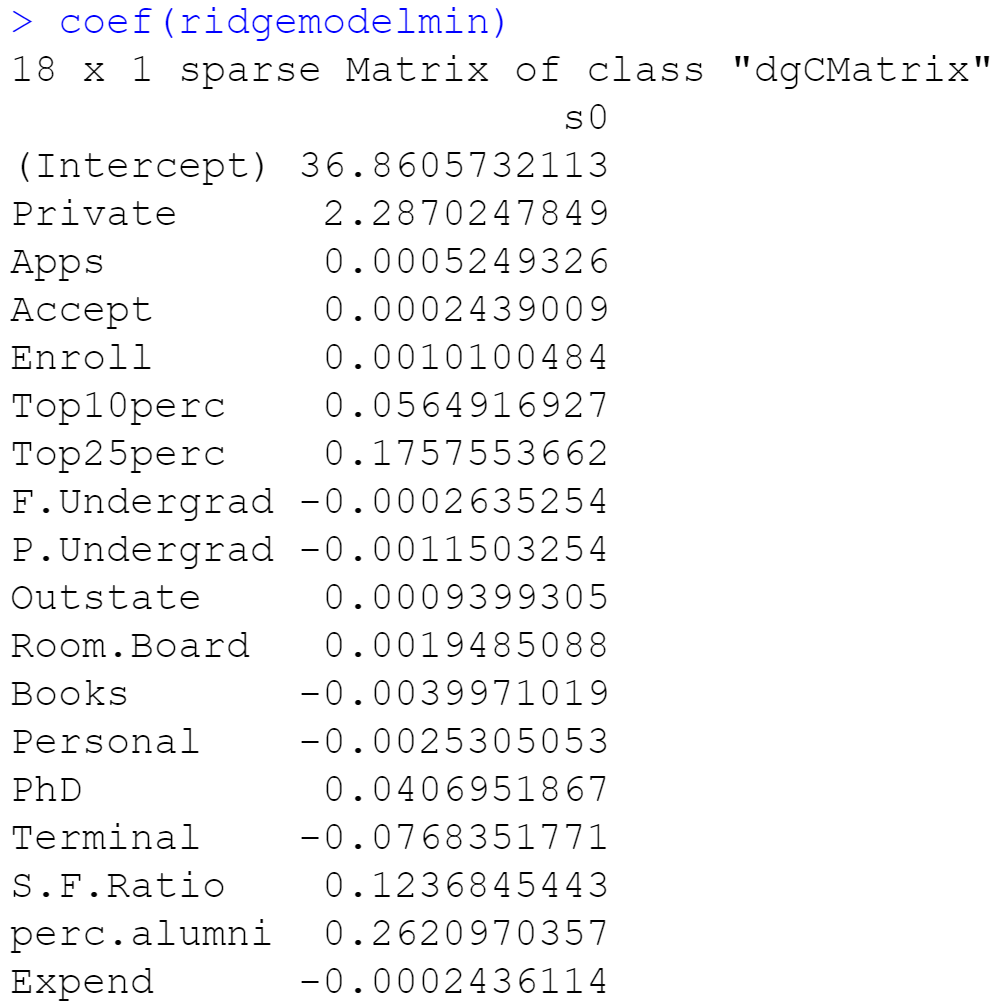
**Ridge Regression**

We started our Ridge Regression by plotting the Mean-Squared Error against the log of λ. The numbers on top of the following chart show how many predictor variables we would use at each value. The grey bars represent the intervals of the estimated variance of the loss metric (red dots) at each value. What is most important, however, are the two vertical dotted lines. The left dotted line is the minimum λ value of .48, which gives minimum average cross-validated error. The right dotted line is the maximum λ value of 3.27, which is the maximum value within one standard error.



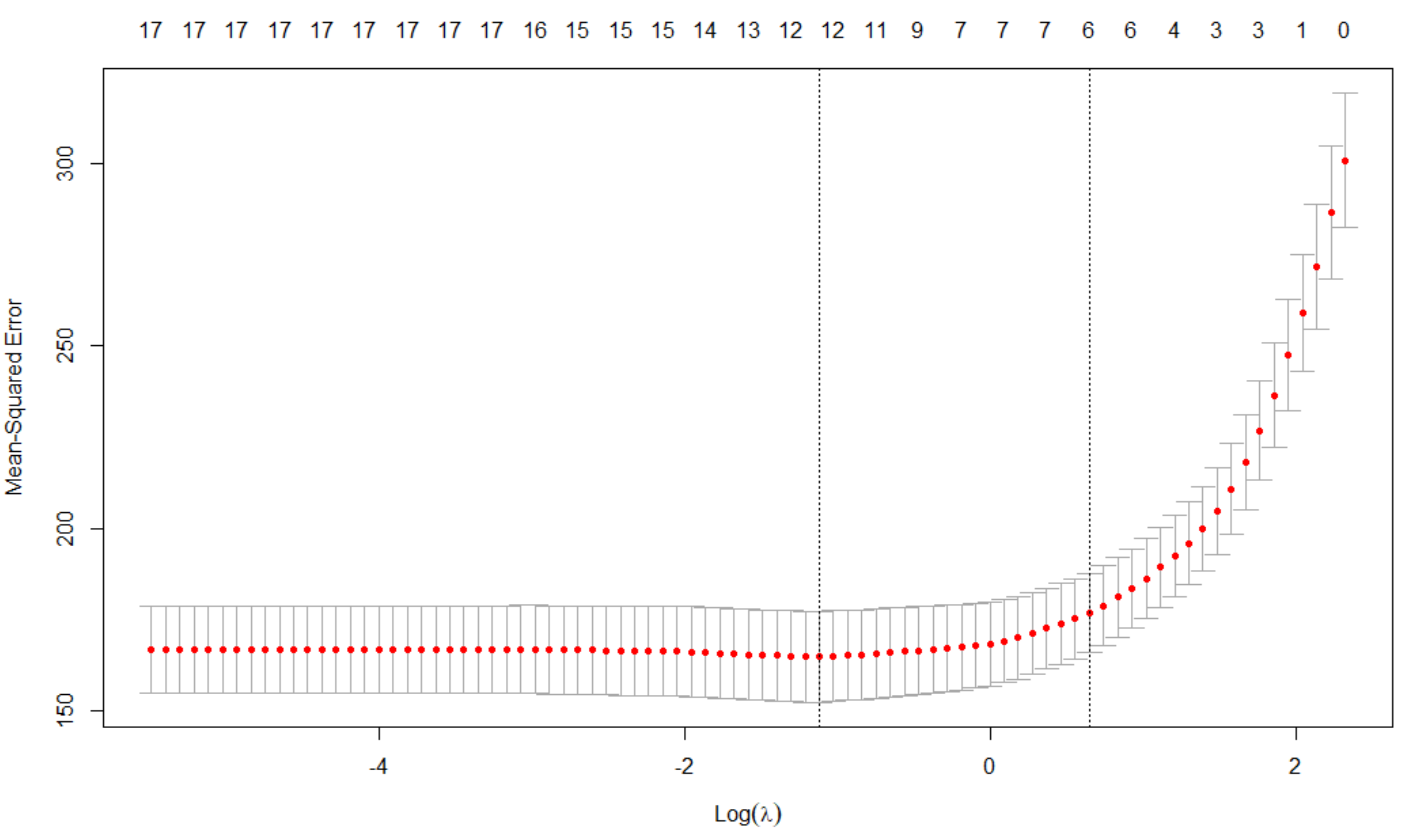
Now that we have two λ values, I calculated the Root Mean Square Error in order to determine which model had the lower error. Both were similar, however the model using the minimum λ of .48 produced a lower RMSE of 12.42 against our training data. Before we can declare that we are choosing the model with the minimum λ value, we have to compare our RMSE of 12.42 to the RMSE of the test dataset. Even though our model performed well against the training data, we need a similar RMSE against the test data to be sure that it has predictive value outside of the training data. The RMSE from the test data was 13.22 so we can say that, in addition to our model having a low RMSE from training the model, the same model also performs similar to the testing data. It is not overfitted to the training data.

Since our model reduced overfitting by regularization, it made changes to the coefficient values in the regression. Ridge Regression does not allow itself to remove any predictors by setting their values to zero, but some of them are close to zero. You can see the new coefficient values in the following chart.

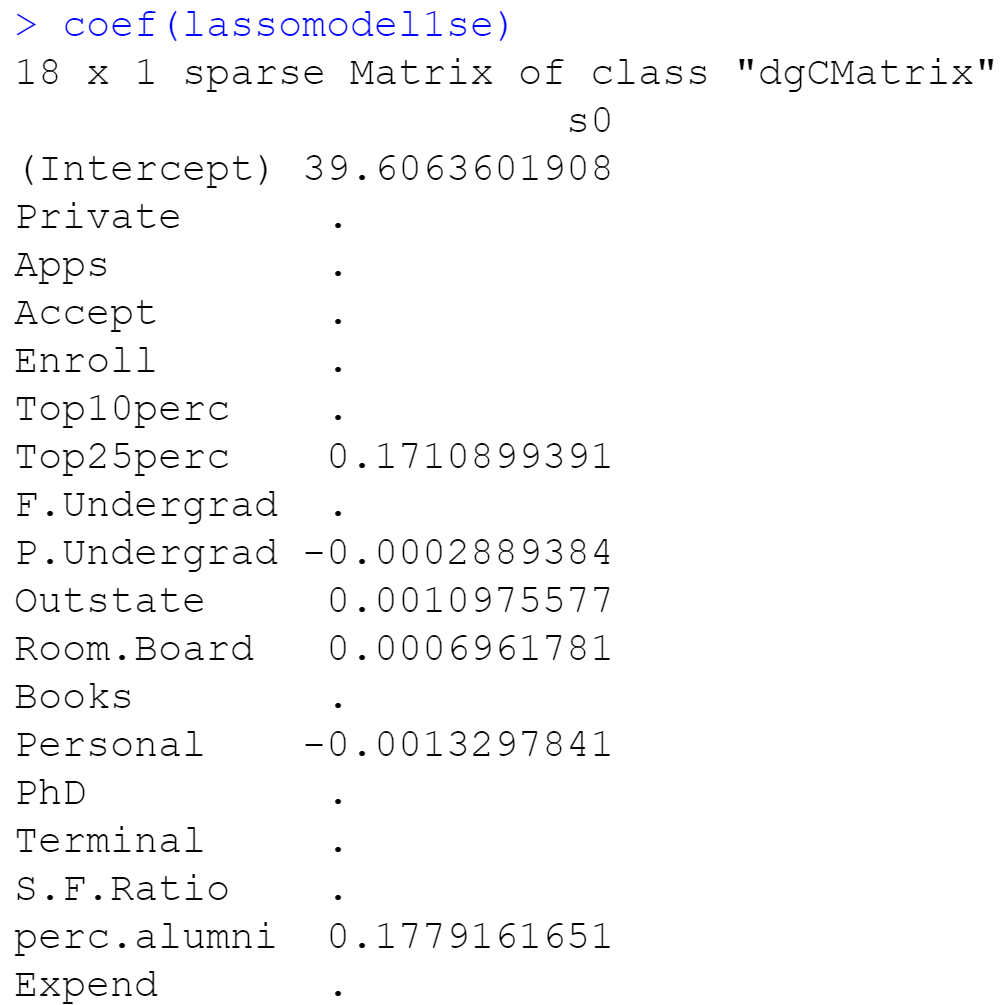
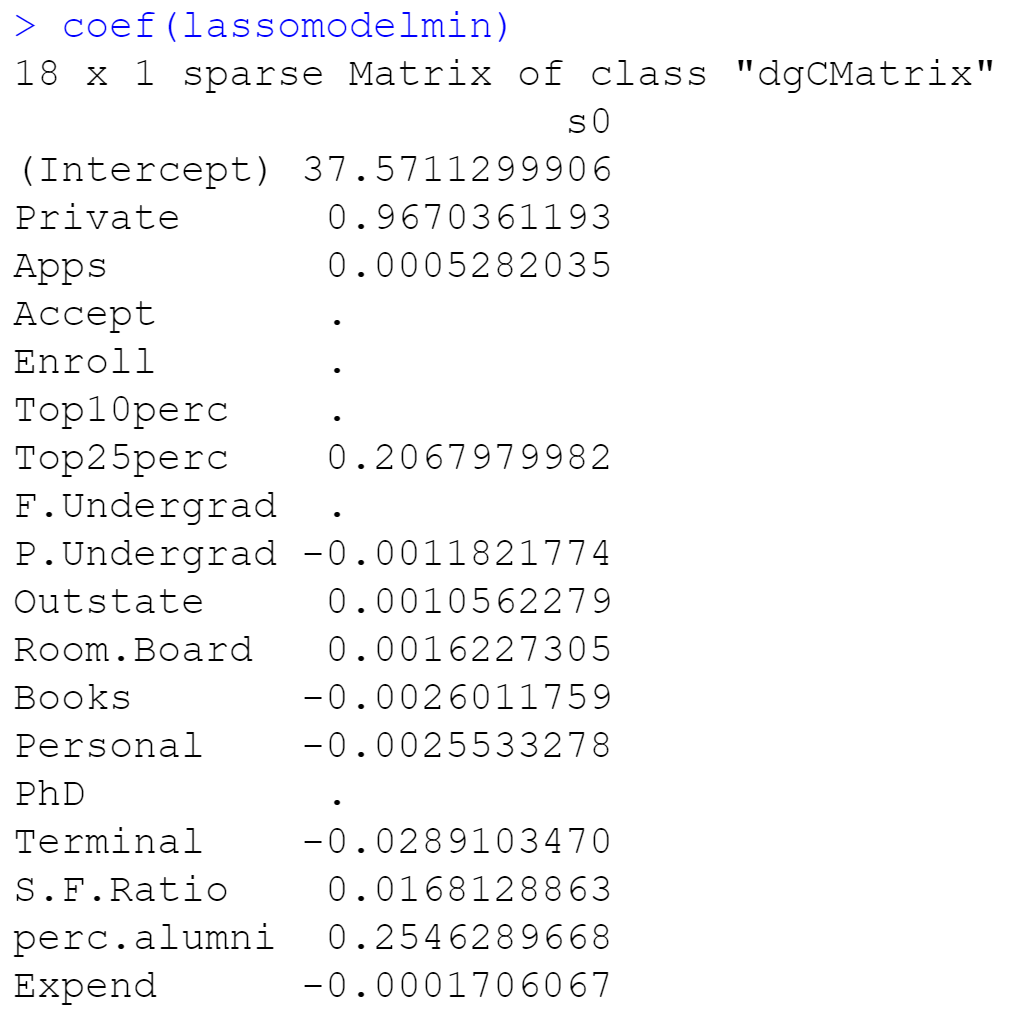


**LASSO**

Since we noticed some coefficients were reduced to almost zero, we ran a LASSO regression to see which coefficients the new regression would actually remove from the model, while also reducing overfitting. The following plot and calculations show us that our minimum λ is -1.12 and our maximum value within one standard error is .65. What is intriguing here compared to our Ridge Regression is that this model removes some predictors. If we use the minimum λ, we should have a model with 12 predictors. If we use the maximum λ, we should have a model with 6 predictors.



Now that we have our λ values, we can create models using each of them and see which predictors were kept and which were removed. In the charts below the model with the minimum λ keeps 12 predictors and the model with the maximum λ keeps 6 predictors, just like the previous chart showed. It is interesting to see that both models removed the Accept, Enroll, Top 10 Percent, Full-Time Undergraduates, and PhD Students variables.

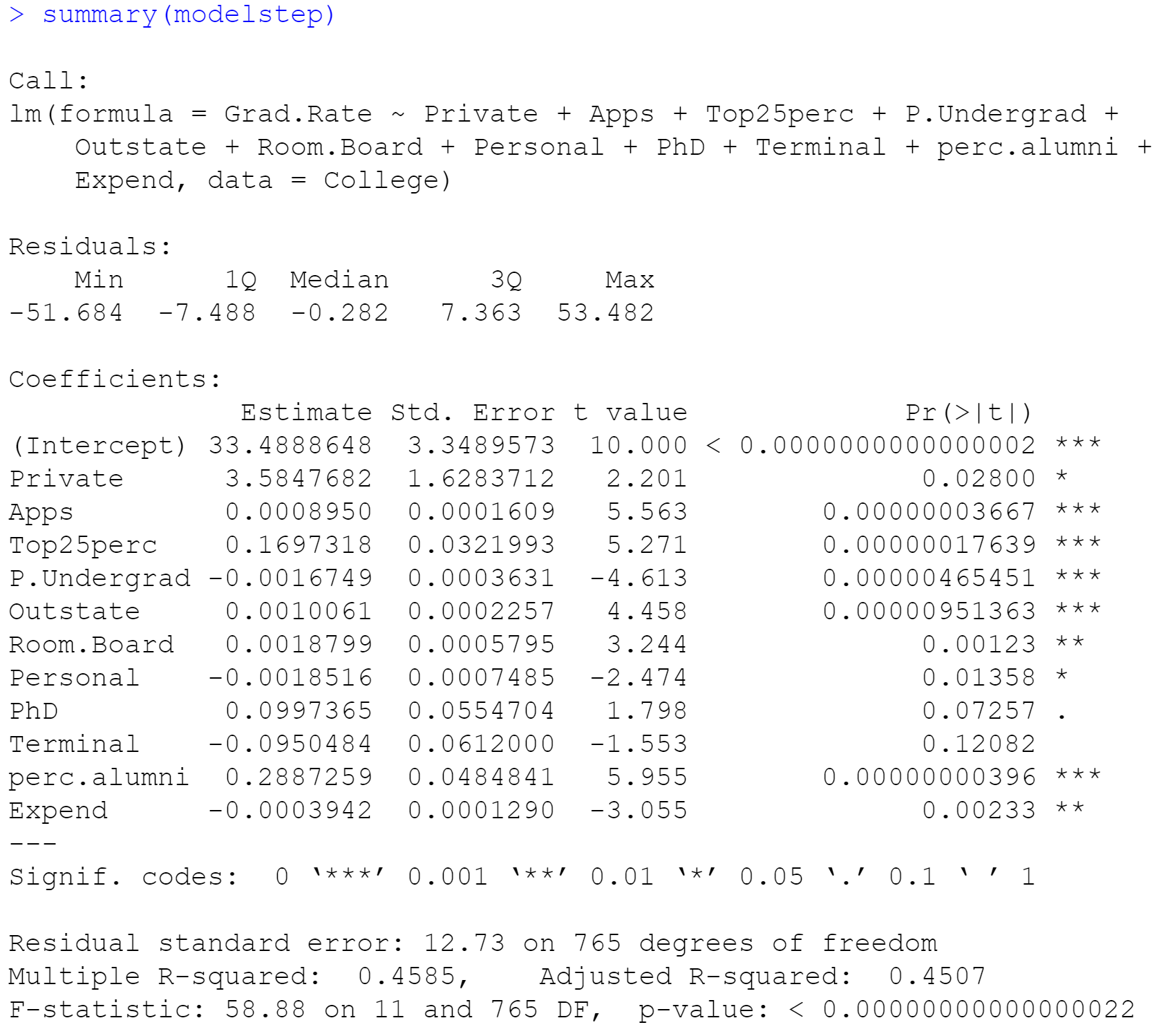


Before we determine which LASSO model to use, we need to calculate the RMSE’s of each model using the training data and compare those RMSE’s to the testing data in order show that our models have low error and do not overfit the training data. Our model using the minimum λ has a RMSE of 12.49 against the training data, compared to an RMSE of 13.10 from our model with the maximum λ. However, when compared to the test data, the minimum λ model has an RMSE of 13.29 and the maximum λ model has an RMSE of 13.49. Even though the model with the minimum λ has lower RMSE’s, the model with the maximum λ has less overfitting. It has less overfitting because its results are more similar between the training and test data so it does not overfit the training data as much.

**Comparison**

Based on our Ridge and LASSO regressions, we feel comfortable moving forward with the Ridge Regression model that used the minimum λ value since it had the lowest RMSE. Even though it did not perform as similar to the test data compared to the other models we looked at, it still did not overfit the data that much to be concerned. It was surprising that the LASSO models were not chosen since those models would have been simpler with the removed variables. If I were presenting this model to executives who were unfamiliar with statistics, I would have chosen the LASSO model with 6 variables since it would be simpler to explain but yet it also has a low RMSE and has little overfitting.

To see if there was a better model, I performed a stepwise selection. As seen below, it did not find the Terminal variable significant but did for the other 10 variables. That is not that simple of a model, however it still has 2 less variables than the LASSO model with the minimum λ. The biggest downside of the stepwise model is that it’s adjust R2 value is only .4507. So only 45.07% of the variation on Grad.Rate is explained by the 11 variables in the model. Since it’s adjusted R2 value is so low I would not choose this model.



**Conclusion**

Normal regression models have 1 goal and that is to reduce the sum of the squared errors (the total distance between the regression line and each data point). It does not penalize for having too many variables and thus can be prone to overfitting. If a model is overfit, it looks very accurate in predicting data from the dataset, but holds little predictive value outside of the dataset. By using Ridge or LASSO Regression, they add a λ value that does penalize for having too many variables. We ran 4 models using Ridge and LASSO Regression and their respective λ minimum values and maximum λ values within one standard error. We calculated their Root Mean Square Errors from the training data and compared them with the test data to see the levels of overfitting. Since there is no right choice between choosing a model with the lowest RMSE or choosing a model with the least overfitting, it was a difficult decision. However, given all of our models had low overfitting, I chose the Ridge model which had the lowest RMSE.

**Citations**

Bhattacharyya, Saptashwa. “Ridge and Lasso Regression: L1 and L2 Regularization.” *Medium*, Towards Data Science, 28 Sept. 2020, https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b.

“Cv.glmnet: Cross-Validation for Glmnet.” *RDocumentation*, https://www.rdocumentation.org/packages/glmnet/versions/4.1-2/topics/cv.glmnet.

“An Introduction to `Glmnet`.” *An Introduction to `Glmnet` • Glmnet*, https://glmnet.stanford.edu/articles/glmnet.html.

MehMeh 1, et al. “What Is Regularization in Plain English?” *Cross Validated*, 1 Feb. 1959, https://stats.stackexchange.com/questions/4961/what-is-regularization-in-plain-english.

Mewtwo, et al. “How to Interpret Cv.glmnet() Plot?” *Cross Validated*, 1 Mar. 1965, https://stats.stackexchange.com/questions/253963/how-to-interpret-cv-glmnet-plot.

“Ridge and Lasso Regression Models.” *Wavedatalab.github.io*, http://wavedatalab.github.io/machinelearningwithr/post4.html#:~:text=The%20tuning%20parameter%20lambda%20is,zero%20can%20be%20thrown%20away.

Sajjad, Bassalat, and Pussi says. “Start Here to Learn R!” *R*, 12 June 2017, https://www.r-exercises.com/2017/06/12/lasso-regression-in-r-exercises/?utm\_source=rss&%3Butm\_medium=rss&%3Butm\_campaign=lasso-regression-in-r-exercises&\_\_cf\_chl\_jschl\_tk\_\_=pmd\_mIZyHX4wQj0WXNL0gB.e48.1rkol8UKHCXrTE5CVEyU-1635120407-0-gqNtZGzNAqWjcnBszQjl.

Shishir PandeyShishir Pandey 1, et al. “What Are Good RMSE Values?” *Cross Validated*, 1 July 1961, https://stats.stackexchange.com/questions/56302/what-are-good-rmse-values#:~:text=If%20the%20RMSE%20for%20the,when%20tested%20out%20of%20sample.

Singh, Seema. “Understanding the Bias-Variance Tradeoff.” *Medium*, Towards Data Science, 9 Oct. 2018, https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229.