Project 1

Justin Ross, Tyler Smith

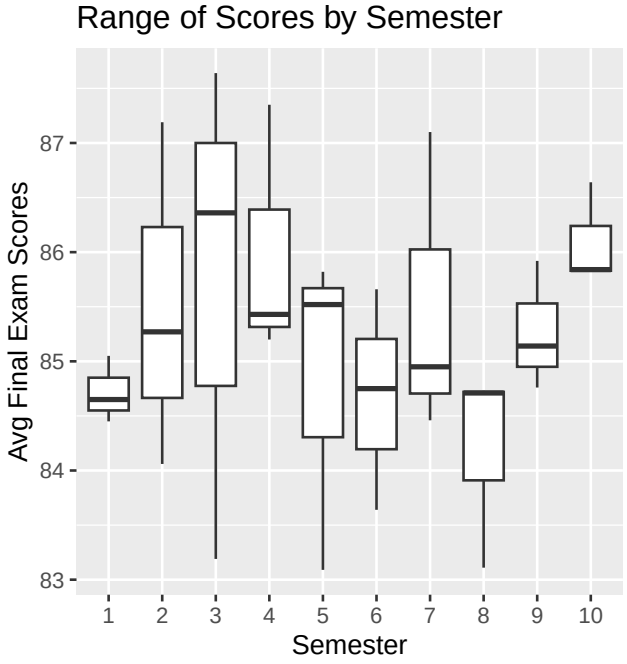
2024-03-26

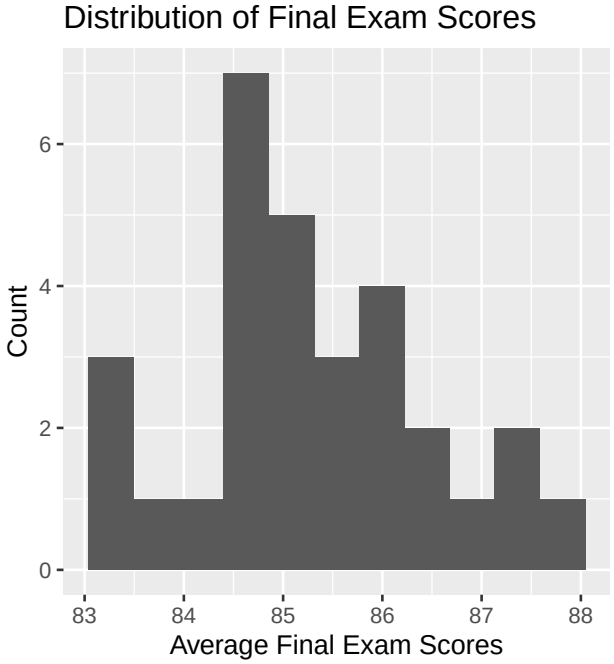
Executive Summary In our analysis, we found that student learning was most associated with performance on midterm exams. Homework scores were also a good indicator of student learning but were not determined to be statistically significant. We determined that there was no semester that was significantly worse or better in terms of student learning. The number of students in a section and quiz scores were not found to be statistically significant.

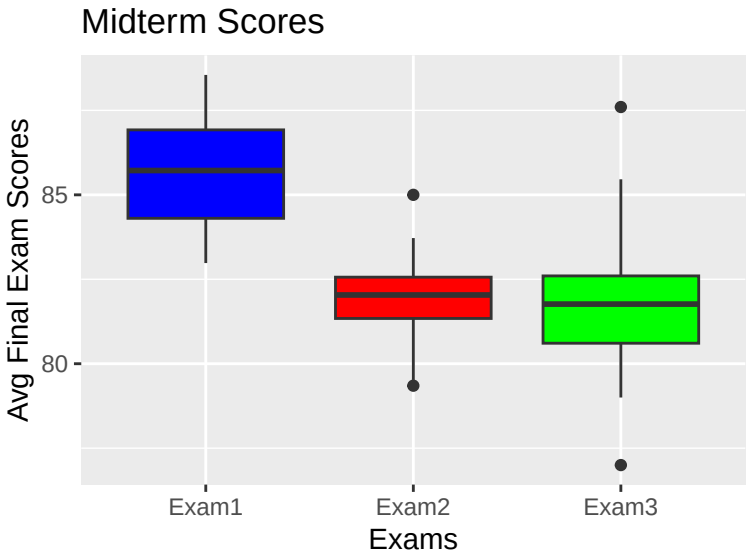
Introduction The purpose of this study is to evaluate the effect of various learning activities on student learning in an introductory statistics class. Our data consists of average scores on 3 unique learning activities administered by a statistics department over several semesters. The learning activities include exams, homework, and quizzes, with student learning measured by their performance on the final exam. From our analysis we hope to determine:

- Which of these three learning activities are most closely associated with student learning
- If these learning activities explain student learning, and
- If there are semesters that had better or worse student learning than average

Graphical Summaries







When examining the data, we found that most of the learning activities have a very low correlation with the average final exam score. Exam 3 has the highest correlation with the final exam with a correlation of 0.84, but the next highest correlation with the final exam is Exam 2's correlation of 0.44. This suggests that some of our explanatory variables may not have a strong linear relationship with average final exam score. We are also concerned about differences in class sizes. Class sizes in this data range from 280 to 873 students, and we are concerned that these differences in class size may affect the variability in the average final exam scores.

If these issues are not addressed, our standard errors will be inaccurate. Inaccurate standard errors lead to biased coefficient estimates that do not capture the true relationship between the explanatory variables and the average final exam score. We can account for differing class sizes by fitting a model that uses a variance function to produce a unique variance for each individual observation. The variance of an individual observation will be the sample variance divided by an observation's value for NStudents.

To analyze this data, we plan to use a multiple linear regression model. Multiple linear regression seems like the best choice because it can include several explanatory variables in the process of explaining the response variable. We will include all of the variables provided in the data as well as interaction terms related to class size. This should improve our model's ability to capture the relationship between learning activities and average final exam scores.

Section 2: Statistical Model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \quad , \quad \mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{D})$$

$$\begin{pmatrix} Final_1 \\ Final_2 \\ \vdots \\ Final_n \end{pmatrix} = \begin{pmatrix} 1 & NStudents_1 & Exam1_1 & \dots & Semester10_1 \\ 1 & NStudents_2 & Exam1_2 & \dots & Semester10_2 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & NStudents_n & Exam1_n & \dots & Semester10_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_{NStudents} \\ \beta_{Exam1} \\ \beta_{Exam2} \\ \beta_{Exam3} \\ \beta_{HW} \\ \vdots \\ \beta_{Semester10} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

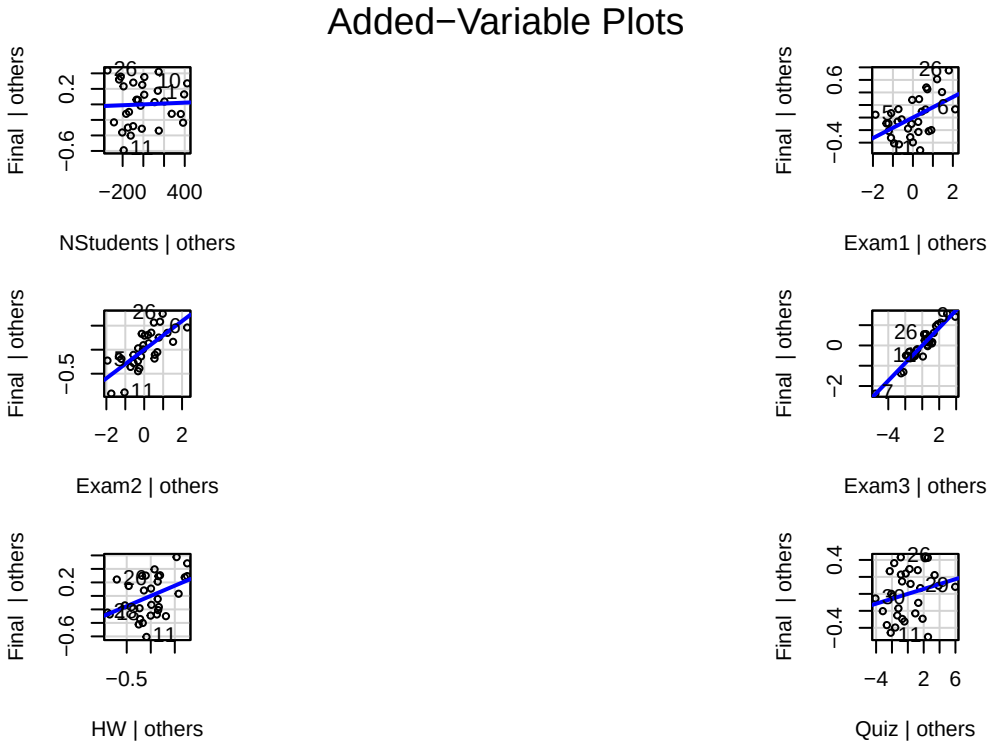
$$\text{Var}(\epsilon_i) = \frac{\sigma^2}{NStudents_i}, \quad D = \begin{pmatrix} d_{11} & 0 & 0 & \cdots & 0 \\ 0 & d_{22} & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & d_{nn} \end{pmatrix}$$

- **y** is a column vector of the response variable observations (average final exam score).
- **X** is a matrix of the explanatory variables. Each row represents an observation and each column represents one of the variables (NStudents, Exam1, Exam2, etc).
- β is a column vector of coefficients. β_0 represents the intercept and the rest correspond to individual explanatory variables in the **X** matrix.
- ϵ is a column vector of residuals (the difference between the observed and predicted values for average final exam score)
- σ^2 represents the variance of the average Final Exam score in the data.
- **D** represents a diagonal matrix where each diagonal element corresponds to the unique variance of each observation in the data. Larger values of d_{ii} represent a larger variance for the ith observation. The off-diagonal elements are all zero because we assume that the errors are not correlated with each other.

For our analysis, there are four assumptions that must be met in order for our inferences to be valid. The assumptions are:

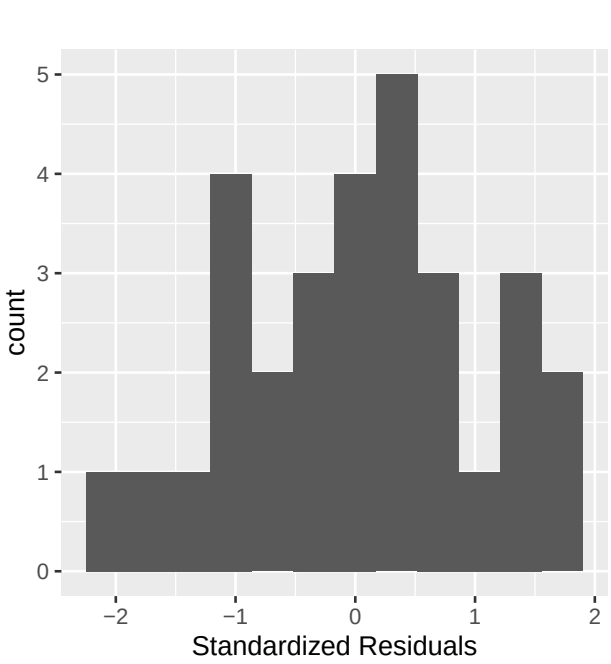
- A linear relationship between the average final exam scores and the numeric explanatory variables in the data.
- Independence of the observations. The average final exam scores from one section should not impact the average final exam scores of another section.
- Normality distributed standardized residuals.
- Equal Variances. The variance of the residuals should be constant across all of the explanatory variables

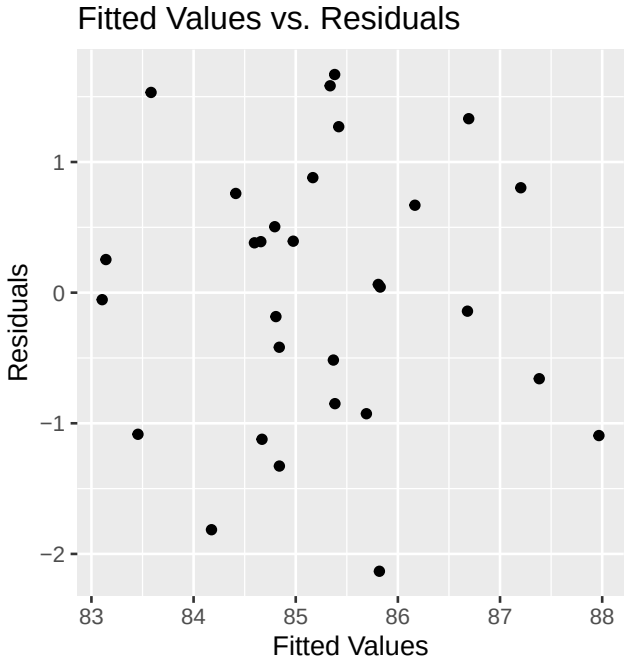
Section 3: Model Validation



These added variable plots look sufficiently linear for the linearity assumption to be met. There are no obvious non-linear trends in any of the plots for numeric variables. In our analysis we are treating Semester as a category, so we are not concerned about the linearity of the Semester variables.

This data was collected from past semesters of an introductory statistics class. There is no reason for us to believe that the average final exam score of one section would affect the average final exam score of another section. Thus, we are confident that the independence assumption is met.





The histogram of standardized residuals looks approximately normal. The Kolmogorov-Smirnov test returns a p-value of 0.978, which confirms our belief that the standardized residuals are normally distributed.

The fitted values vs. residuals plot shows an approximately equal spread of residuals. The lack of obvious patterns or trends confirms our belief that the equal variance assumption is satisfied.

After verifying that the necessary assumptions are met, this is the model that we produced:

	Estimate	Std. Error	t-value	p-value
(Intercept)	-20.678894	14.106414	-1.465921	0.1648
NStudents	0.000084	0.000301	0.279317	0.7841
Exam1	0.174047	0.072629	2.396391	**0.0311**
Exam2	0.298527	0.081761	3.651215	**0.0026**
Exam3	0.444123	0.036699	12.101756	**0.0000**
HW	0.347460	0.165549	2.098834	0.0545
Quiz	0.017499	0.033726	0.518872	0.6120
Semester2	0.314248	0.312913	1.004268	0.3323
Semester3	0.135855	0.358230	0.379239	0.7102
Semester4	0.217420	0.362182	0.600304	0.5579
Semester5	0.153728	0.296367	0.518709	0.6121
Semester6	0.142737	0.322422	0.442702	0.6647
Semester7	0.123949	0.378959	0.327079	0.7484
Semester8	-0.376932	0.323211	-1.166208	0.2630
Semester9	-0.048025	0.366645	-0.130986	0.8977
Semester10	0.414087	0.304966	1.357815	0.1960

To determine how well our model fits the data, we performed leave-one-out cross validation. The mean RPMSE was 0.2266437 and the mean Bias was -0.004780159. These low values indicate that our model

produces predictions that are close to the true values. We also calculated a pseudo R-Squared of 0.942, which indicates that our model's explanatory variables explain much of the variability in the Final Exam scores.

Our primary research questions did not involve predictions, so we did not make predictions for new observations outside of our cross validation.

Section 4: Analysis Results

From hypothesis tests on the coefficients, we have determined that Exams are associated with improved learning. We have determined that their effects are as follows (holding all else constant):

- We are 95% confident that as the average Exam 1 score increases by 1 percent, the average final exam score will increase by between 0.03169714 and 0.316396693 percent on average.
- We are 95% confident that as the average Exam 2 score increases by 1 percent, the average final exam score will increase by between 0.1382784 and 0.458775603 percent on average.
- We are 95% confident that as the average Exam 3 score increases by 1 percent, the average final exam score will increase by between 0.3721938 and 0.516051365 percent on average.

To determine if there were any semesters that were significantly better or worse in terms of student learning, we fit two models: one with all of the variables and one that excluded the semester variable. We performed an F-test on these different models, and the p-value of 0.0918 indicated that there was not a significant difference between these models. There is insufficient evidence to conclude that one semester is significantly better or worse than another in terms of student learning.

Section 5: Conclusions

Our analysis indicates that midterm exams are the best indicators of student performance on the final exam. Homework scores were a good indicator of student performance on the final exam but not as much as the exams. Quiz scores have a very minimal effect final exam scores, so they should not be prioritized. We suggest that the department could create a better learning environment by holding exam reviews both before and after the exams are administered.

Code Appendix

```
# load libraries
library(tidyverse)
library(vroom)
library(nlme)
library(patchwork)

# read in data
class <- vroom("ClassAssessment.txt")
class$Semester <- as.factor(class$Semester)
source("predictgls.R")

# combine plots to save space
p1 <- class %>%
  ggplot(aes(x = Semester, y = Final)) +
  geom_boxplot() +
  theme(aspect.ratio = 1) +
  labs(y = "Avg Final Exam Scores",
       title = "Range of Scores by Semester")

p2 <- class %>%
  ggplot(aes(x = Final)) +
  geom_histogram(bins = 11) +
  theme(aspect.ratio = 1) +
  labs(title = "Distribution of Final Exam Scores",
       x = "Average Final Exam Scores",
       y = "Count")

p1 + p2

# boxplot
class %>%
  ggplot() +
  geom_boxplot(aes(x = "Exam1", y = Exam1), fill = "blue") +
  geom_boxplot(aes(x = "Exam2", y = Exam2), fill = "red") +
  geom_boxplot(aes(x = "Exam3", y = Exam3), fill = "green") +
  labs(title = "Midterm Scores",
       x = "Exams",
       y = "Avg Final Exam Scores")

# correlations
numeric <- class %>% select(-c("Semester"))
corrplot::corrplot(cor(numeric))

# fit gls model
glsmodel <- gls(Final ~ NStudents+Exam1+Exam2+Exam3+HW+Quiz+Semester, class, weights= ~1/NStudents, metl

# AV Plots
model <- lm(Final ~ ., class)
par(pty = "s")
car::avPlots(model, terms=~Exam1+Exam2+Exam3+HW+Quiz+NStudents)

# more plots
```

```r
p3 <- ggplot() +
  geom_histogram(aes(x = resid(glsmodel, type="pearson")), bins = 12) +
  labs(x = "Standardized Residuals") +
  theme(aspect.ratio = 1)

p4 <- ggplot() +
  geom_point(aes(x = fitted(glsmodel), y = resid(glsmodel, "pearson"))) +
  labs(x = "Fitted Values",
       y = "Residuals",
       title = "Fitted Values vs. Residuals") +
  theme(aspect.ratio = 1)

p3 + p4

ks.test(resid(glsmodel, type="pearson"), "pnorm")
n.cv <- nrow(class)
rpmse <- rep(NA, n.cv)
cvg <- rep(NA, n.cv)
bias <- rep(NA, n.cv)

for(cv in 1:n.cv){
  test.obs <- cv

  # Split into test and training sets
  test.set <- class[test.obs, , drop = FALSE]
  train.set <- class[-test.obs, , drop = FALSE]

  # Fit a GLS model using the training data
  train.lm <- gls(Final ~ NStudents+Exam1+Exam2+Exam3+HW+Quiz+Semester,
                  class, weights= varFixed(value = ~(1/NStudents)), method="ML")

  # Generate predictions for the test set
  my.preds <- predictgls(glsobj = train.lm, newdframe = test.set, level = .95)

  # Calculate RPMSE for this iteration
  rpmse[cv] <- sqrt(mean((test.set[['Final']] - my.preds[['Prediction']])^2))

  # Calculate Coverage for this iteration
  cvg[cv] <- mean((test.set[['Final']] > my.preds[['lwr']])
                  & (test.set[['Final']] < my.preds[['upr']]))

  ## Calculate bias
  bias[cv] <- mean(my.preds[,'Prediction']-test.set[['Final']])
}

mean(rpmse)
mean(bias)
residuals_model <- residuals(glsmodel)

nullmodel <- gls(Final ~ 1, class, method="ML")
residuals_null <- residuals(nullmodel)

# Calculate pseudo R-squared
```

```
pseudo_r_squared <- 1 - (var(residuals_model) / var(residuals_null))
pseudo_r_squared
confint(glsmodel)
no_semester <- gls(Final ~ NStudents+Exam1+Exam2+Exam3+HW+Quiz, class,
                   weights= ~1/NStudents, method="ML")

anova(no_semester, glsmodel)
```