Value of a College Education

Joshua Christensen

1/28/2021

Question 1

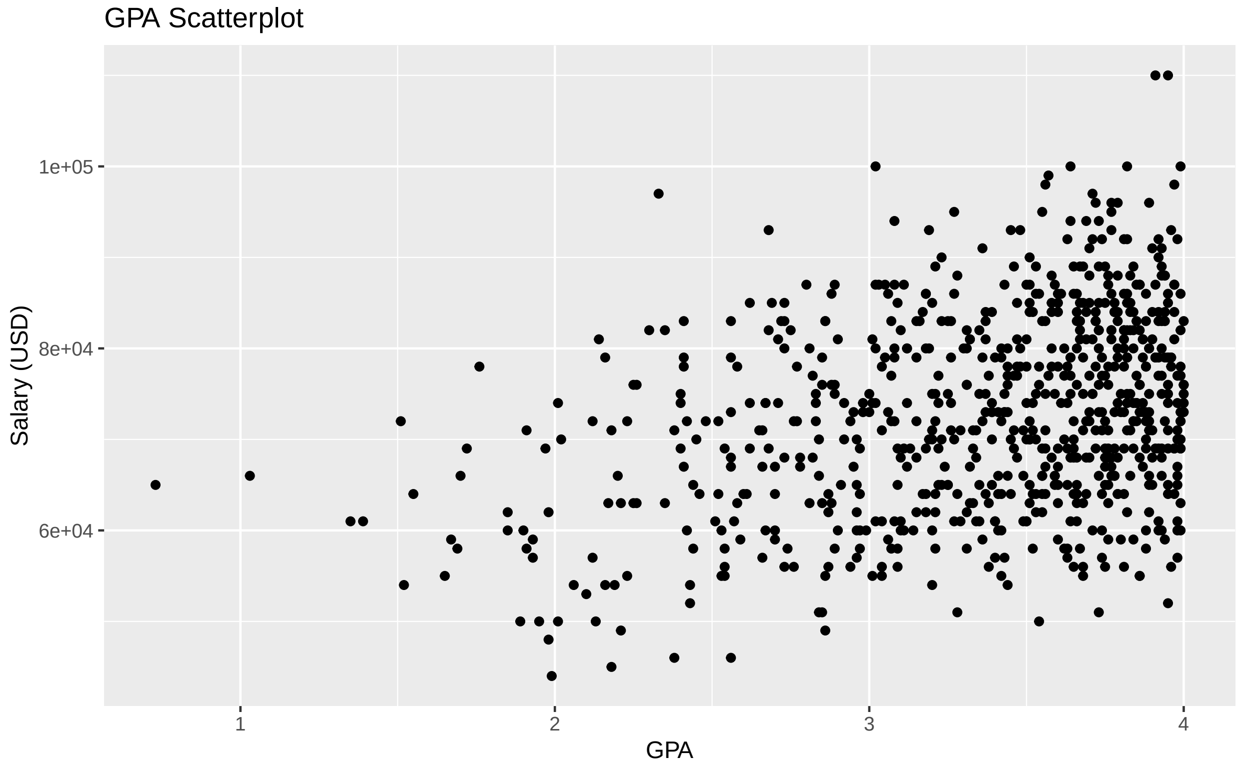


Table 1: Summary Statistics for GPA

	Mean	SD	Correlation
GPA	3.341	0.572	0.339

The scatterplot of GPA and salary shows a weak positive linear relationship. Note that the GPAs are more dense closer to 4.0 because of the upper limit.

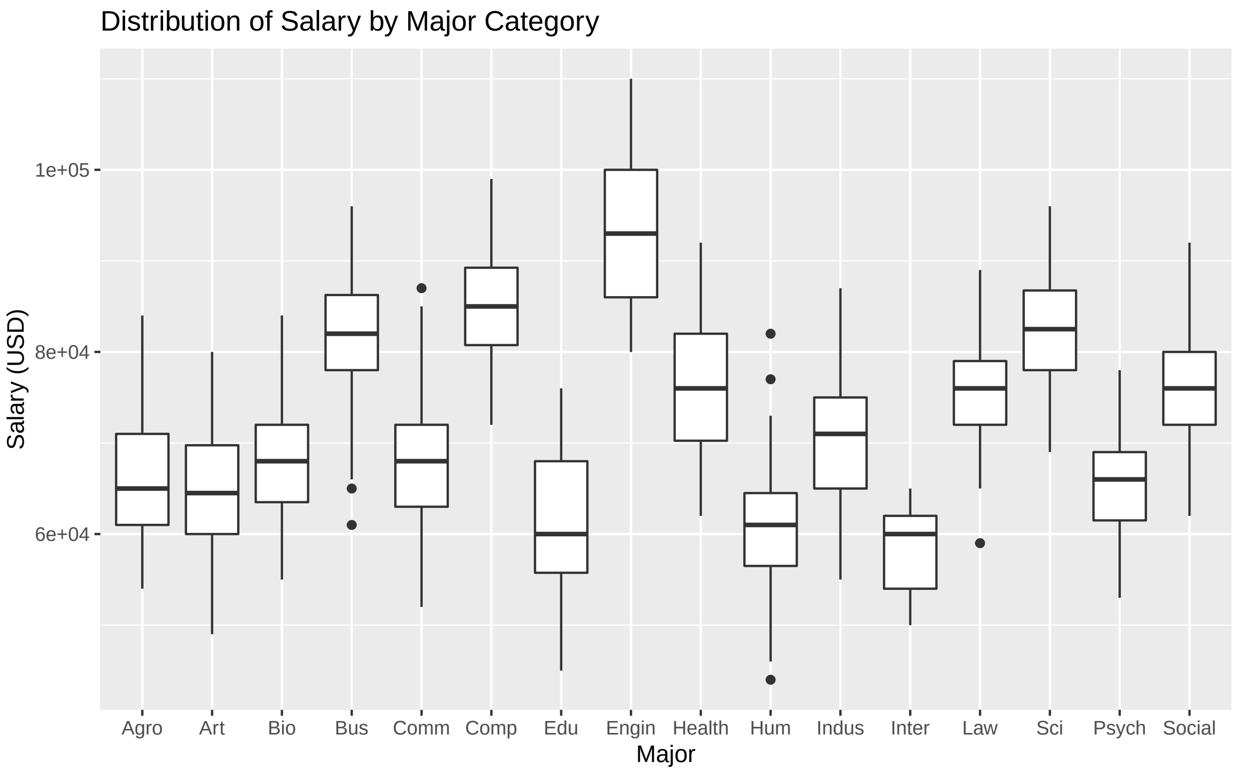


Table 2: Salary Statistics by Major Category

Major Category	Mean	SD
Agriculture & Natural Resources	66923.08	10111.557
Arts	64685.19	6954.906
Biology & Life Science	68263.16	7526.561
Business	81881.94	6454.246
Communications & Journalism	67674.16	7212.324
Computers & Mathematics	85666.67	7087.212
Education	61100.00	7231.522
Engineering	93000.00	8705.715
Health	75907.41	6879.745
Humanities & Liberal Arts	61348.84	7785.468
Industrial Arts & Consumer Services	70704.55	7245.141
Interdisciplinary	58230.77	5019.194
Law & Public Policy	75870.97	5970.717
Physical Sciences	83041.67	7404.342
Psychology & Social Work	65372.09	5223.679
Social Science	76202.90	6110.694

The side-by-side boxplots provide good evidence that there is a difference between different majors in terms of salary. A drastic example is that the engineering and education majors do not have any overlap at all in their salaries. The summary statistics appear to confirm the differences between groups.

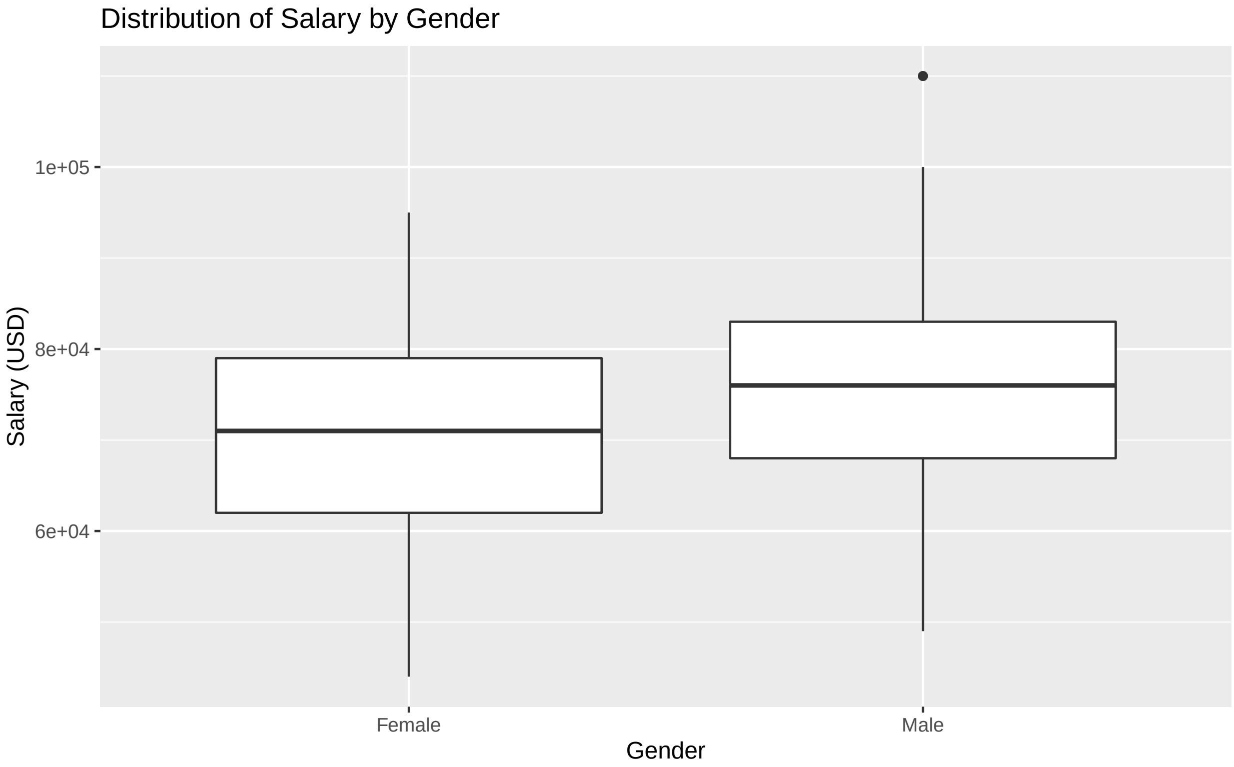
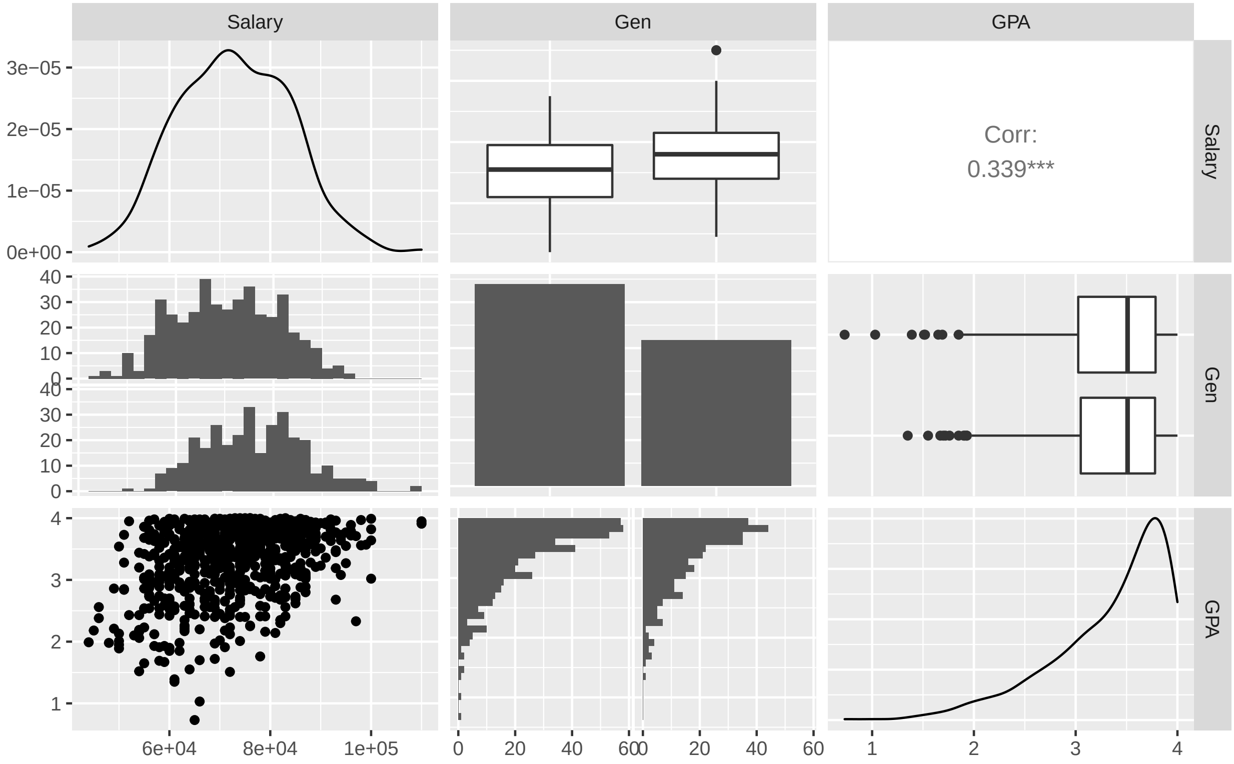


Table 3: Salary Statistics by Gender

Gender	Mean	SD
F	70469.25	10728.20
M	76236.59	10487.67

The boxplots by gender show that males tend to have higher salaries. Males have a higher minimum, maximum and median than females, though there is considerable overlap between the two boxplots.



The pairs plot confirms what the other plots on GPA, gender and salary indicated. Also note the distributions of each variable individually.

Question 2

$$\mathbf{Y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\sigma^2}\mathbf{I})$$

The parameters for this model are β and σ^2. β is the vector of coefficients for our explanatory variables. They can be thought of as partial slopes that relate to each individual explanatory variable. In our case these coefficients would indicate whether gender, major and GPA have an effect on salary and would help us characterize those effects. σ^2 describes the variance of the residuals, or the variability of points about the regression line. It helps us understand how tight our points are on the regression line and informs our ability to make predictions.

To understand the effect of major choice on salary we can look at the coefficients for the explanatory variables related to major. By looking at the direction and magnitude of the effects of the majors we can characterize their impact on salary. In order to look for gender discrimination we can perform a similar analysis on gender. If gender has a statistically significant effect, we would have good evidence that gender discrimination plays a part in salary determination.

Question 3

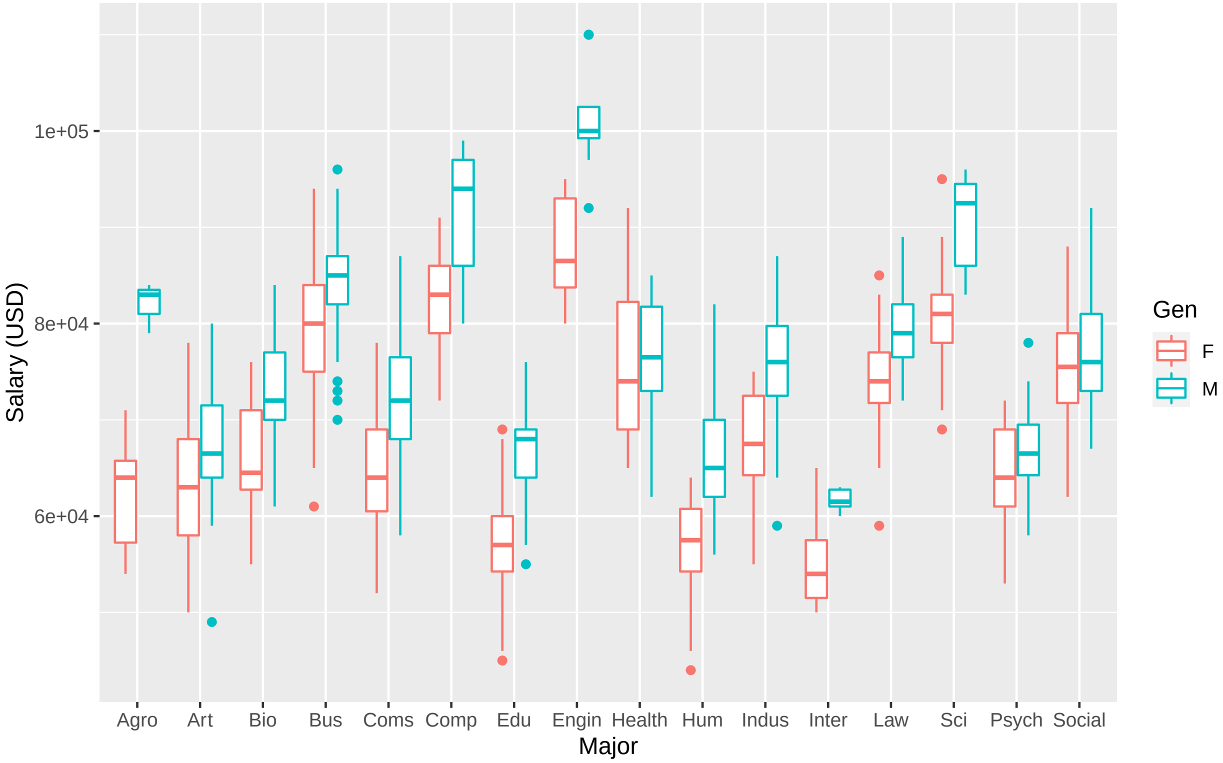
Table 4: Estimates of Model Coefficients

Coefficient	Estimate
(Intercept)	46672.99
MajorCategoryArts	-2551.64
MajorCategoryBiology & Life Science	769.13
MajorCategoryBusiness	14282.15
MajorCategoryCommunications & Journalism	114.60
MajorCategoryComputers & Mathematics	17936.91
MajorCategoryEducation	-5894.85
MajorCategoryEngineering	24406.23
MajorCategoryHealth	8670.16
MajorCategoryHumanities & Liberal Arts	-5972.59
MajorCategoryIndustrial Arts & Consumer Services	2823.53
MajorCategoryInterdisciplinary	-7397.00
MajorCategoryLaw & Public Policy	7664.85
MajorCategoryPhysical Sciences	17118.28
MajorCategoryPsychology & Social Work	-1979.70
MajorCategorySocial Science	7923.38
GenM	5931.63
GPA	5488.74

Table 5: Model Performance measures

Residual SD	R-squared
5406.169	0.7637

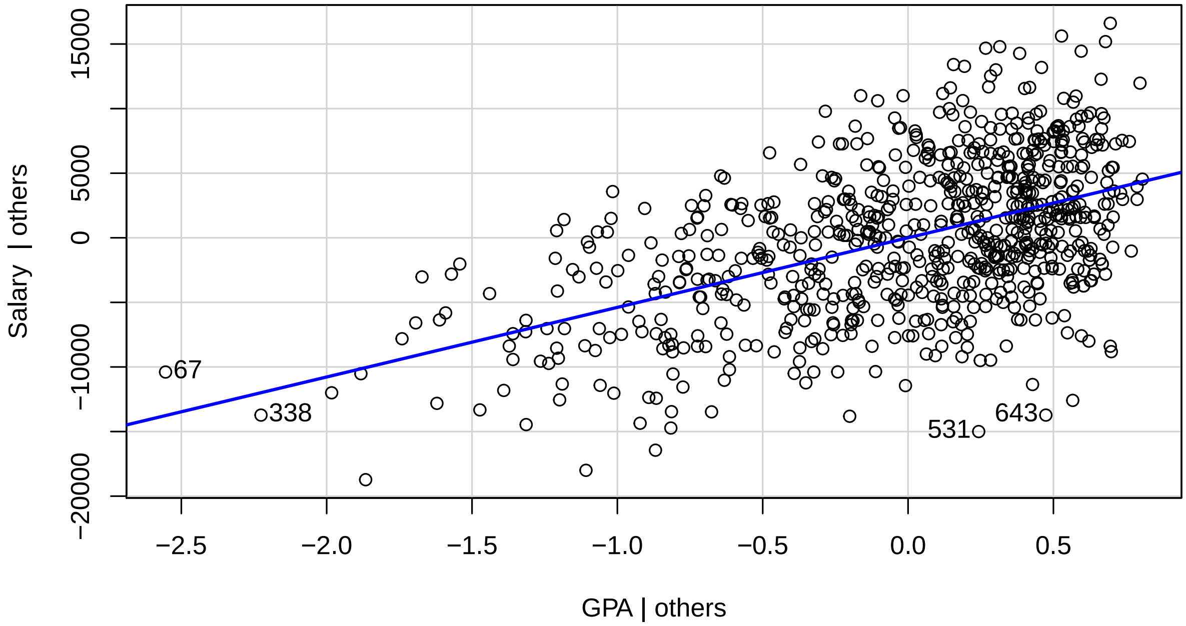
Question 4



The boxplots show that men have higher average salaries in every single major category. Some of the major categories have drastic differences between male and female salaries. Both of these observations support the idea of gender discriminaton as well as the fact that its severity may vary according to the field of study.

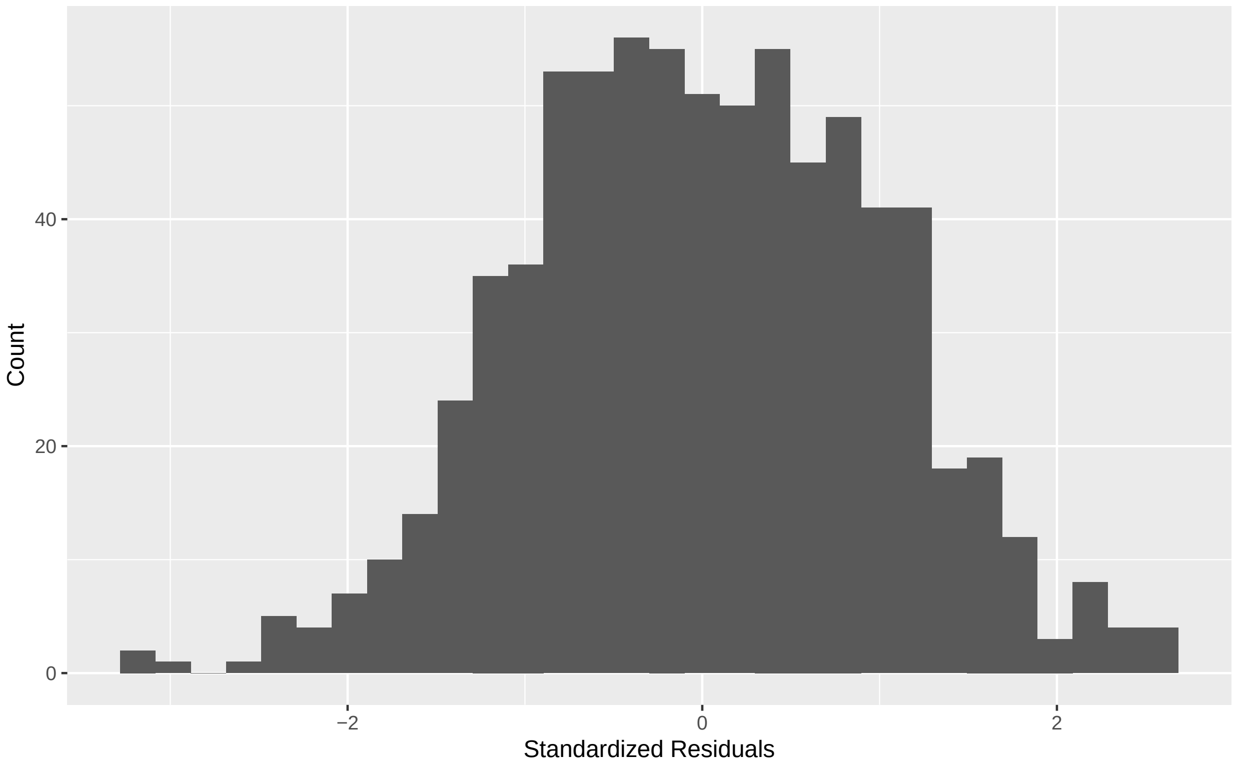
We performed an ANOVA test to determine if gender differences in salary varied by field of study. In order to do so we tested the null hypothesis that gender and major do not have a significant interaction against the alternative hypothesis that gender and major do have a significant interaction. We obtained a test statistic of $F \approx 4.36$ and a P-value of $Pr(> F) \approx 0$. We therefore reject the null hypothesis and conclude that gender and major do have a significant interaction.

Question 5

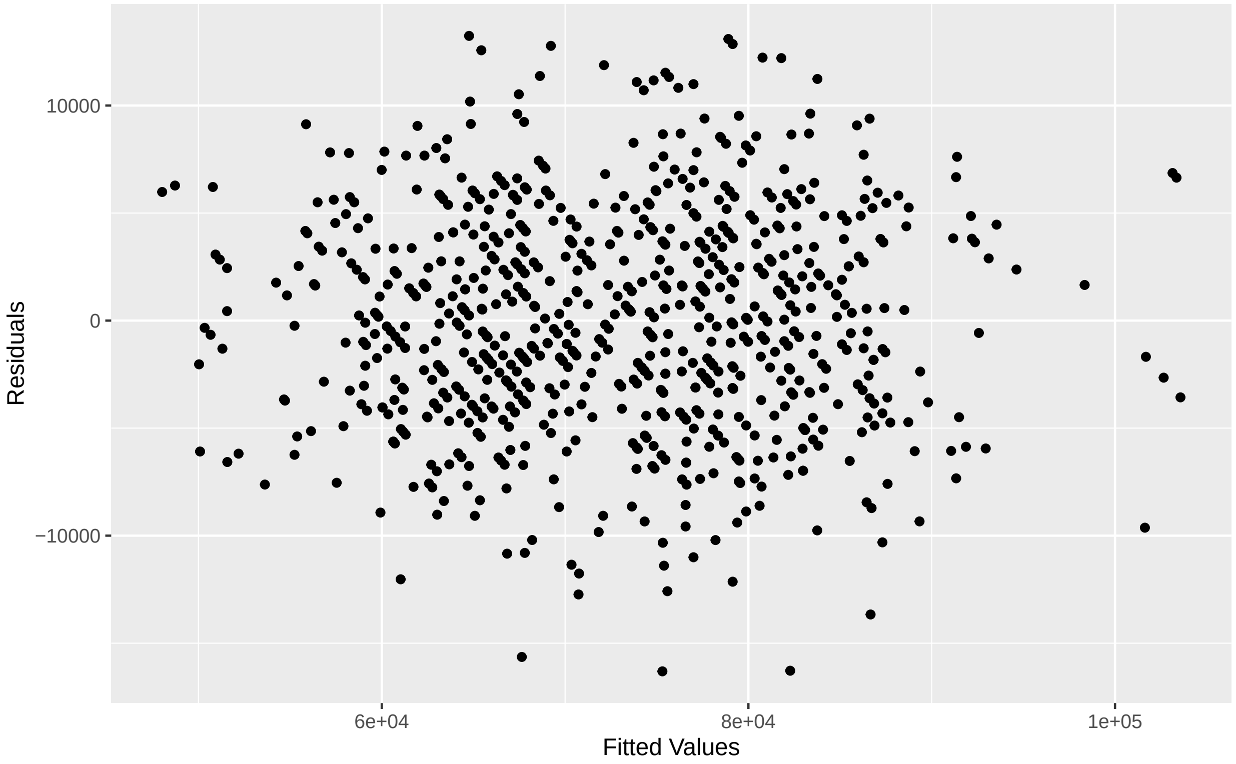


The linearity assumption appears to be met. Our added variable plot for GPA shows a very clear linear relationship between salary and GPA. GPA is our only quantiative variable.

The assumption of independence is reasonable because we would not expect any given graduate's salary to affect another's salary. This is especially true across major categories, since there would be little or no relationship between the jobs they would be applying for. Even for graduates in the same major, it is unlikely that two graduates would be greatly affected by the applications and careers of their classmates.



Our assumption of normality appears to be reasonable. The standardized residuals follow a standard normal distribution fairly closely, with no evidence of outliers. Additionally, a Kolmogorov-Smirnov test for normality gives a p-value of .7421, indicating that we should fail to reject the null hypothesis of normality.



Our assumption of equal variance appears reasonable. There is no clear trend in variance as we move across the x-axis. Additionally a Breusch-Pagan test for equal variance gives a P-value of .6075, indicating that we should fail to reject the null hypothesis of equal variance.

Question 6

Table 6: Confidence Intervals for Variables of Interest

	1.5 %	98.5 %
MajorCategoryComputers & Mathematics	15446.12	24062.71
GenM	9395.57	24387.63
GPA	4646.39	6129.76

We are 97% confident that computer and mathematics majors make between \$14,131.71 and \$21,742.11 more than agriculture and natural resources majors 5 years after graduation on average. We are 97% confident that men make \$5,059.64 and \$6,803.62 more than women 5 years after graduation on average. We are 97% that for every 1 point increase in GPA we could expect between a \$4,727.47 and a \$6,250.00 increase in salary 5 years after graduation on average.

Question 7

We wish to determine whether men in the computers and mathematics major category have significantly larger salaries than women, given the same GPA. Our null hypothesis is $H_\circ : \mu_{women} = \mu_{men}$ and our alternative is $H_a : \mu_{women} < \mu_{men}$ where the respective means are within the same major and GPA. Our hypothesis test gives a P-value of approximately zero. We therefore reject the null hypothesis and conclude that there is a significant difference between men's and women's salaries within the computers and mathematics major category, for student's with the same GPA. We are 95% confident that the difference described above is captured by the interval (\$10,123.43, \$23,659.76).

Question 8

We are 95% confident that my salary 5 years after graduation will be captured by the interval (\$83,040.07, \$104,361.90).

Question 9

In order to assess our predicting power we conducted a leave-one-out cross-validation. The average *RPMSE* for our predictions was \$4,358.08. The average width of our prediction intervals was \$21,013.43.

These metrics indicate good predicting power. Our predictions are within \$5,000 of the truth on average and we can be 95% confident of a range that is only about \$21,000 wide. While this is far from pinpoint precision, it is still very useful to be able to predict a salary between \$60,000 and \$80,000 as opposed to taking a random guess.

Appendix

```
knitr::opts_chunk$set(echo = FALSE, fig.dim = c(8,5))
# Include packages
library(ggplot2)
library(GGally)
library(car)
library(MASS)
library(lmtest)
library(multcomp)
library(knitr)

# Read in the data
college <- read.csv("https://mheaton.byu.edu/docs/files/Stat469/Topics/1%20-%20Independence/1%20-%20IID/
                        header = TRUE)
# 1
# Scatterplot of Salary by GPA
ggplot(college, aes(GPA,Salary)) + geom_point() + labs(title = "GPA Scatterplot", x = "GPA", y = "Salary

# Summary Statistics
kable(data.frame(mean(college[,"GPA"]),
                 sd(college[,"GPA"]),
                 cor(college["GPA"],college["Salary"])),
      col.names = c("Mean","SD","Correlation"),
      caption = "Summary Statistics for GPA",
      digits = 3)
# Side-by-side boxplots of Salary for each category in Major
ggplot(college, aes(MajorCategory,Salary)) + geom_boxplot() + labs(title = "Distribution of Salary by Ma

# Summary Statistics
data.frame(aggregate(Salary ~ MajorCategory, data = college, FUN = mean),
           aggregate(Salary ~ MajorCategory, data = college, FUN = sd)[,2]) %>% kable(col.names = c("Maj
      caption = "Salary Statistics by Major Category")
# Side-by-side boxplots of Salary for each gender
ggplot(college, aes(Gen,Salary)) + geom_boxplot() + labs(title = "Distribution of Salary by Gender", x =

# Summary Statistics
data.frame(aggregate(Salary ~ Gen, data = college, FUN = mean),
           aggregate(Salary ~ Gen, data = college, FUN = sd)[,2]) %>% kable(col.names = c("Gender","Mean
      caption = "Salary Statistics by Gender")
# GG pairs plot
ggpairs(college[-2])
# 3
```

```r
# Construct design matrix
X <- model.matrix(object=Salary~MajorCategory+Gen+GPA, data=college)
# Extract Y vector (birth weights)
Y <- college$Salary
# Calculate P from number of columns
P <- ncol(X)-1
# Calculate number of observations
n <- nrow(X)

# Calculate betahat and s2 using matrix algebra
beta.hat <- solve(t(X) %*% X) %*% t(X) %*% Y
# store row names as vector
names <- rownames(beta.hat)
# eliminate rownames
rownames(beta.hat) <- NULL
# display names and values of coefficients
kable(data.frame(names,beta.hat),
      col.names = c("Coefficient","Estimate"),
      digits = 2,
      caption = "Estimates of Model Coefficients")

# Identify s2 R2
s2 <- (t(Y - (X %*% beta.hat)) %*% (Y - (X %*% beta.hat)))/(n - P - 1)
college.lm <- lm(Salary~MajorCategory+Gen+GPA, data = college)
r2 <- summary(college.lm)$r.squared
kable(data.frame(sqrt(s2),r2),
      col.names = c("Residual SD","R-squared"),
      caption = "Model Performance measures",
      digits = 4)
# 4
# Side-by-side boxplots of salary by major and then by gender
ggplot(college, aes(x = MajorCategory, y = Salary, color = Gen)) + geom_boxplot() + labs(x = "Major", y

# Test for a significant difference between males and females within each major
test.lm <- lm(Salary~MajorCategory*Gen+GPA, data = college)
gen.mod.test <- anova(test.lm,college.lm)
# F = 4.36
#p-value = approx. 0

# Because our tests show significant interaction between major and gender we will replace the original
final.college.lm <- test.lm
# 5
# Create added-variable plots
```

```
avPlots(final.college.lm, terms = "GPA")
# Histogram of the standardized residuals
st.res <- stdres(final.college.lm)
ggplot() + geom_histogram(mapping=aes(x=st.res), bins = 30) + xlab("Standardized Residuals") + ylab("Co

# KS-test for normality
# p-value = .7421
# fail to reject the null hypothesis and conclude that the residuals are normal
#ks.test(st.res, "pnorm")
# Store fitted values and residuals
fit <- fitted(final.college.lm)
res <- residuals(final.college.lm)

# Scatterplot of the fitted values vs. standardized residuals
ggplot(mapping = aes(fit,res)) + geom_point() + xlab("Fitted Values") + ylab("Residuals")

# BP-test for equal variance
# p-value = .6075
# fail to reject the null hypothesis and conclude that the variance is constant with respect to x
#bptest(final.college.lm)
# 6
kable(confint(final.college.lm, level = .97)[c(6,17,18),],
      digits = 2,
      caption = "Confidence Intervals for Variables of Interest")
# 7
# First birth characteristics
a1 <- rep(0,33)
a1[c(1,6,17,23)] <- 1

# Second birth characteristics
a2 <- rep(0,33)
a2[c(1,6,23)] <- 1

# Difference between first and second characteristics
a <- a1-a2

# Test on the difference between those two birth scenarios
gender.test <- glht(final.college.lm, linfct = t(a), alternative="two.sided")
# summary(gender.test)
# confint(gender.test, level = .95)
# Create data.frame with explanatory values for prediction
new.x = data.frame(MajorCategory = "Computers & Mathematics", Gen = "M", GPA = 4.00)
```

```r
# Predict with a 95% prediction interval
# predict.lm(final.college.lm, newdata = new.x, interval="prediction", level=0.95)
n.cv <- nrow(college) #Number of CV studies to run
rpmse <- rep(x=NA, times=n.cv)
wid <- rep(x=NA, times=n.cv)
for(cv in 1:n.cv){
  ## Select test observations
  test.obs <- cv

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

  ## Fit a lm() using the training data
  train.lm <- lm(formula = Salary~MajorCategory*Gen+GPA,
                 data = train.set)

  ## Generate predictions for the test set
  my.preds <- predict.lm(train.lm, newdata=test.set, interval="prediction")

  ## Calculate RPMSE
  rpmse[cv] <- (test.set[['Salary']] - my.preds[,'fit'])^2 %>% mean() %>% sqrt()

  ## Calculate Width
  wid[cv] <- (my.preds[,'upr'] - my.preds[,'lwr']) %>% mean()

}

# mean(rpmse)
# mean(wid)
```