469 Midterm: Pedagogy

Joshua Christensen and Riley Millar

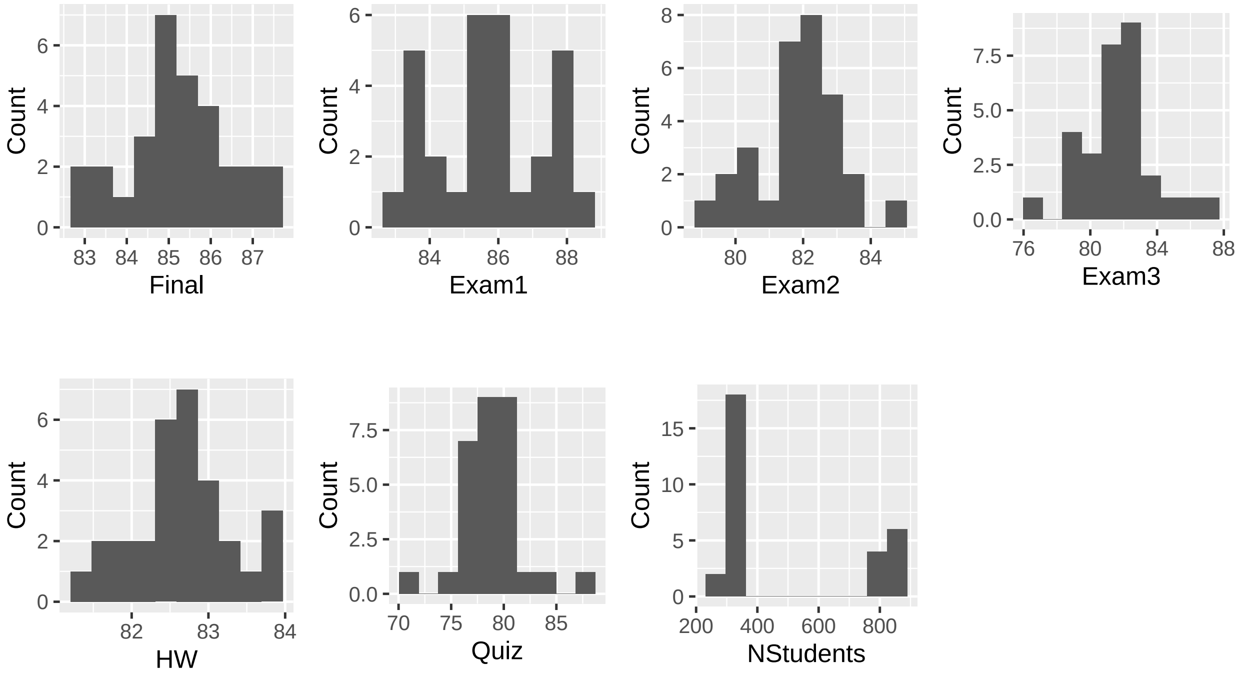
2/25/2021

Section 0: Executive Summary

In an effort to evaluate the effectiveness of learning activities in the STAT 121 course curriculum at BYU, we have run several analyses on a dataset provided by the BYU Statistics Department. This dataset contains the number of students who completed the course during ten previous semesters, and the average percentage scores (based on each section of the class) for the individual learning activities and final exam. From our analysis, we have concluded that exams 1-3 and the homework assignments are associated with improved final exam scores for the class, while the quizzes are not. We recommend that the Statistics Department should alter the STAT 121 course curriculum to focus more on the learning activities that are associated with improved learning, and less on the inconsequential quizzes.

Section 1: Introduction and Problem Background

STAT 121 is the introductory statistics class taught at Brigham Young University. Because there are so many students who take this class each semester, there are multiple options for professors and time sections. The class material and assignments have been standardized across all sections and professors. In an effort to determine if there are any learning activities (e.g., exams, quizzes, or homeworks) that are not associated with improved learning and should be removed from the standardized curriculum, we have run some statistical analyses on a dataset provided by the BYU Statistics Department. This dataset consists of the performance metrics (i.e., the average percent scores on exams 1-3, homeworks, quizzes, and the final exam) of STAT 121 students from the last 10 semesters. For each of our learning activities, the associated values represent average percentage scores. The same is true for our response variable, which is the average percentage score on the final exam. Semester is a number representing one of the ten semesters from the last five academic years, and NStudents represents the number of students who completed the course.



One potential issue with the data is the variability in the number of students who attended each section. The data shows that there was one section each semester that had twice as many students as other sections. This is a potential problem for our analysis because we are dealing with means (average scores), and the larger sections will therefore have slightly less variability than the smaller sections. If we ignored this issue, our standard error calculations would be incorrect. This would in turn cause problems in our confidence intervals and prediction intervals. We account for this issue by using the inverse of the number of students that finished the course each semester as variance weights. By properly accounting for the heteroskedasticity in our data we will improve our ability to predict. Specifically, we will be able to correctly scale the prediction bands according to the number of students that finish the class in a given semester.

Due to the effect of the number of students on the variability of the mean response, we will fit a heteroskedastic linear regression model and determine which of the learning activities are significant in predicting final exam score. After checking for significance and dropping any insignificant variables, we will check our model assumptions again. If our model passes these assumptions, we will evaluate model fit and then perform a Monte Carlo cross validation to determine its prediction ability. This will allow us to provide an accurate representation of how useful our model is in answering our research questions.

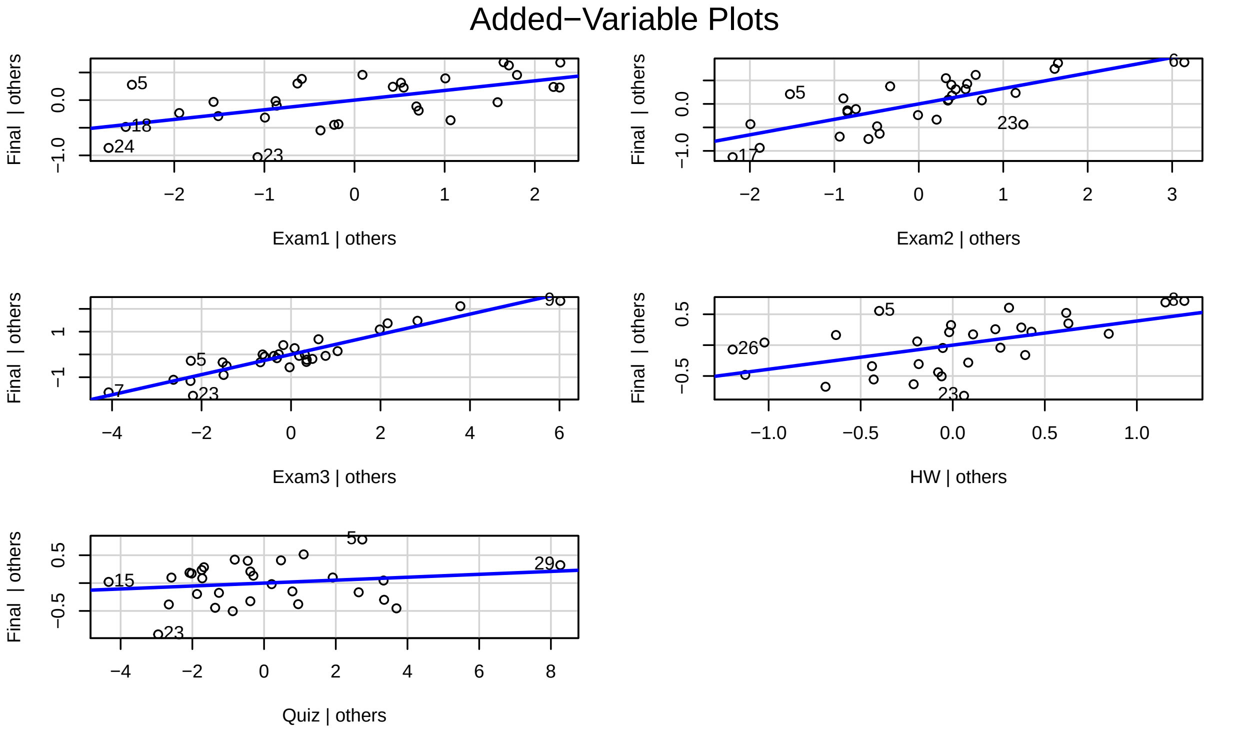
Section 2: Statistical Model

We will use a heteroskedastic linear regression model to analyze the teaching activities data. We define the model for our data as $\boldsymbol{y} \sim MVN(\boldsymbol{X\beta}, \sigma^2\boldsymbol{D})$. In our model $\boldsymbol{y}$ represents the average final exam score (in percent) for students in a given section for a given semester. We have selected the average final score as our response variable to represent student learning. $\boldsymbol{X}$ represents the design matrix which includes our explanatory variables as well as a column of ones for the intercept. We included all of the teaching activities in our final model as part of the $\boldsymbol{X}$ matrix, but did not include semester or number of students after finding that their effect on $\boldsymbol{y}$ was not significant. Our parameters for this model are $\boldsymbol{\beta}$ and σ^2. $\boldsymbol{\beta}$ is a vector of coefficients that quantify the relationship between each respective explanatory variable and the response variable, in addition to setting the intercept. σ^2 represents the variance of our observations about the line prior to accounting for the weights in our $\boldsymbol{D}$ matrix. The $\boldsymbol{D}$ matrix is a diagonal matrix with weights along the diagonal that correspond to each observation. The variance at each observation is defined as $d_{ii}\sigma^2$. We used fixed weights in our model that corresponded to the number of students in each section. We justify this choice based on the theoretical variance of any mean, σ^2/n. Therefore in our $\boldsymbol{D}$ matrix each d_{ii} corresponds to $1/n_i$.

Our model depends on four assumptions. The first is that all quantitative variables have a linear relationship with the response variable. The second is that each response is independent of the other responses. The third is that the residuals are normally distributed with mean 0 and variance $\sigma^2\boldsymbol{D}$. The last assumption is that our $\boldsymbol{D}$ matrix weights combined with σ^2 accurately represents the variability of the data. In other words, while our response is not assumed to have constant variance, we would like our weights to account for the changes in variance so that the underlying σ^2 is still constant.

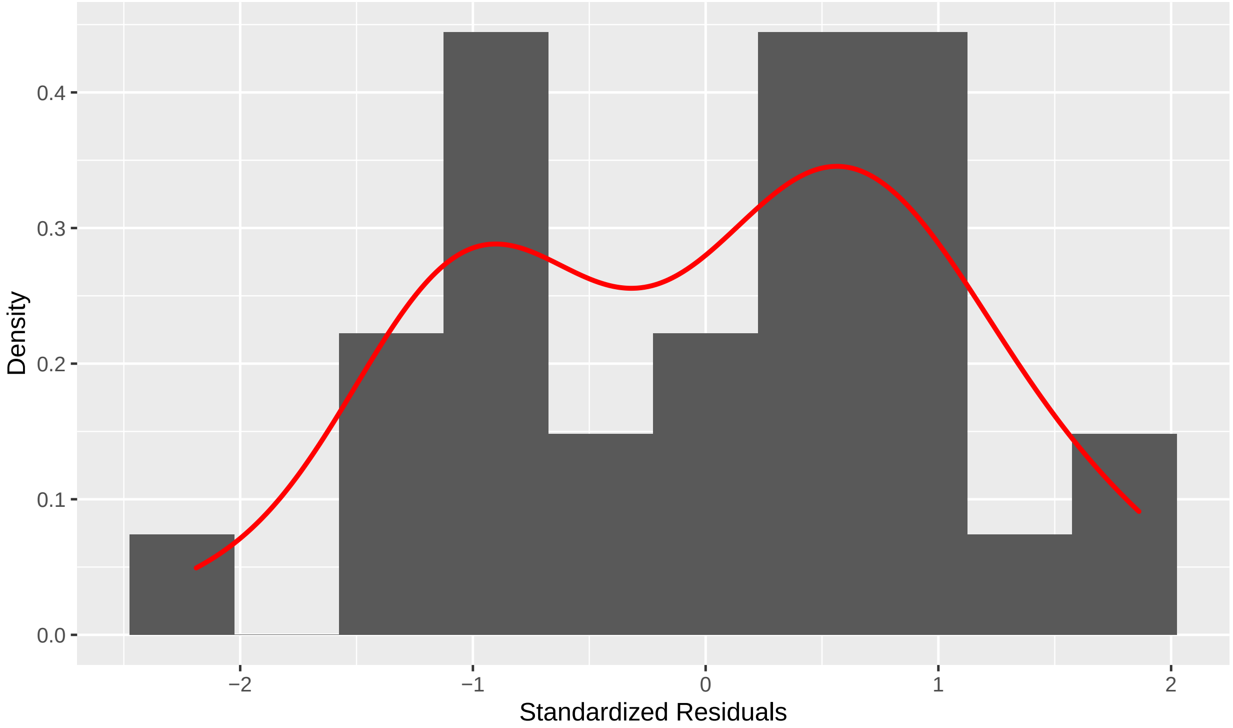
Section 3: Model Validation

We tested the linearity assumption using added-variable plots, which regress both the response and the explanatory variable being tested against all other variables in the model and then display the resulting fitted values plotted against one another. This allows us to evaluate the relationship between the explanatory variable being tested and the response, while accounting for the effects of all other variables in the model. The added-variable plots displayed below all show approximately linear relationships.

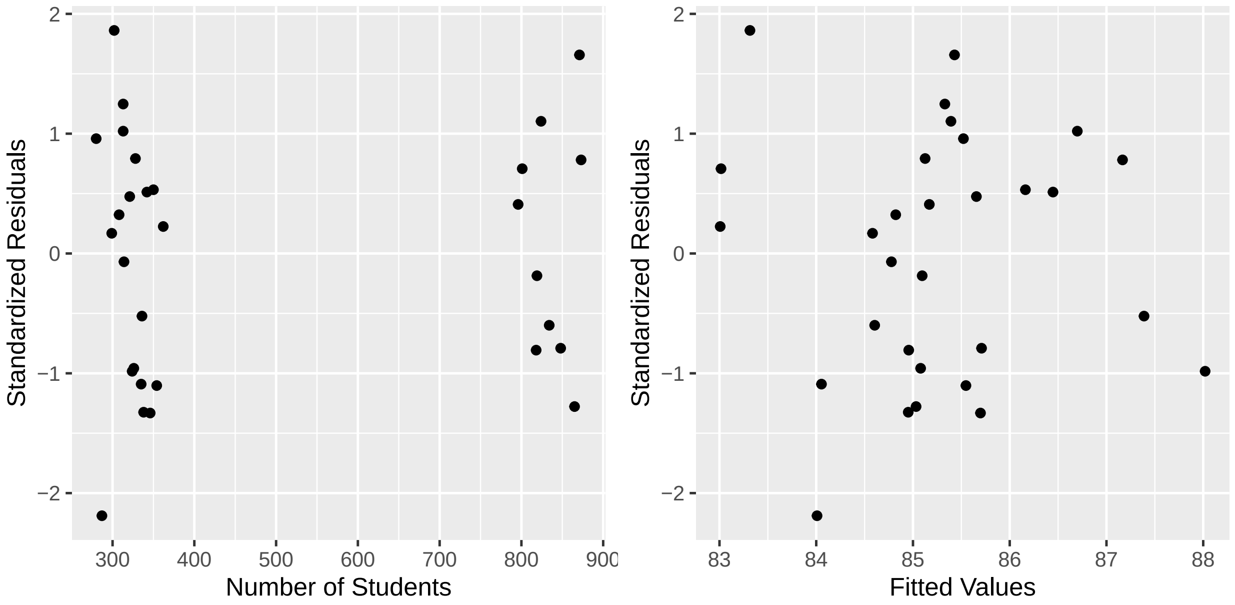


We believe the assumption of independence to be reasonable due to the rotation of students from semester to semester. Professors and teaching assistants could impact performance of different sections, but given the standardized curriculum of introductory courses, we do not believe this correlation is sufficient to invalidate our assumption. Students who retake the class could also create correlation between semesters, but retakes are rare relative to the total enrollment and therefore would have minimal impact on the average responses for the entire semester.

We checked the assumption of normally distributed residuals by creating a histogram of the standardized residuals. While the histogram is not perfectly normal in shape it is concentrated in the center and tapers at the edges. Additionally, there are no evident outliers. We also performed a Kolmogorov-Smirnov hypothesis test for normality. Given our p-value of 0.7475 we failed to reject the null hypothesis, therefore concluding that the standardized residuals came from a normal distribution.



The final assumption (that our fixed weights account for heteroskedasticity of the response) was checked by plotting the standardized residuals against the number of students that finished the course each semester. We also plotted the standardized residuals against the fitted values from our model. The scatter plot of standardized residuals against number of students shows that our accounting for the heteroskedasticity of the mean responses has left us with similar variance between the large class sections and the smaller sections. Additionally the scatter plot of standardized residuals against the fitted values shows constant variance.



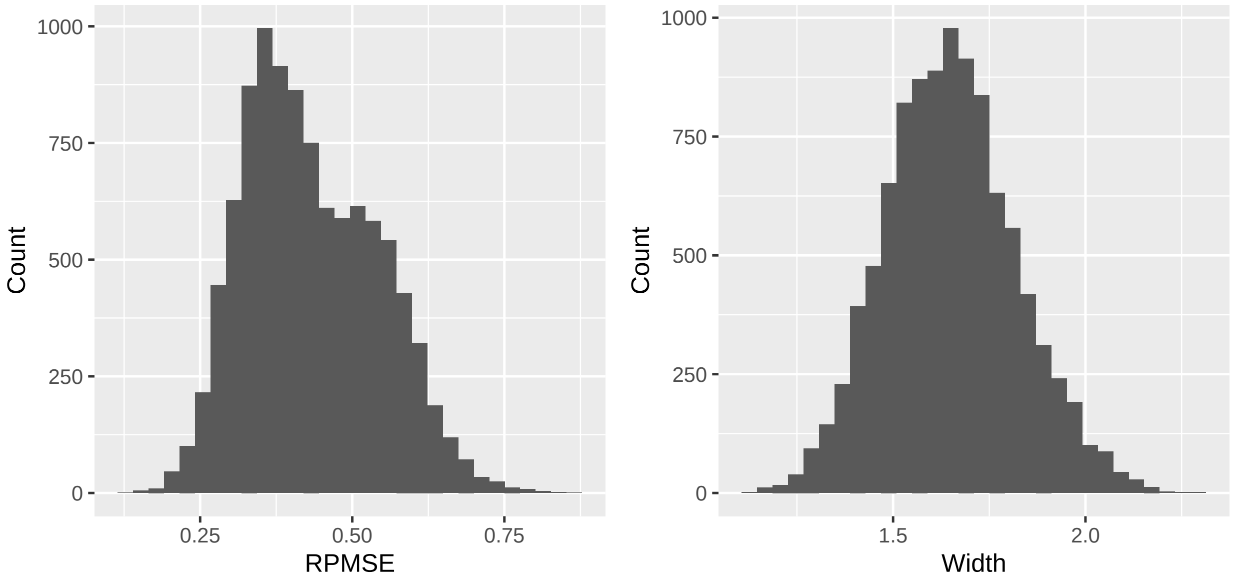
We calculated our coefficient of determination, R^2 as 0.91, indicating that 91% of the variation in mean final scores can be explained using the average teaching activities scores in our model. The root mean square error is 0.356. This is very small relative to our data, which confirms that our model fits the data well.

While we do not need predictions in order to evaluate the effectiveness of learning activities of students, we provide prediction diagnostics so that predictions can be used to evaluate future curriculum changes. For example, if one of the learning activities is modified in a given semester and the average final score in that semester is well outside of our prediction interval we would want to evaluate and possibly adjust our new curriculum.

We evaluate our predictions using root prediction mean square error (RPMSE), coverage, prediction interval width and bias. We evaluate these metrics using Monte Carlo cross-validation. The average of each metric is reported in the table below along with the histograms of RPMSE and interval width.

Table 1: Prediction Diagnostics

	Means
Bias	-0.001
RPMSE	0.432
Coverage	0.939
Width	1.650



Given our negligible bias and relatively low RPMSE we find our predictions to be very accurate. Our coverage is also very close to the expected 0.95. Our prediction interval width indicates that we will generally construct intervals with a margin of error smaller than 1%. We believe this to be sufficient accuracy to track unexpected movements caused by curriculum changes, as suggested above.

Section 4: Analysis Results

We calculated $\boldsymbol{\beta}$ coefficients to determine which learning activities have a significant impact on student learning.

Table 2: Coefficient Table

	Estimate	Lower	Upper
(Intercept)	-29.077	-54.703	-3.451
Exam1	0.177	0.083	0.271
Exam2	0.341	0.222	0.460
Exam3	0.450	0.380	0.520
HW	0.403	0.148	0.658
Quiz	0.014	-0.046	0.074

We are 95% confident that, holding all else constant, for each additional one percentage increase in average score on exam 1, average percentage score on the final exam increases between 0.083 and 0.271 percentage points on average. To keep our report concise, we will not include the formal interpretations of the other intervals. However, this same format applies to the other explanatory variables and confidence intervals in the above table. Based on the intervals in the table above we conclude that the three exams and the homeworks are associated with improved learning. Note that the confidence interval for β_{quiz} includes 0 which indicates that quizzes do not have an effect on student learning. This is confirmed by the t-test for significance of β_{quiz} which produces a test statistic of $t = 0.467$ and a p-value of 0.645. We therefore fail to reject the null hypothesis and conclude that quizzes have no impact on student learning.

Based on the R^2 (0.91) and RMSE (0.356) reported above we believe that using all learning activities, our model explains student learning well.

We performed an ANOVA test comparing a model that included semester with our final model that included only learning activities. We obtained a p-value of 0.091. Based on this p-value we fail to reject the null hypothesis and conclude that there is no signficiant difference between the models. This indicates that the semester variable does not have a significant impact on final score. In other words, we did not identify any semesters that were better or worse in terms of student learning.

Section 5: Conclusions

We found that all learning activities are associated with student learning except for quizzes. In other words we believe that the exams and the homework assist in student learning, but the quizzes have no effect. Exam 3 is the most closely related followed by exam 2, homework and exam 1. By using the scores on these learning outcomes we were able to construct a model that very accurately portrays the variability associated with student learning, as measured by the final exam score. We did not find that any one semester in our dataset impacted student learning positively or negatively. This bodes well since it indicates consistency in curriculum, teaching, and resources between semesters.

We recommend that the department review the structure and content of the quizzes since they were not associated with student learning. Once a suggested change has been tested we recommend that the resulting final exam scores be compared against a prediction from this model to determine if a change took place.

Appendix A: Analysis Code

```
# Include packages
library(ggplot2)
library(car)
library(nlme)
library(MASS)
library(lmtest)
library(magrittr)
library(knitr)
library(gridExtra)

# Define function predictgls
source("predictgls.R")

# Set seed
set.seed(10131008)

# Read in the data
pedagogy <- read.table("https://mheaton.byu.edu/docs/files/Stat469/Topics/1%20-%20Independence/3%20-%20I
    header = TRUE)



# Convert semester to factor
pedagogy$Semester <- as.factor(pedagogy$Semester)

# Exploratory histograms
p1 <- ggplot(pedagogy, aes(x = Final)) + geom_histogram(bins = 10) +
    labs(y = "Count") + theme(aspect.ratio = 1)
p2 <- ggplot(pedagogy, aes(x = Exam1)) + geom_histogram(bins = 10) +
    labs(y = "Count") + theme(aspect.ratio = 1)
p3 <- ggplot(pedagogy, aes(x = Exam2)) + geom_histogram(bins = 10) +
    labs(y = "Count") + theme(aspect.ratio = 1)
p4 <- ggplot(pedagogy, aes(x = Exam3)) + geom_histogram(bins = 10) +
    labs(y = "Count") + theme(aspect.ratio = 1)
p5 <- ggplot(pedagogy, aes(x = HW)) + geom_histogram(bins = 10) +
    labs(y = "Count") + theme(aspect.ratio = 1)
p6 <- ggplot(pedagogy, aes(x = Quiz)) + geom_histogram(bins = 10) +
    labs(y = "Count") + theme(aspect.ratio = 1)
p7 <- ggplot(pedagogy, aes(x = NStudents)) + geom_histogram(bins = 10) +
    labs(y = "Count") + theme(aspect.ratio = 1)

# Combine histograms into a grid
```

```r
grid.arrange(p1, p2, p3, p4, p5, p6, p7, ncol = 4, nrow = 2)

# Fit linear model to generate av plots Can be homoskedastic since
# variance structure doesn't affect the linearity assumption
pedagogy.lm <- lm(Final ~ Exam1 + Exam2 + Exam3 + HW + Quiz, data = pedagogy)

# Generate av plots
avPlots(pedagogy.lm)

# Fit the heteroskedastic linear model
ped4.gls <- gls(model = Final ~ Exam1 + Exam2 + Exam3 + HW + Quiz,
    data = pedagogy, weights = varFixed(~1/NStudents), method = "ML")

# Standardize the residuals
st.res <- resid(object = ped4.gls, type = "pearson")

# Run ks-test for normality
ks <- ks.test(resid(object = ped4.gls, type = "pearson"), "pnorm")

# Histogram and density plot for standardized residuals
ggplot(mapping = aes(x = st.res)) + geom_histogram(aes(y = stat(density)),
    bins = 10) + xlab("Standardized Residuals") + ylab("Density") +
    geom_density(color = "red", lwd = 1)

# Fitted values
fit <- ped4.gls$fitted

# Scatterplot of the NStudents vs. standardized residuals
p1 <- ggplot(mapping = aes(pedagogy$NStudents, st.res)) + geom_point() +
    xlab("Number of Students") + ylab("Standardized Residuals") +
    theme(aspect.ratio = 1)

# Scatterplot of the fitted values vs. standardized residuals
p2 <- ggplot(mapping = aes(fit, st.res)) + geom_point() + xlab("Fitted Values") +
    ylab("Standardized Residuals") + theme(aspect.ratio = 1)

# Combine plots into a grid
grid.arrange(p1, p2, ncol = 2)

# Calculate R2
R2 <- 1 - sum((ped4.gls$fitted - pedagogy$Final)^2)/sum((pedagogy$Final -
    mean(pedagogy$Final))^2)
```

```
# Calculate the RMSE
ped.preds <- predictgls(ped4.gls)
rmse <- (pedagogy[["Final"]] - ped.preds[, "Prediction"])^2 %>% mean() %>%
    sqrt()

# Monte Carlo Cross-Validation
n.cv <- 10000  #Number of CV studies to run
n.test <- round(0.2 * nrow(pedagogy), 0)  #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)
n <- nrow(pedagogy)
for (cv in 1:n.cv) {
    ## Select test observations
    test.obs <- sample(x = 1:n, size = n.test)

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

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

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

    ## 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()
```

```r
}

# Table of cross-validation prediction diagnostics
kable(data.frame(Means = c(mean(bias), mean(rpmse), mean(cvg), mean(wid)),
    row.names = c("Bias", "RPMSE", "Coverage", "Width")), caption = "Prediction Diagnostics",
    digits = 3)

# RPMSE histogram
p1 <- ggplot(mapping = aes(rpmse)) + geom_histogram(bins = 30) + xlab("RPMSE") +
    ylab("Count") + theme(aspect.ratio = 1)

# Width of prediction intervals histogram
p2 <- ggplot(mapping = aes(wid)) + geom_histogram(bins = 30) + xlab("Width") +
    ylab("Count") + theme(aspect.ratio = 1)

# Combine plots into a grid
grid.arrange(p1, p2, ncol = 2)

# Store intervals as an object
ints <- intervals(ped4.gls)

# Reorganize columns and store as a dataframe
frame <- as.data.frame(ints$coef)[c(2, 1, 3)]

# Add desired column names
colnames(frame) <- c("Estimate", "Lower", "Upper")

# Display in kable
kable(frame, caption = "Coefficient Table", digits = 3)

# Fit model with semester
pedsem.gls <- gls(model = Final ~ Exam1 + Exam2 + Exam3 + HW + Quiz +
    Semester, data = pedagogy, weights = varFixed(~1/NStudents), method = "ML")

# Run ANOVA test
anosems <- anova(ped4.gls, pedsem.gls)
```

Appendix B: EDA and Model Creation Code

```r
# GG pairs was used to explore the data, but looks janky
library(GGally)
ggpairs(pedagogy)
```

```r
# This represents the process we went through to find our final
# model The first three models were NOT included in our analysis
ped.gls <- gls(model = Final ~ ., data = pedagogy, weights = varExp(form = ~.),
    method = "ML")
summary(ped.gls)$tTable

ped2.gls <- gls(model = Final ~ Exam1 + Exam2 + Exam3 + HW + Quiz,
    data = pedagogy, weights = varExp(form = ~Exam1 + Exam2 + Exam3 +
        HW + Quiz), method = "ML")
summary(ped2.gls)$tTable

ped3.gls <- gls(model = Final ~ ., data = pedagogy, weights = varFixed(~1/NStudents),
    method = "ML")
summary(ped3.gls)$tTable

ped4.gls <- gls(model = Final ~ Exam1 + Exam2 + Exam3 + HW + Quiz,
    data = pedagogy, weights = varFixed(~1/NStudents), method = "ML")
summary(ped4.gls)$tTable
```