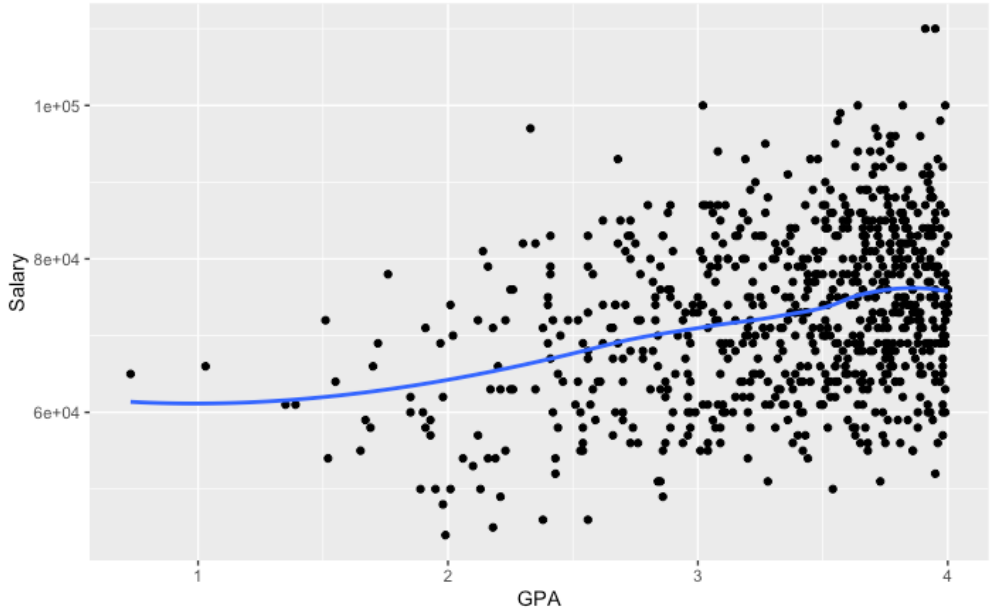
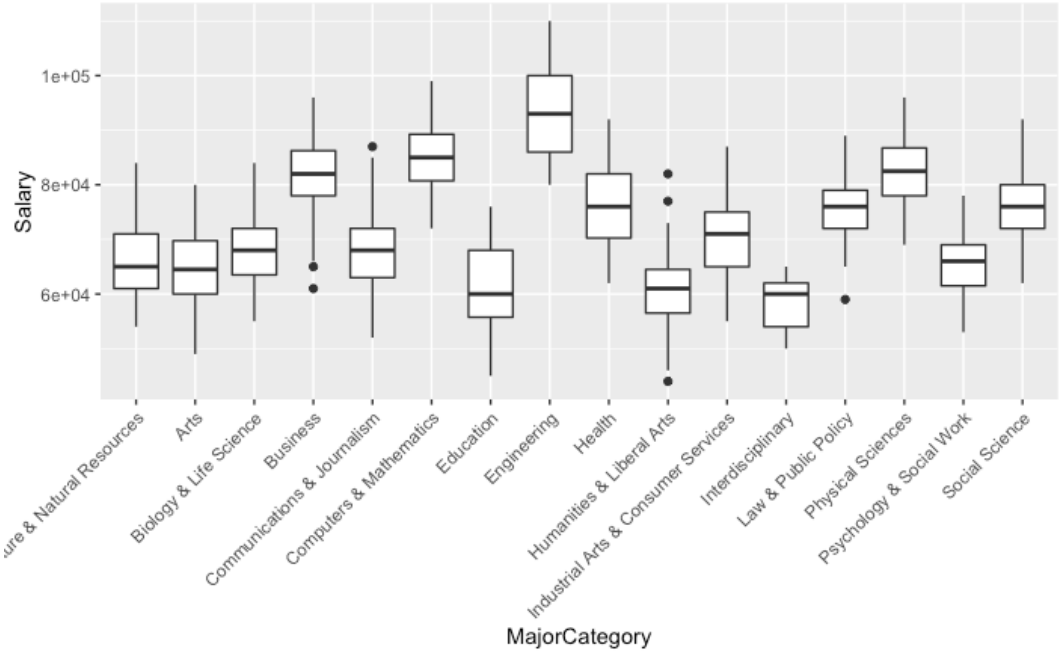
**The Value of a College Education**
Statistics 469: Analysis of Correlated Data
Oscar Briones Ramirez

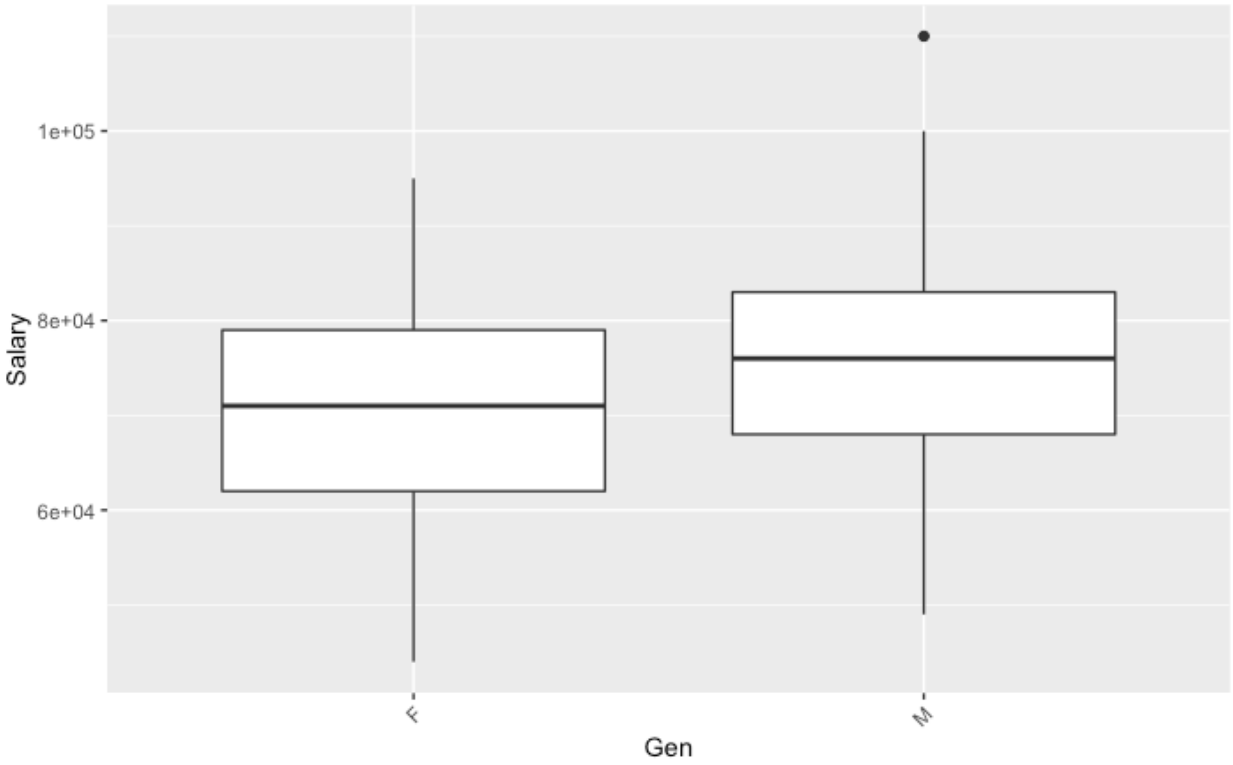
1. Create exploratory plots and calculate summary statistics from the data. Comment on any potential relationships you see from these exploratory plots.




We can see in this scatter plot that, as GPA increases, Salary increases but only slightly. It doesn't seem to have a strong correlation




We can see in this box plot that there are definitely some majors that pay better than others.




We can see on this box plot that men seem to make more than women.

The correlation between GPA and Salary is 0.3387171, which is not too strong.

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.

$y = X\beta + E$
$E \sim MVN(0, \text{sigma}^2 I)$

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

X is the matrix containing the explanatory variables.
$\beta$ is the matrix containing the coefficients associated with each explanatory variable.
E is the matrix containing the residuals, which is the error on how far the predictions are from the actual data. They are normally distributed.
Sigma^2 is the variance
I is the identity matrix

Statistical inference can be used to answer the questions by running tests to see first if our variables have significant effects on salary, and then by modeling we can find how much of an effect does major, gpa and gender have on the salary of a person after 5 years of graduating.

3. Using first principles (i.e. DON'T use lm() but you can check your answer with lm() ), calculate 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 R^2 (you can use lm() to get R^2).

| | V1 |
|---|---|
| (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 |

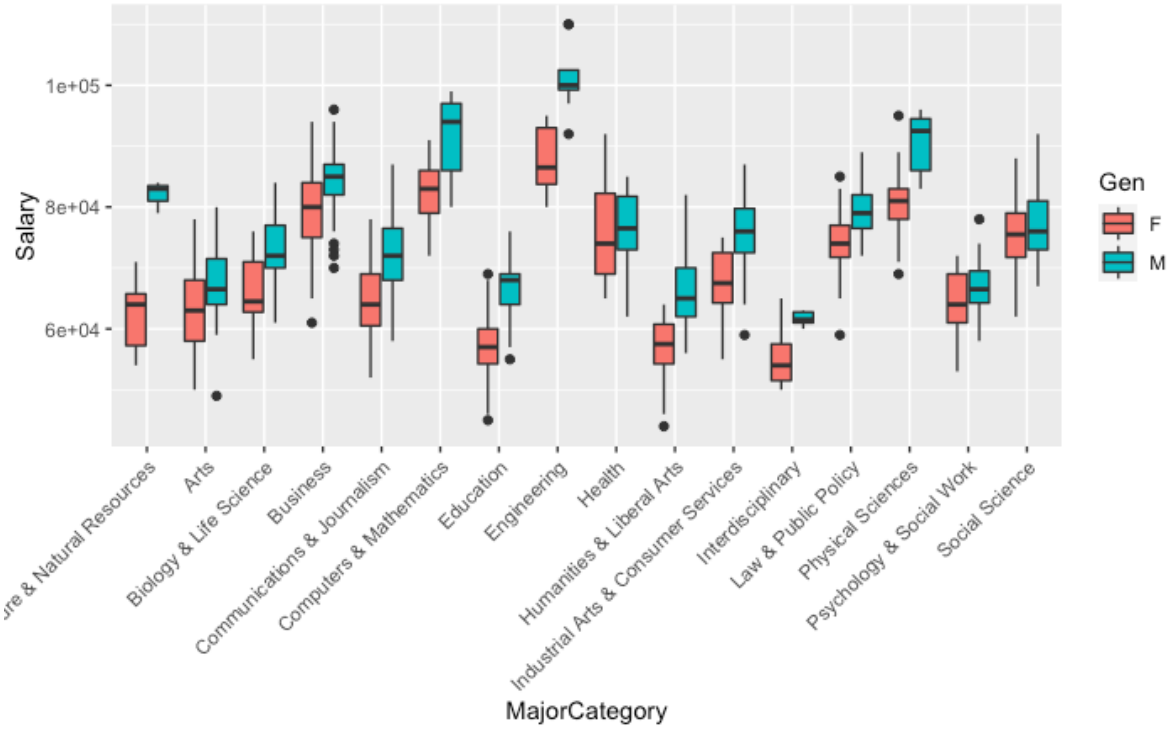
For the categorical variable **GenM.** If the person is male instead of female, and all other variables stay the same, their annual salary will increase by $5931.62 on average.

For **GPA**, as it increases by 1, and all other variables stay the same, their annual salary will increase by $5488.7368 on average.

**Standard deviation**: 5406.17

**R^2**: 0.7637316

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, -value and give your conclusion.

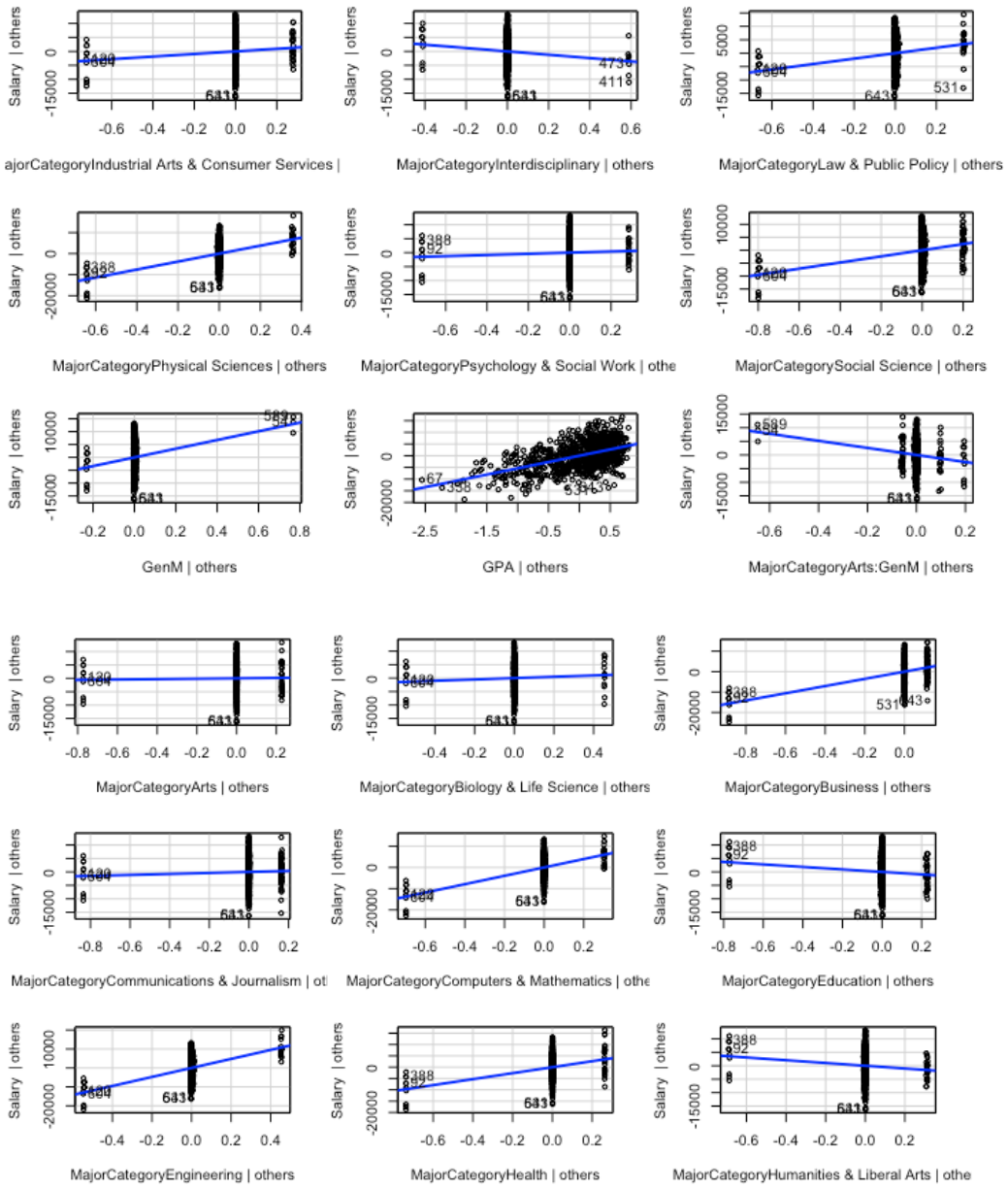


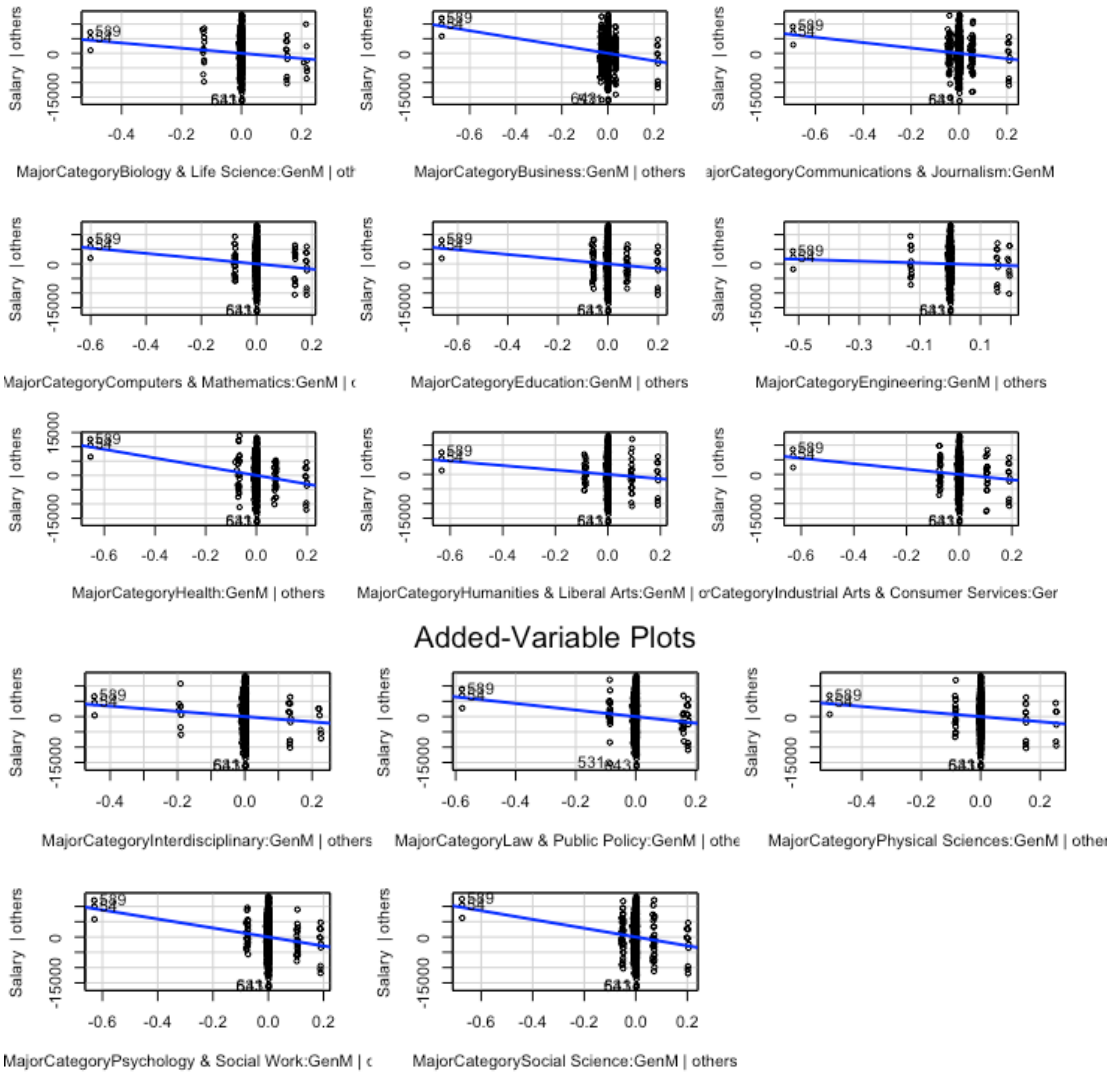

From this box plot, it appears that men make more money than women on average in all majors.

After running an F-test with ANOVA, the p-value of the interaction between major and gender is **7.161e-08,** which confirms that there is a difference in salary in all disciplines if the person is male or female.

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 - 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

**Linearity**:



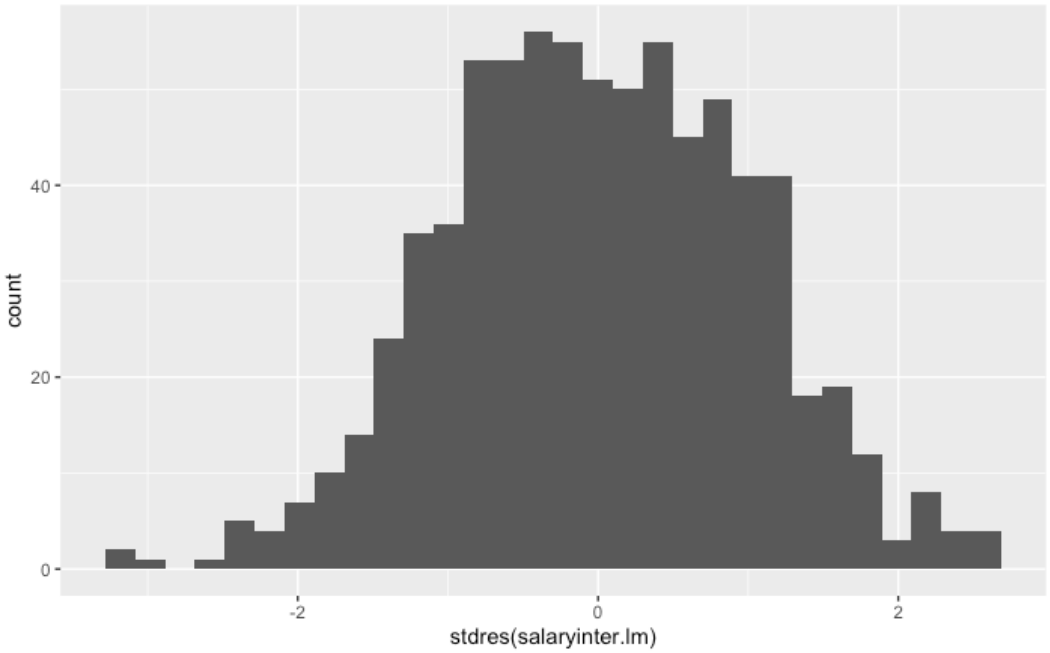



Looking at the added variable plots we can say that the linearity assumption is met because there seems to be a linear relationship among all the variables and salary.

**Independence**:

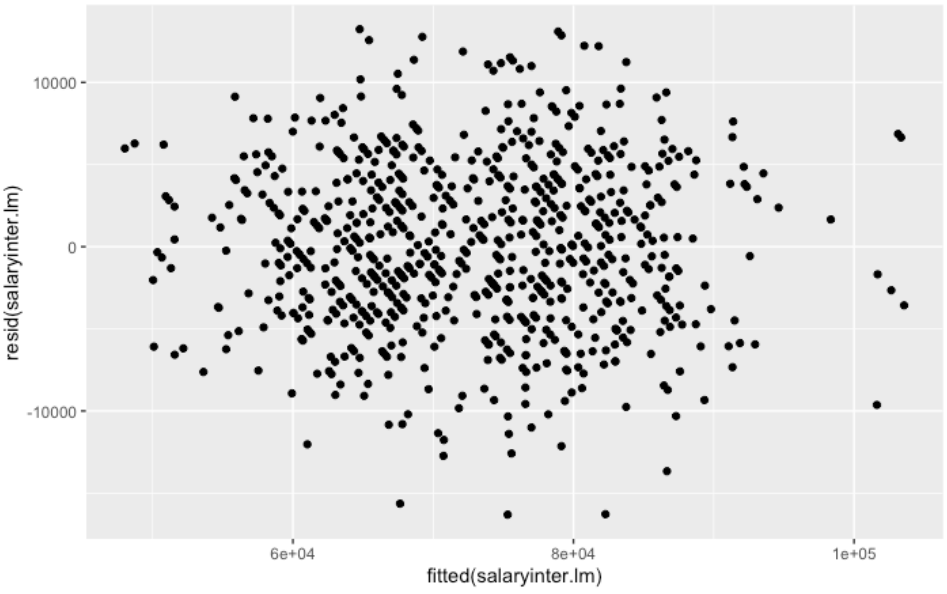
Each observation seems to be independent from the other ones. In other words, the salary of one person does not affect the salary of the next, nor is it affected by the previous.
The independence assumption is met.

**Normality:**




The normality assumption is met because the standardized residuals seem to be normally distributed, and the p-value of the KS test is 0.7421, which checks for normality.

**Equal Variance:**




The equal variance assumption is met because the fitted values vs. residuals plot shows that there is equal variance. The p-value of the BP test is 0.6075, which checks for equal variance.

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

**GPA**:

| | 1.5 % | 98.5 % |
|---|---|---|
| (Intercept) | 46281.909 | 55989.463 |
| GPA | 5078.182 | 7941.895 |

*As GPA increases by 1, Salary will increase 97% of the time between $5078.182 and $7941.895*

**Gender:**

| | 1.5 % | 98.5 % |
|---|---|---|
| (Intercept) | 69366.383 | 71572.114 |
| GenM | 4064.191 | 7470.498 |

*If the person is Male, Salary will increase 97% of the time between $4064.191 and $7470.498*

**Engineering:**

| | 1.5 % | 98.5 % |
|---|---|---|
| MajorCategoryEngineering | 20730.0833 | 31423.7629 |

*If the person's major is in Engineering, Salary will increase 97% of the time between $20730.0833 and $31423.7629*

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, -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.

The hypothesis test is to see if gender has a significant effect on salary if the major is computers and mathematics. The p-value is: <2e-16, which means gender does have a significant effect.

*If the person is Male, and has a Computers & Mathematics major, Salary will increase 95% of the time between $5144.3019 and $6718.9520 with and average of $5931.6270.*

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

My major is Computers & Mathematics, my GPA is 3.22, and I am a Male.
The prediction for my salary after 5 years is 95% of the time between $77442.62 and $98987.89 with an average of $88215.25.

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.

After conducting the leave-one-out cross validation, the RMSE is 4341.785, which compared to the sigma with a value of 5406.17 our model predicts better than by guessing, which means, it is a fairly good model. The width is 21491.08, which also indicates it is a fairly good model, considering the ranges of salary.

**R Code:**

```
library(tidyverse)
library(GGally)
library(car)
library(MASS)
library(lmtest)
library(multcomp)

salary <- read_csv("Salary.csv")

#1.----

#scaterplot of GPA and Salary
ggplot(data=salary,      mapping=aes(x=GPA,      y=Salary))      +
geom_point()+geom_smooth(se=FALSE)

#Boxplot of majors and salary
ggplot(data=salary,  mapping=aes(x=MajorCategory,  y=Salary))  +
geom_boxplot()  +  theme(axis.text.x  =  element_text(angle  =  45,
hjust = 1))

#Boxplot of gender and salary
ggplot(data=salary,      mapping=aes(x=Gen,      y=Salary))      +
geom_boxplot()  +  theme(axis.text.x  =  element_text(angle  =  45,
hjust = 1))

#Correlation of GPA and Salary
cor(salary$Salary, salary$GPA)
```

**#3.----**

```
#Model to verify matrices
salary.lm <- lm(formula=Salary~., data=salary)

#model matrix
X <- model.matrix(object=Salary~., data=salary)
y <- salary$Salary

#Coefficients
Bhat <- solve((t(X)%*%X))%*%t(X)%*%y
coef(salary.lm)
Bhat

#standard deviation
S2 <- (t(y-(X%*%Bhat))%*%(y-(X%*%Bhat)))/(756-17-1)
sigma(salary.lm)
sqrt(S2)

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

**#4.----**

```
#F-test with anova for gender and major interaction
salaryinter.lm     <-     lm(formula=Salary~.+Gen:MajorCategory,
data=salary)
anova(salaryinter.lm)

#boxplot of salary and gender
ggplot(data=salary,    mapping=aes(x=MajorCategory,    y=Salary,
fill=Gen))     +     geom_boxplot()+     theme(axis.text.x     =
element_text(angle = 45, hjust = 1))
```

```
#5.----
#AV Plots
avPlots(salaryinter.lm, ask=FALSE)


#Standardized res
ggplot() + geom_histogram(mapping=aes(x=stdres(salaryinter.lm)))

#KS Test
ks.test(stdres(salaryinter.lm), "pnorm")

#Fitted vals vs. res
ggplot(mapping=aes(x=fitted(salaryinter.lm),
y=resid(salaryinter.lm))) + geom_point()

#BP test
bptest(salaryinter.lm)


#6.----
#Confidence intervals of 97%:


#GPA
gpa.lm <- lm(formula=Salary~GPA, data=salary)
confint(gpa.lm,level=.97)

#Gender
gender.lm <- lm(formula=Salary~Gen, data=salary)
confint(gender.lm,level=.97)

#Engineering
engin.lm <- lm(formula=Salary~MajorCategory, data=salary)
confint(engin.lm,level=.97)
```

**#7.----**

```
#Hypothesis test of gender in the field of math and computers
at <- t((c(1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 4)-
c(1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4)))

my.test <- glht(salary.lm, linfct=at, alternative="two.sided")
summary(my.test)

confint(my.test,level=.95)
```

**#8.----**

```
#Prediction for my personal salary:

new.x = data.frame(MajorCategory="Computers & Mathematics",
Gen="M", GPA=3.22)
predict.lm(salary.lm, newdata=new.x, interval="prediction",
level=0.95)
```

**#9.---**

```
#Leave one out cross validation:

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)
wid <- rep(x=NA, times=n.cv)

for(cv in 1:n.cv){
  ## Select test observations
  test.obs <- sample(x=1:n.cv, size=n.test)

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

  ## Fit a lm() using the training data
  train.lm <- lm(formula=, 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()

}

#RMSE
mean(rpmse)
sigma(salary.lm)

#Width
mean(wid)
```