Statistical Pedagogy

A Study of Learning Activities in a Statistics Course

Josh Bergstrom, Aaron Oustrich, Anna Wolford

February 26, 2024

Executive Summary

The goal of this analysis is to assess if the semester learning activities are associated with mastering the given content. Specifically, we seek to know the learning activities with the strongest effect on the Final Exam grade, how well class activities model student learning, and if there were any semesters that had either better or worse learning than average. We found that the learning activities with the strongest effect on the Final Exam grade were the 3 "mid-term" exams which is consistent across all semesters.

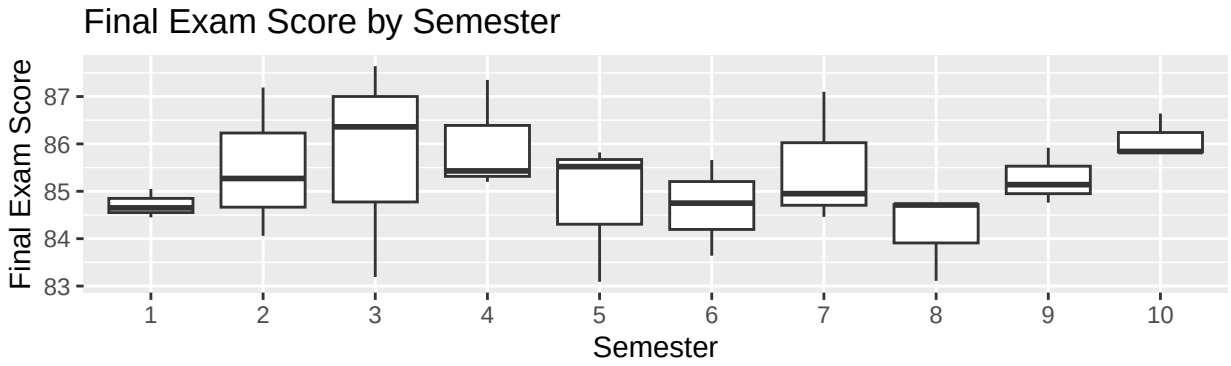
Introduction and Problem Background

To maximize student learning, curriculum creators and college instructors develop instruction to help students master the given content. This instruction should include meaningful learning activities that help the students progress. In this study, a statistics department gathered data on the performance of the students in their introductory statistics course for the most recent 5 academic years (excluding summer semesters).

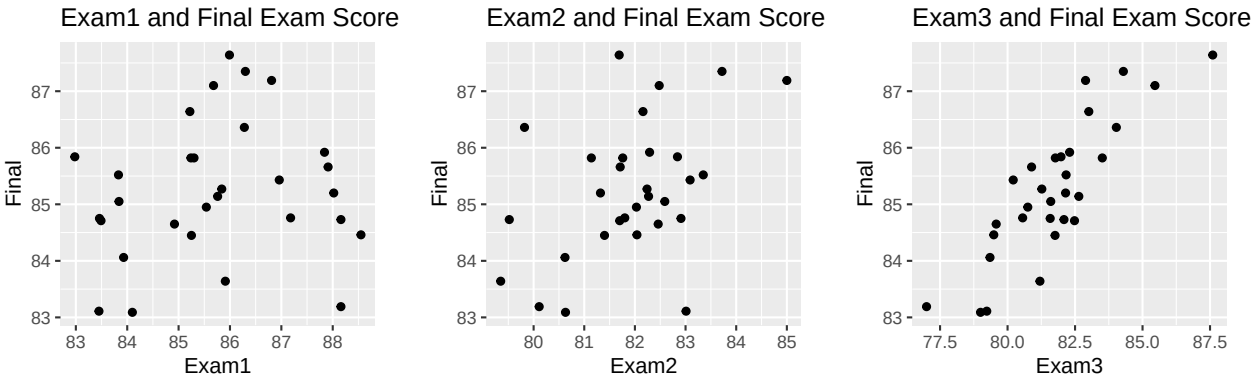
The goal of this analysis is to assess if the semester learning activities are associated with mastering the given content, which said mastery is measured by the final exam score. Specifically, we seek to know if there are any activities associated or not associated, of those associated which have the strongest effect, how well class activities model student learning, and if there were any semesters that had either better or worse learning than average.

The explanatory variables included in the analysis include the number of students who completed the course (NStudents), and the average scores (in percent) of Exam 1, Exam 2, Exam 3, the homework, and the in-class quizzes (respectively). As mentioned, the response variable (FinalExam) includes the average scores on the Final.

To assess the potential difference in semesters for their Final Exam score, we created side-by-side boxplots to see if there were any that appeared extreme. We noticed that Semester 1 and Semester 10 appeared to have the largest difference in medians, however, they only ranged between 2 points. A further analysis is required to determine if this difference is significant.



We also created scatterplots comparing the three different exams with the Final Exam score for any given semester. These scatterplots individually were a bit difficult to draw information from, however, side-by-side it appears that between Exam 1 and then Exam 2 and then Exam 3 there grows a stronger positive association between their respective averages and the Final exam score. From this, we speculate that Exam 3 could have the strongest effect and a positive association with Final Exam scores. Further analysis is required in order to confirm this speculation.



We perceive a potential stumbling block in our analysis due to the variable "NStudents". Across sections and semesters, the number changes by as many as 593 students. We note that all of the learning activities and the Final Exam score are averages. Thus, having averages based on different numbers of students means that the averages are weighted differently. This will especially affect the variability of our model and the standard errors. Failure to account for this difference would result in incorrect standard errors and incorrect predictions. Perhaps it would also result in accepting a model that is unfit for the data and would therein lead to incorrect conclusions.

For our analysis, we are using a generalized least squares regression to account for heteroskedastic errors. This method models the relationship between a response variable and multiple explanatory variables. In our case, the response variable is the average scores on the final exam, and the explanatory variables include various factors such as the number of students in each class section, average exam scores, homework scores, and quiz scores. We chose multiple linear regression because it allows us to examine the impact of least squares regression on the final exam scores simultaneously.

Finally, we wanted to see if any semesters were better or worse in student learning. After experimenting with different models that included interactions between semester and other variables, we found that none of the interaction terms were significant at the 0.05 level. Therefore, because we only have 30 observations in this dataset, we opted to exclude the interaction terms from our final model to make it more parsimonious.

Statistical Model

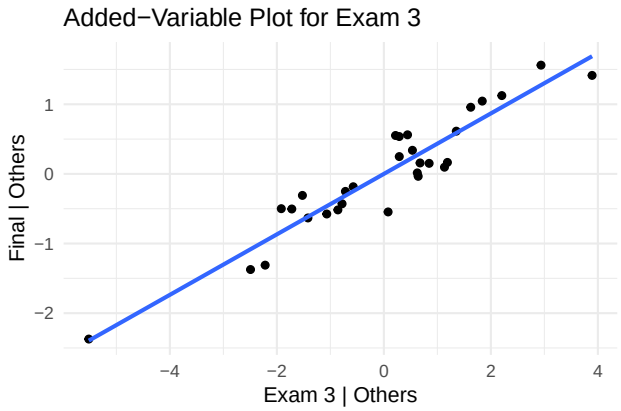
The multiple linear regression model is based on the equation $\mathbf{y} = \mathbf{X}\beta + \epsilon$. However, due to heteroskedasticity in the model, the response vector follows $\mathbf{y} \sim \text{MVN}(\mathbf{X}\beta, \sigma^2\mathbf{D})$ where the diagonal values (d_{ii}) of matrix $\mathbf{D}$ are

equal to the inverse of the number of students in each class $\left(\frac{1}{\text{NStudents}_i}\right)$, and the off-diagonal values are all 0 to since the Final Exam scores are independent. We then weighted our model's variance of average final exam scores (σ^2) by multiplying it to $\mathbf{D}$. Doing so accounts for the issues caused by the varying amounts of students in each section and semester. In our model, $\mathbf{y}$ is a vector of average scores (in percent) on the final exam. $\mathbf{X}$ is the matrix of explanatory variables including a column of 1s for the intercept. We dummy-encoded the Semester variable because we are treating it as a factor. β is the vector of coefficients. Finally, ϵ is the vector of error terms of each observation. We assume that the errors are normally distributed with mean 0 and constant variance (after accounting for it with $\mathbf{D}$).

This type of model must meet four different assumptions in order to be used. First, the observations must have a linear relationship with the predictors (they roughly follow a line when plotted against one another). Next, the responses must be independent of one another (the average final score of one section cannot depend on the average final score of another). Then, the residuals of the responses must follow an approximately normal distribution. Finally, there must be equal variance among the responses.

Model Validation

Each of the added-variable plots follows a linear trend, but only the plot for Exam3 is shown below; therefore, the linearity assumption is met. The assumption of independence is also met since the average scores of one section's final exam does not depend on the average score of another section's final exam. The normality assumption is met as the histogram of the residuals appears approximately normal, and the Kolmogorov-Smirnov test of normality outputs a large p-value of 0.978. This means there is not enough evidence to refute the normality assumption. Lastly, the equal variance assumption is met since this model adjusted for the unequal variance caused by uneven sample sizes per observation, and we looked at a plot of the standardized residuals vs. fitted values to confirm that it was met, as shown below. (Plots and other statistical output not shown here will be found in the appendix.)



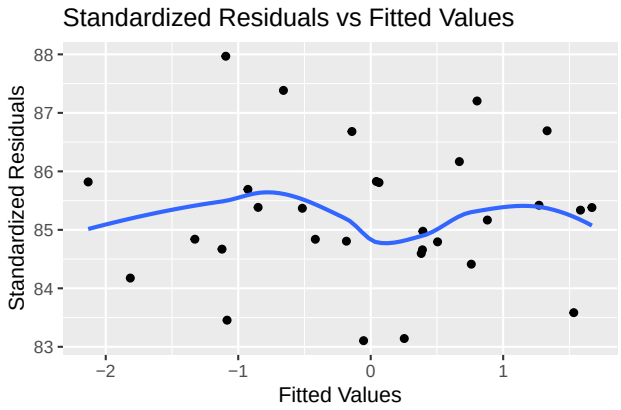


Table 1: Model Validation Metrics

CV.RMSE	Width	Coverage	Bias	In.sample.RMSE
0.462	2.144	0.967	-0.012	0.286

We are confident that our model accurately fits the data. Our model has an out-of-sample root mean square error value of 0.462. This means that, on average, our predicted final scores are off by about 0.46 percentage points. We are using this metric to say that our data fits the model well.

We are also confident that our model accurately predicts values. By using a Leave-one-out Cross-Validation Method, the prediction intervals created by our model correctly contained the true value 96.667% of the time (based on the model's mean coverage value). Considering how the intervals were set to contain the true value 95% of the time, a 96.667% coverage rate is extremely acceptable. Similarly, the mean RPMSE was 0.462

which represents how far off our predictions were on average from the true values. Considering that it's less than a single percentage point off, our model accurately predicts a student's final exam score on average by within a percent.

Analysis Results

Based on the summary output of our Heteroskedastic GLS model, only the coefficients for Exam1, Exam2, and Exam3 are significant at the 0.05 level (p-values of approximately 0.0311, 0.0026, and 0.0000 respectively). In the context of this analysis, we have reason to believe the average scores on the exams have significant positive associations with the average scores on the final exam and therefore are the class activities most associated with improved student learning. Based on the value of these significant coefficients, we believe that for every unit increase in the Exam1 average score and while holding all else constant, there is an average increase of about 0.17 percentage points on the average score of the Final exam. Likewise the average effects for Exam2 and Exam3 are 0.31 and 0.42 percentage points respectively. Thus, Exam3 has the largest effect on the average score of the final exam.

The in-sample RMSE for our model is 0.286.This means that, on average, our predicted final scores are off by about 0.29 percentage points. Because this is a very small margin, we believe our model shows these class activities explain learning very well.

Finally, to see if there were any significant differences between semesters, we conducted a generalized linear hypothesis test on the Semester coefficients all being equal to 0. The resulting p-value of 0.631 means that we do not have enough evidence to say that the average scores on the final exam differ significantly between semesters. This means that the semester in which a student takes the class does not significantly affect their final exam score.

Conclusion

Ultimately, we found that the average scores on the exams are the most important predictors of the average scores on the final exam and are the most important class activities for predicting student learning. We also found that the semester in which a student takes the class does not significantly affect their final exam score. We are confident in these results because we tailored the methods we used to the data we had and validated our model to ensure it was accurate.

We would suggest the curriculum creators for this introductory statistics class experiment with the format of Homework assignments and Quizzes to find ways these learning activities can better explain the average scores on the final exam. We would also suggest they focus on the material covered in Exam3, as it has the largest effect on the average score of the final exam.

Appendix

```
#  Read in Data
ped <- vroom("ClassAssessment.txt")
ped$Semester <- as.factor(ped$Semester)

# # EDA
## Scatterplot for Exams
plot_exam3 <- ggplot(ped, aes(x = Exam3, y = Final)) +
  geom_point() +
  labs(title = "Exam3 and Final Exam Score") +
  theme(plot.margin = margin(0.5, 0.5, 0.5, 0.5, "cm"))

plot_exam2 <- ggplot(ped, aes(x = Exam2, y = Final)) +
  geom_point() +
  labs(title = "Exam2 and Final Exam Score") +
  theme(plot.margin = margin(0.5, 0.5, 0.5, 0.5, "cm"))

# Scatterplot for Exam1
plot_exam1 <- ggplot(ped, aes(x = Exam1, y = Final)) +
  geom_point() +
  labs(title = "Exam1 and Final Exam Score") +
  theme(plot.margin = margin(0.5, 0.5, 0.5, 0.5, "cm"))

# Remove legend from Exam2 and Exam1 plots
plot_exam2 <- plot_exam2 + theme(legend.position = "none")
plot_exam1 <- plot_exam1 + theme(legend.position = "none")
plot_exam3 <- plot_exam3 + theme(legend.position = "none")

# LM
ped.lm <- lm(Final ~ .  , data = ped)
summary(ped.lm)

# GLS Model
ped.gls <- gls(model=Final~. ,
    data=ped,
    weights=varFixed(~1/NStudents),
    method="ML")

summary(ped.gls)

## Linearity
av3 <- car::avPlot(ped.lm, "Exam3")
av3 <- as.data.frame(av3)


av3_plot <- ggplot(av3, aes(x=Exam3, y=Final)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Added-Variable Plot for Exam 3",
       x = "Exam 3 | Others",
       y = "Final | Others") +
  theme_minimal()
```

```r
## Normality
ggplot(data = data.frame(residual = resid(ped.gls, type="pearson")),
       aes(x = residual)) +
  geom_histogram(binwidth = 0.2) +
  labs(title = "Histogram of Standardized Residuals",
       x = "Standardized Residuals",
       y = "Frequency")

normality_assumption <-  ks.test(resid(ped.gls, type="pearson"), "pnorm")

## Equal Variance
gls_ev_plot <- ggplot(data = data.frame(residual = resid(ped.gls, type="pearson")),
       aes(x = residual,
           y = ped.gls$fitted)) +
  geom_point() +
  geom_smooth(se = F) +
  labs(title = "Standardized Residuals vs Fitted Values",
       y = "Standardized Residuals",
       x = "Fitted Values")

# Cross Validataion

n <- nrow(ped)
rpmse <- rep(x=NA, times=n)
wid <- rep(x=NA, times=n)
bias <- rep(x=NA, times=n)
cvg <- rep(x=NA, times=n)
my.preds.df <- data.frame()


for(i in 1:n){
  ## Select test observations


  ## Split into test and training sets
  test.set <- ped[i,]
  train.set <- ped[-i,]

  ## Fit a lm() using the training data
  train.gls <- gls(model=Final ~. ,
    data=train.set,
    weights=varFixed(~1/NStudents),
    method="ML")

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


  ## Calculate RPMSE
  rpmse[i] <- (test.set[['Final']]-my.preds[,'Prediction'])^2 %>% mean() %>% sqrt()
```

```
  ## Calculate Width - width of the interval
  wid[i] <- (my.preds[,'upr'] - my.preds[,'lwr']) %>% mean()


  ## Bias
  bias[i] <- mean(my.preds[,'Prediction']-test.set[['Final']])

  ## Calculate Coverage - num of datapoints within interval (mean)
  cvg[i] <- ((test.set[['Final']] > my.preds[,'lwr']) & (test.set[['Final']] < my.preds[,'upr'])) %>% me

}

## CV Results
# RPMSE
hist(rpmse, main="RPMSE Histogram", xlab="RPMSE")
mean(rpmse) #rpmse is how off you are on average

#standard deviation of pedary = 10996.17
mean(cvg)
mean(bias)

## use for q 3
(var(ped$Final) - mean(rpmse)^2 ) / var(ped$Final)  # 89.5% of overall variance reduction
# with semester, 79%

## can we use R^2? noooooo
# num <- sum((ped$Final - my.preds.df$Prediction)^2)
# denom <-  sum((ped$Final - mean(ped$Final))^2)
# 1-num/denom

cor(my.preds.df$Prediction, ped$Final) ## use for q3
# with semester included: 86.9%

# Width histogram
hist(wid, main="Width Histogram", xlab="Width")
mean(wid)

# CV Prediction Plots
dataPreds <- predictgls(glsobj=ped.gls, level=0.95, newdframe=ped)
ggplot() +
  geom_point(data=dataPreds,
             mapping=aes(x=Exam3, y=Final)) + #Scatterplot
  geom_line(data=dataPreds,
            mapping=aes(x=Exam3, y=Prediction)) + #Prediction Line
  geom_line(data=dataPreds,
            mapping=aes(x=Exam3, y=lwr),
            color="red", linetype="dashed") + #lwr bound
  geom_line(data=dataPreds,
            mapping=aes(x=Exam3, y=upr),
            color="red", linetype="dashed") #Upper bound

ggplot() +
  geom_point(data=dataPreds,
```

```
             mapping=aes(x=Exam2, y=Final)) + #Scatterplot
  geom_line(data=dataPreds,
            mapping=aes(x=Exam2, y=Prediction)) + #Prediction Line
  geom_line(data=dataPreds,
            mapping=aes(x=Exam2, y=lwr),
            color="red", linetype="dashed") + #lwr bound
  geom_line(data=dataPreds,
            mapping=aes(x=Exam2, y=upr),
            color="red", linetype="dashed") #Upper bound

# In sample rmse
final_rmse <- sqrt(mean((dataPreds$Prediction - ped$Final)^2))
final_rmse  # .3557

# GLS Stuff
coef(ped.gls)
summary(ped.gls)
summary(ped.gls)$tTable # t table
summary(ped.gls)$sigma

# GLHT
quiz <- matrix(c(0, #intercept
              0,0,0,0,0,0,0,0,0, # semesters baseline 1
              0, # n students
              0,0,0, #exams
              0, #hw
              1#quiz
              ), nrow=1)
summary(multcomp::glht(ped.gls, linfct=quiz, rhs=0))

hw <- matrix(c(0, #intercept
              0,0,0,0,0,0,0,0,0,0, # semesters baseline 1
              0, # n students
              0,0,0, #exams
              1, #hw
              0#quiz
              ),nrow=1)
summary(multcomp::glht(ped.gls, linfct=hw, rhs=0))

semester <- matrix(c(0,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0), nrow=1)
semesterGLHT <- summary(multcomp::glht(ped.gls, linfct=semester, rhs=0))

# Confidence intervals
confint(ped.gls, level = 0.95)
```