

Analysis of the Value of a College Education

Aaron Oustrich, Josh Bergstrom, Anna Wolford

2024-02-07

1.

```
print("Salary Statistics")
```

```
## [1] "Salary Statistics"
```

```
summary(sal$Salary)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  44000   64750   73000   72888   81000  110000
```

```
print("GPA Statistics")
```

```
## [1] "GPA Statistics"
```

```
summary(sal$GPA)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.730   3.040   3.510   3.341   3.780   4.000
```

```
print("How many Genders in each Major Category?")
```

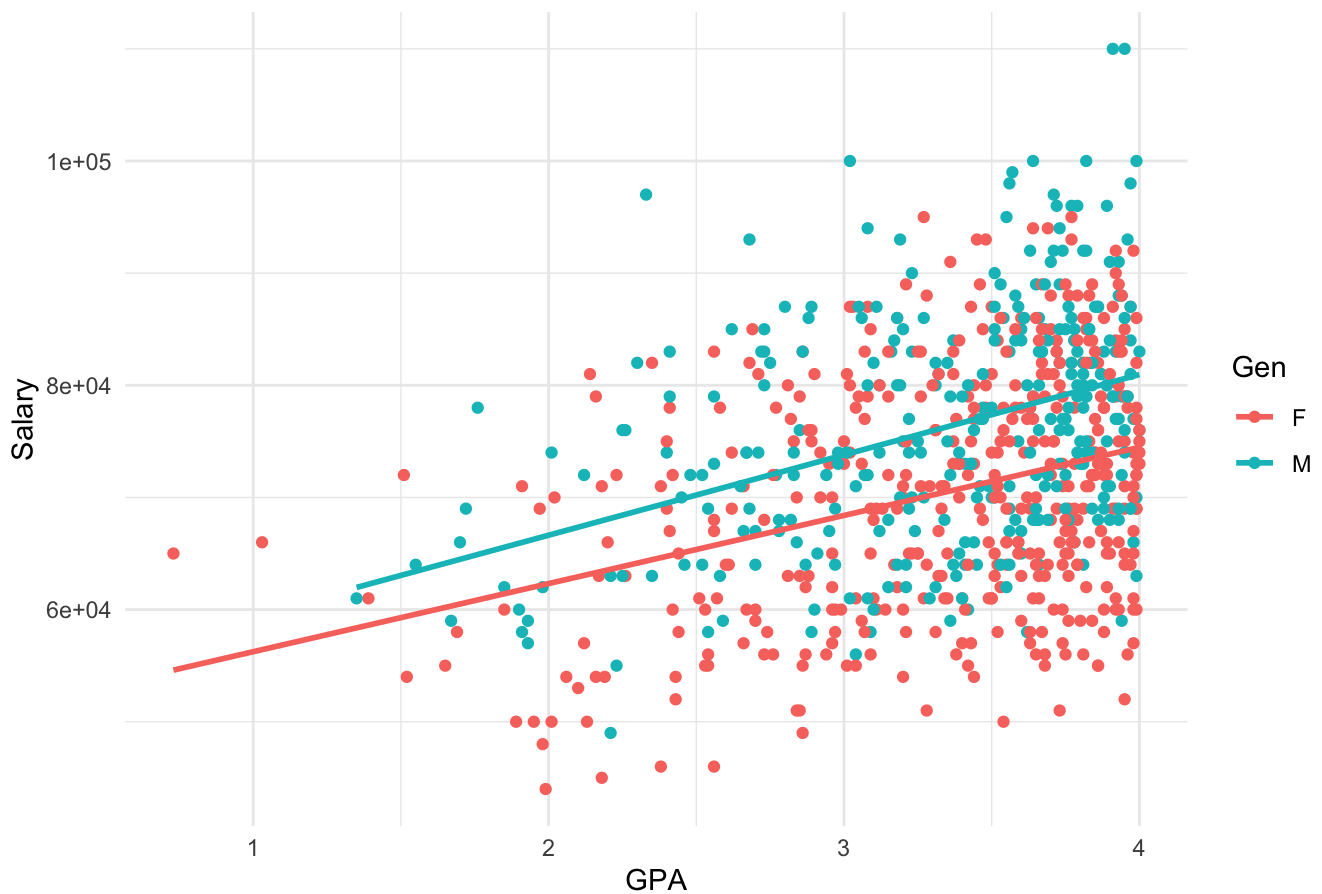
```
## [1] "How many Genders in each Major Category?"
```

```
table(sal$MajorCategory, sal$Gen)
```

```
##
##           F  M
## Agriculture & Natural Resources 10  3
## Arts                             34 20
## Biology & Life Science          12  7
## Business                         77 67
## Communications & Journalism      51 38
## Computers & Mathematics         23 13
## Education                       34 26
## Engineering                      12  8
## Health                           28 26
## Humanities & Liberal Arts       22 21
## Industrial Arts & Consumer Services 26 18
## Interdisciplinary                  7  6
## Law & Public Policy              20 11
## Physical Sciences                18  6
## Psychology & Social Work        25 18
## Social Science                   40 29
```

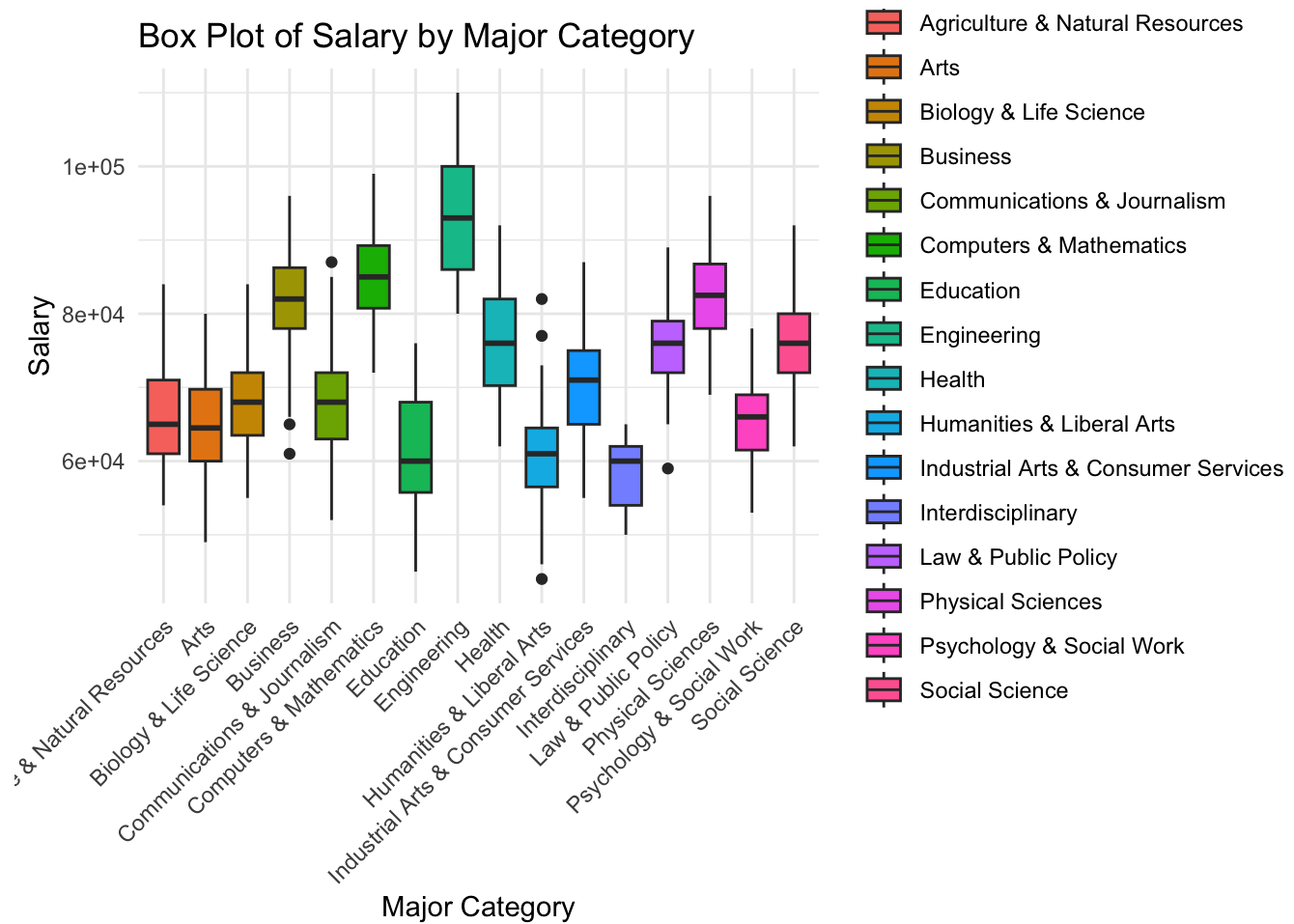
Above are summary outputs of the continuous variables (Salary and GPA). We show the MajorCategory as a table which shows the number of each Gender in that Major.

Scatterplot of GPA and Salary

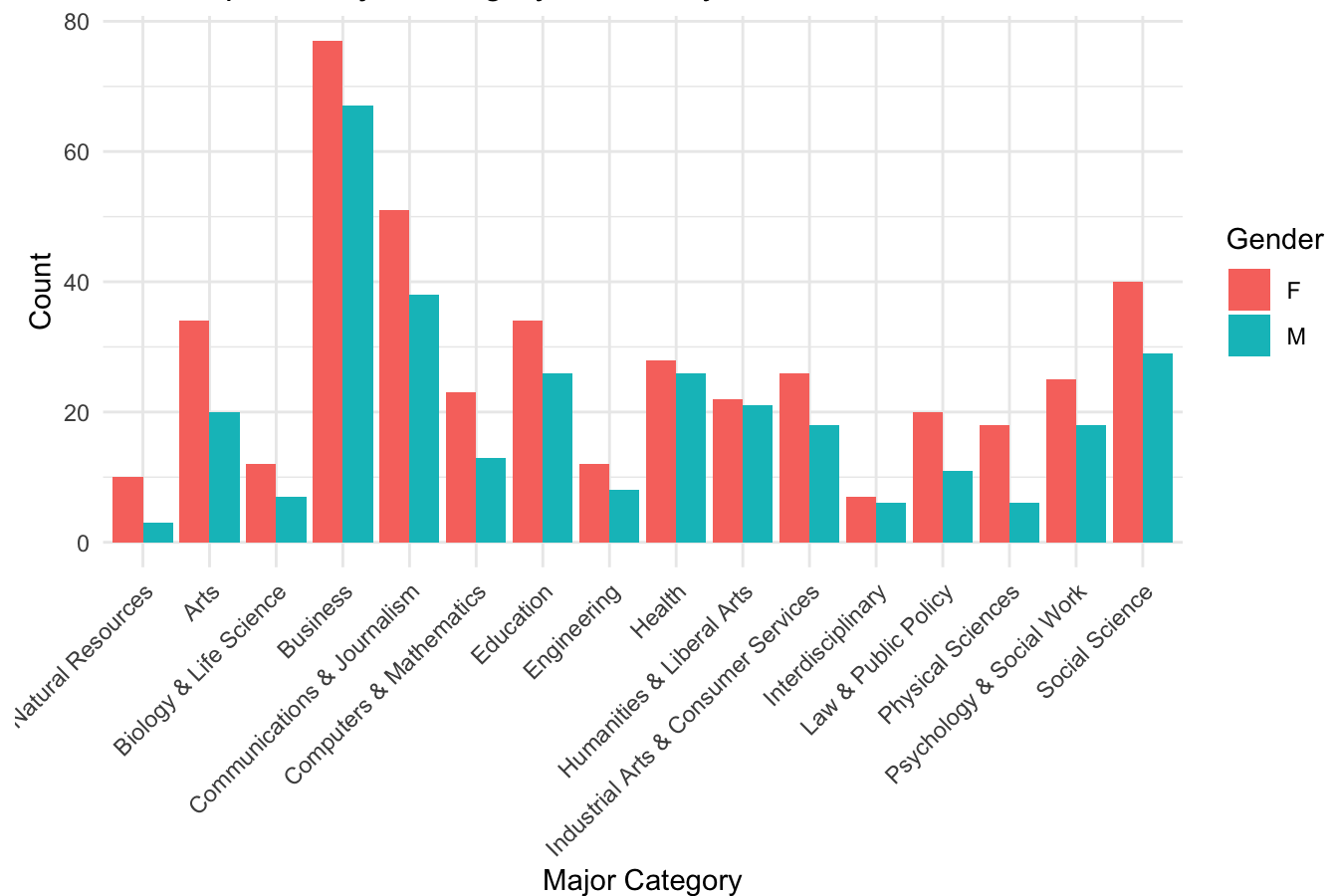


This scatterplot shows a positive relationship between GPA and Salary for both genders. Based on the lines of best fit, Male salary appears higher than Female salary for all GPA levels. It's interesting to note that both genders appear to have similar slope which leads us to believe the effect of GPA is consistent across the genders.

Box Plot of Salary by Major Category



Bar Graph of Major Category Counts by Gender



These last two plots show there are few outliers in salary across all the major categories and that there are more women represented in each major category than men (though they are all relatively equal).

2.

The multiple linear regression model is based on the following equation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where:

- \mathbf{y} is the vector of salaries
- \mathbf{X} is the matrix of explanatory variables which includes a column of 1s for the intercept and dummy-encoded columns for the categorical variables of Major Category and Gender.
- $\boldsymbol{\beta}$ is the vector of coefficients
- $\boldsymbol{\epsilon}$ is the vector of errors which are assumed to be normally distributed with mean 0 and constant variance σ^2 ($\boldsymbol{\epsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I})$)

Thus, the model is specified as:

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

If we fit the data to the specified model, and confirm the model assumptions are sufficiently met, we can determine the effect of major choice and identify any gender discrimination.

3.

```
sal.lm <- lm(Salary ~., data=sal)
summary(sal.lm)
```

```
##
## Call:
## lm(formula = Salary ~ ., data = sal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16468.1  -3643.6   -48.9   3877.8  14811.7
##
## Coefficients:
##                                     Estimate Std. Error t value
## (Intercept)                      46673.0      1925.0  24.246
## MajorCategoryArts                 -2551.6      1671.4  -1.527
## MajorCategoryBiology & Life Science    769.1      1946.7   0.395
## MajorCategoryBusiness               14282.1      1569.1   9.102
## MajorCategoryCommunications & Journalism   114.6      1607.5   0.071
## MajorCategoryComputers & Mathematics  17936.9      1750.1  10.249
## MajorCategoryEducation              -5894.8      1657.5  -3.557
## MajorCategoryEngineering             24406.2      1927.7  12.661
## MajorCategoryHealth                  8670.2      1674.9   5.177
## MajorCategoryHumanities & Liberal Arts  -5972.6      1715.8  -3.481
## MajorCategoryIndustrial Arts & Consumer Services  2823.5      1708.1   1.653
## MajorCategoryInterdisciplinary         -7397.0      2129.3  -3.474
## MajorCategoryLaw & Public Policy        7664.9      1787.4   4.288
## MajorCategoryPhysical Sciences         17118.3      1863.1   9.188
## MajorCategoryPsychology & Social Work  -1979.7      1713.3  -1.155
## MajorCategorySocial Science           7923.4      1636.4   4.842
## GenM                                5931.6       401.0  14.790
## GPA                                5488.7       350.1  15.677
##
##                                     Pr(>|t|)
## (Intercept)                      < 2e-16 ***
## MajorCategoryArts                 0.127284
## MajorCategoryBiology & Life Science 0.692892
## MajorCategoryBusiness               < 2e-16 ***
## MajorCategoryCommunications & Journalism 0.943184
## MajorCategoryComputers & Mathematics < 2e-16 ***
## MajorCategoryEducation              0.000400 ***
## MajorCategoryEngineering             < 2e-16 ***
## MajorCategoryHealth                 2.92e-07 ***
## MajorCategoryHumanities & Liberal Arts 0.000529 ***
## MajorCategoryIndustrial Arts & Consumer Services 0.098752 .
## MajorCategoryInterdisciplinary         0.000543 ***
## MajorCategoryLaw & Public Policy       2.04e-05 ***
## MajorCategoryPhysical Sciences         < 2e-16 ***
## MajorCategoryPsychology & Social Work  0.248275
## MajorCategorySocial Science           1.57e-06 ***
## GenM                                < 2e-16 ***
## GPA                                < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5406 on 738 degrees of freedom
```

```
## Multiple R-squared:  0.7637, Adjusted R-squared:  0.7583
## F-statistic: 140.3 on 17 and 738 DF,  p-value: < 2.2e-16
```

```
X <- model.matrix(Salary ~., data=sal)
y <- sal$Salary
bhat <- solve(t(X)%*%X)%*%t(X)%*%y
bhat_table <- as.table(bhat)
colnames(bhat_table) <- "Beta Estimates"
bhat_table
```

##	Beta Estimates
## (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

```
# Estimate of residual variance
s2 <- t(y-X %*% bhat)%*% (y-X%*%bhat) /(nrow(sal)-ncol(X))
s2
```

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

```
sqrt(s2)
```

```
##           [,1]
## [1,] 5406.17
```

```
sigma(sal.lm)
```

```
## [1] 5406.17
```

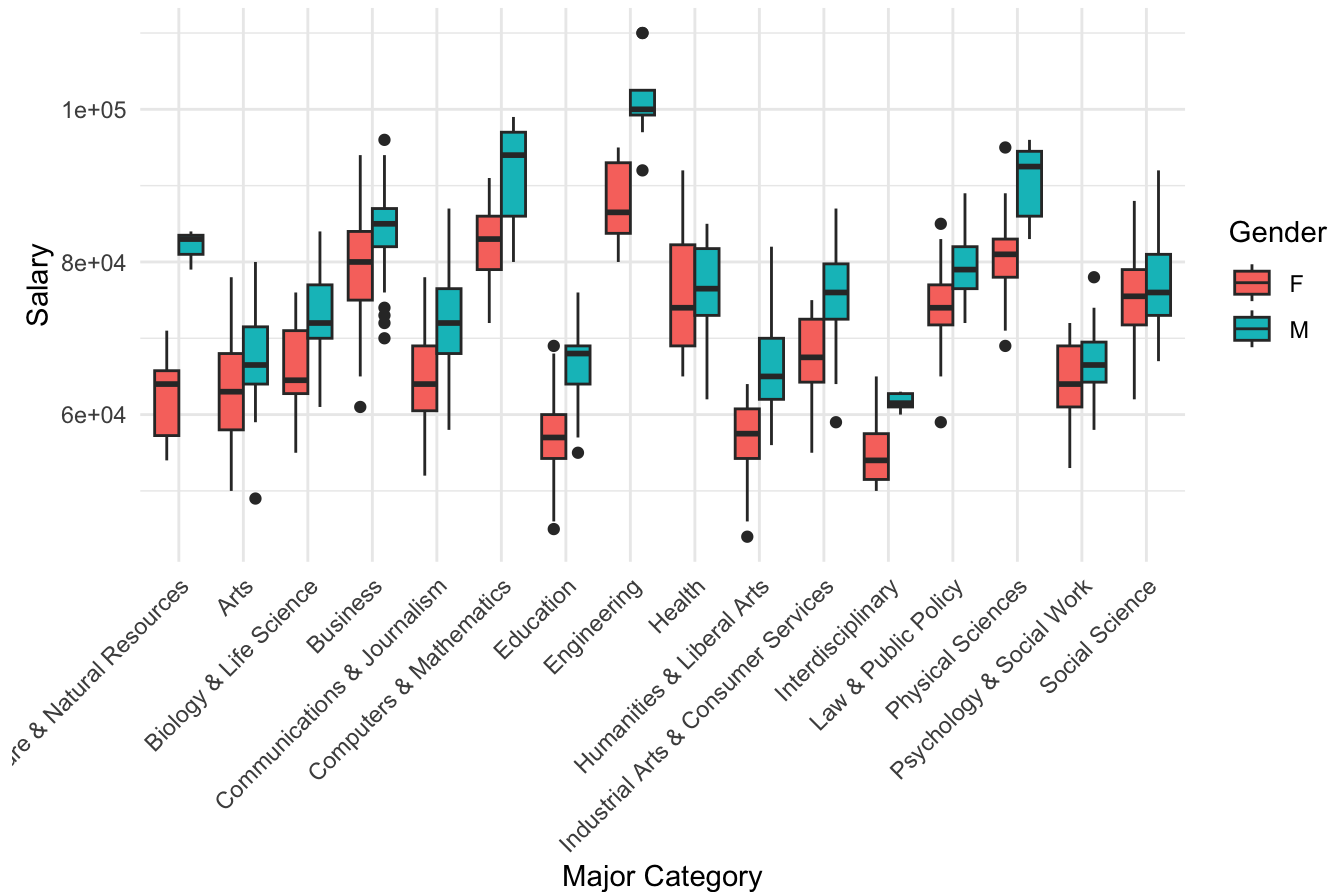
```
# R^2
summary(sal.lm)$r.squared
```

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

Compared to women, on average men make \$5931.63 more, holding all else constant. With every unit increase in GPA, holding all else constant, the average salary increases by \$5,488.74.

4.

Box Plot of Salary by Major Category and Gender



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

anova(sal.full, sal.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Salary ~ MajorCategory + Gen + GPA + MajorCategory:Gen
## 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
```

To see if there is an interaction between MajorCategory and Gender, we created a new model (`sal.full`) which adds interaction terms for these variables of interest. The null hypothesis is that there is no interaction between MajorCategory and Gender (i.e. all the β coefficients for the interaction terms are all 0). The alternative hypothesis is that there is a significant interaction between MajorCategory and Gender.

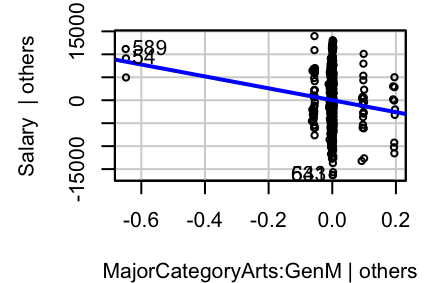
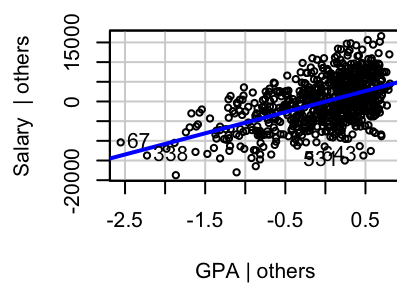
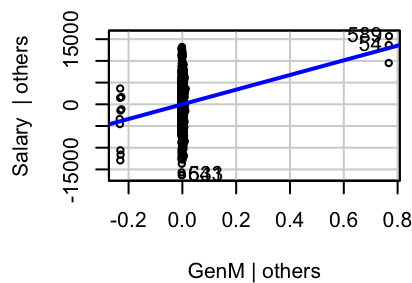
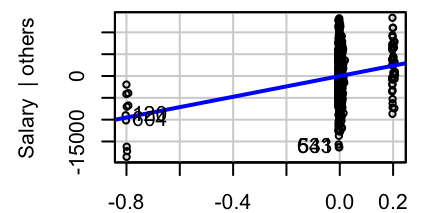
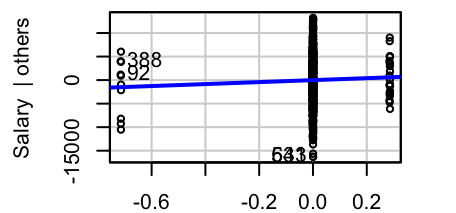
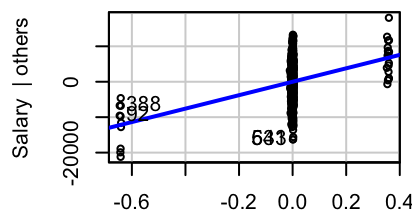
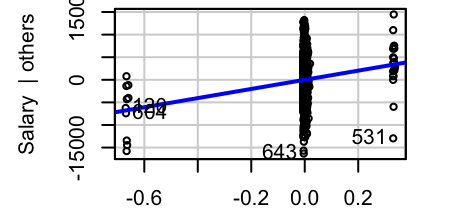
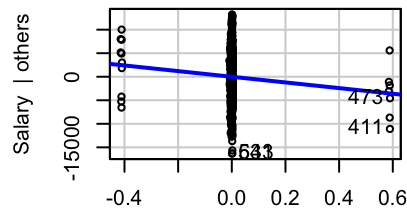
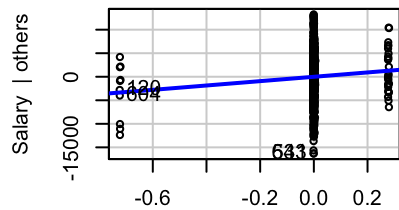
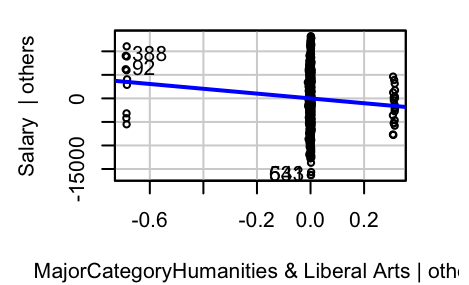
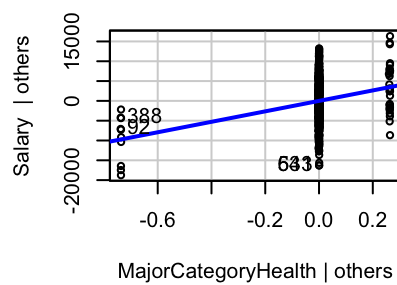
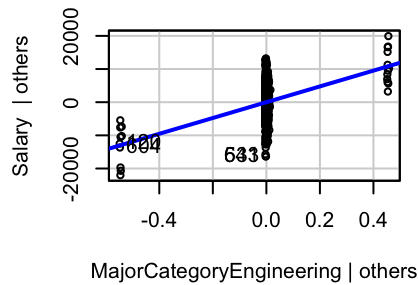
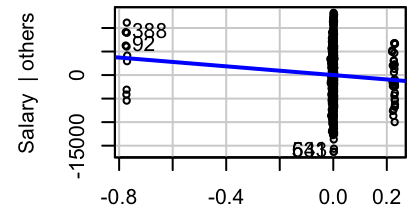
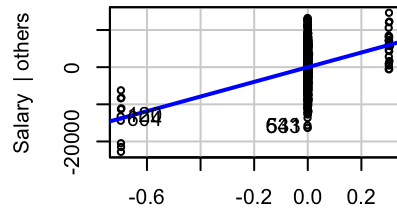
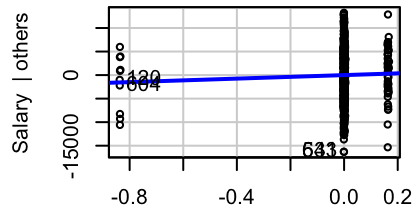
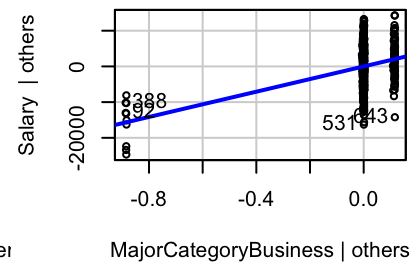
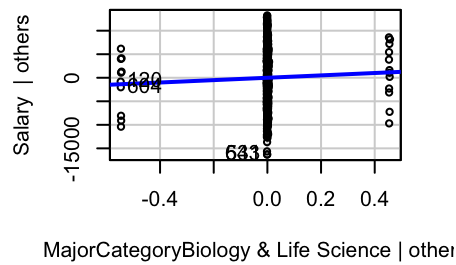
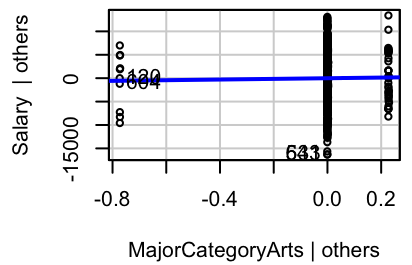
We evaluated this hypothesis using an ANOVA test on the two models and got an F statistic of 4.3 with a p-value less than 0.05. Therefore, we reject the null hypothesis and conclude that there is a significant interaction between MajorCategory and Gender.

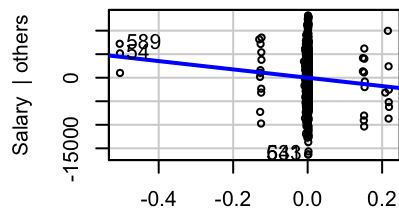
It appears that men have a higher salary across all majors. Personally, we don't feel that we can comment on "gender discrimination" even if there is an apparent difference, because we don't know enough about the data (ex: how these salaries were negotiated or how salaries may be different in the same company).

5.

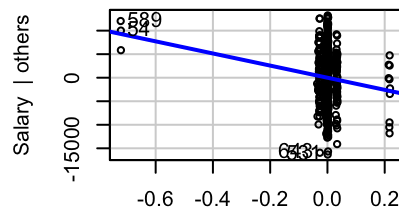
Linearity Assumption

```
avPlots(sal.full)
```

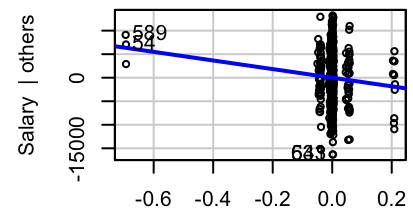





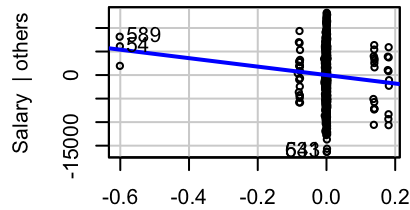
MajorCategoryBiology & Life Science:GenM | others



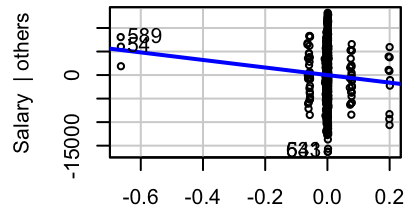
MajorCategoryBusiness:GenM | others



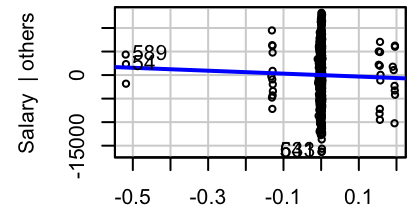
MajorCategoryCommunications & Journalism:GenM | others



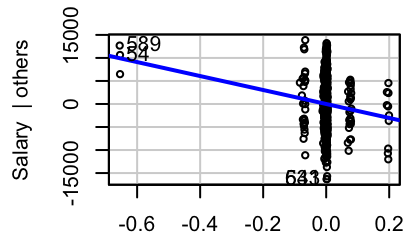
MajorCategoryComputers & Mathematics:GenM | others



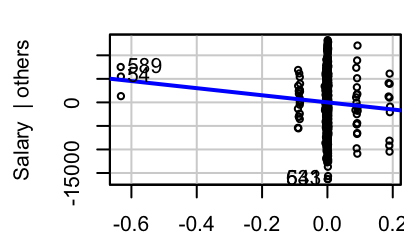
MajorCategoryEducation:GenM | others



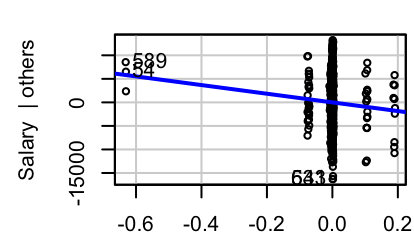
MajorCategoryEngineering:GenM | others



MajorCategoryHealth:GenM | others

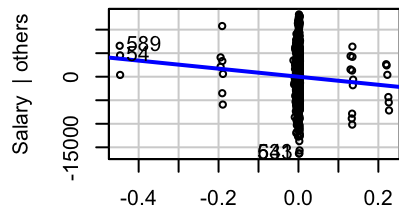


MajorCategoryHumanities & Liberal Arts:GenM | others

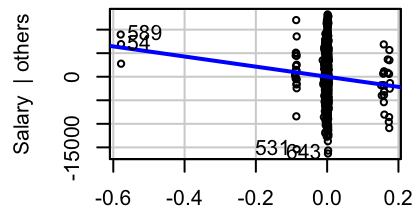


MajorCategoryIndustrial Arts & Consumer Services:GenM | others

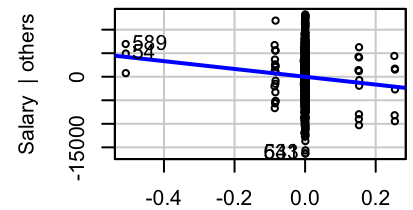
Added-Variable Plots



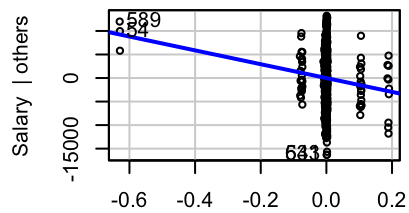
MajorCategoryInterdisciplinary:GenM | others



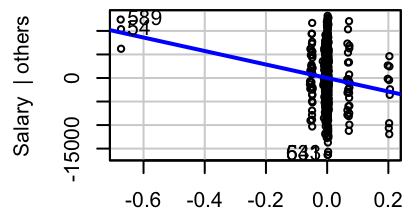
MajorCategoryLaw & Public Policy:GenM | others



MajorCategoryPhysical Sciences:GenM | others



MajorCategoryPsychology & Social Work:GenM | others



MajorCategorySocial Science:GenM | others

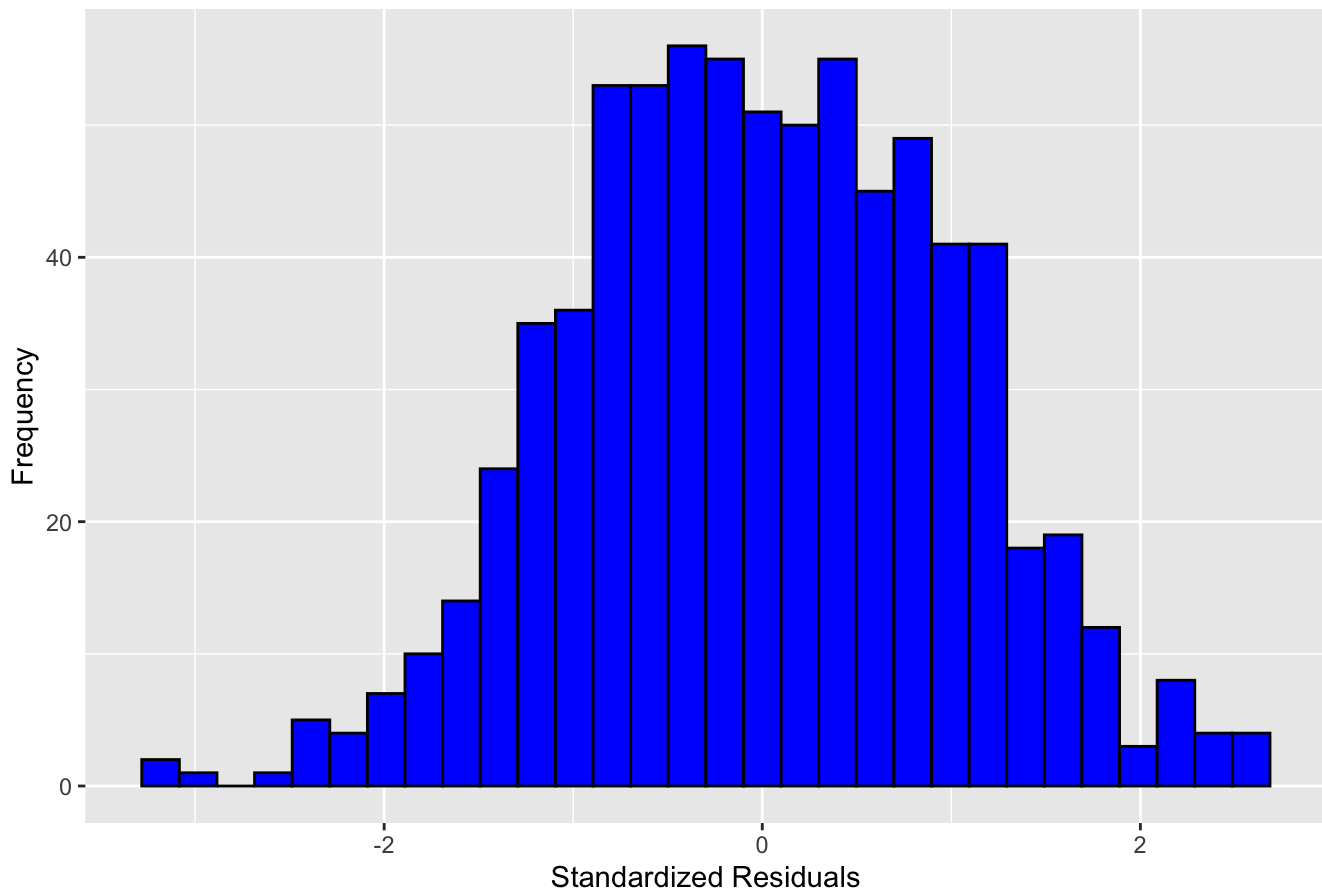
Based on all the added variable plots above, we believe the Linearity assumption is met. Looking at the added variable plot of the only continuous variable (GPA) is linear. All other categorical plots look weird, but there's nothing in them that suggest a non-linear relationship.

Independence Assumption

We don't know exactly how the data was collected, so we can't say for sure if the data were randomly sampled and independent of one another. However, we know that a salary for one person doesn't typically depend on the salary of another person, so we can assume independence. This assumption is met, and we can proceed with the analysis by checking the other assumptions.

Normality Assumption

Histogram of Standardized Residuals

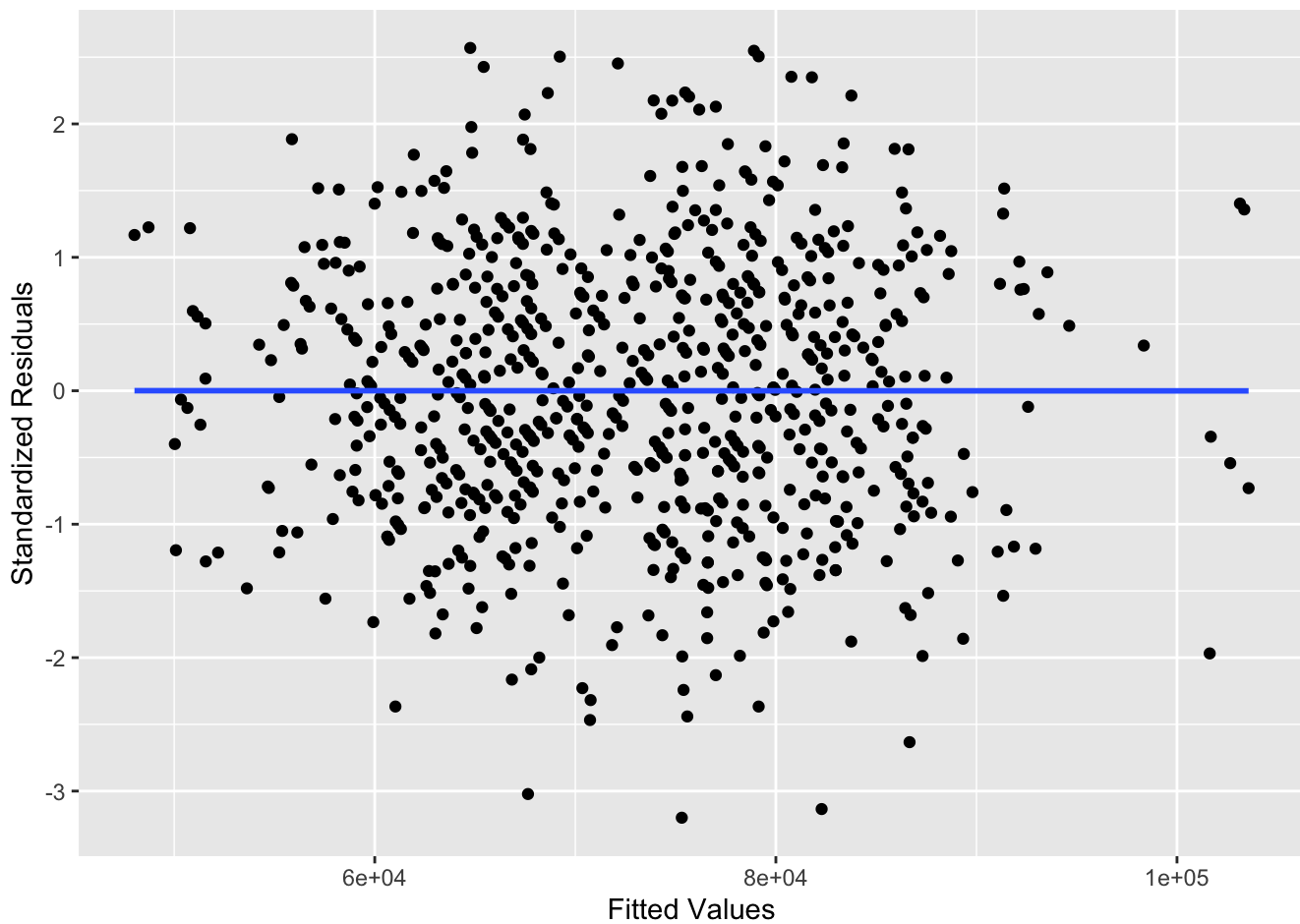


```
ks.test(std_resids, "pnorm")
```

```
##  
## Asymptotic one-sample Kolmogorov-Smirnov test  
##  
## data: std_resids  
## D = 0.024779, p-value = 0.7421  
## alternative hypothesis: two-sided
```

Because the histogram of the standardized residuals looks roughly normal and the p-value of the Kolmogorov-Smirnov test is greater than 0.05, we fail to reject the null hypothesis and conclude that the standardized residuals are normally distributed. Therefore, we will say this assumption is met.

Equal Variance Assumption



```
lmtest::bptest(sal.full)
```

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

We can see from the fitted values vs. standardized residuals scatter plot that the spread of the points at any value along the x-axis is roughly the same. Also, the Breusch-Pagan test produces a large p-value. This means we fail to reject the null hypothesis and have insufficient information to say that the variance of the data is unequal. Therefore, we will say this assumption is met.

6.

```
confint(sal.full,"GPA", level = 0.97)
```

```
##          1.5 %    98.5 %  
## GPA 4646.385 6129.755
```

```
confint(sal.full, "GenM", level = 0.97)
```

```
##          1.5 %    98.5 %  
## GenM 9395.567 24387.63
```

```
confint(sal.full, "MajorCategoryArts", level = 0.97)
```

```
##          1.5 %    98.5 %  
## MajorCategoryArts -3377.766 4805.189
```

The 97% confidence interval for the coefficient of GPA is (4646.385, 6129.755). This means that we are 97% confident that for every unit increase in GPA, while holding all other variables constant, average salary increases between \$4,646.39 and \$6,129.76.

The 97% confidence interval for the coefficient of GenM is (939.567, 24387.63). This means that we are 97% confident that a male's salary on average is between \$4,646.39 and \$6,129.76 higher than a female's while holding all other variables constant.

The 97% confidence interval for the coefficient of MajorCategoryArts is (-3377.766, 4805.189). This means that we are 97% confident that the average salary for the Arts major category is between \$3,377.77 lower and \$4,805.19 higher than the average salary for the baseline major (Agriculture and Natural Resources), while holding all else constant.

7.

```

a1 <- c(1, #intercept
        0,0,0,0,
        1, #computer and math
        0,0,0,0,0,0,0,0,0,0,
        1, #genM
        2.5, #gpa
        0,0,0,0,
        1, # math interaction man
        0,0,0,0,0,0,0,0,0,0)

a2 <- c(1, #intercept
        0,0,0,0,
        1, #computer and math
        0,0,0,0,0,0,0,0,0,0,
        0, #genF
        2.5, #gpa
        0,0,0,0,
        0, # math interaction man
        0,0,0,0,0,0,0,0,0,0)

# my.test <- multcomp::glht(sal.full, linfct=t(a1-a2), alternative="two.sided")
my.test <- multcomp::glht(sal.full, linfct=t(a1-a2), alternative="greater")
summary(my.test)

```

```

##
## Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Salary ~ MajorCategory + Gen + GPA + MajorCategory:Gen,
## data = sal)
##
## Linear Hypotheses:
## Estimate Std. Error t value Pr(>t)
## 1 <= 0 7904 1816 4.353 7.67e-06 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)

```

```

my.test.ci <- multcomp::glht(sal.full, linfct=t(a1-a2), alternative="two.sided")
confint(my.test.ci, level = 0.95)

```

```
##
## Simultaneous Confidence Intervals
##
## Fit: lm(formula = Salary ~ MajorCategory + Gen + GPA + MajorCategory:Gen,
## data = sal)
##
## Quantile = 1.9633
## 95% family-wise confidence level
##
## Linear Hypotheses:
## Estimate      lwr      upr
## 1 == 0  7904.2434  4339.6592 11468.8276
```

H0: Women's salary for computer and math = Men's salary for computer and math
 HA: Women's salary for computer and math < Men's salary for computer and math

Based on the results of our general linear hypothesis test, we have a p-value of 7.67e-06 so we reject the null hypothesis and conclude that women's salary for computer and math not equal to men's salary for computer and math major. The 95% confidence interval for the difference in average salary between men and women in the computer and math major category is (4339.6592,11468.8276). This means that we are 95% confident that the average salary of a man who majored in computer and math is between \$4,339.66 and \$11,468.83 higher than the average salary for women of the same major, while holding all else constant.

8.

```
new.josh <- data.frame(MajorCategory='Computers & Mathematics', GPA = 3.98, Gen="M")
predict.lm(sal.full, new.josh, interval="prediction", level=0.95)
```

```
##      fit      lwr      upr
## 1 93593.23 82932.69 104253.8
```

```
new.aaron <- data.frame(MajorCategory='Computers & Mathematics', GPA = 3.81, Gen="M")
predict.lm(sal.full, new.aaron, interval="prediction", level=0.95)
```

```
##      fit      lwr      upr
## 1 92677.26 82019.24 103335.3
```

```
new.anna <- data.frame(MajorCategory='Computers & Mathematics', GPA = 3.5, Gen="F")
predict.lm(sal.full, new.anna, interval="prediction", level=0.95)
```

```
##      fit      lwr      upr
## 1 83102.71 72612.75 93592.68
```

For Josh, who is a male computers & mathematics major with a 3.98 GPA we are 95% confident that his average salary lies somewhere between \$82,932.69 and \$104,253.80.

For Aaron, who is a male computers & mathematics major with a 3.81 GPA we are 95% confident that his average salary lies somewhere between \$82,019.24 and \$103,335.30.

For Anna, who is a female computers & mathematics major with a 3.5 GPA we are 95% confident that her average salary lies somewhere between \$72,612.75 and \$93,592.68.

9.

```
n <- nrow(sal)
rpmse <- rep(x=NA, times=n)
wid <- rep(x=NA, times=n)

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

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

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

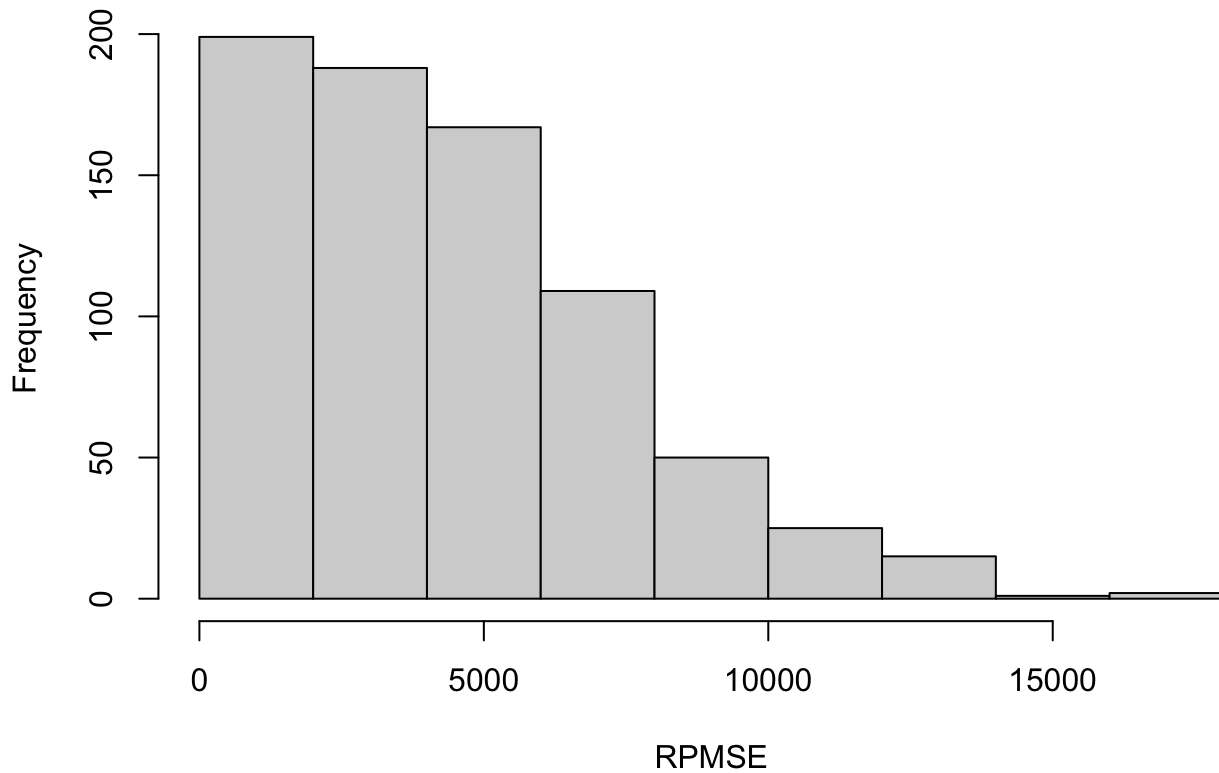
  ## Generate predictions for the test set
  my.preds <- predict.lm(train.lm, newdata=test.set, interval="prediction")

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

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

# RPMSE
hist(rpmse, main="RPMSE Histogram", xlab="RPMSE")
```


RPMSE Histogram



```
mean(rpmse) #rpmse is how off you are on average
```

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

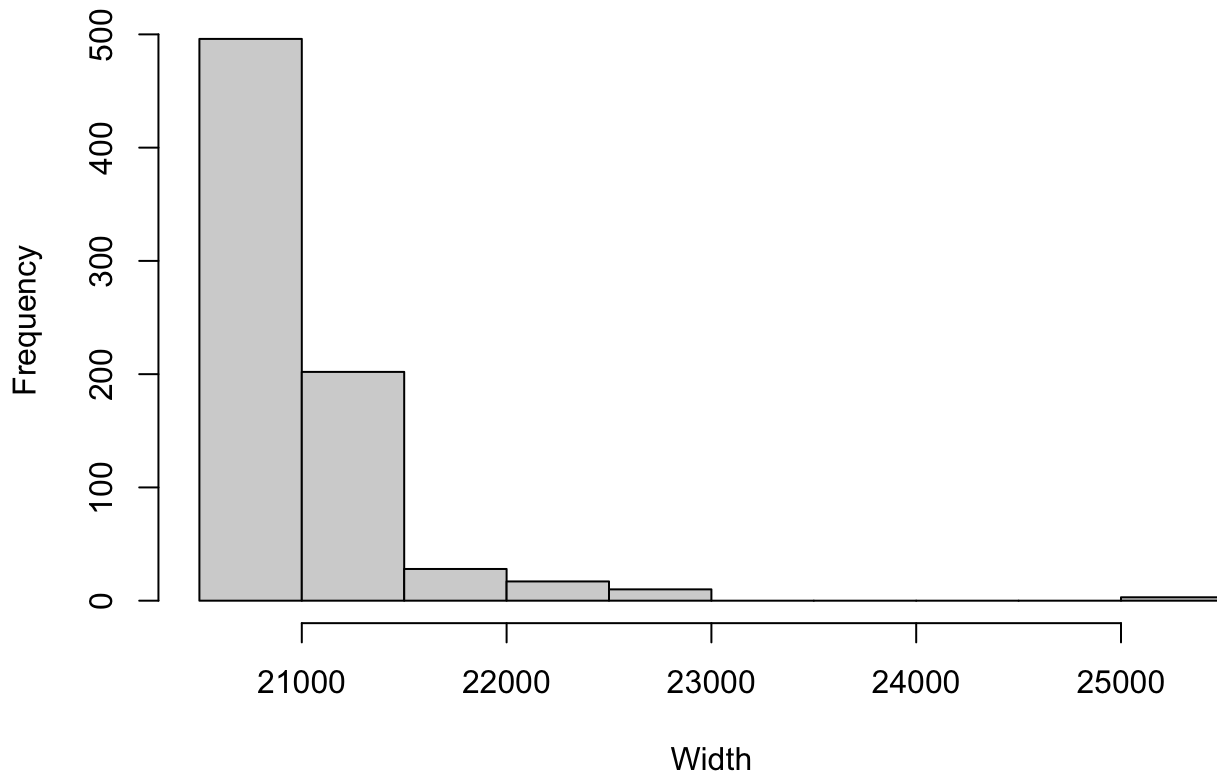
```
#standard deviation of salary = 10996.17
```

```
(var(sal$Salary) - mean(rpmse)^2 ) / var(sal$Salary) # 84.29% of overall variance reduction
```

```
## [1] 0.8429246
```

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

Width Histogram



```
mean(wid)
```

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

Using the average RPMSE from our Leave One Out Cross-validation and the variance of the salary in the dataset, we calculated an overall variance reduction of 84.29%.

The average width of the predicted intervals is \$21,013.43. This means that on average, the predictions are about \$10,506.72 over or under the actual salary.

Both the width and RPMSE histograms above are right-skewed so the average values we report above are made higher than most of the cross-validated values. Overall, we think our model is fairly accurate at predicting new salaries.