The Value of a College Education

Nick Hass

1/26/2022

Homework 1

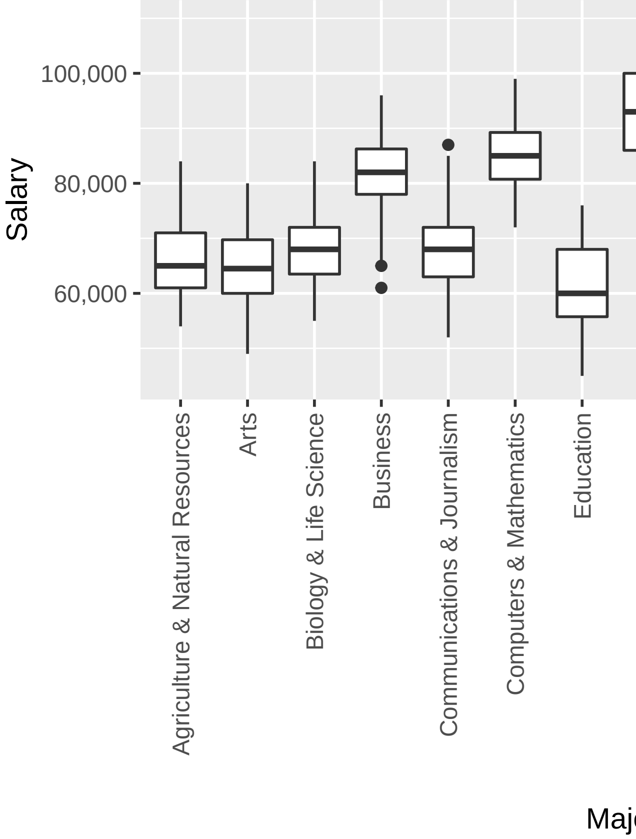
Analysis Questions

```
salary <- read.csv("salary.csv")
#head(salary)

salaryByMajor <- salary %>%
  group_by(MajorCategory) %>%
  summarise(AvgSalary = mean(Salary))

# Plot a boxplot of the salary distribution for each major
ggplot(data = salary, mapping = aes(x = MajorCategory, y = Salary)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  scale_y_continuous(labels = scales::comma)
```

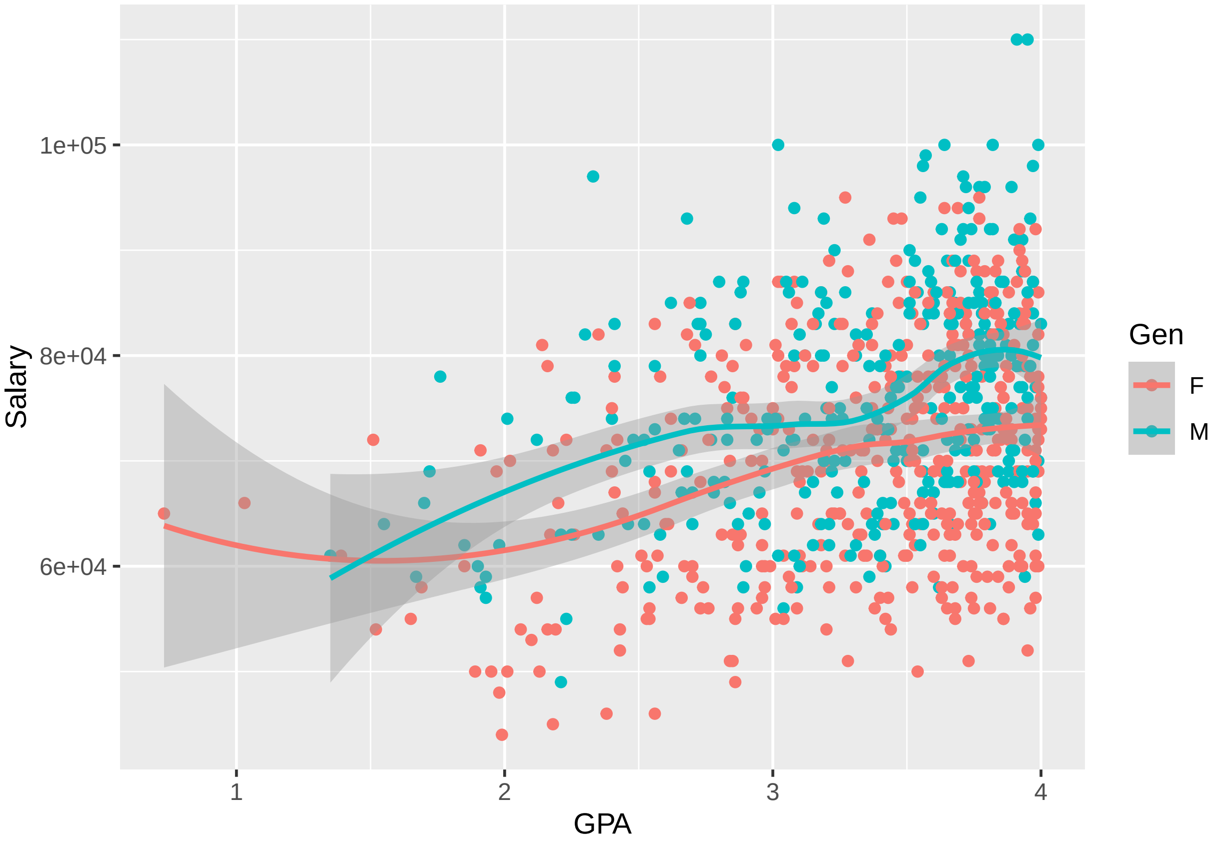
1. Create exploratory plots and calculate summary statistics from the data. Comment on any potential relationships you see from these exploratory plots.



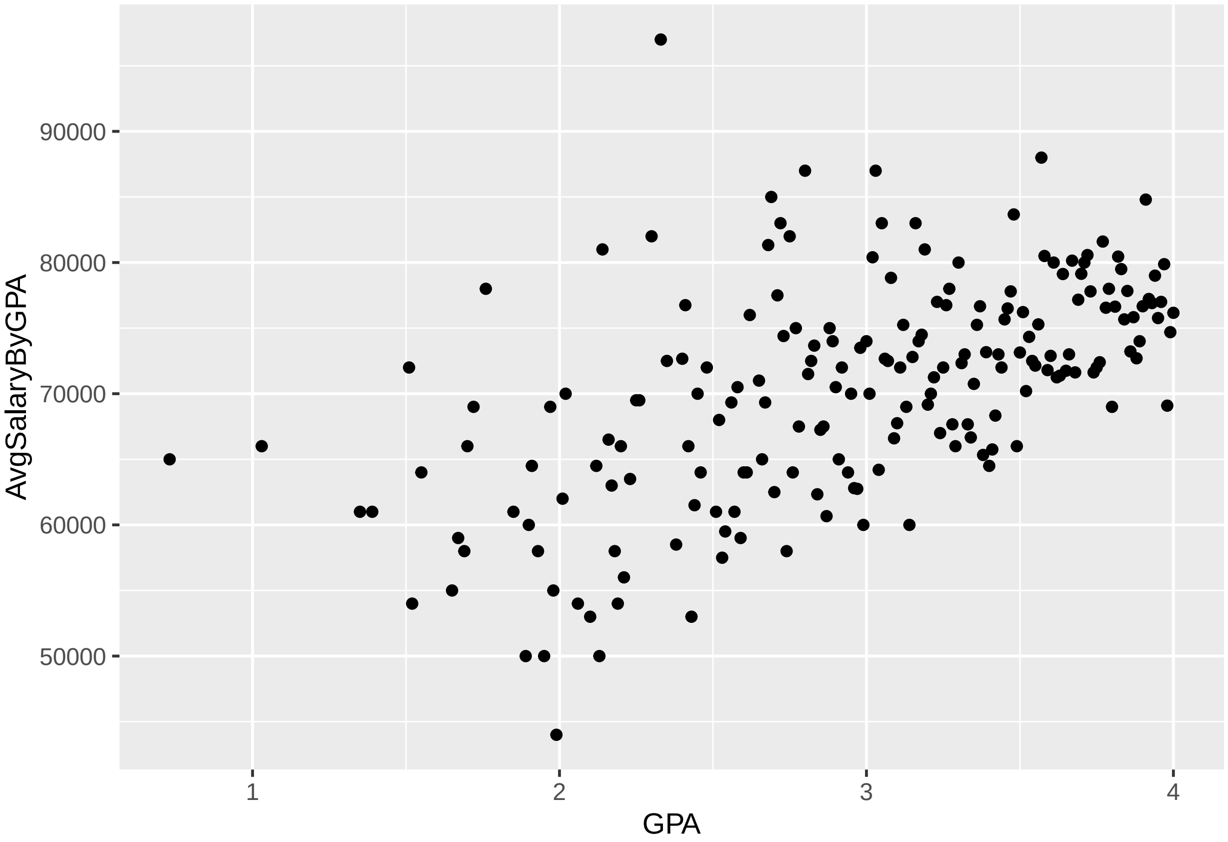
```
salaryByGPA <- salary %>%
  group_by(GPA) %>%
  summarise(AvgSalaryByGPA = mean(Salary))

ggplot(data = salary, mapping = aes(x = GPA, y = Salary, colour = Gen)) +
  geom_point() + geom_smooth()
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



```
ggplot(data = salaryByGPA, mapping = aes(x = GPA, y = AvgSalaryByGPA)) +
  geom_point()
```

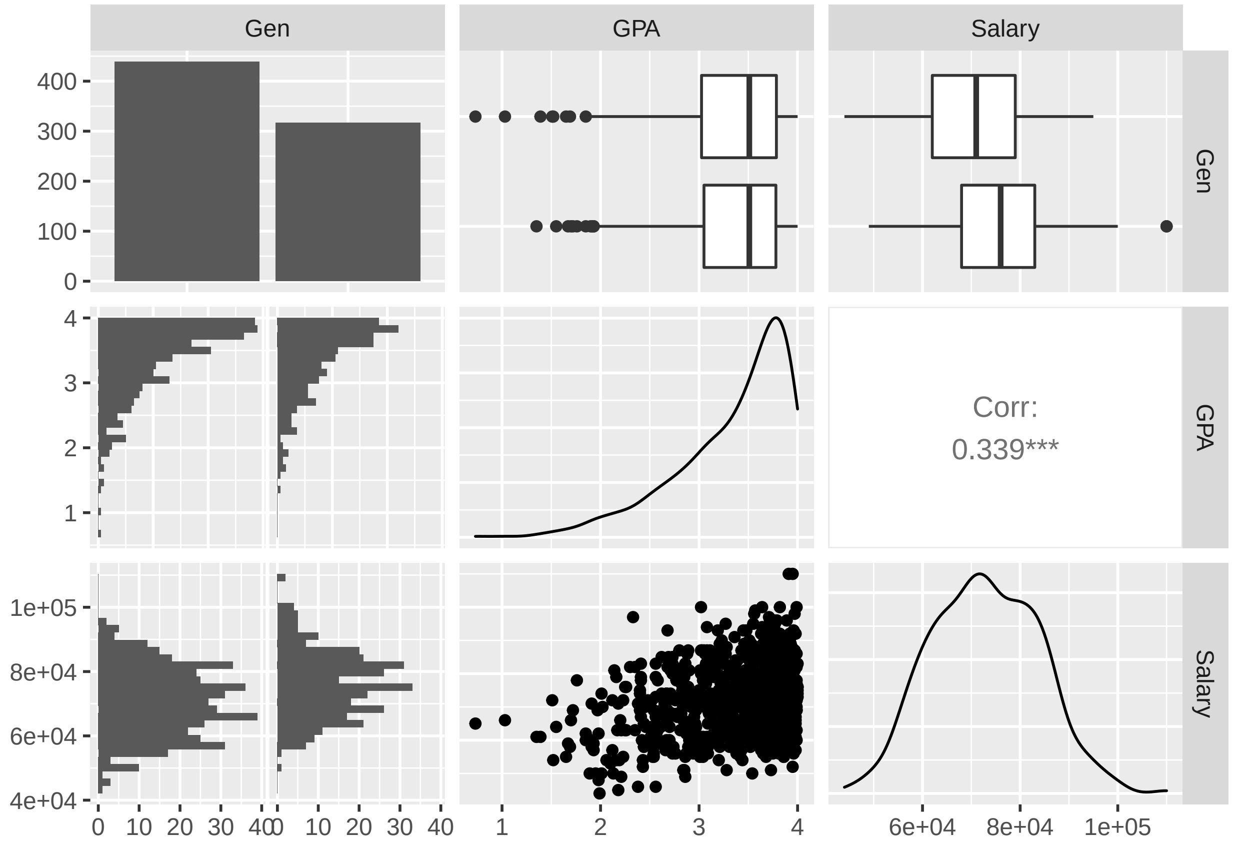


```
newSalary <- salary[,-1]
response <- salary[1]
newerSalary <- newSalary[-1]

newererSalary <- cbind(newerSalary, response)
GGally::ggpairs(newererSalary)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Relationships I see from EDA: Salaries are very distinct when grouped by `MajorCategory`. From the scatterplot of GPA by Average Salary (by gpa), it looks like there is a positive correlation between GPA and Average Salary. It looks like Males make on average, more money than females.

2. Write down a linear regression model (in matrix and vector form) in terms of parameters. Explain the meaning of any parameters in your model. Explain how statistical inference for your model can be used to answer the effect of major choice and identify any gender discrimination. $Y = X\beta + \epsilon, \quad \epsilon \sim \text{MVN}(0,\sigma^2)$

X is the Design matrix with 1s on the first column β is a vector of the coefficients for each explanatory variable ϵ is the random vector

Statistical inference can be used to answer the effect of major choice by testing if their is a difference between expected salary by major choice and using a confidence interval to see how large that difference is.

Statistical inference can be used to identify any gender discrimination by determining if the slope for expected salaries are different amongst the two genders.

```
X <- model.matrix(Salary~., data = salary)
P <- 17 # GPA + gender + 15 major categories
n <- dim(salary)[1]

beta_hat <- solve(t(X)%*%X) %*% t(X) %*% salary$Salary

resid_variance <- t(salary$Salary - X %*% beta_hat) %*% (salary$Salary - X %*% beta_hat) / (n-P-1)
salary.lm <- lm(Salary ~ ., data=salary)
```

```r
print(lm(Salary ~ ., data=salary))
```

3. Using first principles (i.e. DON'T use lm() but you can check your answer with lm()), calculate beta hat and report the estimates in a table. Interpret the coefficient for 1 categorical explanatory variable and the coefficient for GPA. Also calculate the estimate of the residual variance (or standard deviation) and R2 (you can use lm() to get R2).

```
## 
## Call:
## lm(formula = Salary ~ ., data = salary)
## 
## Coefficients:
##                                      (Intercept)  
##                                          46673.0  
##                                MajorCategoryArts  
##                                          -2551.6  
##              MajorCategoryBiology & Life Science  
##                                            769.1  
##                            MajorCategoryBusiness  
##                                          14282.1  
##         MajorCategoryCommunications & Journalism  
##                                            114.6  
##             MajorCategoryComputers & Mathematics  
##                                          17936.9  
##                           MajorCategoryEducation  
##                                          -5894.8  
##                         MajorCategoryEngineering  
##                                          24406.2  
##                              MajorCategoryHealth  
##                                           8670.2  
##           MajorCategoryHumanities & Liberal Arts  
##                                          -5972.6  
## MajorCategoryIndustrial Arts & Consumer Services  
##                                           2823.5  
##                   MajorCategoryInterdisciplinary  
##                                          -7397.0  
##                 MajorCategoryLaw & Public Policy  
##                                           7664.9  
##                   MajorCategoryPhysical Sciences  
##                                          17118.3  
##            MajorCategoryPsychology & Social Work  
##                                          -1979.7  
##                      MajorCategorySocial Science  
##                                           7923.4  
##                                             GenM  
##                                           5931.6  
##                                              GPA  
##                                           5488.7  
```

```r
print(beta_hat)
```

```
##                                                         [,1]
## (Intercept)                                    46672.9855
## MajorCategoryArts                              -2551.6387
## MajorCategoryBiology & Life Science              769.1305
## MajorCategoryBusiness                          14282.1484
## MajorCategoryCommunications & Journalism         114.6014
## MajorCategoryComputers & Mathematics           17936.9081
## MajorCategoryEducation                         -5894.8466
## MajorCategoryEngineering                       24406.2278
## MajorCategoryHealth                             8670.1623
## MajorCategoryHumanities & Liberal Arts         -5972.5852
## MajorCategoryIndustrial Arts & Consumer Services  2823.5261
## MajorCategoryInterdisciplinary                 -7396.9963
## MajorCategoryLaw & Public Policy                7664.8538
## MajorCategoryPhysical Sciences                 17118.2762
## MajorCategoryPsychology & Social Work          -1979.6997
## MajorCategorySocial Science                     7923.3790
## GenM                                            5931.6270
## GPA                                             5488.7368
```

```
print(resid_variance)
```

```
##          [,1]
## [1,] 29226669
```

```
sigma(salary.lm)^2
```

```
## [1] 29226669
```

```
summary(salary.lm)$r.squared
```

```
## [1] 0.7637316
```

```
# math and computers, male, 3.0 gpa prediction
# pred <- matrix(c(1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,3.0), ncol=1)
# t(pred) %*% beta_hat
```

Q: Interpret the coefficient for 1 categorical explanatory variable and the coefficient for GPA.

A: Since the base case is for a female, holding everything else constant (major and GPA), a male can expect Salary to go up by 5931.60 dollars on average.

Q: Explain the coefficient for the GPA

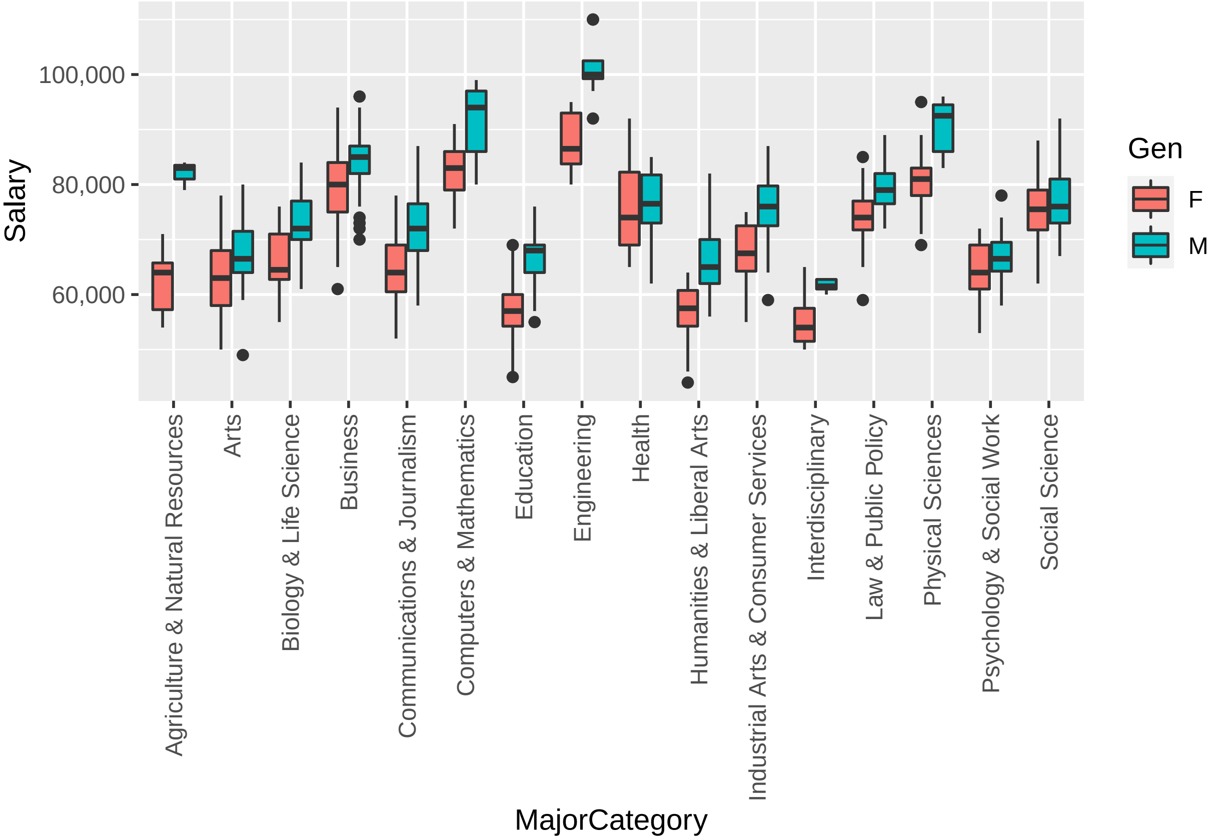
A: Holding everything constant (major and gender), as GPA increases by 1, the average salary will increase by 5488.7 dollars.

Q: Also calculate the estimate of the residual variance (or standard deviation) and R2 (you can use lm() to get R2).

A: R squared is 0.7637316

```r
# Plot side-by-side boxplots of the salary distribution for each major and gender combination
ggplot(data = salary, mapping = aes(x = MajorCategory, y = Salary, fill= Gen)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  scale_y_continuous(labels = scales::comma)
```

4. One common argument is that some disciplines have greater biases (in terms of lower salaries) towards women than others. To verify this, check for interactions between major and gender by (i) drawing side-by-side boxplots of salary for each major category and gender combination and (ii) running an appropriate hypothesis test (either t or F) to check for significance. Comment on potential gender discrimination from your boxplot. For your hypothesis test, state your hypotheses, report an appropriate test statistic, p-value and give your conclusion.



```r
# Hypothesis test to check for significance (F test)
full.lm <- salary.lm
reduced.lm <- lm(Salary ~ MajorCategory + GPA, data = salary)
anova(full.lm,reduced.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Salary ~ MajorCategory + Gen + GPA
## Model 2: Salary ~ MajorCategory + GPA
##   Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
## 1    738 2.1569e+10
## 2    739 2.7963e+10 -1 -6393530395 218.76 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova(full.lm,reduced.lm)["Pr(>F)"]
```

```
##      Pr(>F)
## 1
## 2 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
bigger.lm <- lm(Salary ~ MajorCategory + GPA + Gen + Gen:MajorCategory, data = salary)
anova(bigger.lm, full.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Salary ~ MajorCategory + GPA + Gen + Gen:MajorCategory
## Model 2: Salary ~ MajorCategory + Gen + GPA
##   Res.Df        RSS  Df   Sum of Sq      F    Pr(>F)
## 1    723 1.9780e+10
## 2    738 2.1569e+10 -15 -1789058098 4.3595 7.161e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

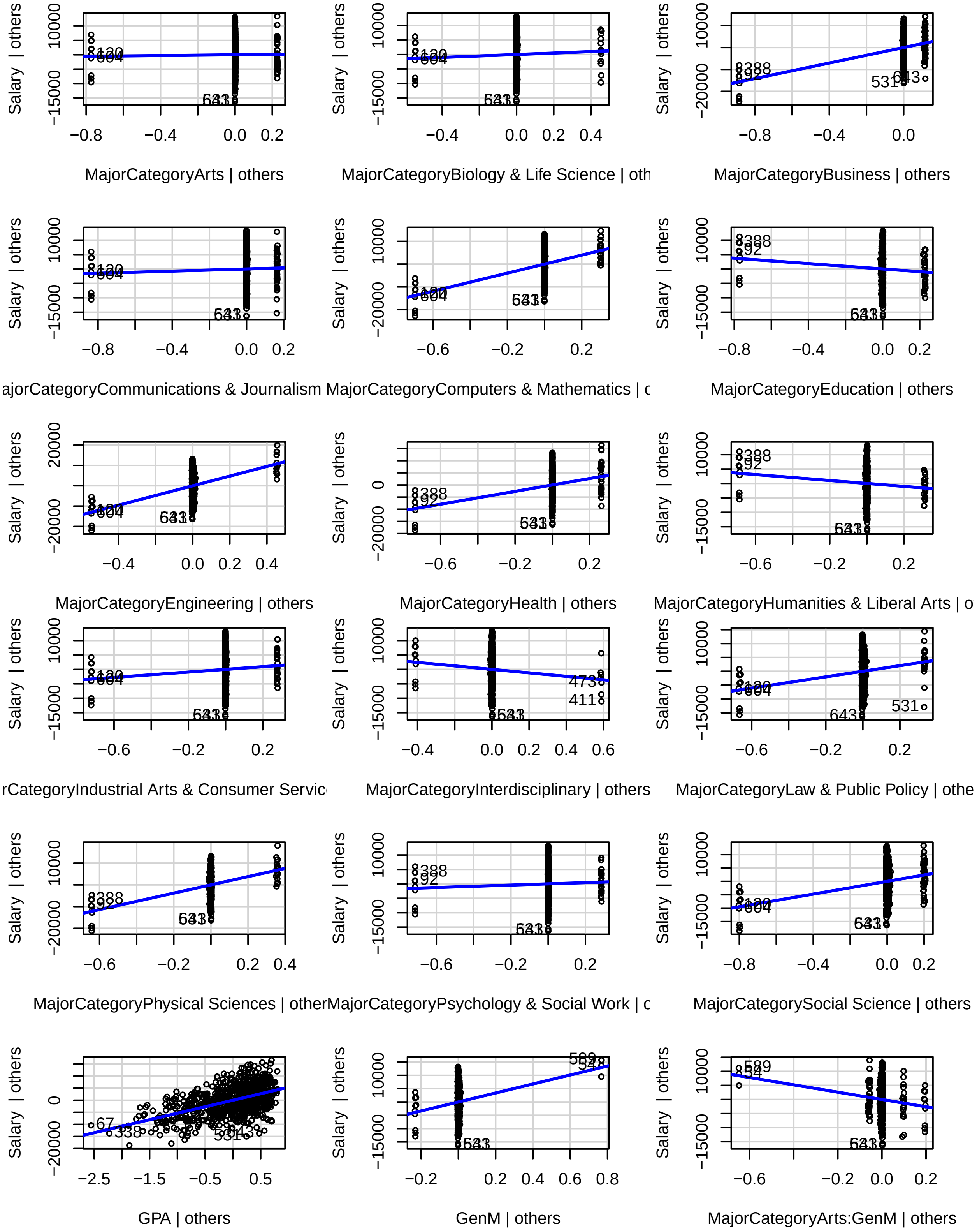
Q: Comment on potential gender discrimination from your boxplot. For your hypothesis test, state your hypotheses, report an appropriate test statistic, p-value and give your conclusion.

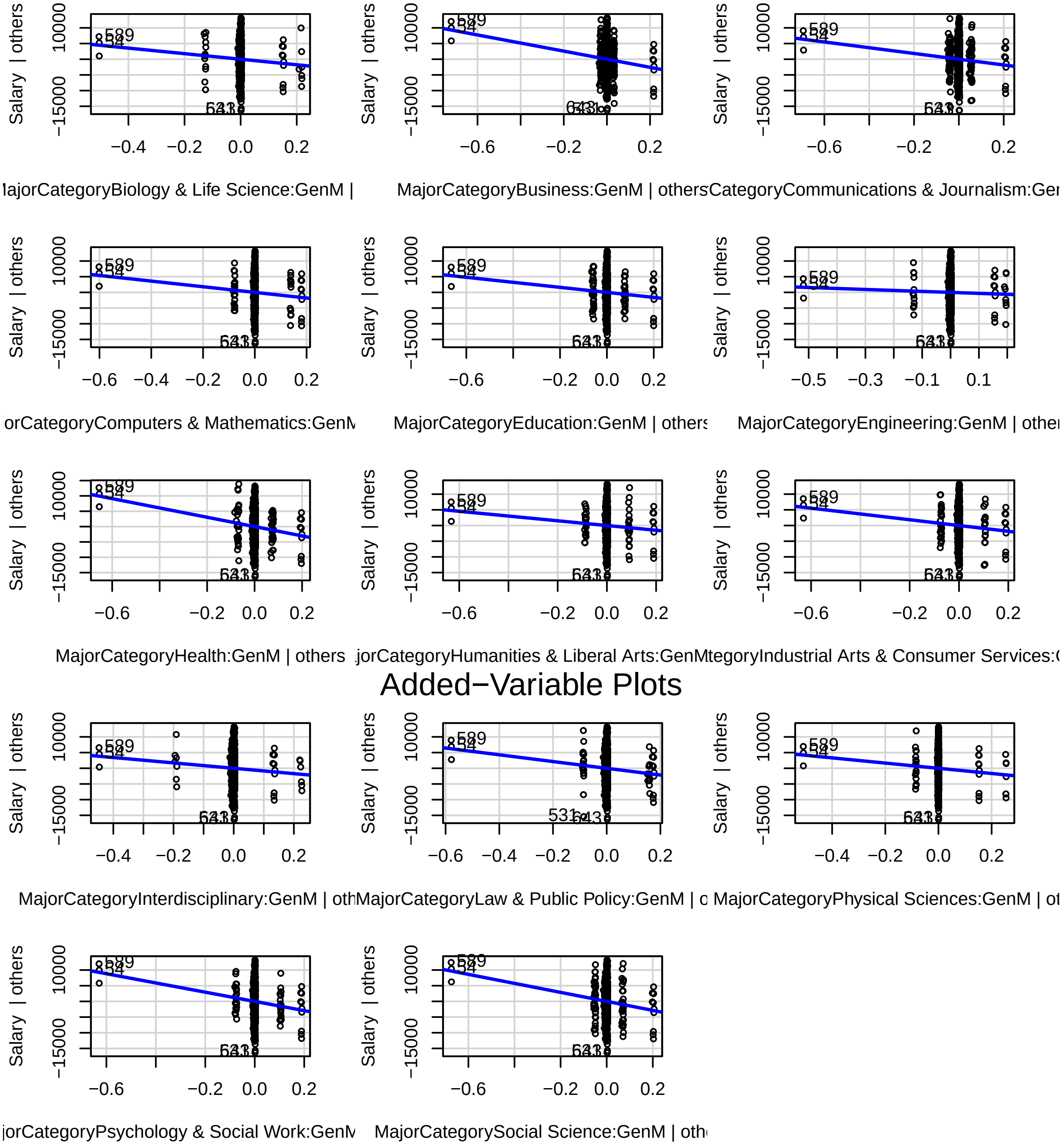
A: The boxplot above shows a significant interaction between salary and gender for all majors. I conducted a F test to test whether there was a significant difference between a model that included gender as an explanatory variable and found that there is a significant interaction (with p-value $<$ 2.2e-16).

I conducted another F test to test whether there was interaction between Gender and Major category and found that there was (because p-value $<$ 7.161e-08).

```
# Linearity - Added variable plots
car::avPlots(bigger.lm)
```

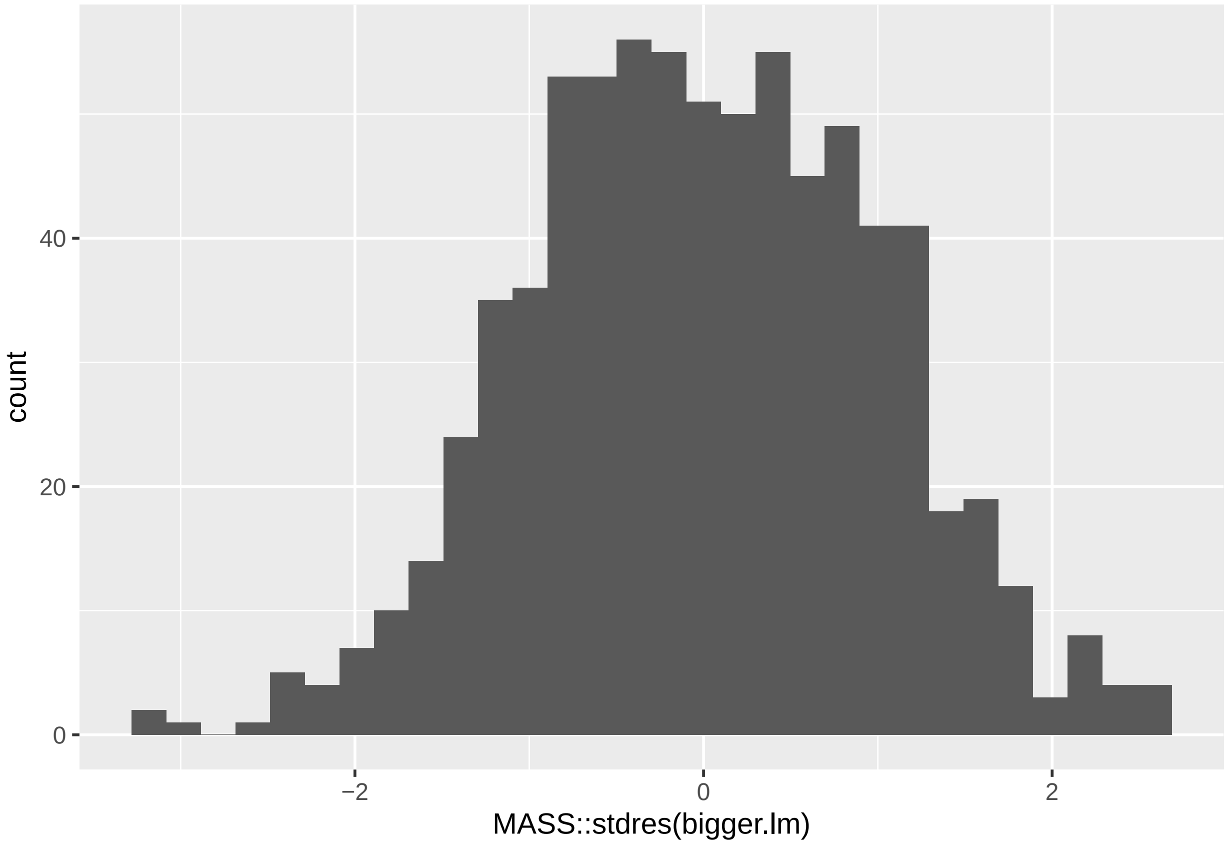
5. The validity of the tests from #4 depend on the validity of the assumptions in your model (if your assumptions are violated then the p-values are likely wrong). Create graphics and/or run appropriate hypothesis tests to check the L-I-N-E assumptions associated with your multiple linear regression model including any interactions you found in #4. State why each assumption does or does not hold for the salary data.


Salary | others
MajorCategoryArts | others
MajorCategoryBiology & Life Science | oth
MajorCategoryBusiness | others
ajorCategoryCommunications & Journalism
MajorCategoryComputers & Mathematics | c
MajorCategoryEducation | others
MajorCategoryEngineering | others
MajorCategoryHealth | others
MajorCategoryHumanities & Liberal Arts | o
rCategoryIndustrial Arts & Consumer Servic
MajorCategoryInterdisciplinary | others
MajorCategoryLaw & Public Policy | othe
MajorCategoryPhysical Sciences | other
MajorCategoryPsychology & Social Work | c
MajorCategorySocial Science | others
GPA | others
GenM | others
MajorCategoryArts:GenM | others
−15000
10000
−20000
20000
0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
−2.5
−1.5
−0.5
0.5
8388
92
531
643
604
120
4736
4118
67
338
589
548
854



```r
# Independence - think about it
# Normality - Histogram of standardized residuals
standardized.residuals = MASS::stdres(bigger.lm)
ggplot() +
  geom_histogram(mapping = aes(x = MASS::stdres(bigger.lm) ))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

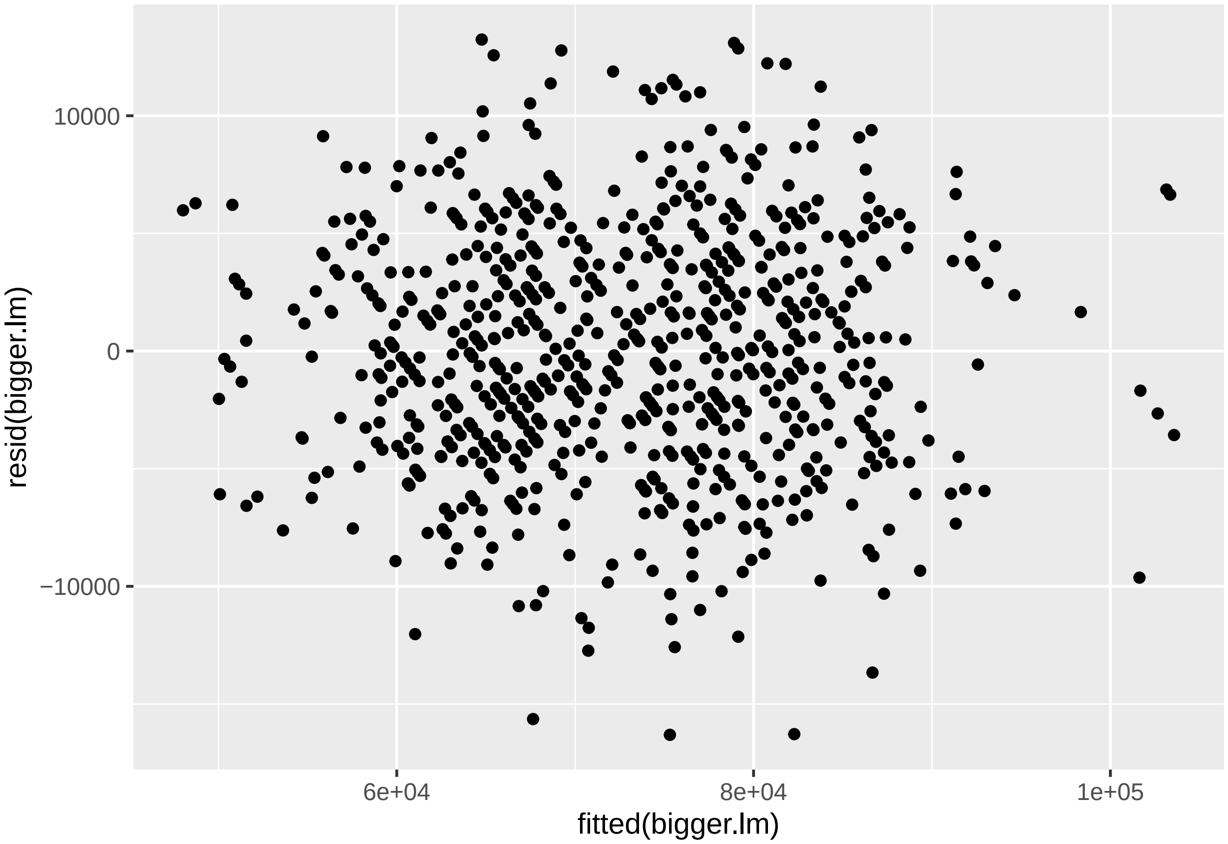


```
# KS Test
ks.test(standardized.residuals, "pnorm")
```

```
## Warning in ks.test(standardized.residuals, "pnorm"): ties should not be present
## for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  standardized.residuals
## D = 0.024779, p-value = 0.7421
## alternative hypothesis: two-sided
```

```
# Equal Variance - BP test or scatterplot of the standardized fitted vs. residuals
ggplot() +
  geom_point(mapping = aes(x = fitted(bigger.lm), y = resid(bigger.lm) ))
```



```
lmtest::bptest(bigger.lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bigger.lm
## BP = 29.23, df = 32, p-value = 0.6075
```

Linearity - looks good.

Independence - Can assume independence.

Normality - looks normal. Results from the KS test indicate that the standardized residuals follow a normal distribution because of the hi p-value (failing to reject that it is not normal).

Equal Variance - The fitted values vs standardized residuals plot looks good. The BP test reveals that the equal variance assumption holds because of the high p-value (the test fails to reject the null hypothesis that it is normal).

```
new.x = data.frame(GPA = 3.5, Gen = "M", MajorCategory = "Computers & Mathematics")
predict.lm(salary.lm, newdata=new.x, interval="prediction", level=0.97)
```

6. Calculate 97% confidence intervals for the coefficients for GPA, Gender and one major category. Interpret each interval.

```
##       fit      lwr      upr
## 1 89752.1 77822.28 101681.9
```

For a Male Computers & Mathematics major with a 3.5 GPA can expect to earn 89752.10 dollars, with a 95% confidence interval between 77822.28 and 101681.90 dollars.

```
# math and computers, male, 3.0 gpa prediction
men <- matrix(c(1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,3.0), nrow = 1)
women <- matrix(c(1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,3.0), nrow = 1)
diff <- men - women  # if positive, then men make more money
ht <- multcomp::glht(full.lm, linfct=diff, alternative="two.sided")
summary(ht)
```

7. For the Computers and Mathematics major category, perform a general linear hypothesis test that women, on average, earn less salary than men (for the same GPA). State your hypotheses, p-value and conclusion. If this test is significant, report and estimate a 95% confidence interval for how much more men earn than women in that major category.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Salary ~ ., data = salary)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0     5932        401   14.79   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## (Adjusted p values reported -- single-step method)
```

```
confint(ht, level=0.95)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = Salary ~ ., data = salary)
##
## Quantile = 1.9632
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##        Estimate  lwr       upr
## 1 == 0 5931.6270 5144.3019 6718.9520
```

```
# t(pred) %*% beta_hat
```

Null Hypothesis: Women and men of the Math & Computers Major with the same GPA will on average have the same salary.

Alternative Hypothesis: Men on average will have a higher salary than women (if both are a Math & Computers major with the same GPA).

P-value: <2e-16

I conclude that because of the low p-value of <2e-16, men on average in the Math and Computers major earn a statistically significant higher salary than women of the same major and GPA.

95% Confidence interval: I am 95% confident that on average, males will earn between $5,144.30 and $6,718 more than their female counterpart from the Math & Computers major with the same GPA.

```
predict.lm(bigger.lm, newdata = data.frame(MajorCategory = "Computers & Mathematics", Gen = "M", GPA = 3
```

8. Using predict.lm() and your fitted model, predict your salary and report an associated 95% prediction interval. Interpret this interval in context.

```
##        fit     lwr      upr
## 1 89983.22 79325.6 100640.8
```

I predict that my salary will be $89,983.22. I am 95% confident that my salary will fall between $79,325.60 and $100,640.80.

```
# Leave-one-out cross validation

n <- nrow(salary)
# placeholder for storing the i-th prediction
preds <- rep(0, n)
wid <- rep(0, n)
bias <- rep(0, n)
rpmse <- rep(0, n)
cvg <- rep(0, n)
for(i in 1:n) {
    dataf.train <- salary[-i, ]
    dataf.test <- salary[i, ]
    lm.fit <- lm(Salary ~ MajorCategory + GPA + Gen + Gen:MajorCategory, data=dataf.train)

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

    ## Calculate bias
    bias[i] <- mean(my.preds[,'fit'] - dataf.test[['Salary']])

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

    ## Calculate Coverage
    cvg[i] <- ((dataf.test[['Salary']] > my.preds[,'lwr']) & (dataf.test[['Salary']] < my.preds[,'upr']

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

9. If we wish to use our model for prediction as we did in #8, we should verify how accurate our predictions are via cross-validation. Conduct a leave-one-out cross validation of the salary data. Report your average RPMSE along with the average prediction interval width. Comment on whether you think your predictions are accurate or not.

```
## [1] 4358.084
```

```
mean(wid)
```

```
## [1] 21013.43
```

Average RPMSE: $4,358.084

Average prediction interval width $21,013.43

I think my predictions are accurate because the root predicted mean square error is only $4,358.084, which is the average distance in amount of money that my predictions were from the actual salary. When thinking of a yearly salary, $4,358 is a very small amount and equates to an error of only $2 an hour, when working a 40 hour work week. This makes my predictions very accurate.