The Value of a College Education Data Analysis

Tetsuya Chau

1/25/2022

###Problem Background

What is the value of a college education? The answer very much depends on factors such as your major and GPA. Unfortunately, it can also unfairly depend upon external factors such as gender and race. For this analysis, you will be analyzing data from the American Community Survey Public Use Microdata Series on graduates and their annual salary 5 years after graduation. The dataset Salary.csv contains the following variables:

###Variable Name and Description

*Salary - Graduate's annual salary 5 year after graduation

*MajorCategory - Graduate's primary major listed on diploma

*Gen - Graduate's gender

*GPA - Graduate's overall grade point average

Analysis Questions:

1. Create exploratory plots and calculate summary statistics from the data. Comment on any potential relationships you see from these exploratory plots.
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.
3. Using first principles (i.e. DON'T use lm() but you can check your answer with lm()), calculate $\hat{\beta}$ 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).
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.
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.
6. Calculate 97% confidence intervals for the coefficients for GPA, Gender and one major category. Interpret each interval.

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

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.6     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.4     ✓ stringr 1.4.0
## ✓ readr   2.1.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#install.packages("MASS")
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
#install.packages("car")
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
#install.packages("multcomp")
library(multcomp)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

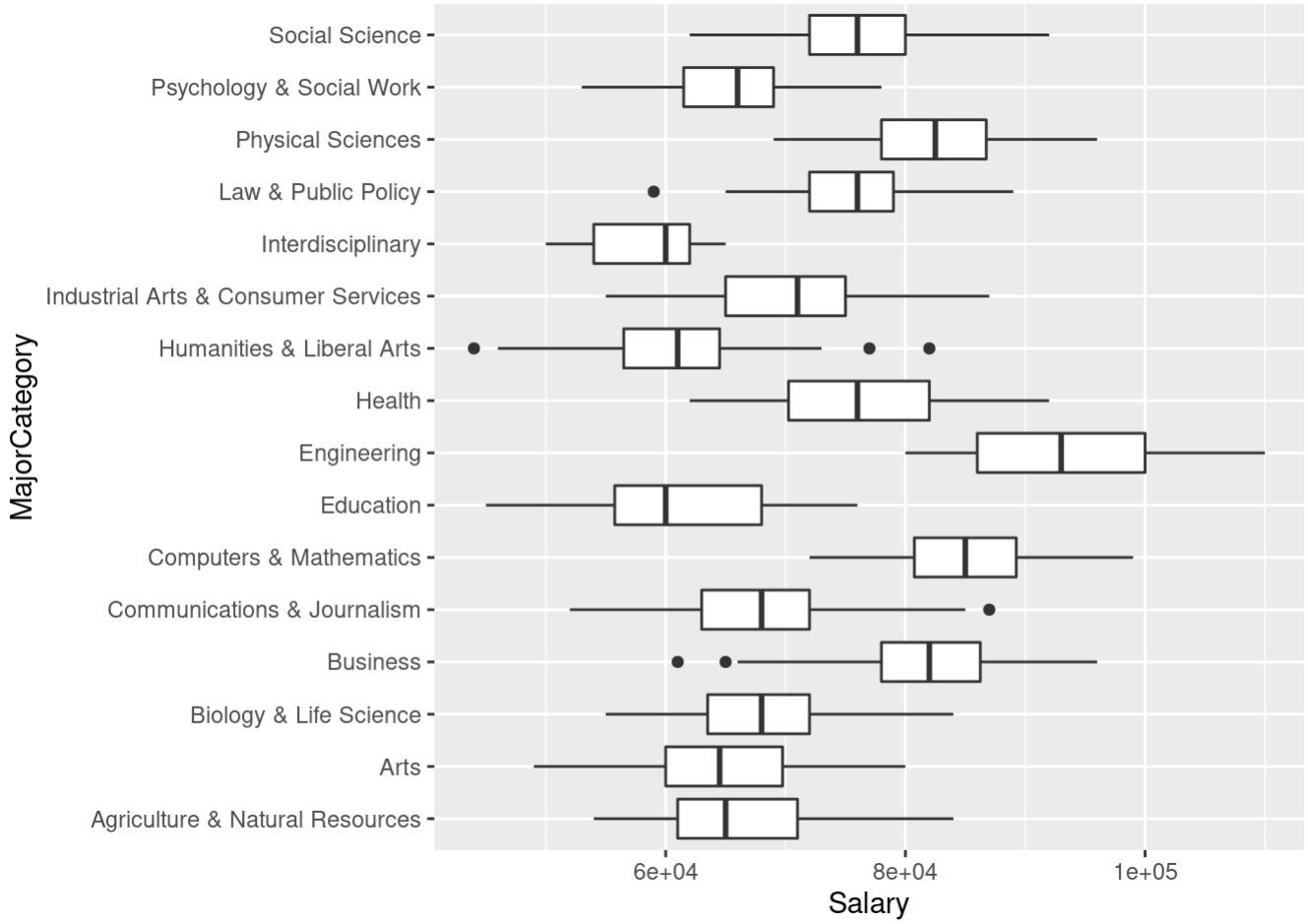
```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
myData <- read.csv("/cloud/project/Salary.csv")
```

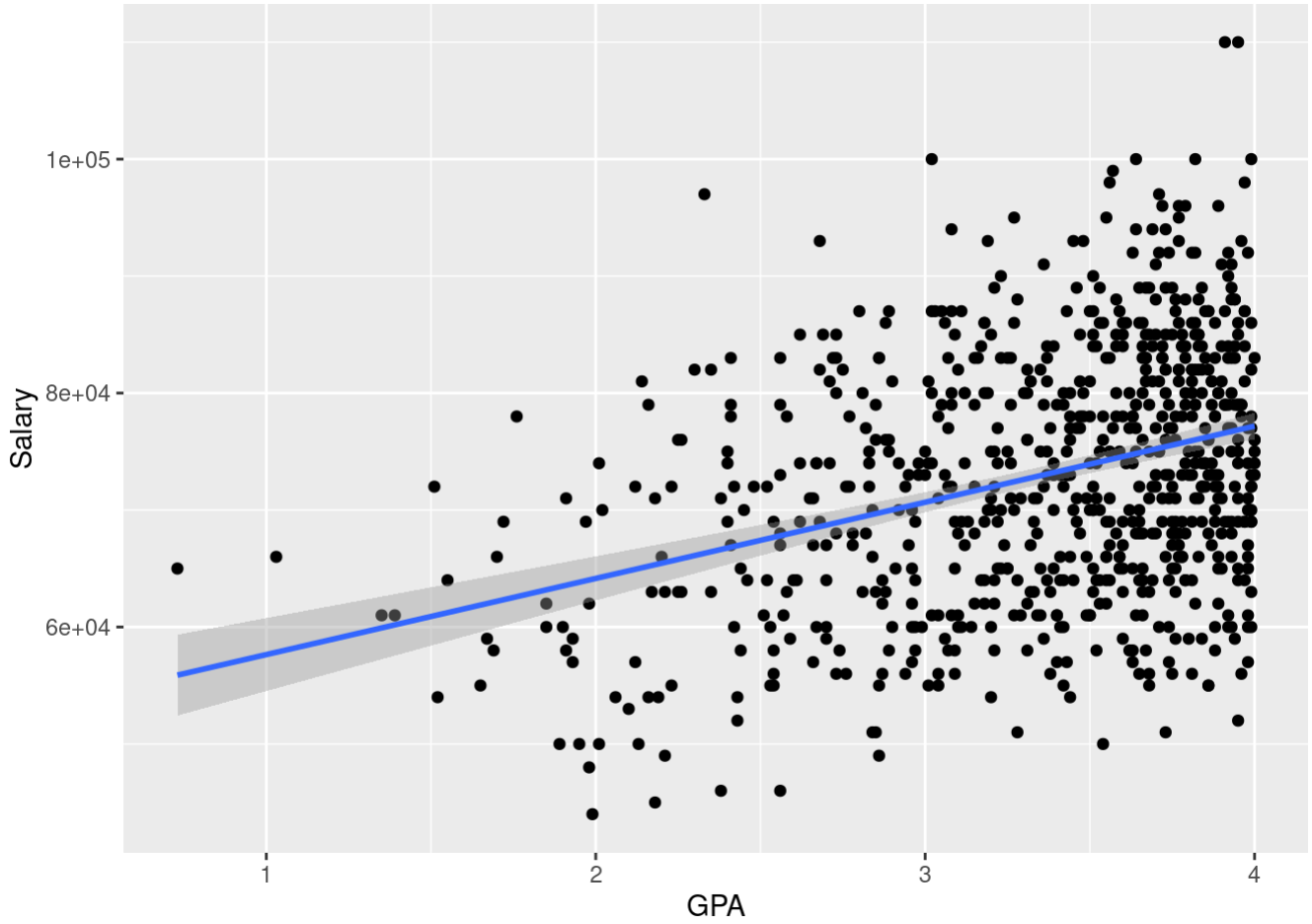
1. Create exploratory plots and calculate summary statistics from the data. Comment on any potential relationships you see from these exploratory plots.

```
#Salary vs MajorCategory side-by-side boxplots
ggplot(data=myData, mapping = aes(x=Salary, y=MajorCategory))+
 geom_boxplot()
```

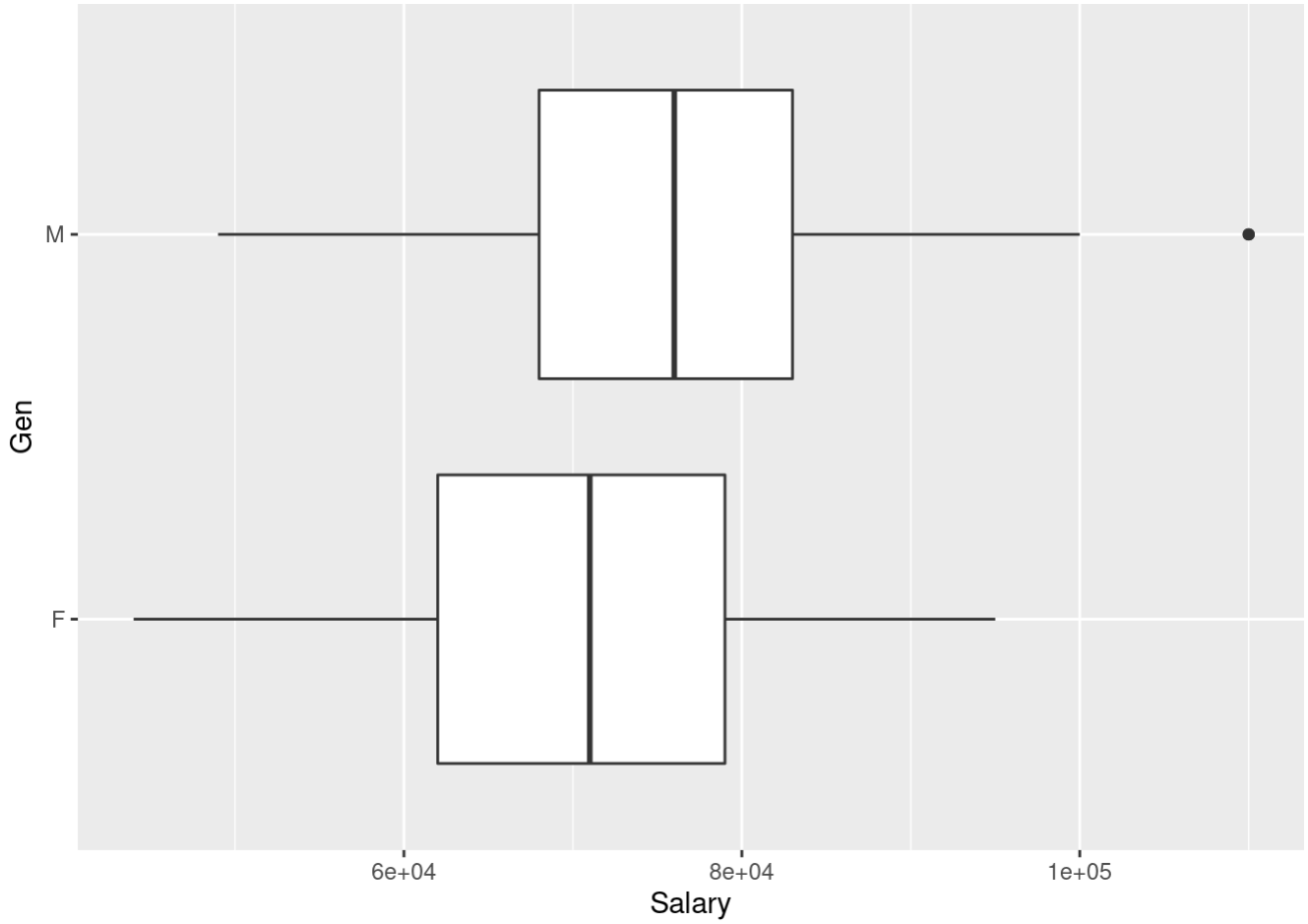



```
#GPA vs Salary scatterplot
ggplot(data=myData, mapping = aes(x=GPA, y=Salary))+
 geom_point()+geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Salary vs Gender side-by-side boxplots
ggplot(data=myData, mapping = aes(x=Salary, y=Gen))+
 geom_boxplot()
```




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

```
##
## Call:
## lm(formula = Salary ~ ., data = myData)
##
## 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
```

Based on the analyzing the graphs of the relationships between the salary and the various explanatory variables, I was able to observe that people that studied engineering, math/computers, and business were paid the most whereas people that studied interdisciplinary, humanities & liberal arts, and education were the paid the least. Also, GPA did not have a very strong correlation with the salary since many people that had a high GPA did not necessarily have a higher salary than people with a lower GPA. Lastly, I observed that males are generally paid a little more than females on average based on the side-by-side box plots that I created which shows the median, lower quartile, upper quartile, minimum, maximum, and the outlier.

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.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim MVN(0, \sigma^2 I)$$

*y is the vector of the true values

*X is the design matrix

*beta is the vector of coefficients

*epsilon is the error vector

We can run tests such as a t-test and ANOVA in order to find the p-values that tell us whether or not the choice of major and gender affect the salary of the individual with a 95% confidence interval (alpha = 0.05).

3. Using first principles (i.e. DON'T use lm() but you can check your answer with lm()), calculate βˆ 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).

```
#find the beta_hat without using lm()
X <- model.matrix(object=Salary~., data=myData)
beta_hat <- solve(t(X)%*%X)%*%t(X)%*%myData$Salary
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
```

```
#verify result using lm().
print(myData.lm)
```

```
##
## Call:
## lm(formula = Salary ~ ., data = myData)
##
## 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
```

```
#find the variance without using lm().
sigma_squared <- ((t(myData$Salary-X%*%beta_hat))%*%(myData$Salary-X%*%beta_hat))/(756-17-1)
print(sigma_squared)
```

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

```
#verify variance using lm().
sigma(myData.lm)**2
```

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

R-squared values:

Multiple R-squared: 0.7637, Adjusted R-squared: 0.7583 (this was completed in step 1 using summary())

Interpretations of the coefficients:

*Suppose the x-variable for beta 1 is a categorical variable, gender.

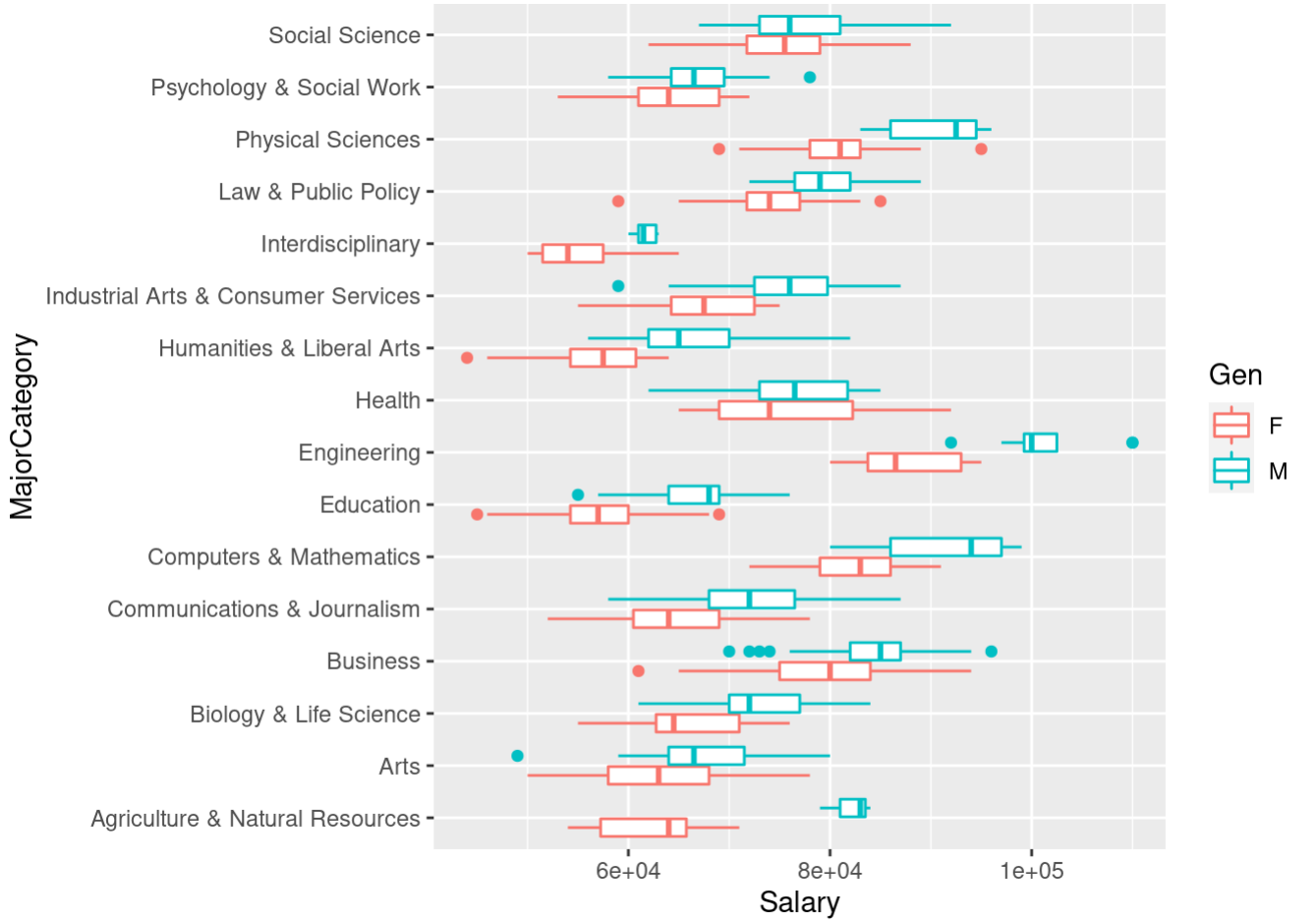
If the gender is male, the salary will increase by beta 1 on average.

*Suppose the x-variable for beta 2 is a numeric variable, GPA.

If the GPA goes up by 1 unit, the salary will increase by beta 2 on average.

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.

```
ggplot(data=myData, mapping = aes(x=Salary, y=MajorCategory, color=Gen))+
 geom_boxplot()
```



Based on the graph above, men are paid more than women on average.

```
noint.lm <- lm(formula = Salary ~ MajorCategory + Gen + GPA, data = myData)
int.lm <- lm(formula = Salary ~ MajorCategory * Gen + GPA, data = myData)
anova(noint.lm,int.lm)
```

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

```
#summary(int.lm)
```

Null hypothesis: There's no interaction between gender and discipline

Alternative hypothesis: There is an interaction between gender and discipline

With a F-test statistic of 4.3595 and a p-value of 7.161e-08, we can reject the null hypothesis and conclude that there is an interaction between gender and discipline.

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.

Assumptions:

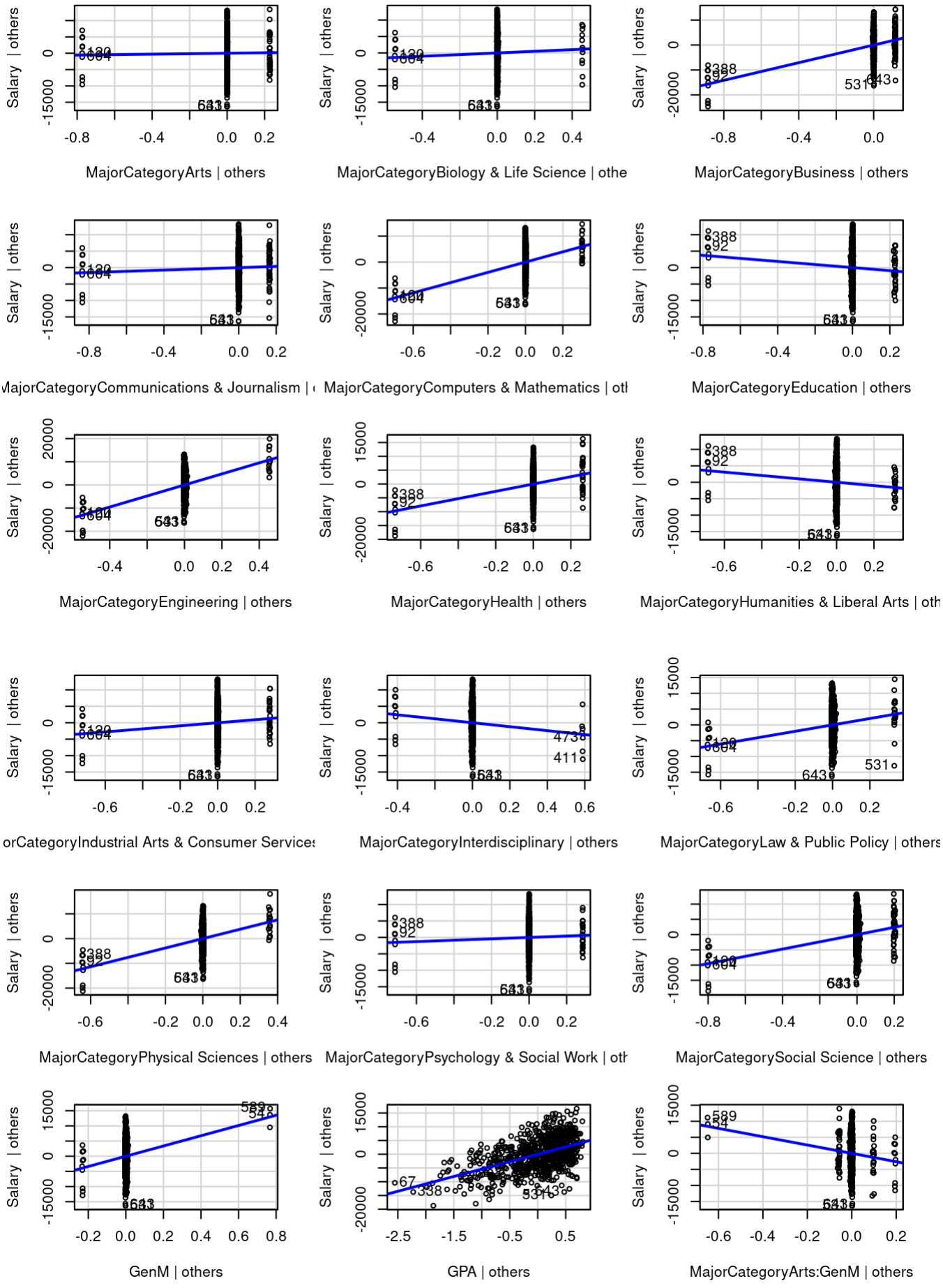
Linearity-It is assumed by the linear relationship between salary and GPA in the AV plot.

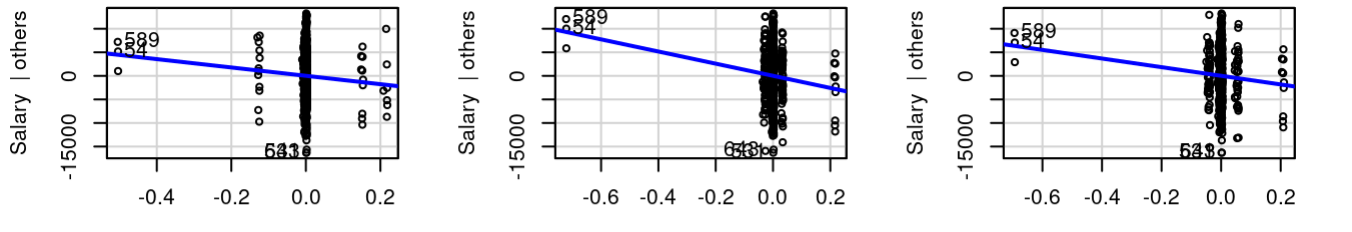
Independence-We can assume that the salary of one individual does not affect the salary of another individual.

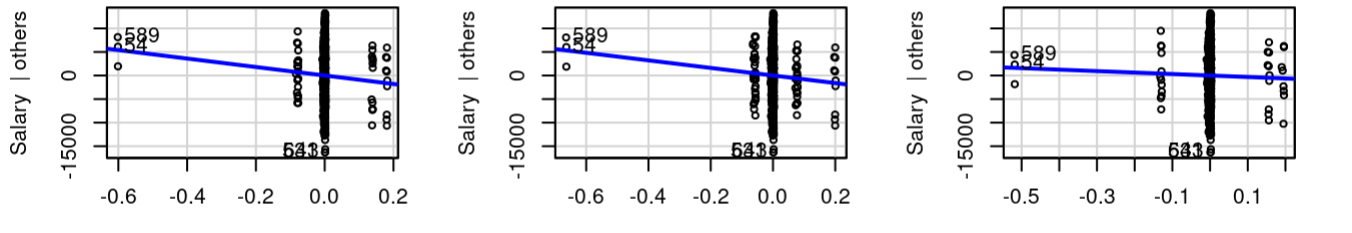
Normality-The histogram has a bell curve which implies that the data is normal.

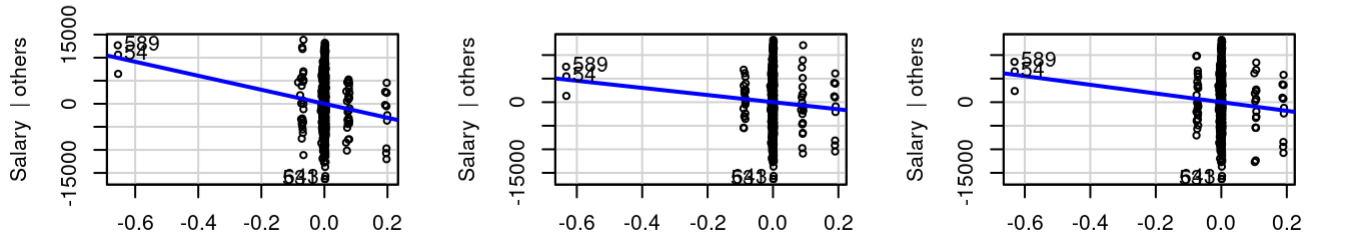
Equal Variance-The dots in the fitted vs residual plot are relatively evenly scattered throughout the graph which implies equal variance.

```
#AV plot to check for linearity
avPlots(int.lm, ask=FALSE)
```

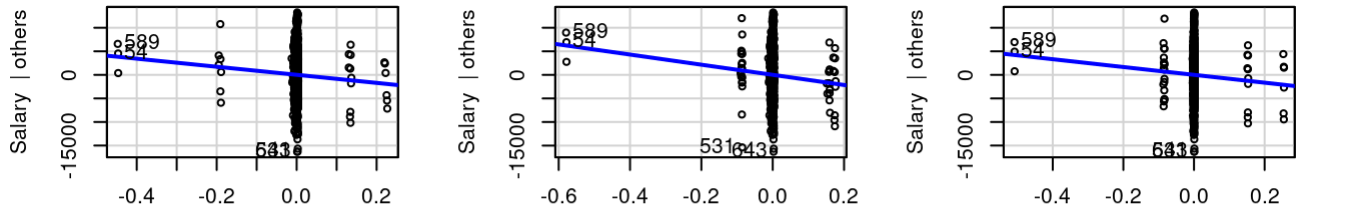

Salary | others
MajorCategoryArts | others
MajorCategoryBiology & Life Science | othe
MajorCategoryBusiness | others
MajorCategoryCommunications & Journalism |
MajorCategoryComputers & Mathematics | otl
MajorCategoryEducation | others
MajorCategoryEngineering | others
MajorCategoryHealth | others
MajorCategoryHumanities & Liberal Arts | oth
orCategoryIndustrial Arts & Consumer Service
MajorCategoryInterdisciplinary | others
MajorCategoryLaw & Public Policy | others
MajorCategoryPhysical Sciences | others
MajorCategoryPsychology & Social Work | otl
MajorCategorySocial Science | others
GenM | others
GPA | others
MajorCategoryArts:GenM | others

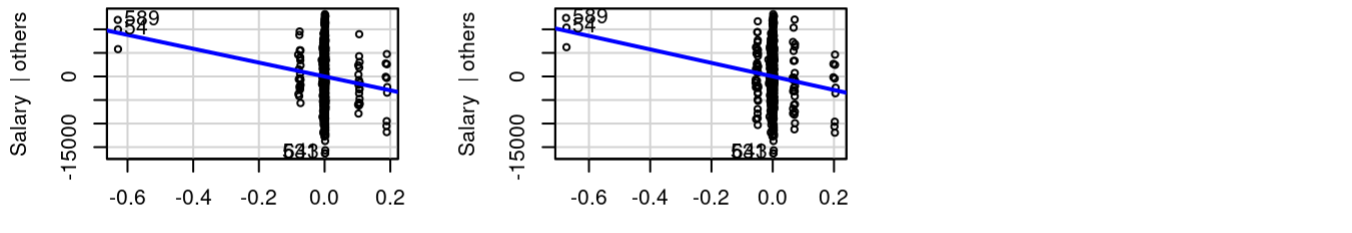




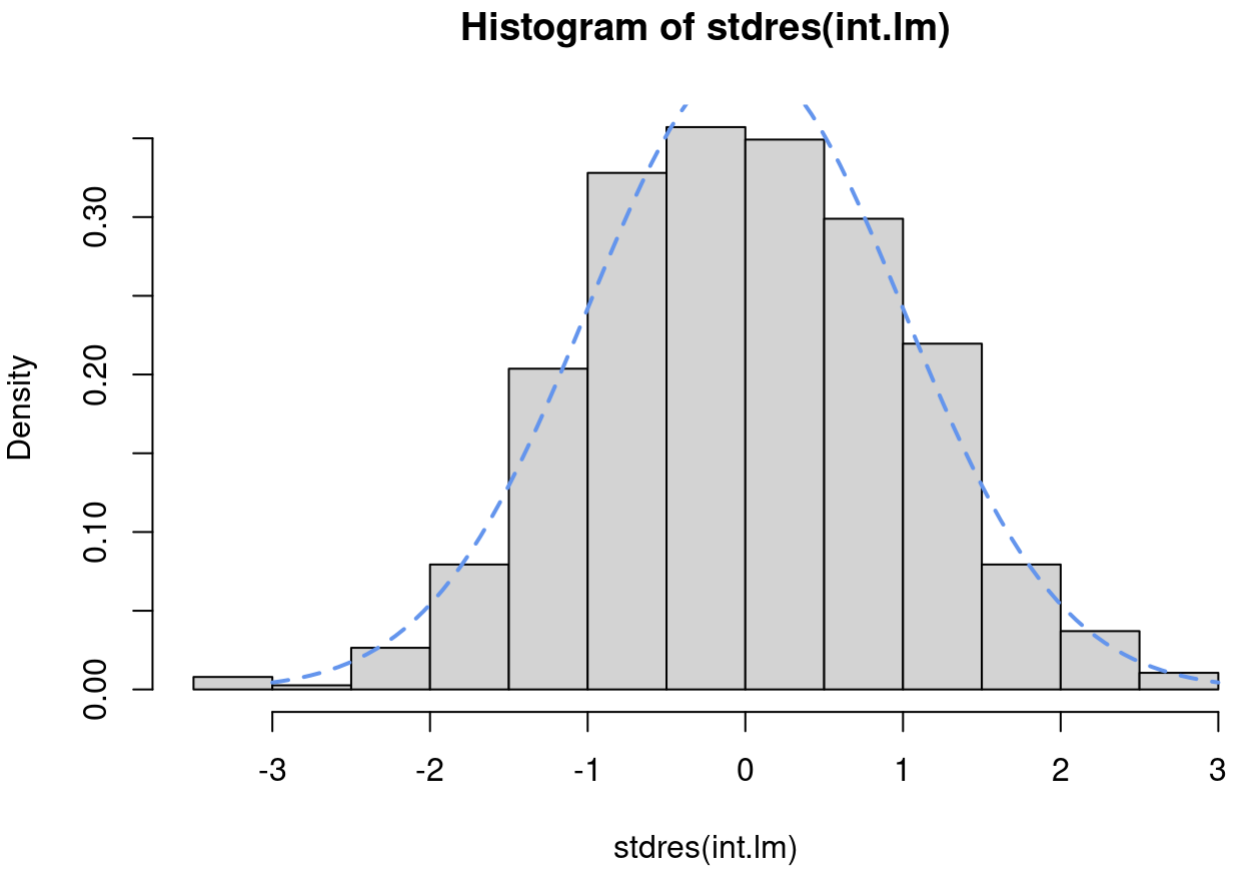


Added-Variable Plots

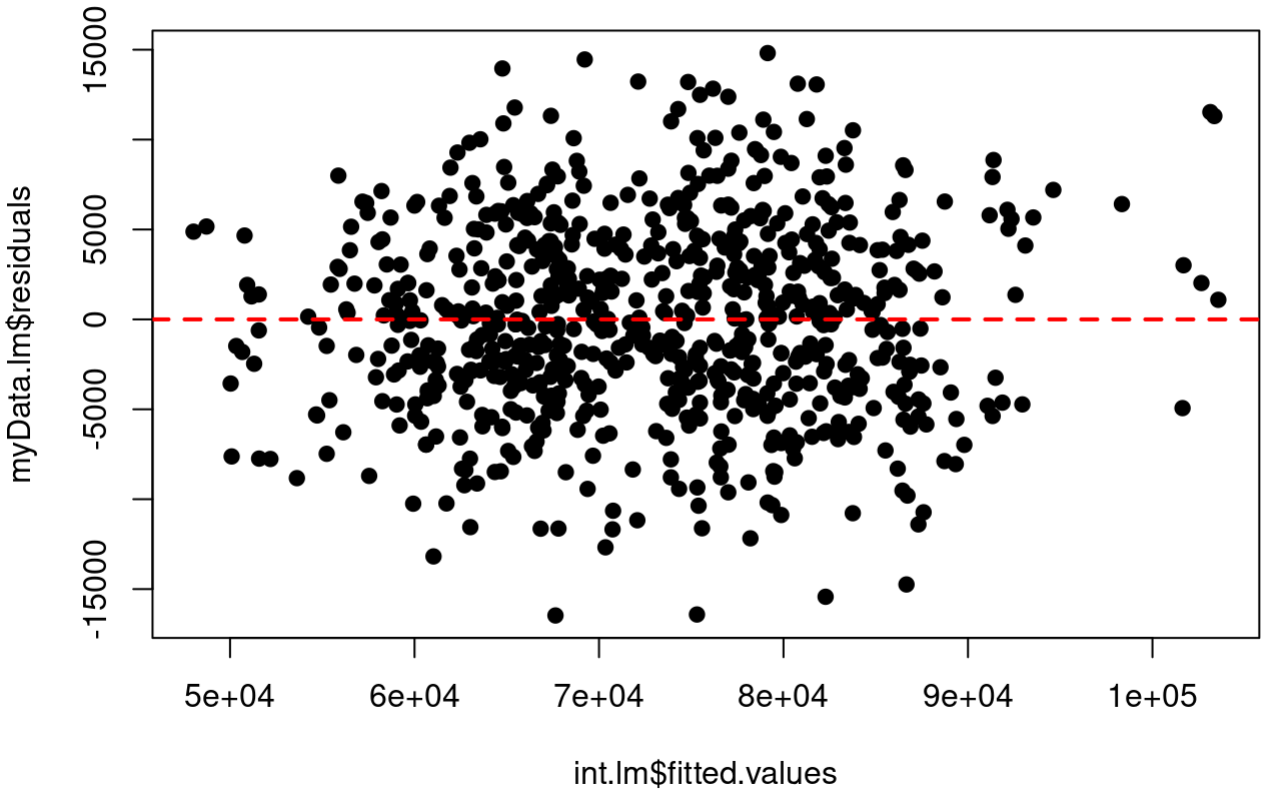




```
#Histogram to check for normality
hist(stdres(int.lm),freq = FALSE,breaks = 20)
curve(dnorm,from = -3,to = 3,col = "cornflowerblue",lwd = 2,
lty = 2,add = TRUE)
```




```
#Fitted vs Residual Plot to check for equal variance
plot(int.lm$fitted.values,myData.lm$residuals,pch=19)
abline(0,0,col = "red",lwd = 2,lty = 2)
```



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

```
confint(int.lm,level=0.97,parm="GPA")
```

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

```
confint(int.lm,level=0.97,parm="GenM")
```

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

```
confint(int.lm,level=0.97,parm="MajorCategorySocial Science")
```

```
##                                1.5 %   98.5 %
## MajorCategorySocial Science 7854.447 15903.25
```

We are 97% confident that the true mean increase in salary for every GPA unit gained is between 4646.385 and 6129.755.

We are 97% confident that mens' true mean salary is greater than womens' true mean salary by between 9395.567 24387.63 on average.

We are 97% confident that the true mean of salary for social science is between 7854.447 and 15903.25 greater than the agriculture and natural resources on average.

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.

```
male <- c(1,numeric(4), 1, numeric(10), 1, 1, numeric(4), 1, numeric(10))
female <- c(1,numeric(4), 1, numeric(10), 0, 1, numeric(4), 0, numeric(10))

a.transpose <- t(male - female)
my.test <- glht(int.lm, linfct = a.transpose, alternative = "greater")
summary(my.test)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Salary ~ MajorCategory * Gen + GPA, data = myData)
##
## 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)
```

```
confint(my.test, level = 0.95)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = Salary ~ MajorCategory * Gen + GPA, data = myData)
##
## Quantile = -1.647
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##        Estimate  lwr       upr
## 1 <= 0 7904.2434 4913.9264       Inf
```

Null Hypothesis: Holding all else constant, women earn the same as men.

Alternative Hypothesis: Holding all else constant, men earn more than women.

With a t-test statistic of 4.353 and a p-value 7.67e-06, we can reject the null hypothesis and conclude that men are paid more than women on average.

Holding all else constant, we are 95% confident that the true mean of men's salary is between 4913.9264 and infinity greater than women's true salary.

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

```
pData = data.frame(MajorCategory="Computers & Mathematics", Gen="M", GPA = 3.5)
predict.lm(int.lm, newdata=pData, interval="prediction", level=0.95)
```

```
##        fit      lwr      upr
## 1 91006.96 80350.42 101663.5
```

I predict that as a male student who majored in Computers & Mathematics with a 3.5 GPA, my salary will be 91006.96 in USD with a 95% prediction interval of (80350.42, 101663.5) in USD.

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.

```
set.seed(456)
n=756
n.cv <- 756 #Number of CV studies to run
n.test <- 1 #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:n, size=n.test)

 ## Split into test and training sets
 test.set <- myData[cv,]
 train.set <- myData[-cv,]

 ## 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.pred <- predict.lm(train.lm, newdata=test.set, interval="prediction")

 ## Calculate bias
 bias[cv] <- mean(my.pred[,'fit']-test.set[['Salary']])

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

 ## Calculate Coverage
 cvg[cv] <- ((test.set[['Salary']] > my.pred[,'lwr']) & (test.set[['Salary']] < my.pred[,'up
r'])) %>% mean()

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

}

mean_rpmse <- mean(rpmse)
print(mean_rpmse)
```

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

```
mean_width <- mean(wid)
print(mean_width)
```

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

The root squared mean error turned out to be 4358.084 which means that my predictions missed the mark by an average of 4358 usd. Considering the scale of the data, 4358 isn't a very significant error, I can conclude that the predictions are pretty accurate. We also predicted the width to be 21013.43 so we can say that my predictions are accurate or at least close to accurate.