Pedagogy

Jacob Eliason and Gabriel Adams

February 2019

0 Executive Summary

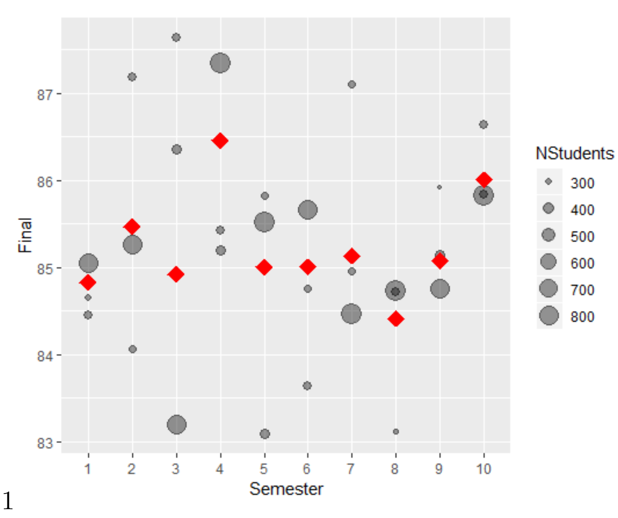
Average scores for midterm exams and homework are associated with final exam scores for a given course in the statistics department at Brigham Young University. The association between the third midterm exam score and the final exam score is particularly strong. Quiz scores and the semester to which a section pertains do not have a statistically significant effect on the final score. We advise university faculty to focus more on midterm exams and homework in their introductory statistics classes.

1 Introduction and Problem Background

Learning is an integral part of the university experience. Researchers are interested in determining which class activities are the most effective at promoting learning. With this knowledge, more time and effort could be dedicated to class activities deemed effective, while activities deemed ineffective could be discarded or reduced. Our primary research goal is to use our data to determine a quantifiable relationship between various explanatory variables (scores for three exams, a score for the quiz category, and a score for the homework category) and the response variable (overall learning as measured by Final score). Through this relationship, we will be able to determine which activities are truly associated with improved learning and what the specific effects of those activities are. We will also be able to determine how well all of the activities together explain overall learning. Finally, we will be able to see whether there were any semesters that had either better or worse student learning than average.

To explore this topic, we will use a set of data collected from a statistics department. The data consists of average scores from various sections of a single course, taken from different historical semesters. For each section, the specific semester (Semester) and the number of students (NStudents) in the course were recorded. In addition, the average scores for three midterm exams (Exam1, Exam2, Exam3), the average grade for all quizzes (Quiz), the average grade for all homework (HW), and the average grade on the final exam (Final) were recorded. There are 30 total observations in the data (representing three sections from each of ten semesters). The average grades in the various categories range (approximately) from 71 to 88.

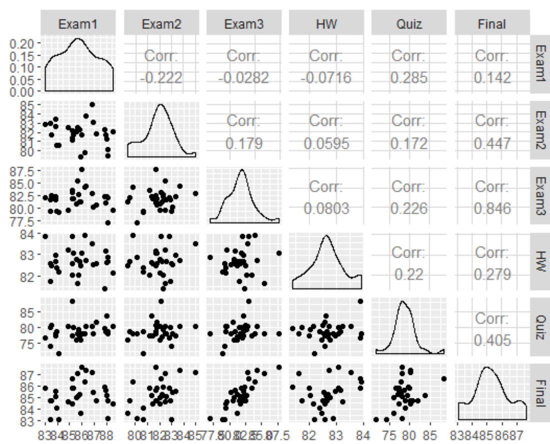
The plot to the right shows the relationships between Semester and Final. Gray dots represent individual sections. The size of each dot reflects the number of students enrolled in that section. Red diamonds represent the average final score for all students from a given semester. Notably, we observe that larger classes have a bigger "pull" on this average than smaller classes. Should we treat big and small classes the same when we fit a model? We should not. The discrepancies in class size create complications for traditional modeling techniques which assume equal variance in the data. We will address this issue later in our paper.



A pairs plot of our quantitative variables is shown below. We can see roughly linear relationships between the explanatory variables and the response (Final). In this analysis, we have chosen to disregard potential concerns relating to collinearity.

We will use a heteroskedatistic multiple linear regression model to determine a quantifiable relationship between our explanatory variables and our response variable (Final). This model will properly account for the differing variance that exists because our observations are actually averages of many individual scores. It will do so by using a diagonal covariance matrix instead of a standard identity matrix. Sections with more students will be more influential in this model, and sections with fewer students will be less influential. By accounting for this characteristic in our model, we will better utilize all of the information contained in our data set. Failing to do so would undervalue class sections with more students and overvalue sections with fewer students.

Using this model, we will be able to determine which activities are or are not associated with improved learning. If any activities are so associated, we will be able to determine which ones have the strongest effects and quantify what those effects are. We will also be able to explain how well the various class activities (as described by our explanatory variables) explain overall student learning. Finally, we will be able to determine whether there were any semesters that had either better or worse learning outcomes than average.



2 Statistical Model

To reiterate, in order to accomplish the goals of our analysis set forth in the previous section we will use a heteroskedatistic multiple linear regression model. A heteroskedatistic model is necessary because each of our observations is actually an average of many students' grades. It is fairly intuitive that our model should give more credence to averaged grades from larger classes (≥ 800 students) than to those of smaller classes (≤ 200 students). The variability of an average grade will be much smaller in a larger class versus a smaller class. Using a weighted regression model will account for this difference in variability.

Our model is as follows:

$$\boldsymbol{y} \sim MVN(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{D}) \tag{1}$$

- $\boldsymbol{y}$ is a vector of all the Final values (our response variable).
- $\boldsymbol{X}$ is the design matrix containing relevant information from our observations. Its first column is a column of 1's, corresponding to the intercept. The following nine columns of $\boldsymbol{X}$ correspond to semesters 2 through 10 (the first semester is absorbed into the intercept). In a given column, a 1 indicates that observation (section) belongs to that semester, while a 0 indicates it does not. The following five columns correspond to Exam1, Exam2, Exam3, HW, and Quiz, respectively. These columns contain average scores corresponding to each of the sections. Note that we will not be including NStudents as a explanatory variable in our model. It will only be used for weighting purposes.

- $\boldsymbol{\beta}$ is a vector of coefficients. The first element of $\boldsymbol{\beta}$ (β_0) is the intercept term and it corresponds with the first column of $\boldsymbol{X}$. The next nine elements (β_1 through β_9) correspond to the effects of the second through tenth semesters on Final. The last five elements of $\boldsymbol{\beta}$ (β_{10} through β_{14}) correspond to the effects of Exam1, Exam2, Exam3, HW, and Quiz, respectively. A theoretical section in the first semester with scores of 0 for all categories will have a Final score of β_0, on average. Holding all else constant, we expect a theoretical section in the second semester to have a final score β_1 different than a course in the first semester, on average. Similar relationships hold for β_2 through β_9. Holding all else constant, as Exam1 increases by 1, we expect Final to increase by β_{10}, on average. Similar relationships hold for β_{11} through β_{14}.

- $\sigma^2\boldsymbol{D}$ indicates that the residuals have an unequal variance about the line (as explained earlier in this report) and that our observations are independently distributed (0's in the off-diagonal). We will justify the independence assumption later.

  $\mathbf{D}$ is a diagonal matrix with different values (d_{ii}) going down the diagonal to indicate different variances for the various observations. Because our observations are averages, we will define

$$d_{ii} = \frac{1}{NStudents_i} \tag{2}$$

  These values will be multiplied by a constant variance (σ^2) to give us a proper heteroskedastic model. This will allow us to give more value to sections with more students and less value to sections with fewer students.

To use this model, several assumptions will need to be justified: linearity, independence, normality, and equal variance (after standardizing). Our data must have linear relationships. Our responses should be independent of each other. The standardized residuals of our model should be normally distributed. Finally, after standardizing, there should be equal variance about the line. We will justify these assumptions in the following section.

3 Model Validation

Our model is only valid if the following assumptions hold: linearity, independence, normality, and equal variance (after standardizing). We justify these assumptions now.

1. Linearity can be verified through the use of an added-variables plot (next page). We can clearly see linear trends in the quantitative variables. We do not observe any obvious curvature or other nonlinear relationships. Thus, our assumption of linearity holds.

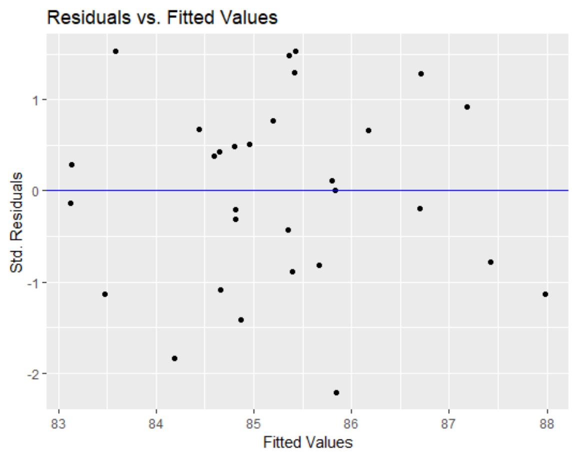
2. It is reasonable to assume that the average final score of one semester does not affect the average final score of another. This means that we can reasonably assume independence.

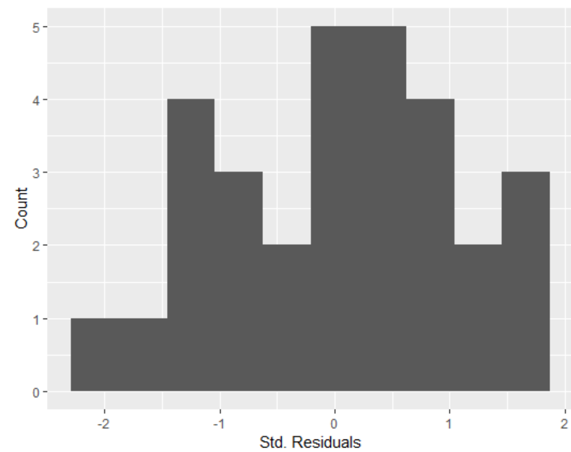
3. Normality can be verified by evaluating a histogram of standardized residuals (next page, upper right). The standardized residuals are few in number (30), so it is relatively difficult to see a distribution. They range between -2 and 2 and we conclude that a normal distribution would probably fit them better than any other distribution. We will therefore assume normality. Running a KS test for normality in R returns a p-value of 0.8782. This p-value is larger than 0.05, our chosen alpha level, so we fail to reject the null hypothesis and further conclude that the normality assumption is met.

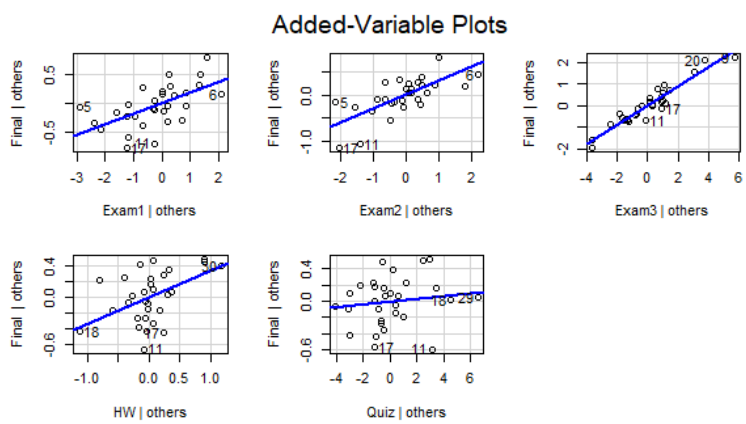
4. Equal variance can be verified by looking at a modified plot of standardized residuals vs. fitted values (next page, upper left). The spread of the points about the fitted line is roughly the same as we move from left to right, so equal variance is reasonable to assume.

Our four assumptions hold, so our model is valid to use.

We analyze the fit of our model by calculating R^2. The R^2 value for our model is 0.9541. This means that 95.41% of the variation in the average final exam scores is explained by our model. Acknowledging that the inclusion of the semester variables in our model will increase the value of R^2, we fit a model without those variables and find that R^2 is still greater than 90%. This means that our model fits the data well.


Residuals vs. Fitted Values
Std. Residuals
Fitted Values


Count
Std. Residuals


Added-Variable Plots
Final | others
Exam1 | others
Exam2 | others
Exam3 | others
HW | others
Quiz | others

4 Analysis Results

After fitting our model in R, we determine the effects, along with corresponding confidence intervals, of our various explanatory variables. These findings are included in the following table.

	Estimate	Lower	Upper
(Intercept)	-21.2198	-47.7511	5.3115
Semester2	0.2986	-0.2859	0.8830
Semester3	0.1177	-0.5512	0.7866
Semester4	0.1937	-0.4748	0.8622
Semester5	0.1555	-0.4071	0.7181
Semester6	0.1250	-0.4752	0.7252
Semester7	0.0917	-0.5936	0.7770
Semester8	-0.3856	-0.9965	0.2252
Semester9	-0.0745	-0.7470	0.5981
Semester10	0.4189	-0.1593	0.9970
Exam1	0.1844	0.0657	0.3031
Exam2	0.3037	0.1526	0.4549
Exam3	0.4437	0.3741	0.5133
HW	0.3403	0.0298	0.6509
Quiz	0.0165	-0.0472	0.0801

Although it is not particularly important for our analysis, we note that our estimate for σ is 5.4149, with a 95% confidence interval from 4.3271 to 7.2380.

The values in the table above can be interpreted as follows. A theoretical course will have a Final score of -21.22, on average, when it is in the first semester and has scores of 0 for all the different categories. We are 95% confident that this theoretical course will have a final score between -47.75 and 5.31. Holding all else constant, we expect a theoretical course in the second semester to have a final score 0.299 different than a course in the first semester, on average, with a 95% confidence interval from -0.286 to 0.883. Similar relationships hold for the other semesters. Holding all else constant, as Exam1 increases by 1, we expect Final to increase by 0.184, on average. We are 95% confident that as Exam1 increases by 1, Final will increase by between 0.066 and 0.303. Similar relationships hold for the other midterm exams, the quiz scores, and the homework scores.

From this table, we can answer the research questions we posed at the beginning of this report. Our first research question was which, if any, of the class activities were associated with learning (as measured by Final). Intervals not containing zero indicate effects that are statistically significant. In the table above, we can see that the midterm exams and the homework category have positive associations with Final, while the semester variables and quiz category do not appear to have a statistically significant effect on Final. Increases in performance on the exams and homework generally correspond with increases in performance on the final exam. These increases are explained in the following paragraph.

Holding all else constant, as Exam1 increases by 1, we expect Final to increase by 0.184, on average. We are 95% confident that as Exam1 increases by 1, Final will increase by between 0.066 and 0.303. Holding all else constant, as Exam2 increases by 1, we expect Final to increase by 0.304, on average, with a 95% confidence interval from 0.153 and 0.455. Holding all else constant, as Exam3 increases by 1, we expect Final to increase by 0.444, on average, with a 95% confidence interval from 0.374 to 0.513. That is, as the average score on the third midterm increases by 1 point, we expect the average score on the final exam to increase by between 0.374 and 0.513 points. Finally, as HW increases by 1, Final will increase, on average, by 0.34, with a 95% confidence interval from 0.030 to 0.651.

Thus, it appears that the largest effect is Exam3, followed closely by HW and Exam2, and finally by Exam1. It makes intuitive sense that the last midterm exam is highly associated with the final exam, as the last midterm exam is typically taken fairly late in the semester, after most of the learning has taken place. It also makes sense that the score for the first midterm has a relatively small effect on Final because it is much closer to the beginning of the semester and students generally are still familiarizing themselves with the class structure and content. The effect sizes of Exam2 and HW are somewhere in between those of Exam1 and Exam3.

We are also interested in how well the class activities explain learning outcome. We calculate an R^2 value of about 0.9541. This means that approximately 95.41% of the variation in Final is explained by our model. As mentioned previously, we fit a model without the semester variables to isolate the effects of class activities on Final. This model still has an R^2 value of 0.9243. In other words, you can know quite a bit about the average final score given the scores on the previous assignments, exams, and quizzes.

Our final research question was whether or not there were any semesters with a significantly different Final score than the average. We can answer this by looking at the table above. All of the confidence intervals for the semester variables contain zero. This means that none of their effects are significant at an alpha level of 0.05. To further confirm this result, we fit a reduced model without the semester variables and conduct an F test comparing that reduced model to our original model. This test returns a F statistic of 1.0822 and a p-value of 0.4284. We can therefore conclude that none of the semesters had a Final score significantly different than the average.

5 Conclusions

Through the use of a heteroskedastic multiple linear regression model, we were able to answer the questions posed at the beginning of this analysis. Average scores for midterm exams and homework are associated with final exam scores for this specific course. The association between the third midterm exam score and the final exam score is particularly strong. Quiz scores and the semester to which a section pertains do not have a statistically significant effect on the final score.

In deciding where to allocate resources for struggling students, administrators should focus on students who have lower exam scores (and to a lesser extent, those with lower homework scores), and not necessarily those with low quiz scores. A potential change to be considered for the course would be to eliminate quizzes and add a fourth midterm exam, to better identify and aid struggling students.