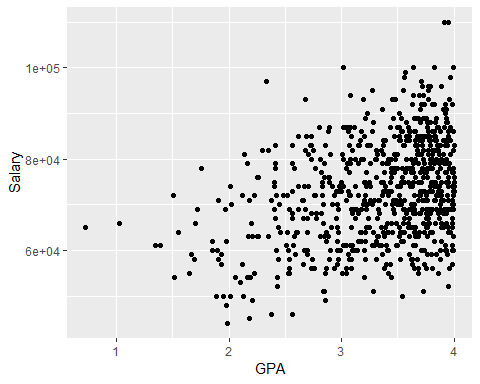
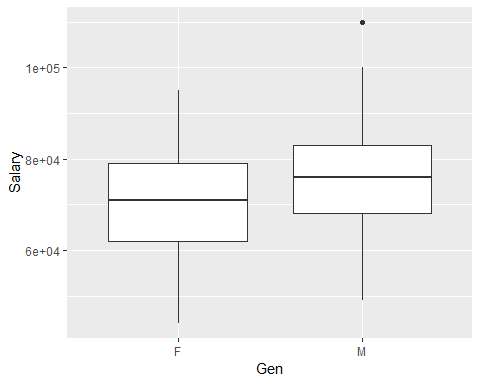
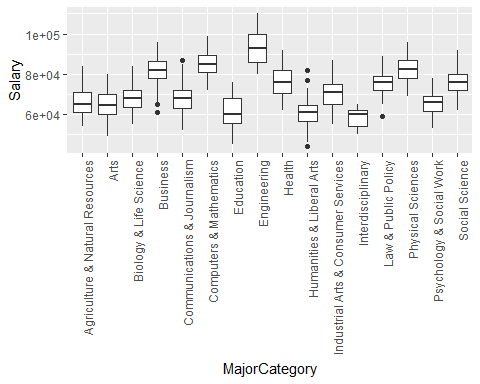
HW 1

Jonah Meherg

# 1







There appears to be some correlation between GPA and Salary after graduation. There appears that men on average make more than females, there also appears that some majors make more than others.

# 2

The linear regression model is y = X B + e, where y is a n x 1 vector of the salaries, X is a n x (P + 1) matrix of the explanatory variables, B is the (P + 1) x 1 vector of coefficients, and e is the n x 1 vector of residuals. Where e is distributed normally with a mean of 0 and variance of sigma-squared \* I, where I is the n x n identity matrix.

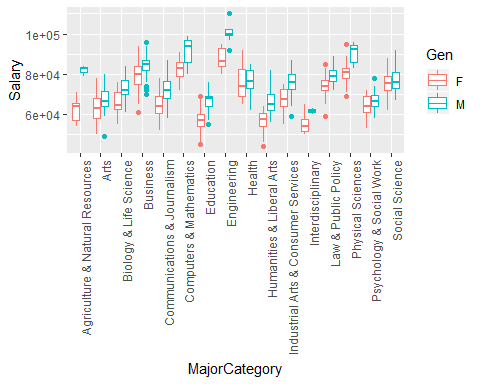
By making statistical inference on the B we will see how GPA, gender and major affect salary 5 years after graduation.

# 3

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

The left column shows the betahat coefficients using first principles, the right is the betahats using lm(), (code given at end). The effect of Biology and Life science majors on the salary differs by $769.1305 on average compared to Agriculture which is the baseline, or they make on average $769.1305 more than agriculture students, while holding all other variables constant. As GPA goes up by 1, the average salary for students, holding all else constant, goes up by $5488.7368.

# 4

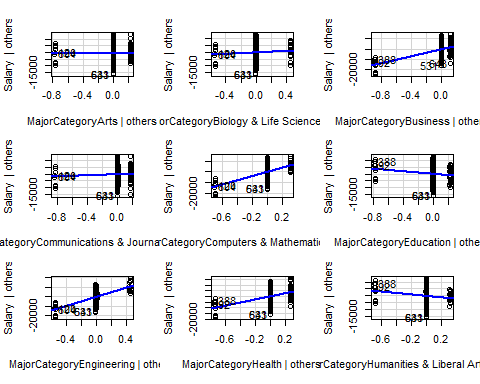


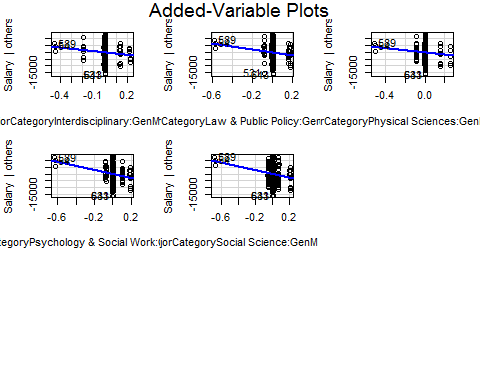
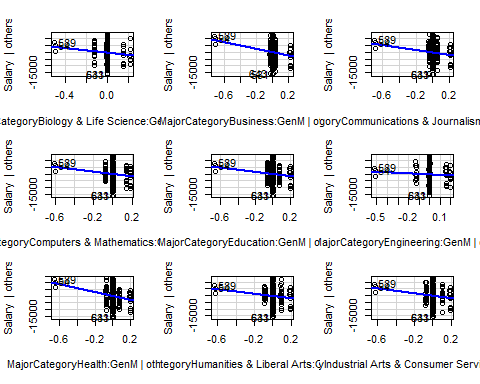
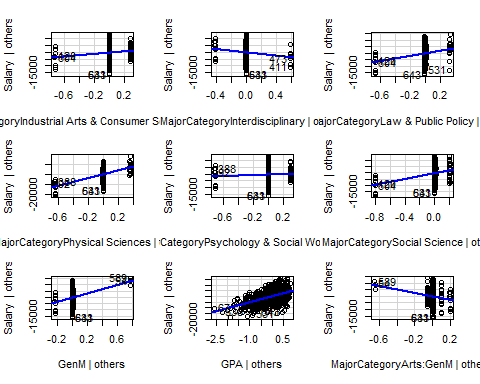
The boxplot shows a clear possible interaction between gender and major affecting salary, as there are some majors that have differing salaries depending on gender, while other majors have similar average salaries across gender.

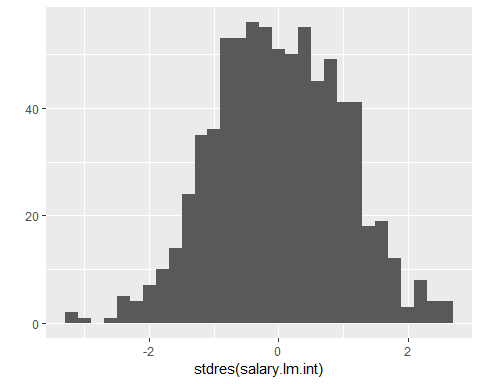
## Analysis of Variance Table  
##   
## Model 1: Salary ~ MajorCategory + Gen + GPA  
## Model 2: Salary ~ MajorCategory \* Gen + GPA  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 738 2.1569e+10   
## 2 723 1.9780e+10 15 1789058098 4.3595 7.161e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001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

To determine this interaction I ran an F test between MajorCategory and Gender interaction and their affect on salary to determine if there was an interaction. The null hypothethis is that there is no interaction between major and gender, and the alternative hypothesis is that there is at least one interaction between major and gender. The interaction f test gives an f-statistic of 4.3595 with a p-value of essentially zero, rejecting the null hypothesis and determining that there is an interaction between at least one major and gender. Thus there is some form of gender discrimination.

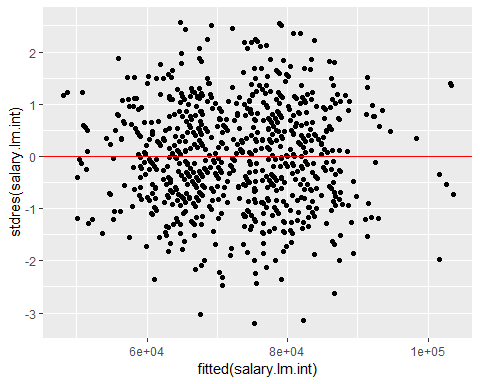
# 5







## One-sample Kolmogorov-Smirnov test  
##   
## data: stdres(salary.lm.int)  
## D = 0.024779, p-value = 0.7421  
## alternative hypothesis: two-sided



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

Linearity - the relationship between y and all explanatory variables is linear (only applies to quantitative explanatory variables). Independence - the salary of one person does not impact the salary of another. Normality - the residuals e are normall distributed. Equal variance - the residuals have equal variation across the whole fitted line.

Linearity - the av plots appear to be linear across the variables. Independence - There is no information showing that there is dependence across the different salaries, thus we can assume that there is independence. Normality - the histogram of standardized residuals appears to be normal and the ks test passes. Equal variance - fitted values vs. standardized residuals appears to have same spread and the BP test passes.

# 6

## 1.5 %  
## (Intercept) 40129.6692  
## MajorCategoryArts -3377.7656  
## MajorCategoryBiology & Life Science -2350.4808  
## MajorCategoryBusiness 13870.5355  
## MajorCategoryCommunications & Journalism -2066.1082  
## MajorCategoryComputers & Mathematics 15446.1187  
## MajorCategoryEducation -8752.2253  
## MajorCategoryEngineering 18939.6047  
## MajorCategoryHealth 9016.9184  
## MajorCategoryHumanities & Liberal Arts -9463.9242  
## MajorCategoryIndustrial Arts & Consumer Services 401.9554  
## MajorCategoryInterdisciplinary -11629.6010  
## MajorCategoryLaw & Public Policy 5719.5865  
## MajorCategoryPhysical Sciences 14474.3521  
## MajorCategoryPsychology & Social Work -2150.1863  
## MajorCategorySocial Science 7854.4475  
## GenM 9395.5671  
## GPA 4646.3854  
## MajorCategoryArts:GenM -21127.0156  
## MajorCategoryBiology & Life Science:GenM -18106.2633  
## MajorCategoryBusiness:GenM -20614.1306  
## MajorCategoryCommunications & Journalism:GenM -17050.2062  
## MajorCategoryComputers & Mathematics:GenM -17454.5833  
## MajorCategoryEducation:GenM -16078.8759  
## MajorCategoryEngineering:GenM -12242.5310  
## MajorCategoryHealth:GenM -23289.3746  
## MajorCategoryHumanities & Liberal Arts:GenM -15815.9165  
## MajorCategoryIndustrial Arts & Consumer Services:GenM -17473.7806  
## MajorCategoryInterdisciplinary:GenM -18386.6301  
## MajorCategoryLaw & Public Policy:GenM -19380.9872  
## MajorCategoryPhysical Sciences:GenM -17507.5252  
## MajorCategoryPsychology & Social Work:GenM -22984.3081  
## MajorCategorySocial Science:GenM -22342.1237  
## 98.5 %  
## (Intercept) 48850.4391  
## MajorCategoryArts 4805.1888  
## MajorCategoryBiology & Life Science 7390.2497  
## MajorCategoryBusiness 21517.0675  
## MajorCategoryCommunications & Journalism 5801.0814  
## MajorCategoryComputers & Mathematics 24062.7079  
## MajorCategoryEducation -567.1345  
## MajorCategoryEngineering 28686.3302  
## MajorCategoryHealth 17396.7626  
## MajorCategoryHumanities & Liberal Arts -786.4691  
## MajorCategoryIndustrial Arts & Consumer Services 8866.6584  
## MajorCategoryInterdisciplinary -416.0322  
## MajorCategoryLaw & Public Policy 14538.5271  
## MajorCategoryPhysical Sciences 23448.3907  
## MajorCategoryPsychology & Social Work 6360.8855  
## MajorCategorySocial Science 15903.2493  
## GenM 24387.6262  
## GPA 6129.7553  
## MajorCategoryArts:GenM -4818.9441  
## MajorCategoryBiology & Life Science:GenM 387.8554  
## MajorCategoryBusiness:GenM -5155.2507  
## MajorCategoryCommunications & Journalism:GenM -1278.5001  
## MajorCategoryComputers & Mathematics:GenM -520.1231  
## MajorCategoryEducation:GenM 35.8744  
## MajorCategoryEngineering:GenM 5996.1670  
## MajorCategoryHealth:GenM -7058.5824  
## MajorCategoryHumanities & Liberal Arts:GenM 698.6455  
## MajorCategoryIndustrial Arts & Consumer Services:GenM -942.7193  
## MajorCategoryInterdisciplinary:GenM 1260.1467  
## MajorCategoryLaw & Public Policy:GenM -2115.9150  
## MajorCategoryPhysical Sciences:GenM 913.1989  
## MajorCategoryPsychology & Social Work:GenM -6429.0842  
## MajorCategorySocial Science:GenM -6344.5053

As GPA goes up by 1 I am 97% confident that average salary increases between $4646.39854 and $6129.7553. For men in agriculture(the baseline), I am 97% confident that average salary is between $9395.5671 and $24387.6262 higher than for women. For women majoring in social science, I am 97% confident that the average salary increases between $7854.4475 and $15903.2493, compared to women in the agriculture major.

# 7

For the null hypothesis, B(GenM) + B(CompMath:GenM) = 0, and alternative, B(GenM) + B(CompMath:GenM) > 0, we perform a t test. The p-value is close to zero and thus we can reject the null hypothesis and determine that men on average make more than women in the Computers and Mathematics major category.

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

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

For a 95% confidence interval for how much men make more than women, it is: (4339.6592, 11468.8276) dollars.

# 8

## fit lwr upr  
## 1 87774.11 77108.96 98439.27

My predicted salary is $87,774.11 5 years after graduation. My 95% prediction interval for salary 5 years after graduation is: (77108.96,98439.27)

# 9

With a leave one out cross validation the RPMSE is about 4358 with a width of 21013.43, although the rpmse is only off by the range of salary by a small amount, the width suggests that there is much uncertainty that suggest the predictions are not very accurate.

.

Code

#1  
salarydata <- read.csv("Salary.csv", header = TRUE)  
ggplot(data=salarydata, mapping=aes(x=GPA,y=Salary)) + geom\_point()

ggplot(data=salarydata, aes(x=Gen,y=Salary)) + geom\_boxplot()

ggplot(data=salarydata, mapping=aes(x=MajorCategory,y=Salary)) + geom\_boxplot() +   
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

#3  
X <- model.matrix(Salary~.,data=salarydata)  
y <- salarydata[['Salary']]  
beta.hat <- solve(t(X)%\*%X)%\*%t(X)%\*%y  
lm.betahat <- coef(lm(Salary~.,data=salarydata))  
rownames(beta.hat) <- colnames(X)  
cbind(beta.hat,lm.betahat)

#4

ggplot(data=salarydata, mapping=aes(x=MajorCategory, y=Salary, color=Gen)) + geom\_boxplot() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

salary.lm.int <- lm(Salary~MajorCategory\*Gen+GPA,data=salarydata)  
salary.lm.no.int <- lm(Salary~MajorCategory+Gen+GPA,data=salarydata)  
anova(salary.lm.no.int,salary.lm.int)

#5

avPlots(salary.lm.int)

qplot(stdres(salary.lm.int), geom="histogram")

ks.test(stdres(salary.lm.int), "pnorm")

qplot(fitted(salary.lm.int), stdres(salary.lm.int), geom="point") +  
 geom\_hline(yintercept=0, color="red")

bptest(salary.lm.int)

#6

confint(salary.lm.int, level=0.97)

#7

a.glht <- matrix(0, nrow=length(coef(salary.lm.int)), ncol=1)  
a.glht[names(coef(salary.lm.int))%in%c("GenM","MajorCategoryComputers & Mathematics:GenM")] <- 1  
summary(glht(salary.lm.int, linfct=t(a.glht), alternative="greater"))

confint(glht(salary.lm.int, linfct=t(a.glht)))

#8

predict.lm(salary.lm.int, newdata=data.frame(Gen="M",MajorCategory="Computers & Mathematics", GPA=2.9),  
 interval="prediction")

#9

rpmse <- rep(NA,nrow(salarydata))  
wid <- rep(NA,nrow(salarydata))  
for(cv in 1:nrow(salarydata)){  
 train.lm <- lm(Salary~MajorCategory\*Gen+GPA,data=salarydata[-cv,])  
 preds <- predict.lm(train.lm, newdata=salarydata[cv,], interval="prediction")  
 rpmse[cv] <- (salarydata[cv,'Salary']-preds[,'fit'])^2 %>% mean() %>% sqrt()  
 wid[cv] <- preds[,'upr'] - preds[,'lwr'] %>% mean()  
}  
mean(rpmse)

mean(wid)