Teaching Pedagogy for an Introductory Statistics Course

Mitchell Pudil and Tom Jensen

February 2019

Executive Summary

We use a generalized least squares (GLS) model robust to heteroskedasticity to determine the economic and statistical significance of various assignments, such as homework assignments or midterms, that teachers in the BYU Introductory Statistics department oftentimes use to both assess student learning and prepare them for the final exam. Using this GLS model, we find that homework assignments and exams improve student learning (p-value < 0.01), while we were unable to reject the hypothesis that quizzes have no effect on final exam scores (p-value > 0.6). This leads us to conclude that professors of the statistics department should not place as much weight on quizzes compared to midterms, and perhaps rethink using quizzes as a way of assessing student understanding.

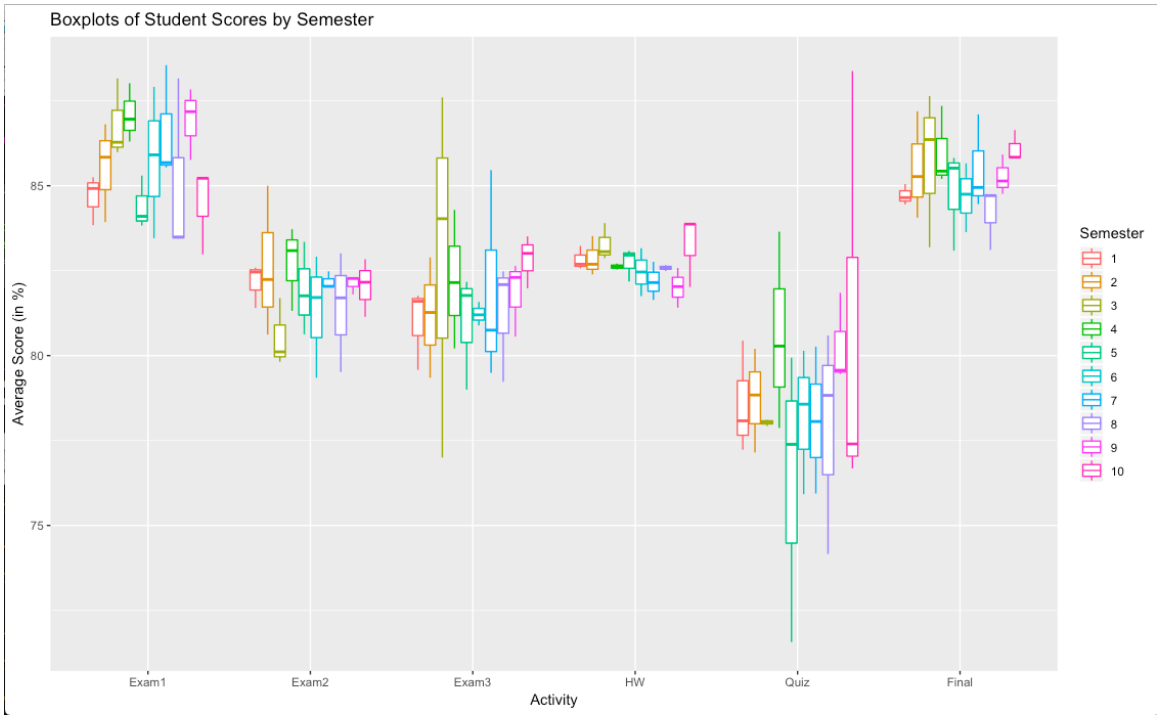
1 Introduction and Problem Background

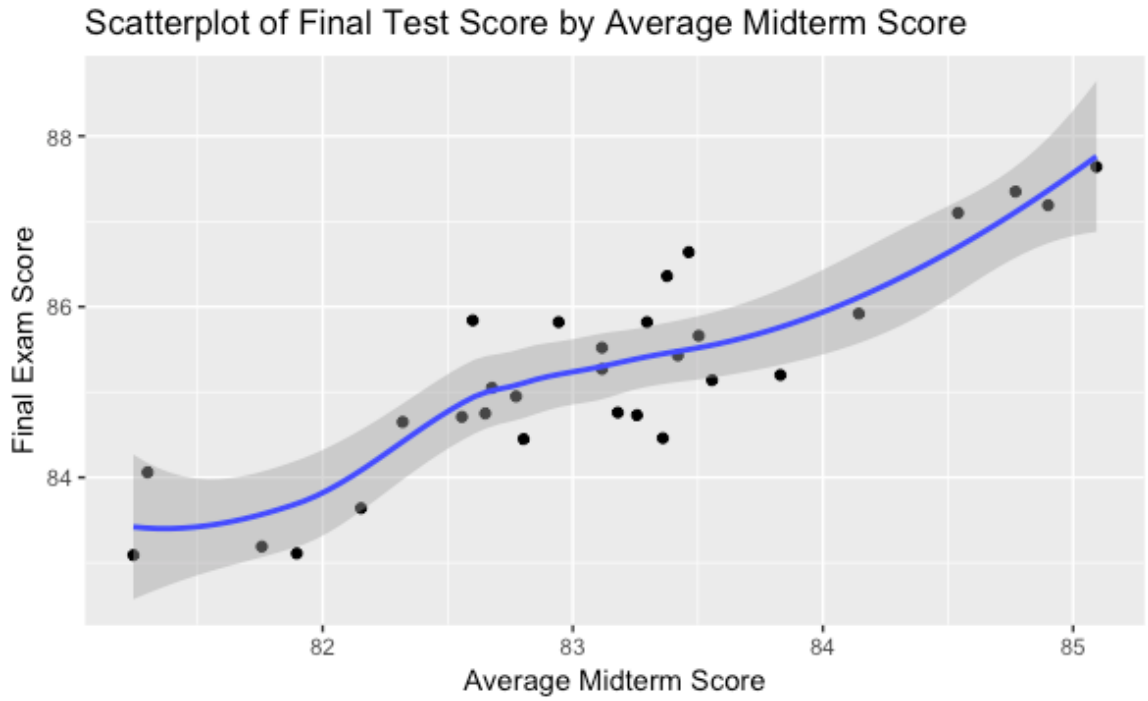
The statistics department at Brigham Young University is interested in determining the effect that homework assignments and midterm test scores have on the final exam score of their introductory-level course. The goal of the project is to determine which, if any, of these activities are associated with improved learning, and which are not. Further, we will determine which of the activities have the strongest effect of improved learning and estimate the size of these effects. We will also examine how well the class activities explain student learning and if there have been any semesters that have had either better or worse student learning than average.

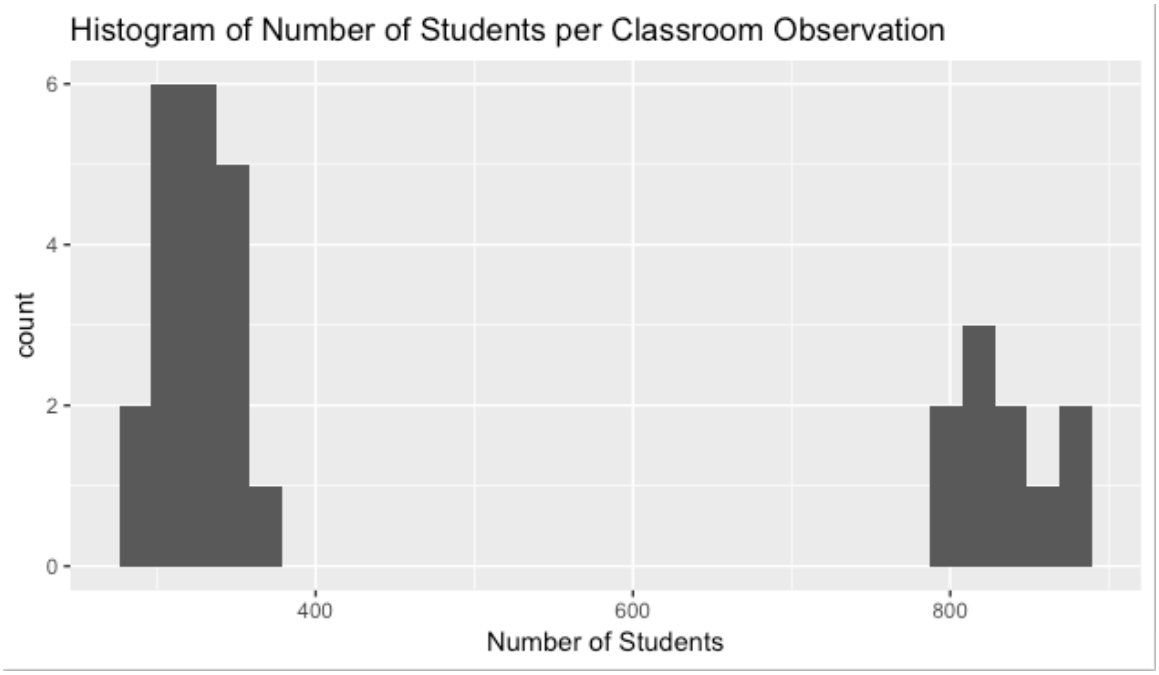
The data for this project have been gathered by BYU and include the average scores for each semester and section over the past five years of Fall and Winter semesters. Specifically, the scores

included in the dataset are from three midterms, homework, quizzes, and, of course, the final exam score. Additionally, we have the number of students who completed the course.

Shown below are a couple of graphics to summarize the data. The first graphic is a boxplot that shows the average scores for each activity, including the final, by semester. The second is a scatterplot of the final test score by the average midterm score, which is simply the sum of each of the three midterm scores for each semester-section observation. This scatterplot includes a 95% confidence interval band that captures the relationship between the average midterm score and the final exam score. The third graphic is a histogram of the number of students per classroom.







From the first second graphic, there appears to be an expected positive correlation between the average midterm score of a class and the final exam score. Specifically, the correlation between the average midterm score and the final exam score is 0.88. Each successive midterm has a greater

correlation with the final exam than the last. The first has a correlation of 0.14, the second 0.45, and the last has a correlation of 0.85.

Unfortunately, there are a few problems with the data. The greatest issue we have here, which will prevent us from using a basic linear regression model, is the fact that the final test scores are not individual test scores, but rather an average of varying class sizes. As shown by the histogram above, the number of students is unequal among each classroom. Thus, the variance of our dependent variable, final test scores, will be: $\frac{\sigma^2}{n}$. By accounting for this heteroskedasticity, our standard errors for the model will change to incorporate the unequal variance among class sizes, and we will be able to perform inference.

Also, as shown in the boxplot, the variation is small for final exam score and homework. The explanatory variable homework has an especially small variance which may make it difficult to determine the relationship between homework scores and the final exam score. Additionally, we have a very small sample size. While thousands of students have taken the introductory statistics course, we unfortunately do not have data of each student, only the average score in the class. This drops the number of observations we have down to only 30. Since we have several regressors, this means that the degrees of freedom will be very low. Thus, more information will need to be collected to determine a more precise relationship between the activities and the final test score.

Also, while the number of students taking the exam technically does have a large variation, there are really only a couple different class sizes: large (around 250-350 students) and really large (800-900 students). This means that any parameter estimate for the number of students will not be able to be generalized to class sizes under 300, around 500, or over about 1,000 or 1,100. Finally, any model produced will have limited external validity since we will not be using data from outside of BYU.

Another problem with the dataset is related to the idea of selection bias. This could occur because students decide which class to join, and part of many students' decisions in class selection is based by the number of people who attend. To an extent, teachers may also decide which class to teach, so a similar selection bias could occur and potentially be endogenous with the quality or teaching style of the teacher.

Unfortunately, without additional data, many of the problems mentioned above will not be accounted for. However, by understanding that there are few observations, we can maintain higher degrees of freedom by considering the semester number as a numeric value instead of a factor. The

coefficient generated for semester could possibly be interpreted as the improved overall teaching performance with time for the professors who taught the course (assumming the professors were constant over time). We determined that the only violation of the LINE assumptions for a generalized least squares model (Linearity, Independence, Normality of residuals, and Equal variance) was the equal variance assumption. The final model, then, will be a GLS model which includes all activities, the number of students in each class, and the semester number.

2 Statistical Model

Because we have determined that our data has heteroskedasticity due to the different class sizes, we use a fixed weights model to adjust for the heteroskedasticity. In matrix terms, the model is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim\ N(0, \sigma^2\mathbf{D})$$

where

$$\mathbf{Y} = \begin{bmatrix} Final_1 \\ Final_2 \\ \vdots \\ Final_{30} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & Semester_1 & \dots & Quiz_1 \\ 1 & Semester_2 & \dots & Quiz_2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Semester_{30} & \dots & Quiz_{30} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_{Semester} \\ \vdots \\ \beta_{Quiz} \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{30} \end{bmatrix},$$

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & 0 & 0 & \dots & 0 \\ 0 & d_{22} & 0 & \dots & 0 \\ 0 & 0 & d_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_{nn} \end{bmatrix}$$

In other words, the $\mathbf{Y}$ matrix is a list of the outcomes, in this case the final test scores, of each class. The $\mathbf{Y}$ matrix begins with a column of 1's (for the intercept), and then shows the value of the explanatory variables for each class. Recall that the explanatory variables here are: the semester

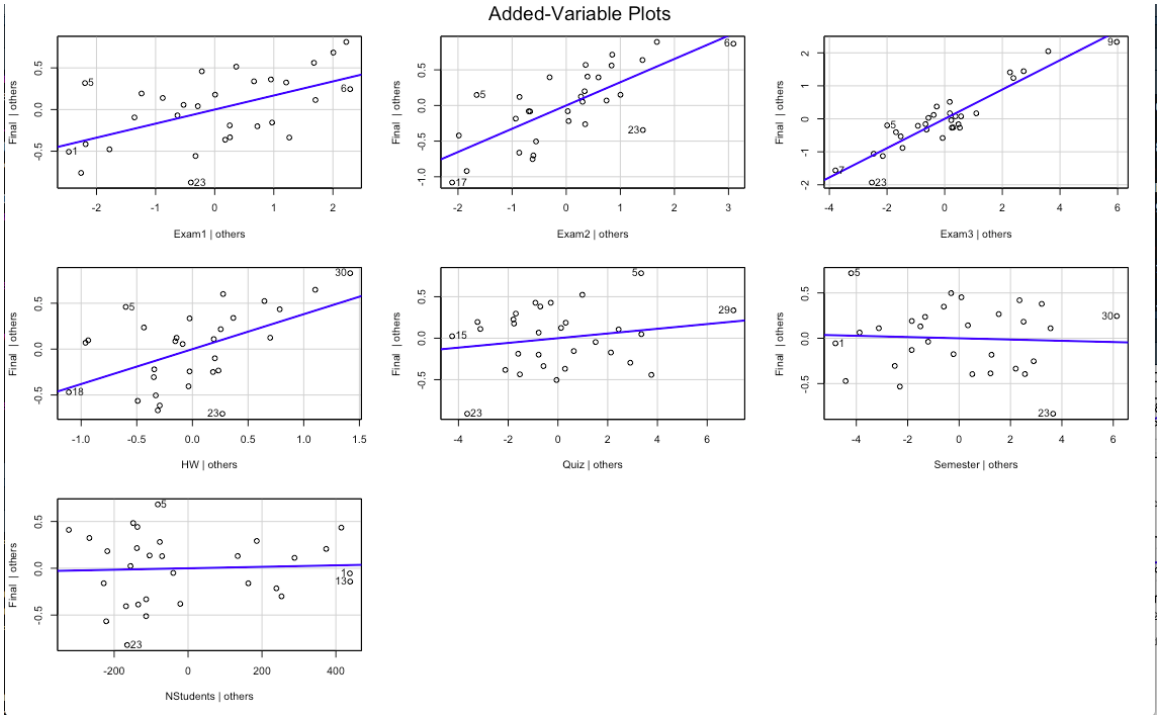
number, number of students, all three midterm scores, quiz scores, and homework scores. The $\boldsymbol{\epsilon}$ matrix is the error term for each class, which is essentially how different the actual predictions were from the predicted values we had calculated by matrix multiplying $\mathbf{X}$ and $\boldsymbol{\beta}$. Finally, for the $\mathbf{D}$ matrix, $d_{ii} = \frac{1}{n}$, where n is the sample size of the i^{th} classroom. By using these fixed weights, we control for the fact that each classroom has a different number of students.

The assumptions of this model are: linearity of parameters, independence of observations, and normality of residuals. Note that the equal variance assumption is resolved through using a fixed weights model, so homoskedasticity is not one of the assumptions we need to prove.

3 Model Validation

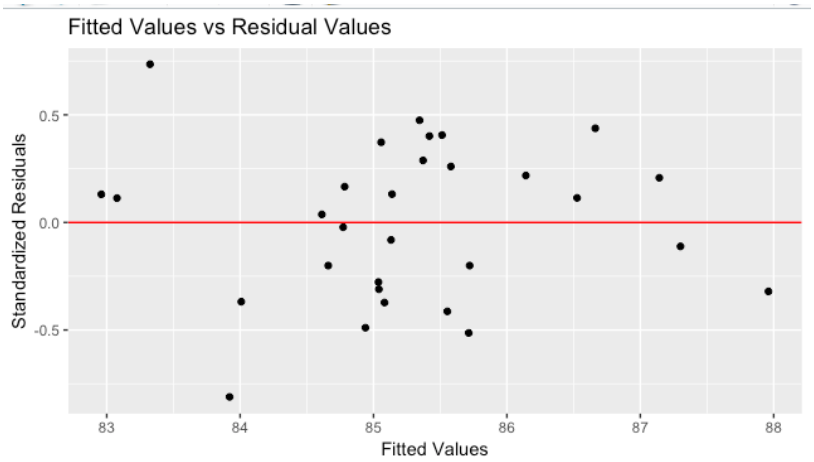
As mentioned, the use of the linear model shown above requires several assumptions to be fulfilled. Such assumptions include linearity of parameters, independence of observations, and normality of residual terms

The first assumption to check is linearity. The idea behind this assumption is that we need to make sure that our setup using matrices, as we outlined in Section 2, is an appropriate model, which suggests that each of the explanatory variables must be linearly related to the Y (or if not, that we have included a function of that variable as another column in the $\mathbf{X}$ matrix). A useful tool to examine the linearity of each (non-categorical) explanatory variable is an added variable (AV) plot, which holds the other variables constant to look at the relationship between each individual effect of the explanatory variable on the outcome (final exam score). Such a graphic is shown below:

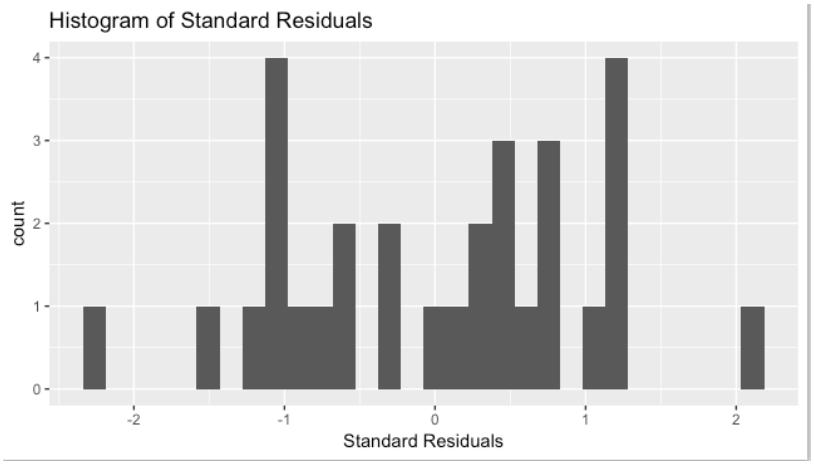


From the added variable plot, it is obvious that a couple of the explanatory variables have a linear relationship with final exam score, such as the second and third midterm exams. On the other hand, a couple of the explanatory variables do not seem to have a strong relationship with the final test score. Overall, however, it is apparent that if there is any sort of relationship in any of the explanatory variables, that the relationship is linear (or at least that there's not enough evidence to suggest that the relationships are anything other than linear). Because of this, we conclude that the linearity assumption does sufficiently hold.

The second assumption we need to explore is independence. Since the observational unit is the entire class, not just each student, we can reasonably assume that each class is uncorrelated with the others. The one caveat may be the fact that only BYU classes were surveyed, so the results of our model are unrepresentative of introductory statistics classes oustide of BYU. That is, the results we get from this model likely won't extend to any other university. Additionally, we see from the plot of fitted values against standardized residuals that there isn't any distinguishable pattern among the residual terms. This plot is shown below:



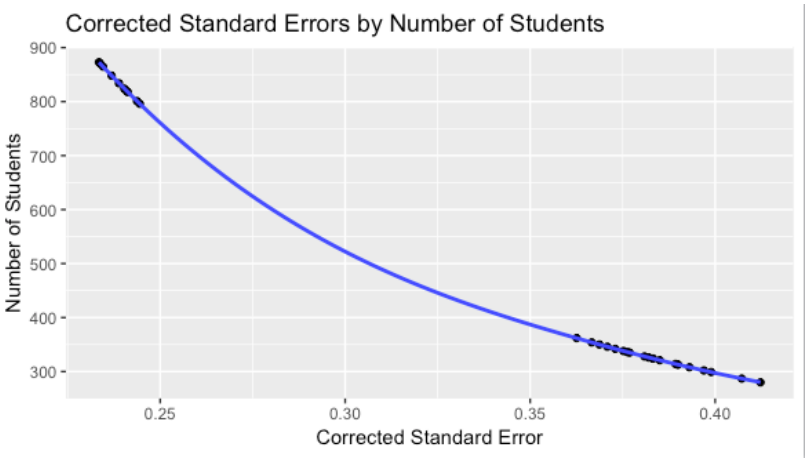
The third assumption we must consider is normality of residual terms. Two techniques for determining the normality of error terms are plotting a histogram of the residuals and performing a K-S test. We will begin with showing the histogram of the standardized residual terms. This graphic is presented below:



The error terms appear to possibly follow a normal distribution. However, normality is difficult to graphically determine in this case where we only have 30 observations. We also performed a K-S test, which tests a null hypothesis that error terms are normally distributed. The p-value for this test is 0.6138, which tells us that the standardized residuals, in fact, reasonably follow a normal distribution. Thus, the assumption of normality of residuals is satisfied.

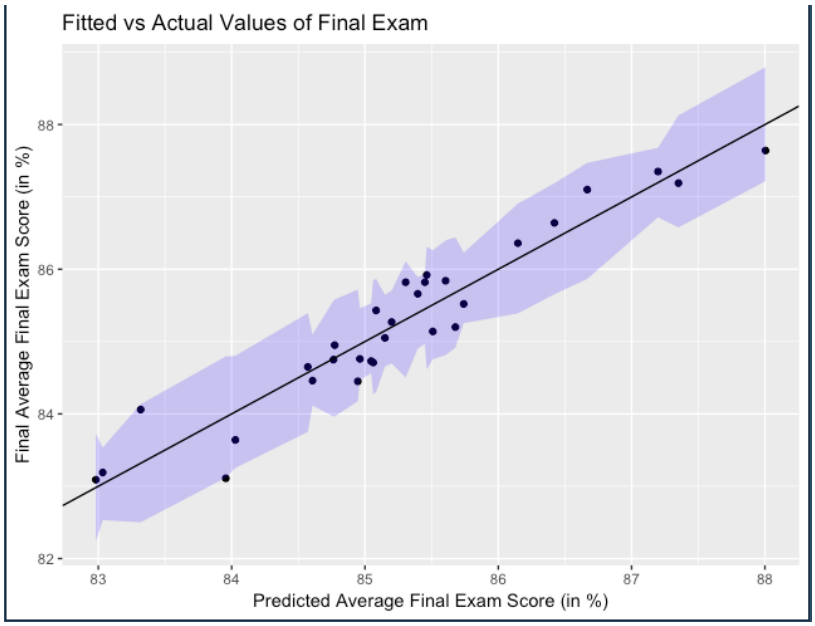
While OLS models would have the additional assumption of homoskedasticity, we know intuitively that our model is heteroskedastic, so we account for the unequal variance by using a fixed

weights GLS model. Thus, the error terms are inversely related to the number of students for each classroom. A graphic of the error terms against the number of students is shown below:



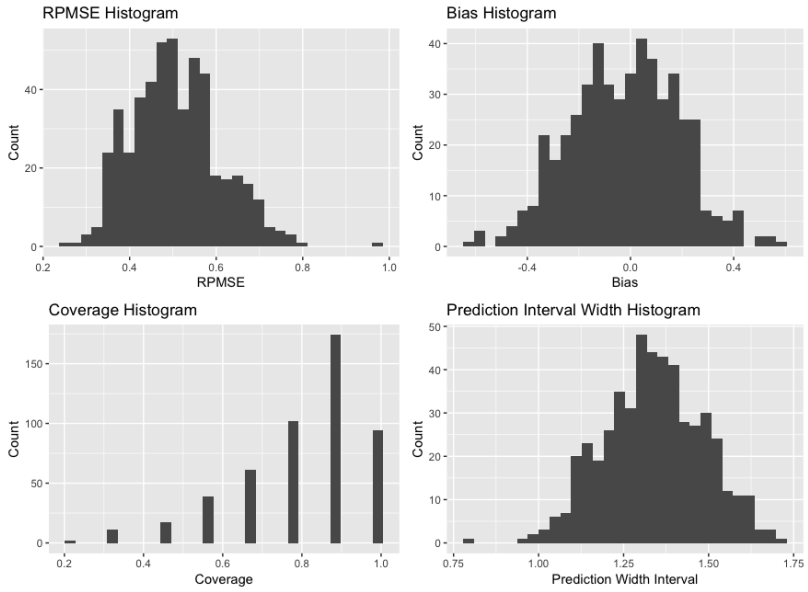
From this plot, it is easy to tell that we indeed have standard errors that are inversely related to the number of students in the classroom.

In assessing how well our model fits the data, we produced the following graphic displaying the actual Final scores on the x-axis and the predicted Final scores along the y-axis. In theory, if our model fits the data well, the scatterplot of fitted final scores against the actual final scores should have an approximate slope of 1 and intercept of 0. The following graphic suggests that the model fits our data well.



From the above graphic, we can see that the fit of our model is exceptionally satisfactory.

After confirming assumptions and verifying the fit of our model, we employed Monte Carlo cross validation methods to assess our model's ability to perform prediction. We used 500 Monte Carlo iterations and split the data so that 30 percent of the observations compiled the testing set and the remaining 70 percent formed the training set. For each iteration, we recorded the root predicted means squared error, bias, coverage, and prediction interval width of the model generated. The reported average root predicted means squared error was 0.507, average bias was -0.023, average coverage was 0.804, and average prediction interval width was 1.339. The following four graphics display histograms of each respective statistic.



As apparent with the histograms listed above, our model's ability to predict is moderately strong, with the exception of a low coverage rate. Calculating 0.950 prediction intervals, we should ideally expect to see our prediction bound estimates to capture the actual values 95 percent of the time as well, which is far superior to the actual coverage of 0.804.

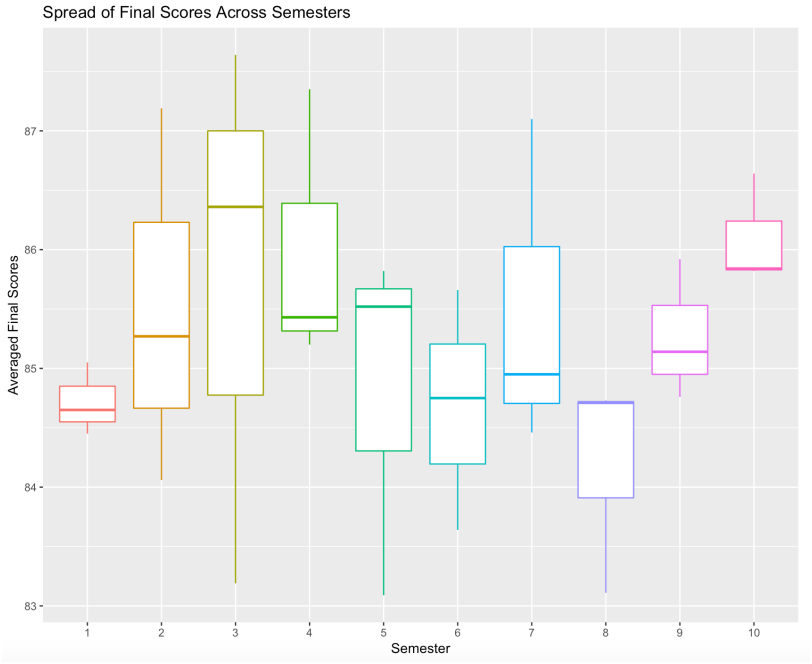
4 Analysis Results

Producing a summary table of our model suggests that performance on midterm exams 1-3 and homework are significant indicators of final exam scores. The p-value associated with the coefficient for exam 1 is 0.0028, the p-values for exam 2 and exam 3 are 0.0000, and the p-value for homework is 0.0077. All the mentioned p-values are less than 0.05, which suggests that the observed coefficients are very unlikely due to chance.

The exact effect estimates for each of the significant coefficients are presented in 0.95 confidence intervals. For each one point increase on Exam 1 score, we expect an increased final score of somewhere between 0.065 to 0.275 points. Exam 2 yields an increase between 0.207 to 0.464. Exam 3 yields an increase between 0.378 to 0.526. Lastly, Homework yields an increase between 0.116 to 0.677. From the effect intervals, Exam 3 clearly has the greatest effect on the final exam score. Although the upper bound for homework is higher than that of exam 3, the large spread of homework scores makes it less reliable than exam 3 as a predictor.

In order to assess how much variance of the data our model accounts for, we calculated a psuedo R-squared statistic of 0.911. The psuedo R-squared value is greater than 0.90, which says a large portion of the variation in the data is explained by our model.

A remaining question lingering in our analysis is whether or not one or more semesters' students performed significantly better than others semesters' students. To tackle this question, we first produced a boxplot graphic to display the spread of scores across each semester.



At first glance, the spread of the averaged final scores may appear to be unequal across semesters as apparent by the displayed boxplot. We decided to mathematically calculate the difference of average final scores across all semesters using one way ANOVA testing. We took three different approaches to the ANOVA testing. The first test we performed compared the values of our entire dataset and yielded a probability of 0.705 that the semesters had equal averages, which is quite high. The second test we ran considered a reduced model that only contained semester as an explanatory variable of the average final scores and yielded a probability of 0.907 that the semesters had equal averages, which is also quite high. The final test we ran considered the amount of variance in final scores that was actually explained by the semester variable, but this still yielded a probability of 0.827 suggesting the semester variable did not account for a significant portion of the variance in averaged final scores.

As none of the ANOVA tests run produced a noteworthy probability to suggest the averages across each semester were unequal, we can reasonably conclude that no one semester performed better than any of the other semesters.

5 Conclusions

After examining the data and encountering violations of the statistical models, we overcame the issues by accounting for the differing class sizes across the sections. The assumptions were appropriately satisfied by using a GLS model. Upon completing cross validation and analysis of our model of averaged semester final scores, we have identified the key activities to promote learning to be all three of the midterm exams and the homework. The observed average score of Exam 3 appears to have the largest effect on the average final score of classes. One possible explanation for the large effect Exam 3 had on the final scores could be that the final exam mostly includes Exam 3 content. Although our model was built with few data points and a sample representative only of BYU, the validation process of our model indicated the effectiveness of our model's ability to predict learning outcomes of internal future semesters.

Given the adequate information provided by our GLS model analysis, we suggest that the department continue to put greater emphasis on the traditional elements of homework and midterm exams in each section to promote successful learning. One interesting possibility to consider in the future may be to consider all midterm and final exams averaged together as a whole as a measure of successful learning. Although doing so would decrease the number of predictors for future analysis and complicate modeling, considering all test scores pooled together would produce a more holistic approach to measuring successful learning throughout the semester.