Project 1: Pedagogy

Naomi Zubeldia and Cole Bowen

2023-02-17

Section 0: Exective Summary

Learning in schools and colleges is extremely important and the task at hand for teachers and professors is to find class activities that create the biggest chance for improved learning during the course of the class. This analysis is over data collected from an introductory statistics class over 5 years. We want to analyze this data in order to find what activities lead to improved student learning. In doing so, we found that Exams and Homework lead to improved student learning while Quizzes do not. Out of the activities, we found to be beneficial, we found that Exam 3 was the most beneficial to improved learning. We concluded that the department should try and make a bigger focus on Exams and Homework while limiting the amount of Quizzes given throughout the class because they do not encourage growth as much as Exams and Homework does.

Section 1: Introduction and Problem Background

Learning is important in today's world to help get a better job and make more money. Getting a college degree is essential for this goal. Often, to effectively learn in college, it is dependent on the type of learning activities in the class. The problem is that there are many activities being used that are not effective, which can inhibit higher learning. These ineffective activities should be identified and discarded to promote the best learning.

The main goal of this analysis is to find the best activities for improved learning. Once those are identified, we would like to know out of those better activities, which are the ones with the strongest influence on learning and large are those influences. It is also important to know how well the activities explain student learning. Lastly, we want to understand if there are certain semesters that the learning was either better or worse on average.

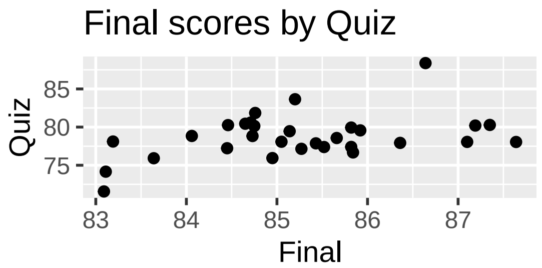
To accomplish these goals, we will be observing data gathered over 5 academic years from multiple sections of Statistics 121 students. The variable we will be observing to identify the better activities is the students' average final scores. The variables that we are comparing to see, which are the most influential on final scores, are semester, number of students who completed the course, avg. score of exam 1, avg. score exam 2, avg. score of exam 3, avg. score of the homework assignments, and the avg. score of class quizzes. To get a quick summary, we included the summary statistics of the data in Table 1. below.

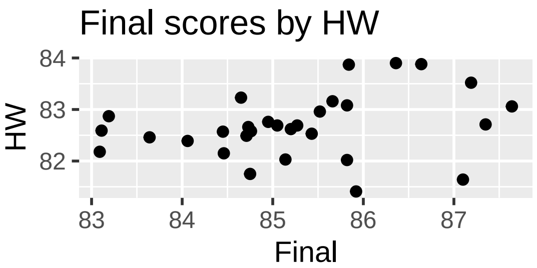
Table 1: Summary Table

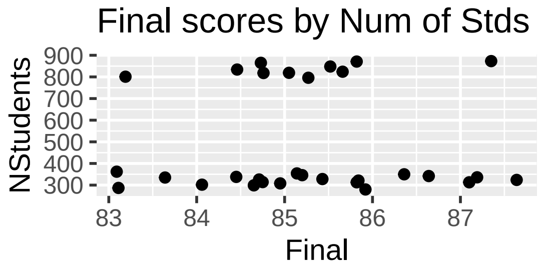
Semester	NStudents	Exam1	Exam2	Exam3	HW	Quiz	Final
Min. : 1.0	Min. :280.0	Min. :82.98	Min. :79.35	Min. :77.00	Min. :81.41	Min. :71.57	Min. :83.09
1st Qu.: 3.0	1st Qu.:315.8	1st Qu.:84.31	1st Qu.:81.34	1st Qu.:80.61	1st Qu.:82.41	1st Qu.:77.39	1st Qu.:84.67
Median : 5.5	Median :340.0	Median :85.72	Median :82.03	Median :81.77	Median :82.64	Median :78.34	Median :85.17
Mean : 5.5	Mean :494.2	Mean :85.73	Mean :81.90	Mean :81.73	Mean :82.68	Mean :78.75	Mean :85.25
3rd Qu.: 8.0	3rd Qu.:813.8	3rd Qu.:86.92	3rd Qu.:82.56	3rd Qu.:82.60	3rd Qu.:83.03	3rd Qu.:80.19	3rd Qu.:85.83
Max. :10.0	Max. :873.0	Max. :88.55	Max. :85.00	Max. :87.60	Max. :83.90	Max. :88.38	Max. :87.64

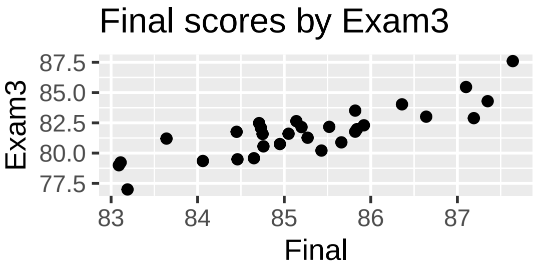
It appears in the averages for the exams, homework, and quizzes that the ranges are very close at widths around ten starting in the high 70s to low 80s and ending in the mid to high 80s, which seems like all the students performed similarly across all assignments. However, the variable of number of students completing the course has a huge range of 280 to 873, which shows that maybe each semester is not the same and there might be an influence there. For the final variable, it seems the range of the average scores is very close with a range width of only 4.5 difference.

In the figures below, we created scatter plots to see the relationship with the final scores and different variables and to see how the data behaves. The only variable that has a sort of linear relationship with final scores is exam 3 in the Final scores by Exam3 table, which is something interesting to look further into. The plot of Final score by Num of Students has the most unusual graph where Nstudents seems to either be below 400 or above 800 students over all the scores. Overall, most of the graphs exhibit a spread that is unusual where it starts close and spreads out in a cone-sort of shape.

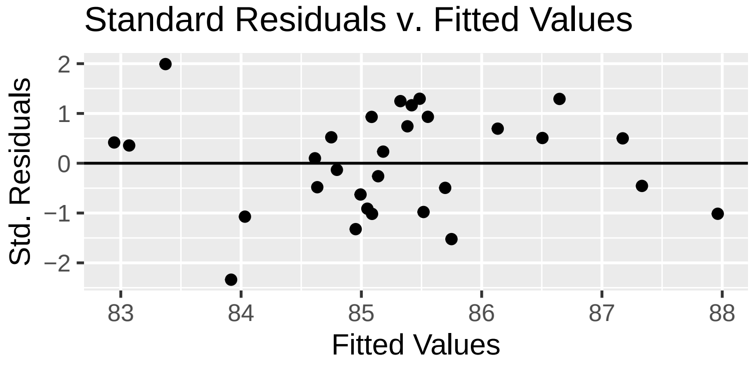

Final scores by Quiz


Final scores by HW


Final scores by Num of Stds


Final scores by Exam3

The cone spreading behavior of the data may be a sign of unequal variance, heteroskedasticity, in the data, which would violate the model assumption of equal variance. To check if the issue is really present, we applied the typical model of multiple simple regression to the data to check the variance assumption, which is in the Standard Residuals vs Fitted Values plot below. It looks the distribution of the fitted values and residuals is not spread very evenly showing that the assumption is not met.


Standard Residuals v. Fitted Values

If this equal variance assumption not being met is ignored, we cannot tell how far our estimates are truly off from the true values due to the standard errors being wrong. We are at risk of giving out false predictions due to the model not meeting this assumption as well. Recognizing the heteroskedasticity not only also allows for better predictions but less biased ones as well.

To combat this problem of the unequal variance in the typical linear model, we will be using a heteroskedastic model. It follows the same distribution and mean as the normal model, but the variance is a bit different, which helps to combat the unequal variance problem previously mentioned. This variance, which uses a diagnol matrix, has parameters that are estimated from the data. This way of estimating the variances helps take into account that each variance per semester, test, assignment, etc is different for each final score avg, and thus meeting the equal variance assumption now.

Section 2. Statistical Model

We determined a normal homoskedastic linear model is not right for this data and we need to use the heteroskedastic model to account for unequal variance. The model is detailed below:

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

The variables and parameters in our model are explained below.

y: is the vector of our response variable which is average score on the final exam. It represents the 30 final exam scores of our subjects.

$\mathbf{X}$: is the design matrix, or matrix beginning with ones and containing the explanatory variables of Number of Students, Exams 1,2,3 scores, Homework scores, and Quiz scores.

$\boldsymbol{\beta}$: is the vector containing all of the coefficients for the explanatory variables.

σ^2: is the variance of the residuals in our model.

$\mathbf{D}$: is the diagonal matrix of weights that changes the variance around our predicted regression line.

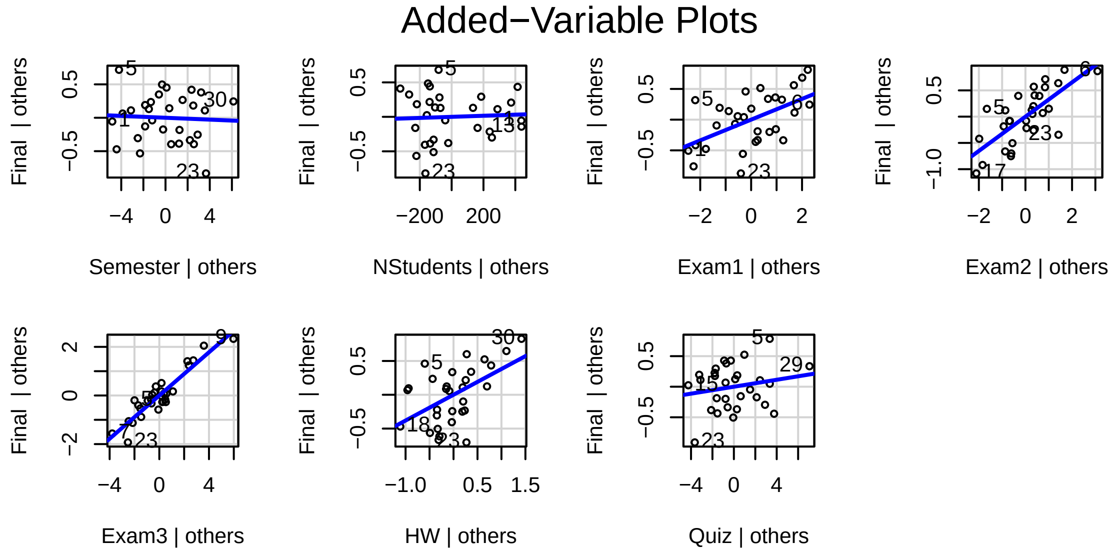
θ: is the coefficient that represents how the variance of the observations changes.

Now that we have determined the best model to use, we will check with the LINE assumptions to check if it is truly valid. We hope that our assumptions are met so we can move forward in analysis.

Section 3. Model Validation

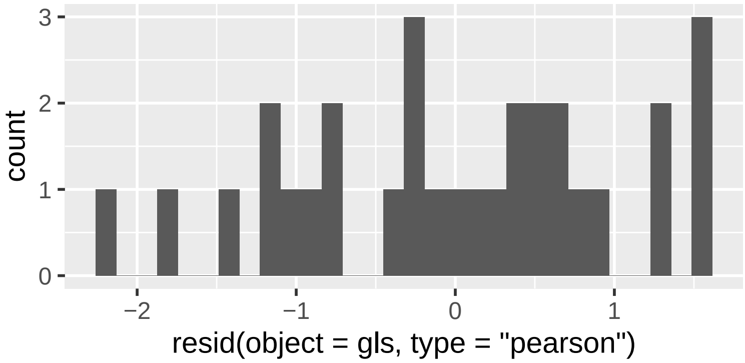
To make sure the heteroskedastic model is valid, we are checking the LINE assumptions of Linearity, Independence, Normality, and Equal Variance below using various graphs and tests.

The linearity assumption is met because there are no extreme curves in the Added-Variable plots below.

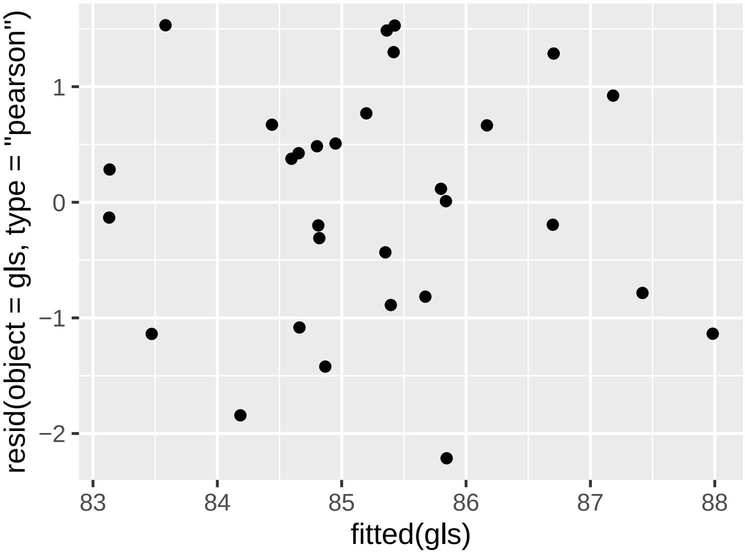


When checking the Independence assumption, we can say that the assumption is met because each observation is not reliant on each other and nothing in the data indicates that they are.

Normality was first check using the histogram below, which the plot was difficult to tell if there is normality. We then checked using the KS test since the histogram was not clear. We determine that normality is met because the p-value came back large, so we fail to reject the null hypothesis and assume normality.



Equal variance assumption is met because the plot looks to have a very steady equal variance across it and the BP test came back with a high p-value, so we can fail to reject the null hypothesis and assume equal variance.



After checking the LINE assumptions, we also checked the fit of the model on the data through maximum likelihood estimates. The beta estimates are -21.2197996, 0.1843731, 0.3037413, 0.4436652, 0.3403041, 0.0164838, 0.2985546, 0.1177107, 0.1936856, 0.1554749, 0.1250016, 0.091697, -0.3856482, -0.0744609, 0.4188662. The last estimate we examined was sigma, which is 5.4149401. These estimates show that the model fits the data well since it was made using the 'ML' parameter of maximum likelihood estimate.

Lastly, we check the ability of our model in making predictions. We calculated the average RPMSE from a Monte Carlo cross validation of the class data to be 0.2857626, which is the average distance of our predictions to the actual values. We compared the RPMSE to the standard deviation of the average Final scores, which is 1.210739. The RPMSE is smaller than the standard deviation, which shows that using the model to predict is better than using the mean of the response to. The average prediction interval width measured as 1.561705. To check the validity of it, I compared it to the range of the average Final scores, which is 4.55 and it was smaller indicating that the prediction is good.

Section 4. Analysis Results.

From the analysis, we were able to determine the best class activities that had a positive association on learning. The most influential class activities were all three exams. The HW did have a slightly significant influence on the Final scores with a pvalue of .0485, as well.

To understand how much of an impact each activity has on student learning, we observed the change in the r squared value of the model when we removed the positive associated ones. We found that the three exams did have a significant impact and the hw slightly. We, then, used confidence intervals to quantify that impact. For each activity, we listed the interpretations below:

Interpretation of Exam1: We are 95% confident that, holding all else constant, as average score of Exam 1 increases by one, the average Final score increases between 0.06567 and 0.30308 points.

Interpretation of Exam2: We are 95% confident that, holding all else constant, as average score of Exam 2 increases by one, the average Final score increases between 0.15260 and 0.45489 points.

Interpretation of Exam3: We are 95% confident that, holding all else constant, as average score of Exam 3 increases by one, the average Final score increases between 0.37405 and 0.51328 points.

Interpretation of HW: We are 95% confident that, holding all else constant, as average score of Homework increases by one, the average Final score increases between 0.02975 and 0.65086 points.

To understand how well the class activities, with this model, currently explain learning, we found the r-squared value for the data set. The value we got was 0.942475, which means that about 94% of the variability observed in the Final scores is explained by the heteroskedastic model.

The semesters, or time, that students took the class also seemed to have an impact on student learning. The semesters that seemed to be better for student learning are semesters 10 and 2. The base semester we are comparing them against is semester 1. The average Final scores of students in these semesters are .41884, semester 10, and .29855, semester 2, greater than those in in Semester 1. The semester that were worse for student learning are 8 and 9. The average Final scores of students in these semesters are .3856, semester 8, and .0744, semester 9, less than those of students in semester 1.

Section 5: Conclusions

Overall, it seems that the class activities that are best for improving learning are the three exams. Homework is slightly significant, but it seems like it might depend on what type of homework is given, which this may need further investigating. These activities have different ranges of positive impact of student learning. Altogether, their ranges go from .06 to .65 of a point of change on the average final scores. The class activities in the model currently explain learning fairly well with 94% of the variability being interpreted. Lastly, certain semesters appeared to be better for student learning, like semester 10, and others were worse, which was semester 8. The department should look into having professors make less of an emphasis on Quizzes in their class curriculum. It did not have a significant effect on learning. They should stick to using exams to help improve learning. Homework also was influential, but not as much as the exams, which may be because certain types of it are not as effect. There should be more investigation on what homework influences learning, so that activity can have a higher positive impact on learning and students can have more activities to improve their learning.

Appendix

```r
#All libraries needed
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(fig.pos = "H", out.extra = "")
library(tidyverse)
library(kableExtra)
library(grid)
library(gridExtra)
library(GGally)
library(car)
library(MASS)
library(multcomp)
library(lmtest)
library(knitr)
library(nlme)
#Reading in data and summary table
score<- read.table('ClassAssessment.txt',header=T)
knitr::kable(summary(score), caption='Summary Table', format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options=c("HOLD_position","scale_down"))
#Observing relationships with Final score

ggplot(mapping=aes(x=Final,y=Quiz), data=score)+
  geom_point()+ggtitle('Final scores by Quiz')->p1
ggplot(mapping=aes(x=Final,y=HW), data=score)+
  geom_point()+ggtitle('Final scores by HW')->p2
ggplot(mapping=aes(x=Final,y=NStudents), data=score)+
  geom_point()+ggtitle('Final scores by Num of Stds')->p3
ggplot(mapping=aes(x=Final,y=Exam3), data=score)+
  geom_point()+ggtitle('Final scores by Exam3')->p4

grid.arrange(p1,p2,p3,p4, ncol=2)

#Checking variance on normal lm
basic.lm<- lm(Final~.,data=score)

fit.vals <- predict.lm(basic.lm)
residuals <- stdres(basic.lm)
ggplot()+
  geom_point(mapping = aes(x = fit.vals, y = residuals))+
  xlab("Fitted Values") + ylab("Std. Residuals") +
  ggtitle("Standard Residuals v. Fitted Values")+
  geom_hline(mapping = aes(yintercept=0))


#Creating our heteroskedastic model
score$Semester<-as.factor(score$Semester)


gls<-gls(Final~Exam1+Exam2+Exam3+HW+Quiz+Semester,data=score,
         weights=varFixed(value=~1/NStudents), method="ML")
summary(gls)
```

```r
# Checking linearity assumption
avPlots(basic.lm, layout=c(2,4))
# Checking Normality of Standardized Residuals
# Histogram of new standardized residuals
ggplot() +
  geom_histogram(mapping=aes(x=resid(object=gls, type="pearson")))
# KS Test
ks.test(resid(object = gls, type="pearson"), "pnorm")
# Check Equal Variance of the Standardized Residuals
# Scatter plot of the fitted values vs the new standardized residuals
ggplot() +
  geom_jitter(mapping = aes(x = fitted(gls),
                           y = resid(object=gls, type="pearson")))


#Fitting the Model
b<-gls$coefficients # Beta hats
s<-gls$sigma # Sigma
#Monte Carlo CV test

source("../STAT469/glstools-master/glstools-master/predictgls.R")
n.cv <- 25 #Number of CV studies to run
n.test <- 30 #Number of observations in a test set
rpmse <- rep(x=NA, times=n.cv)
bias <- rep(x=NA, times=n.cv)
wid <- rep(x=NA, times=n.cv)
cvg <- rep(x=NA, times=n.cv)
for(cv in 1:n.cv){
  ## Select test observations
  test.obs <- sample(x=1:nrow(score), size=n.test)

  ## Split into test and training sets
  test.set <- score[test.obs,]
  train.set <- score[-test.obs,]

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

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

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

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

  ## Calculate Coverage
  cvg[cv] <- ((test.set[['Final']] > my.preds[,'lwr']) &
                (test.set[['Final']] < my.preds[,'upr'])) %>% mean()
  ## Calculate Width
  wid[cv] <- (my.preds[,'upr'] - my.preds[,'lwr']) %>% mean()
}

## Reports mean bias, rpmse, coverage, and width
# Bias
mean(bias)
# RPMSE
mean(rpmse)
# Coverage
mean(cvg)
# Width
mean(wid)
#GLS model Confidence Intervals
round(confint(gls, level = .95),5)[11:14,]

#Checking impact of each of the class activities with r2

#quiz
glsq<-gls(Final~Exam1+Exam2+Exam3+HW+Semester,data=score,
         weights=varFixed(value=~1/NStudents), method="ML")

q <- cor(score$Final,predict(glsq))^2
q

#HW
glsh<-gls(Final~Exam1+Exam2+Exam3+Quiz+Semester,data=score,
         weights=varFixed(value=~1/NStudents), method="ML")

h <- cor(score$Final,predict(glsh))^2
h
#exam1
glse<-gls(Final~Quiz+Exam2+Exam3+HW+Semester,data=score,
         weights=varFixed(value=~1/NStudents), method="ML")

e1 <- cor(score$Final,predict(glse))^2
e1
#exam2
glsee<-gls(Final~Exam1+Quiz+Exam3+HW+Semester,data=score,
         weights=varFixed(value=~1/NStudents), method="ML")

e2 <- cor(score$Final,predict(glsee))^2
e2
#exam3
glseee<-gls(Final~Exam1+Exam2+Quiz+HW+Semester,data=score,
         weights=varFixed(value=~1/NStudents), method="ML")

e3 <- cor(score$Final,predict(glseee))^2
e3

#Getting the R2 to see how the model explains the data
R2 <- cor(score$Final,predict(gls))^2
R2
```