COMM 581 Assignment 6:
Multiple Linear Regression – Categorical Variables

Prepared for

Spectra Technologies

Date: October 20, 2016

by Gurpal Bisra
Student Number: 69295061

Table of Contents

Table of Figures

Table of Tables

1 Introduction

Spectra Technologies is a premier provider of spectroscopy services and products for researchers to identify their compounds. At present, Spectra Technologies employs personnel to fill positions in engineering, human resources, marketing, finance, quality assurance, manufacturing, and higher corporate roles.

Purpose of Analysis

My services were requested to determine whether Spectra Technologies is providing equal pay to its male and female employees after controlling statistically for their level of experience. If there is a wage gap based on an employee's gender, Spectra Technologies has asked me to determine the amount of difference and whether it depends on the level of experience (i.e. if there is an interaction between gender and experience).

Several sources cite that women, on average, earn $0.79 for every $1 a man makes; this disparity is worse for women of colour. However, equal pay for men and women after accounting for their experience is important for several reasons including: (1) lower earnings make it more difficult for women to care for their families; (2) social security and other benefits are associated with one's level of income; (3) females must work longer hours which keeps them away from their children longer; and (4) provincial and federal programs funded by income tax are adversely affected by the wage gap. One may argue that women are paid less than men because they tend to be more submissive during salary negotiations, and have more time-consuming child-nurturing responsibilities, so they appear absent when successions are outlined, and are then forced to assume worse-paying part-time roles.

Spectra Technologies is concerned about their own potential gender wage gap because the aforementioned problems associated with gender wage gap will negatively impact an employee's morale resulting in lower work productivity. Foremost, equal pay is that it's not a transparent issue. For instance, if a female employee learns her male colleagues earn more money for doing the same job, they may feel demoralized and be less encouraged to work harder or leave Spectra Technologies altogether. Consequently, Spectra Technologies wants to know what changes in salaries and hiring practices can be made to reduce such outcomes.

Models Used in Analysis

This report discusses four models that were formulated to perform the analysis to determine whether Spectra Technologies is providing equal pay to men and women after controlling statistically for their level of experience. The four models are the: (1) Full Interaction Model; (2) Experience Only Model; (3) Gender Only Model; and Gender and Experience Only Model (i.e. Final Model). The full interaction model, which is a multiple linear regression (MLR) model, is depicted in Equation 1 below:

$$\boldsymbol{y_i = B_0 + B_1 * x_{1i} + B_2 * x_{2i} + B_3 * x_{1i} * x_{2i} + \varepsilon_i} \qquad \textbf{(Equation 1)}$$

where:

$\boldsymbol{y_i}$ = Salary [$/year]

$\boldsymbol{B_0}$ = Continuous variable's (i.e. experience) intercept term [$]

$\boldsymbol{B_1}$ = Categorical variable's (i.e. gender) intercept adjustment term

$\boldsymbol{x_{1i}}$ = Dummy-variable for categorical variable (i.e. indicator term) where female and male are coded as -1 and 1, respectively

$\boldsymbol{B_2}$ = Continuous variable's slope term

$\boldsymbol{x_{2i}}$ = Continuous variable's term

$\boldsymbol{B_3 * x_{1i}x_{2i}}$ = Interaction term

$\boldsymbol{\varepsilon_i}$ = Errors term

I developed a simple linear regression (SLR) model which I refer to as the Experience Only Model, as characterized by Equation 2, to test the significance of the variable gender. In general, I can determine the significance of a missing term if I apply the ANOVA or drop1 command in R against the Full Interaction Model.

$$\boldsymbol{y_i = B_0 + B_2 * x_{2i} + \varepsilon_i} \qquad \textbf{(Equation 2)}$$

Likewise, another SLR model, the Gender Only Model is depicted by Equation 3.

$$\boldsymbol{y_i = B_0 + B_1 * x_{1i} + \varepsilon_i} \qquad \textbf{(Equation 3)}$$

Moreover, the Gender and Experience Only Model, a second MLR model, is explained by Equation 4.

$$\boldsymbol{y_i = B_0 + B_1 * x_{1i} + B_2 * x_{2i} + \varepsilon_i} \qquad \textbf{(Equation 4)}$$

Furthermore, I made the following assumptions to conduct my analysis:

- The data was collected at one point in time, from the single location for Spectra Technologies, so I did not need to investigate if the assumption of independence of errors was met.
- Experience was self-reported, so there is no reason to think that there is error associated with it, and salary was obtained from the company records.

Predicted Effect on Salary

I predict that, on average, males will be paid more than females for the same experience level because Spectra Technologies is a technology company. For example, Spectra Technologies may employ several engineers - a historically male-dominated profession. Such occupational segregation may favour hiring more males as engineers yet more female employees in departments including human resources and marketing. However, roles within non-engineering departments may, on average, pay less. Furthermore, I predict an employee with more experience is likely to earn a higher wage than someone with less work experience. One who accrues more experience over time is more likely to be hired into a managerial, or corporate, role which generally pay more.

Terminology

Wage Gap is expressed as a percentage and is calculated by dividing the median annual earnings for women by the median annual earnings for men.

Occupational Segregation is the distribution of people across and within occupations and jobs, based upon demographic characteristics, most often gender.

2 Methods & Results

2.1 Stats Package Used

I performed all of my analysis in R version 3.3.0 using the RStudio, version 0.99.896, graphical user interface. All of my R code can be found in the Appendix at the end of this report. Apart from using base R, I used the “nortest” package to determine whether the errors of my predicted models were normally distributed. In particular, I used the citation command to determine the following information about my statistical software:

```
To cite R in publications use:

  R Core Team (2016). R: A language and environment for statistical
  computing. R Foundation for Statistical Computing, Vienna, Austria. URL
  https://www.R-project.org/.

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {R: A Language and Environment for Statistical Computing},
    author = {{R Core Team}},
    organization = {R Foundation for Statistical Computing},
    address = {Vienna, Austria},
    year = {2016},
    url = {https://www.R-project.org/},
  }
```

2.2 Data Set Used

Spectra Technologies provided me a dataset consisting of 48 observations of three variables. A detailed data dictionary is found in Table 1 below.

TABLE 1: Data dictionary of dataset used for analysis.

Column Name	Format	Length	Data Elements [Range]	Description
experience	int	48	[0 - 35] years	Years of related-experience to role
gender	factor	48	2 levels: male, female	Employee’s self-identification
salary	int	48	[28300 – 45000] $/year	Yearly earnings in dollars

Summary Statistics [Questions 1-4]

I performed explanatory data analysis with the provided dataset which included 26 and 22 observations of female and male employees, respectively. I tabulated summary statistics, based on gender, in Table 2.

TABLE 2: Summary statistics of years of experience and yearly salary based on gender at Spectra Technologies.

Gender	Range of Experience [years]	Mean Experience [years]	Range of Salary [$/year]	Mean Salary [$/year]
Both genders	0 - 35	12.75	28,300 - 45,000	35,564
Males only	1 - 35	18.95	34,550 - 45,000	39,659
Females only	0 - 29	7.50	28,300 - 38,200	32,098

Next, I plotted salary as a function of experience as illustrated in Figure 1. I used two different plotting characters, one for men and one for women, and observed predicted salary as a function of experience appeared linear upon first glance.

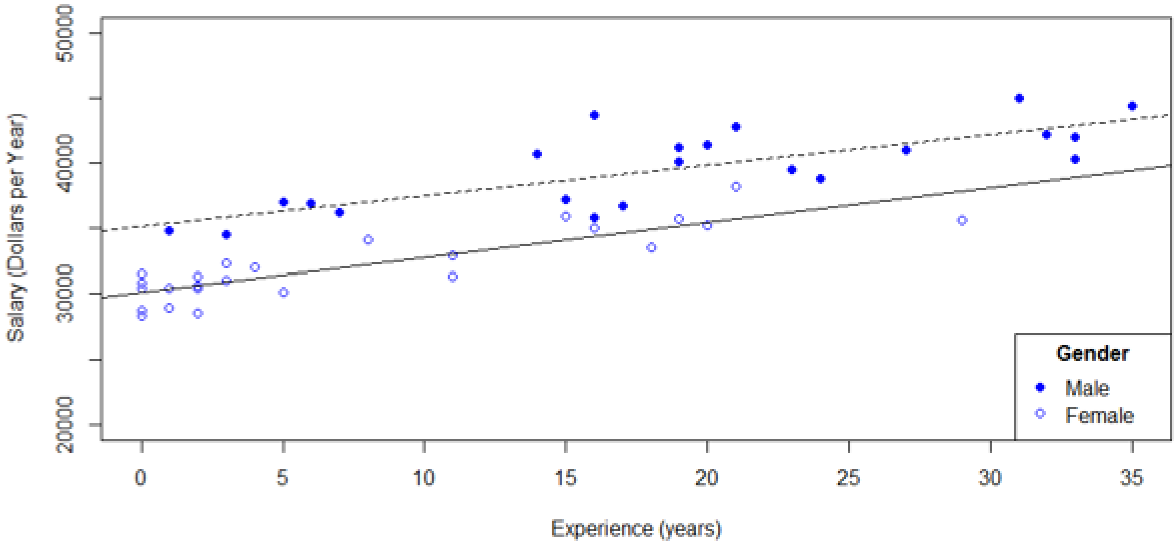


FIGURE 1: Salary as a function of experience. The data for both males and females appears linear, clustered for ages 0 - 5 and 15 - 20 years, with no obvious outliers, and in the positive direction.

2.3 Models Used for Analysis

2.3.1 Full Interaction Model [Question 5]

Fitting Full Interaction Model in R: Assessing Assumptions [Question 6]

Based on the dataset provided, I first developed the Full Interaction Model as explained by Equation 1 in the Introduction of this report. The model statement for the Full Interaction Model, including the interaction between gender and experience, predicts the salary according to the equivalent equation to equation 1 as found below:

$$\widehat{salary}_i = B_0 + B_1(gender)_i + B_2(experience)_i + B_3 * (gender)_i(experience)_i \quad \textbf{(Equation 6)}$$

Here, the value for gender is coded to be -1 or 1 in R if the gender is male or female, respectively. Salary and experience are described in the units dollars per year and years, respectively.

In order to use the Full Interaction Model, to fit a linear regression line into the data well, the following assumptions must be met: (1) linearity; (2) equal variance; and (3) normality of errors. First, I determined whether the assumption of linearity was met. For instance, I plotted the residuals of my MLR model against my predicted salary values as seen in Figure 2. I divided my plot of the residuals into 4 segments labelled a, b, c, and d as seen below.

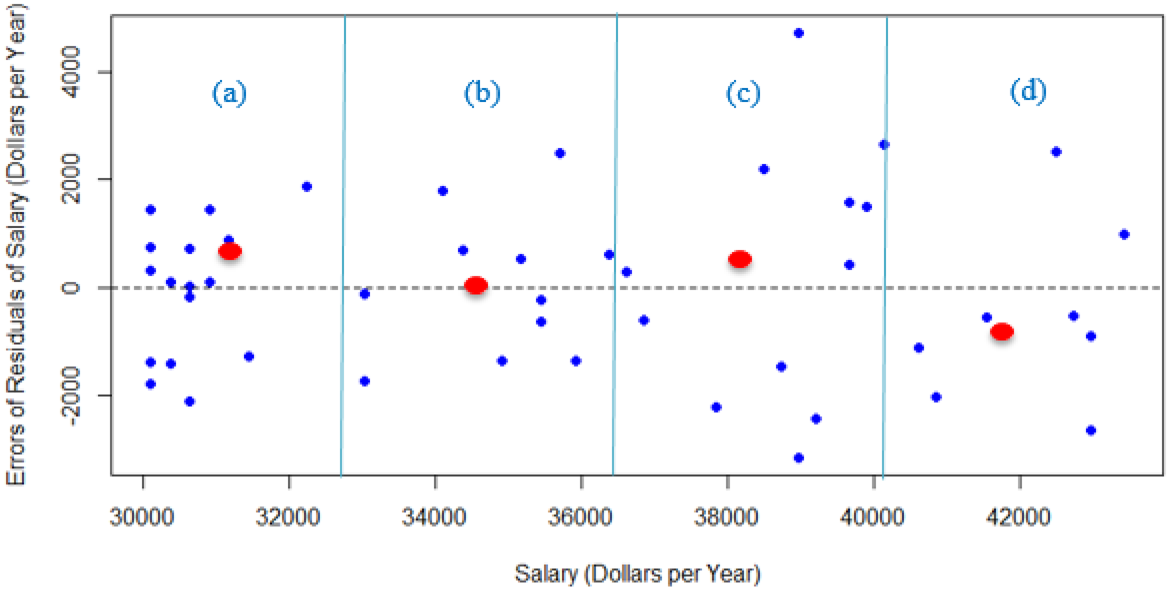


FIGURE 2: Residuals of the Full Interaction Model. The assumption of linearity is met.

I visually predicted the mean values of each segment and plotted a red dot of that mean in the segments with data points. There does appear to be an even spread of points above and below the line of zero residuals. If I were to draw a line connecting the red points, the line would deviate close to the line of zero residuals. Hence, I concluded that the assumption of linearity has been met. This means the regression line would fit into my data well and the estimates of my coefficients and standard errors would not be biased.

Next, I tested whether the assumption of equal variance, or the measure of spread, was met. This assumption can be verified if the residuals fall above and below zero residuals evenly as seen in Figure 3.

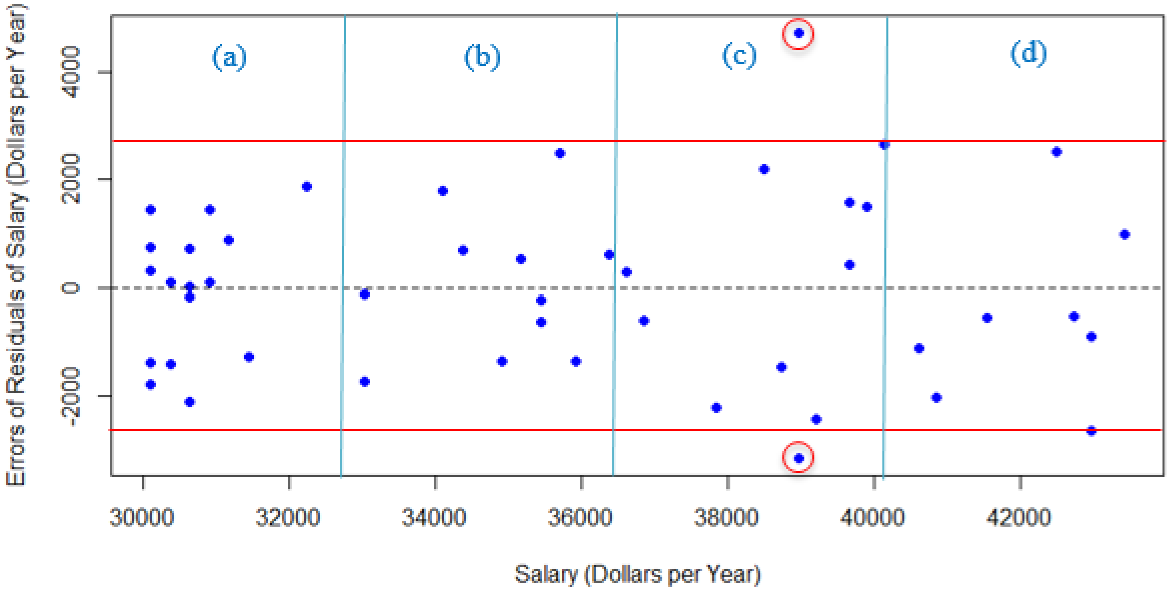


FIGURE 3: Residuals of the Full Interaction Model. The assumption of equal variance is met. I circled outliers in red which do not contribute to determining whether the assumption of equal variance is met.

I drew two straight red lines in the graph to illustrate approximate equal variance. Given my analysis above, I concluded the assumption of equal variance can be met. This means I can calculate the confidence intervals (CI's) and test the significance of the explanatory variable. In addition, the co-efficients of my regression and estimates of standard errors of co-efficients should not be biased.

Afterwards, I determined whether the errors of my residuals were normally distributed. For instance, I plotted a histogram of the residual errors from my predicted model, as seen below in Figure 4, and it appears approximately right-skewed upon visual assessment.

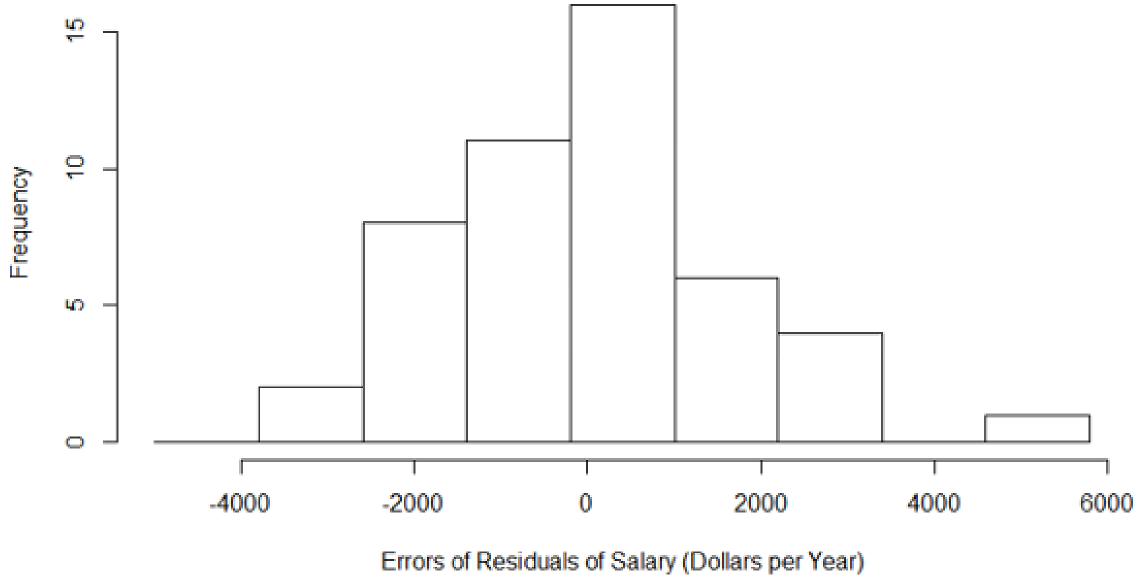


FIGURE 4: Histogram of the Full Interaction Model. The histogram appears right-skewed, not symmetric, to have a gap on the right-hand side, and with no clear outliers. The plot adds proof to my claim that the errors are normally distributed.

Next, I plotted the Q-Q plot of my log-log data as seen below in Figure 5. Since the standardized residuals change linearly by the theoretical quantile, there is further evidence that the residuals errors are normally distributed for the explanatory variables. In particular, the Q-Q plot exhibits a right-skew.

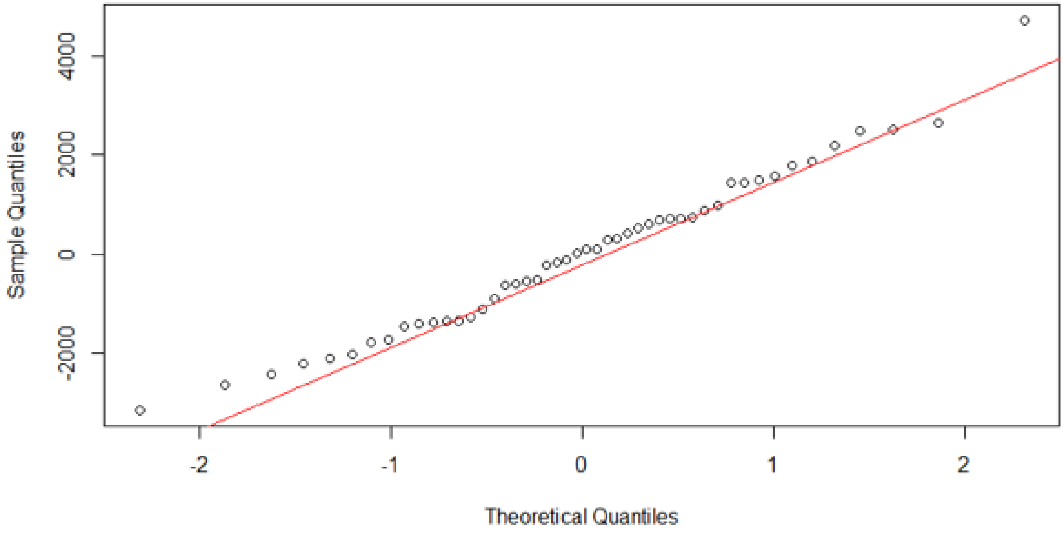


FIGURE 5: Normal Q-Q Plot of Full Interaction Model. The plot adds proof to my claim that the errors are normally distributed.

Furthermore, I performed four normality tests whose results are summarized below in Table 3. My hypothesis is:

H0: the errors of predicted model are normally distributed.

H1: the errors of my predicted are not normally distributed.

TABLE 3: Normality tests applied to the Full Interaction Model using the "nortest" package in R. All four normality tests fail to reject the null hypothesis. The table adds proof to my claim that the errors are normally distributed.

Normality Test	Statistic		p-value	Accept or Reject H0
Shapiro-Wilk	W = 0.98268	p < W	p = 0.6931	Fail to Reject H0
Lilliefors (Kolmogorov-Smirnov)	D = 0.075177	p > D	p = 0.7128	Fail to Reject H0
Cramer-von Mises	W = 0.026208	p > W	p = 0.8906	Fail to Reject H0
Anderson-Darling	A = 0.20288	p < A	p = 0.8701	Fail to Reject H0

In my testing, I am using an α-value of 0.05. Given that all four of the tests fail to reject the null hypothesis, more evidence is given that the errors of the predicted model are normally distributed. Therefore, I conclude the residuals of my errors are normally distributed. Hence, all assumptions are satisfied for the Full Interaction Model so I can fit a linear regression line into the data well.

ANOVA Table for Full Interaction Model [Question 8]

After constructing my Full Interaction Model, I wrote out the ANOVA table for the model using Type III sum of squares. I collected my sum of square error values of the variables gender, experience, and experience*gender using the drop1 command of my Full Interaction Model in R. In addition, I used R commands to determine values for the sum of squares for the regression (SSreg), error (SSE), and total (SSy). My R commands can be found in the Appendix. My ANOVA table is shown in Table 4 below.

Table 4: ANOVA table for the Full Interaction Model. Since there were 48 observations, n = 48.

Source	df	SS	MS	F	p-value
experience	1	126208943			3.688e-08
gender	1	94175493			7.921e-07
experience*gender	1	1006458	MSC = SSC/1 = 1006458/1 = 1006458	F = MSC/MSE = 1006458 / 2850897.95455 = 0.3530319275	0.5554
Error	n – m - 1 = 48 – 3 - 1 = 44	SSE = 125439510	MSE = SSE/(n-4) = 125439510/44 = 2850897.95455		
Total	n – 1 = 48 – 1 = 47	SSY = 1055563698			

Testing Significance of Full Interaction Model [Question 9]

After determining the assumptions for fitting a linear model were met for the Full Interaction Model, I tested the significance of the regression model. The steps I performed are outlined as follows:

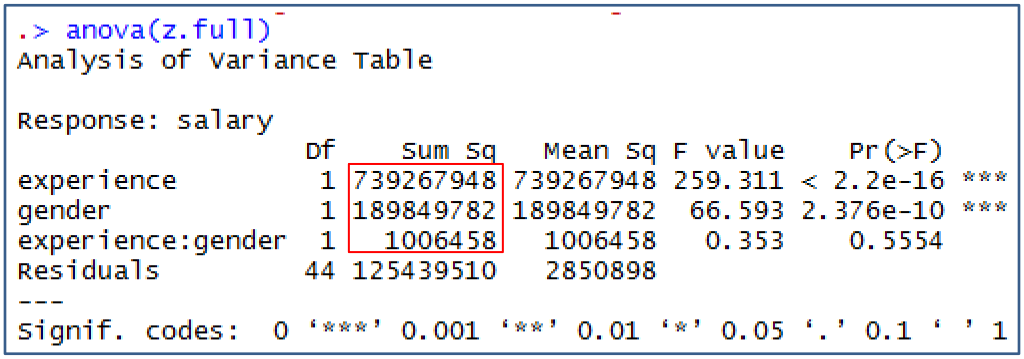
Step 1: Hypothesis for MLR for the Full Interaction Model

H0: the regression is not significant (all slopes B_0, B_1, B_2, B_3 = 0)

H1: the regression is significant (not all slopes B_0, B_1, B_2, B_3 = 0)

Step 2: Determine my F_{global} and p-values

I determined my p-value and F_{global} to be $2.2 * 10^{-16}$ and 109.8, respectively. I then calculated my F_{global} using the output of my ANOVA command, in R, as described in Figure 6.



```
> anova(z.full)
Analysis of Variance Table

Response: salary
                  Df    Sum Sq   Mean Sq F value    Pr(>F)    
experience         1 739267948 739267948 259.311 < 2.2e-16 ***
gender             1 189849782 189849782  66.593 2.376e-10 ***
experience:gender  1   1006458   1006458   0.353    0.5554    
Residuals         44 125439510   2850898                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Output of ANOVA command on full interaction model. The red box indicates the values which were used to calculate F_{global}.

Calculation of F_{global}:

$SSreg$ = 739267948 + 189849782 + 1006458 = 930124188

SSE = 125439510

$$F_{statistic} = \frac{MSreg}{MSE} = \frac{(\frac{SSreg}{m})}{(\frac{SSE}{n-m-1})} = \frac{(\frac{930124188}{3})}{(\frac{125439510}{48-3-1})} = 108.752189992$$

Step 3: Compare my F_{global} and p-values to their critical values

I calculated my F-critical value in R, since $F_{m,\, n\text{-}m\text{-}1,1\text{-}a}$ = $F_{3,\, 44,\, 0.95}$ to be 2.816466 and compared the statistical values as follows:

($F_{critical}$ = 2.816466) < (F_{global} = 109.8)

(p_{value} = 2.2 * 10^{-16}) < (α = 0.05)

Step 4: Therefore, I reject the null hypothesis and the regression is significant. Hence, not all the slopes in my Full Interaction Model, as described by Equation 1, are equal to 0.

2.3.2 Model with Experience Only

Testing Significance of Variable Gender: Partial F-test [Question 10]

After determining the regression for the Full Interaction Model was significant, I tested for the significance of the variable gender. To perform this test, I calculated my partial F-statistic by comparing **Equations 1 and 2**. More specifically, I compared the following "full" and "reduced" models:

Full Model: salary ~ experience + gender + experience*gender

Reduced Model: salary ~ experience (i.e. dropped gender)

The steps I performed are outlined as follows:

Step 1: Hypothesis for MLR model for testing the gender term

H0: the variable gender is not significant given the other x-variables in the model

(at least one of B_1 and B_3 = 0)

H1: the variable gender is significant given the other x-variables in the model

(neither B_1 or B_3 = 0)

Step 2: Determine my Partial F and p-values

I determined my p-value and $F_{partial}$ to be 1.457 * 10^{-9} and 66.946, respectively, using the output of my ANOVA command, in R.

Calculation of $F_{partial}$ using data from Figure 6:

$SSreg_{full}$= 739267948 + 189849782 + 1006458 = 930124188

SSE_{full}= 125439510

$SSreg_{partial}$= 739267948

$$F_{partial} = \frac{(n-m-1)*(SSreg_{full}- SSreg_{reduced})}{r*(SSE_{full})} = \frac{(48-3-1)*(930124188-739267948\,)}{(1)*(125439510)}$$

= 66.9460089568

<u>Step 3: Compare my Partial F and p-values to their critical values</u>

I calculated my F-critical value in R, since $F_{r,\, n\text{-}m\text{-}1,1\text{-}a}$ = $F_{2,\, 44,\, 0.95}$ to be 3.209278 and compared the statistical values as follows:

($F_{critical}$ = 3.209278) < ($F_{partial}$ = 66.946)

(p_{value} = 1.457 * 10^{-9}) < (α = 0.05)

<u>Step 4</u>: I reject the null hypothesis and variable gender is significant. Therefore, I will keep the B_1 and B_3 terms in my MLR model since the interaction predicts the response variable, salary, better when it is included in the model.

2.3.3 Model with Gender Only

Testing Significance of Variable Experience: Partial F-test [Question 11]

After determining the regression for the Full Interaction Model was significant, I tested for the significance of the variable experience. To perform this test, I calculated the partial F-statistic by comparing **Equations 1 and 3**. More specifically, I compared the following “full” and “reduced” models:

Full Model: salary ~ experience + gender + experience*gender

Reduced Model: salary ~ gender (i.e. dropped experience)

The steps I performed are outlined as follows:

<u>Step 1: Hypothesis for MLR model for testing the gender term</u>

H0: the variable experience is not significant given the other x-variables in the model

(at least one of B_2 and B_3 = 0)

H1: the variable experience is significant given the other x-variables in the model

(neither B_2 or B_3 = 0)

Step 2: Determine my Partial F and p-values

I determined my p-value and $F_{partial}$ to be 3.587 * 10^{-11} and 43.646, respectively, using the output of my ANOVA command, in R.

Calculation of $F_{partial}$ using data from Figure 6:

$SSreg_{full}$= 739267948 + 189849782 + 1006458 = 930124188

SSE_{full}= 125439510

$SSreg_{partial}$= 739267948

$$Parital\ F = \frac{(n-m-1)*(SSreg_{full} - SSreg_{reduced})}{r*(SSE_{full})} = \frac{(48-3-1)*(930124188-189849782\ )}{(1)*(125439510)}$$

= 259.663592946

Step 3: Compare my Partial F and p-values to their critical values

I calculated my F-critical value in R, since $F_{r,\ n\text{-}m\text{-}1,1\text{-}a} = F_{2,\ 44,\ 0.95}$ to be 3.209278 and compared the statistical values as follows:

($F_{critical}$ = 3.209278) < ($F_{partial}$ = 259.664)

(p_{value} = 3.587 * 10^{-11}) < (α = 0.05)

Step 4: I reject the null hypothesis and variable experience is significant. Therefore, I will keep the B_2 and B_3 terms in my MLR model since the interaction predicts the response variable, salary, better when it is included in the model.

2.3.4 Final Model: Model with Gender and Experience Only (Option A)

Fitting Gender and Experience Only Model in R: Assessing Assumptions [Question 12a]

After determining both the gender and experience variables are significant for determining the response variable salary better, I attempted to see whether I could simplify my model. Foremost, I developed a new model, which I will refer to as my Final Model, as described by Equation 4 (i.e. or equivalently Equation 7 below), which is a MLR which excludes the interaction term as follows:

$$\widehat{salary}_i = B_0 + B_1(gender)_i + B_2(experience)_i$$ **(Equation 7)**

Here, the value for gender is coded to be -1 or 1 in R if the gender is male or female, respectively.

In order to use the Final Model, and fit a linear regression line into the data well, the following assumptions must be met: (1) linearity; (2) equal variance; and (3) normality of errors. First, I determined whether the assumption of linearity was met. For instance, I plotted the residuals of my MLR model against my predicted salary values as seen in Figure 7. I divided my plot of the residuals into 4 segments labelled a, b, c, and d as seen below.

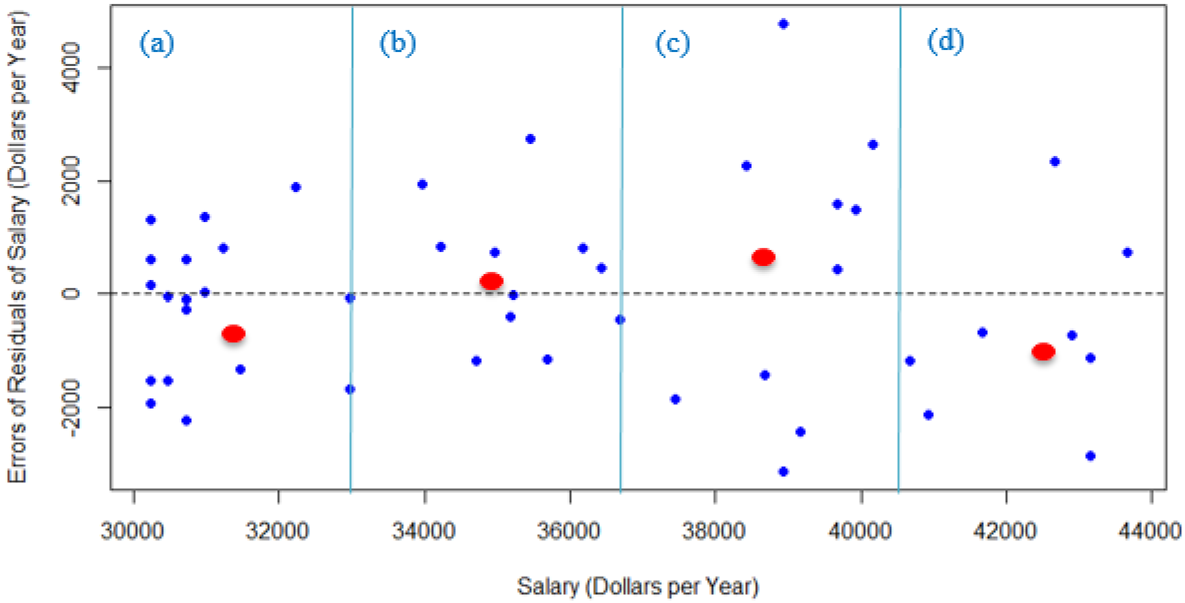


FIGURE 7: Residuals of the Final Model based on gender and experience only. The assumption of linearity is met.

I visually predicted the mean values of each segment and plotted a red dot of that mean in the segments with data points. There does appear to be an even spread of points above and below the line of zero residuals. If I were to draw a line connecting the red points, the line would deviate close to the line of zero residuals. Hence, I conclude that the assumption of linearity has been met. This means the regression line would fit into my data well and the estimates of my coefficients and standard errors would not be biased.

Next, I tested if the assumption of equal variance, or the measure of spread, was met. This assumption can be verified if the residuals fall above and below zero residuals evenly as seen in Figure 8.

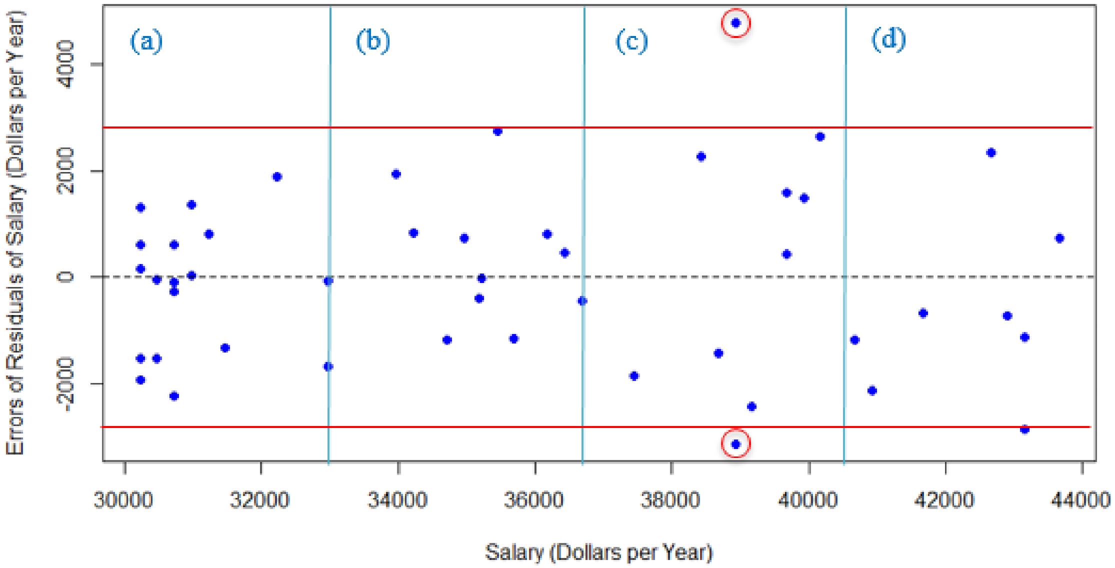


FIGURE 8: Residuals of the Final Model based on gender and experience only. The assumption of equal variance is met. I circled outliers in red which do not contribute to determining whether the assumption of equal variance is met.

I drew two straight red lines in the graph to illustrate approximate equal variance. Given my analysis above, I concluded that the assumption of equal variance can be met. This means I can calculate the confidence intervals (CI's) and test the significance of the explanatory variable. In addition, the co-efficients of my regression and estimates of standard errors of co-efficients should not be biased.

Afterwards, I determined whether the errors of my residuals were normally distributed. For instance, I plotted a histogram of the residual errors from my predicted model, as seen below in Figure 9, and it appears approximately right-skewed upon visual assessment.

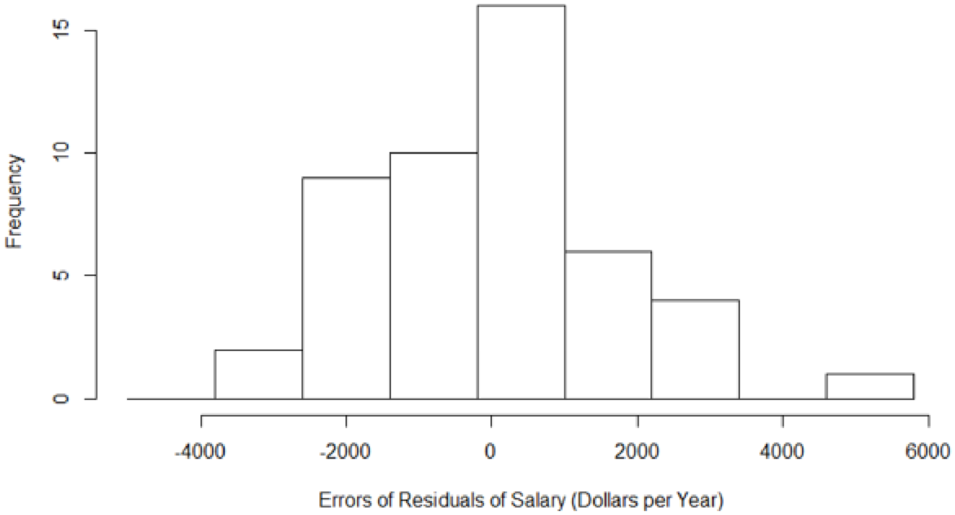


FIGURE 9: Histogram of the Final Model based on gender and experience only. The histogram appears right-skewed, not symmetric, to have a gap on the right-hand side, and with no clear outliers. The plot adds proof to my claim that the errors are normally distributed.

Next, I plotted the Q-Q plot of my log-log data as seen below in Figure 10. Since the standardized residuals change linearly by the theoretical quantile, there is further evidence that the residuals errors are normally distributed for the explanatory variables. In particular, the Q-Q plot exhibits a right-skew.

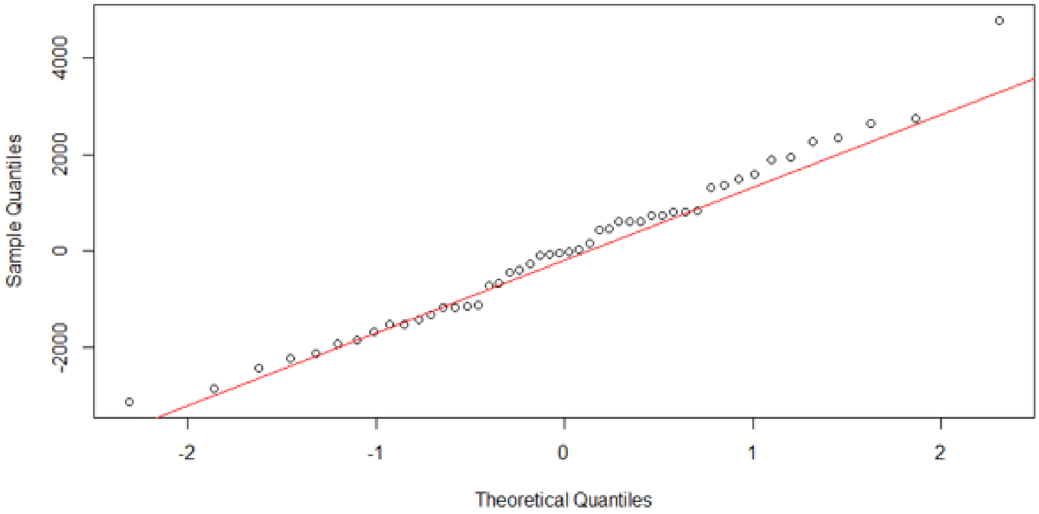


FIGURE 10: Normal Q-Q Plot of Final Model based on gender and experience only. The plot adds proof to my claim that the errors are normally distributed.

Furthermore, I performed four normality tests whose results are summarized below in Table 5. My hypothesis is:

H0: the errors of final predicted model are normally distributed.

H1: the errors of my final predicted are not normally distributed.

TABLE 5: Normality tests applied to the Final Model, based on gender and experience only, using the "nortest" package in R. All four normality tests fail to reject the null hypothesis. The table adds proof to my claim that the errors are normally distributed.

Normality Test	Statistic		p-value	Accept or Reject H0
Shapiro-Wilk	W = 0.98254	p < W	p = 0.6871	Fail to Reject H0
Lilliefors (Kolmogorov-Smirnov)	D = 0.083533	p > D	p = 0.5497	Fail to Reject H0
Cramer-von Mises	W = 0.030055	p > W	p = 0.8424	Fail to Reject H0
Anderson-Darling	A = 0.20933	p < A	p = 0.8541	Fail to Reject H0

In my testing, I am using an α-value of 0.05. Given that all four of the tests fail to reject the null hypothesis, more evidence is given that the errors of the predicted model are normally distributed. Therefore, I conclude the residuals of my errors are normally distributed. Hence, all assumptions are satisfied for the Final Model based on gender and experience only so I can fit a linear regression line into the data well.

Testing Significance of MLP with Interaction Term: Partial F-test [Question 12b]

In order to confirm I could predict the salaries using the Final Model, I tested for the significance of the interaction term. To perform this test, I calculated the partial F-statistic by comparing **Equations 1 and 4**. More specifically, I compared the following "full" and "reduced" models:

Full Model: salary ~ experience + gender + experience*gender

Reduced Model: salary ~ experience + gender (i.e. dropped interaction)

The steps performed are outlined as follows:

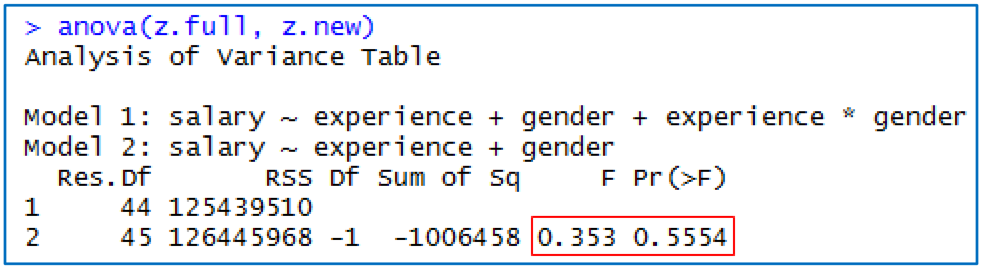
Step 1: Hypothesis for MLR model for testing the interaction term

H0: the interaction is not significant given the other x-variables in the model (B_3 = 0)

H1: the interaction is significant given the other x-variables in the model ($B_3 \neq 0$)

Step 2: Determine my Partial F and p-values

I determined my $F_{partial}$ and p-values to be 0.5554 and 0.353, respectively, using the output of my ANOVA command, in R as shown in Figure 11.



```
> anova(z.full, z.new)
Analysis of Variance Table

Model 1: salary ~ experience + gender + experience * gender
Model 2: salary ~ experience + gender
  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1     44 125439510
2     45 126445968 -1   -1006458 0.353 0.5554
```

Figure 11: Output of ANOVA command on the Full Interaction and Final Models. The red box indicates the values for $\boldsymbol{F_{partial}}$ and p-value.

Step 3: Compare my Partial F and p-values to their critical values

I calculated my F-critical value in R, since $F_{r,\, n\text{-}m\text{-}1,1\text{-}a} = F_{1,\, 46,\, 0.95}$ to be 4.061706 and compared the statistical values as follows:

($F_{partial}$ = 0.353) < ($F_{critical}$ = 4.061706)

(α = 0.05) < (p_{value} = 0.5554)

Step 4: I fail to reject the null hypothesis and interaction term is not significant. Therefore, I will not keep the B_3 term in my MLR model since the interaction does not predict the response variable, salary, better when it is included in the model. Consequently, it appears my Final Model is the best model and the slopes for both male and female genders are the same.

Testing Significance of MLP with Categorical Variable and No Interaction Term: Partial F-test

I next tested the significance of the categorical variable in the Final Mode. To perform this test, I calculated the partial F-statistic by comparing **Equations 2 and 4**. More specifically, I compared the following “full” and “reduced” models:

Full Model:	salary ~ experience + gender
Reduced Model:	salary ~ experience (i.e. dropped gender)

The steps I performed are outlined as follows:

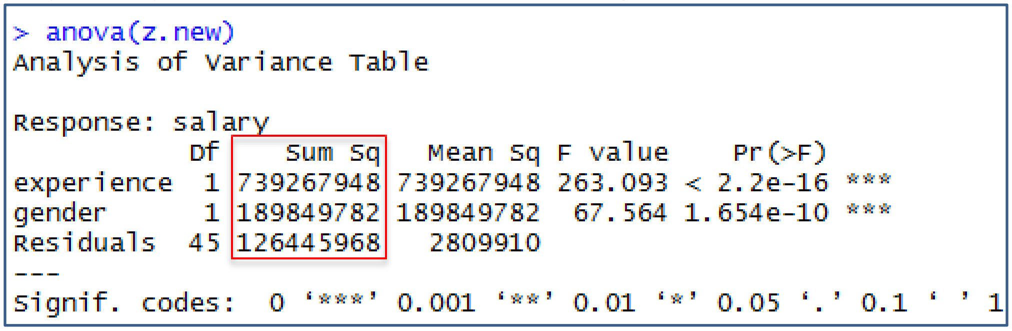
Step 1: Hypothesis for MLR Final Model for testing the gender term

H0: the variable gender is not significant given the other x-variables in the model (B_1 = 0)

H1: the variable gender is significant given the other x-variables in the model (B_1 ≠ 0)

Step 2: Determine my Partial F and p-values

I determined my p-value and $F_{partial}$ to be 1.654 * 10^{-10} and 67.564, respectively, using the output of my ANOVA command, in R as shown in Figure 12.



```
> anova(z.new)
Analysis of Variance Table

Response: salary
           Df    Sum Sq   Mean Sq F value    Pr(>F)    
experience  1 739267948 739267948 263.093 < 2.2e-16 ***
gender      1 189849782 189849782  67.564 1.654e-10 ***
Residuals  45 126445968   2809910                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12: Output of ANOVA command on Final Model. The red box indicates the values which were used to calculate $\boldsymbol{F_{partial}}$.

Calculation of $\boldsymbol{F_{partial}}$ using data from Figure 12:

$SSreg_{full}$= 739267948 + 189849782 = 929117730

SSE_{full}= 126445968

$SSreg_{partial}$= 739267948

$$Parital\ F = \frac{(n-m-1)*(SSreg_{full} - SSreg_{reduced})}{r*(SSE_{full})} = \frac{(48-2-1)*(929117730 - 739267948)}{(1)*(126445968)}$$

= 67.5643543652

Step 3: Compare my Partial F and p-values to their critical values

I calculated my F-critical value in R, since $F_{r,\ n\text{-}m\text{-}1,1\text{-}a} = F_{1,\ 45,\ 0.95}$ to be 4.056612 and compared the statistical values as follows:

($F_{critical}$ = 4.0566) < ($F_{partial}$ = 67.564)

(p_{value} = 3.587 * 10^{-11}) < (α = 0.05)

Step 4: I reject the null hypothesis and the gender is significant. Therefore, I will keep my B_1 term in my MLR Final Model since the gender term predicts the response variable better when it is included in the model.

Testing Significance of MLP with Continuous Variable and No Interaction Term: Partial F-test

Furthermore, I tested the significance of the continuous variable in the Final Mode. To perform this test, I calculated the partial F-statistic by comparing **Equations 3 and 4**. More specifically, I compared the following “full” and “reduced” models:

Full Model: salary ~ experience + gender

Reduced Model: salary ~ gender (i.e. dropped experience)

The steps I performed are outlined as follows:

Step 1: Hypothesis for MLR model for testing the experience term

H0: the variable experience is not significant given the other x-variables in the model (B_2 = 0)

H1: the variable experience is significant given the other x-variables in the model ($B_2 \neq 0$)

Step 2: Determine my Partial F and p-values

I determined my p-value and $F_{partial}$ to be 3.757 * 10^{-12} and 263.093, respectively, using the output of my ANOVA command, in R.

Calculation of $F_{partial}$ using data from Figure 12:

$SSreg_{full}$= 739267948 + 189849782 = 929117730

SSE_{full}= 126445968

$SSreg_{partial}$= 189849782

$$Parital\ F = \frac{(n-m-1)*(SSreg_{full} - SSreg_{reduced})}{r*(SSE_{full})} = \frac{(48-2-1)*(929117730 - 189849782)}{(1)*(126445968)}$$

= 263.093068021

Step 3: Compare my Partial F and p-values to their critical values

I calculated my F-critical value in R, since $F_{r, n-m-1,1-a} = F_{1, 45, 0.95}$ to be 4.056612 and compared the statistical values as follows:

($F_{critical}$ = 4.0566) < ($F_{partial}$ = 263.093)

(p_{value} = 3.757 * 10^{-12}) < (α = 0.05)

Step 4: I reject the null hypothesis and the experience is significant. Therefore, I will keep my B_2 term in my MLR Final Model since the experience term predicts the response variable better when it is included in the model.

3 Discussion & Recommendations

3.1 Final Model Equation [Question 13]

Several reasons can account for the differences observed between experience and salary statistics for males and females as depicted in Table 2. In general, males have a higher salary and range of experience than females. First, males may have a higher salary because this company is a technology company. For example, Spectra Technologies may employ several engineers which is a historically male-dominated profession. Such occupational segregation may favour hiring more males as engineers yet more female employees in departments including human resources and marketing. However, roles within these departments may on average pay less than an engineer's annual salary. Next, males have a higher range of experience than females because women may have taken maternity leave. For instance, males may be less likely to take paternity leave than a female taking maternity leave because paternity leave is rarely paid. Furthermore, male-dominated roles such as engineering require several years of experience to perform as compared to other female-dominated roles such as human resources. Since Spectra Technologies is a technology company, they need to hire several engineers. Consequently, male employees appear to earn a higher mean salary, and have a higher mean years of experience, than female employees. *[Questions 1 and 2]*

After including gender as a variable, the relationship between salary and experience is linear. For instance, there is no significant interaction between gender and experience which determines the predicted salary for employees at Spectra Technologies. *[Question 3]*

Upon inspecting Figure 1, which is a scatter plot of salary as a function of experience where I used two different plotting characters, one for men and women, and observed the predicted salary as a function of experience appeared linear upon first glance. Other qualitative observations I made were that the data for both males and females appears linear, clustered for ages 0 - 5 and 15 - 20 years, with no obvious outliers, and in the positive direction. As predicted in the Introduction of this report, males are, on average, paid more than females for the same years of experience. *[Question 4]*

In the Methods section, I determined the variable gender was significant using a partial F-test, given the other x-variables in the model, for the Full Interaction Model. This means that I rejected the null hypothesis and I kept the B_1 term in my MLR model since the interaction predicts the response variable, salary, better when it is included in the model. In terms of the overall purpose of the study, this implies there is a gender wage gap at Spectra Technologies – males are paid more, on average, than females for the same level of experience. Any violations of this claim could be explained by possible females in corporate roles. *[Question 10]*

Likewise, I determined the variable experience was significant using a partial F-test, given the other x-variables in the model, for the Full Interaction Model. This means that I rejected the null hypothesis and I kept the B_2 term in my MLR model since the interaction predicts the response variable, salary, better when it is included in the model. In terms of the overall purpose of the study, this implies one's years of experience in their role allows them to earn a higher salary at Spectra Technologies when statistically taking gender into account. *[Question 11]*

Based on my analysis performed in the Methods section of this report, I concluded the MLR model, based on gender and years of experience only, best predicts an employee's annual income. This Final Model is depicted in Equation 8 below:

$$\boxed{\widehat{\boldsymbol{Salary}}_{\boldsymbol{i}} = \mathbf{32584.11} + \mathbf{2354.03} * (\boldsymbol{gender_i}) + \mathbf{249.07} * (\boldsymbol{experience_i})} \quad \textbf{(Equation 8)}$$

where:

$\widehat{\boldsymbol{Salary}}_{\boldsymbol{i}}$ = Predicted annual income [$\frac{\$}{year}$]

$\mathbf{32584.11}$ = intercept term [$]

$\mathbf{2354.03}$ = Categorical variable's (i.e. gender) intercept adjustment term [$\frac{\$}{year}$]

$\boldsymbol{gender_i}$ = categorical variable (i.e. indicator term) where females and males are coded as -1 and 1, respectively.

$\mathbf{249.07}$ = Continuous variable's slope term [$\frac{\$}{year^2}$]

$\boldsymbol{experience_i}$ = Continuous variable's term [year]

After controlling statistically for an employee's level of experience, I determined men are paid approximately $4708.06 more than women at Spectra Technologies. The difference does not depend on the interaction between gender and experience. In particular, I performed a partial F-test and determined that the slopes for males and females are the same. This means I failed to reject the null hypothesis and the interaction is not significant given the other x-variables in the model. Hence, I set B_3 to be zero in my Final Model. In terms of the overall purpose of the study, this means there is a gender wage gap at Spectra Technologies – males are paid more than females for the same level of experience. *[Question 12]*

Predicted Salary for Men [Question 14]

The equation to calculate the predicted salary for male employees at Spectra Technologies is described by Equation 9.

$$\boxed{\widehat{\boldsymbol{Salary}}_{\boldsymbol{men}} = \mathbf{34938.14} + \mathbf{249.07} * (\boldsymbol{experience})} \qquad \textbf{(Equation 9)}$$

Predicted Salary for Women

The equation to calculate the predicted salary for female employees at Spectra Technologies is described by Equation 10.

$$\boxed{\widehat{\boldsymbol{Salary}}_{\boldsymbol{female}} = \mathbf{30230.08} + \mathbf{249.07} * (\boldsymbol{experience})} \qquad \textbf{(Equation 10)}$$

Measures of Goodness and Fit [Question 15]

In order to assess how well my Final Model could predict an employee's salary at Spectra Technologies, I determined the measure of goodness of fit. In order to perform these calculations, I first calculated the values for SSreg, SSE, and SSy in R. Next, I determined the estimates of R^2 value and standard error of the estimate (root MSE) as follows:

$$R^2 = \frac{SSReg}{SSy} = \frac{930124188}{1055563698} = 0.88116348616$$

$$SE_E = \sqrt{\frac{SSE}{n-m-1}} = \sqrt{\frac{125439510}{48-2-1}} = 1669.59416226$$

In general, a good measure of fit would have a high R^2 value and low SE_E values. First, a good value for R^2 depends on the field of study performed and what values are acceptable to allow one to publish their results. That said, given my experience, a R^2 value of 0.8812 is good considering R^2 can be any value between 0 and 1. My analysis indicates that 88.12% of the total variation is explained by the regression of my Final Model. In addition, the calculated SE_E value of 1669.59 is low. In fact, this value is the standard deviation of my residuals which means that the salary points, as graphed in Figure 13, do not deviate far from my regression line of it.

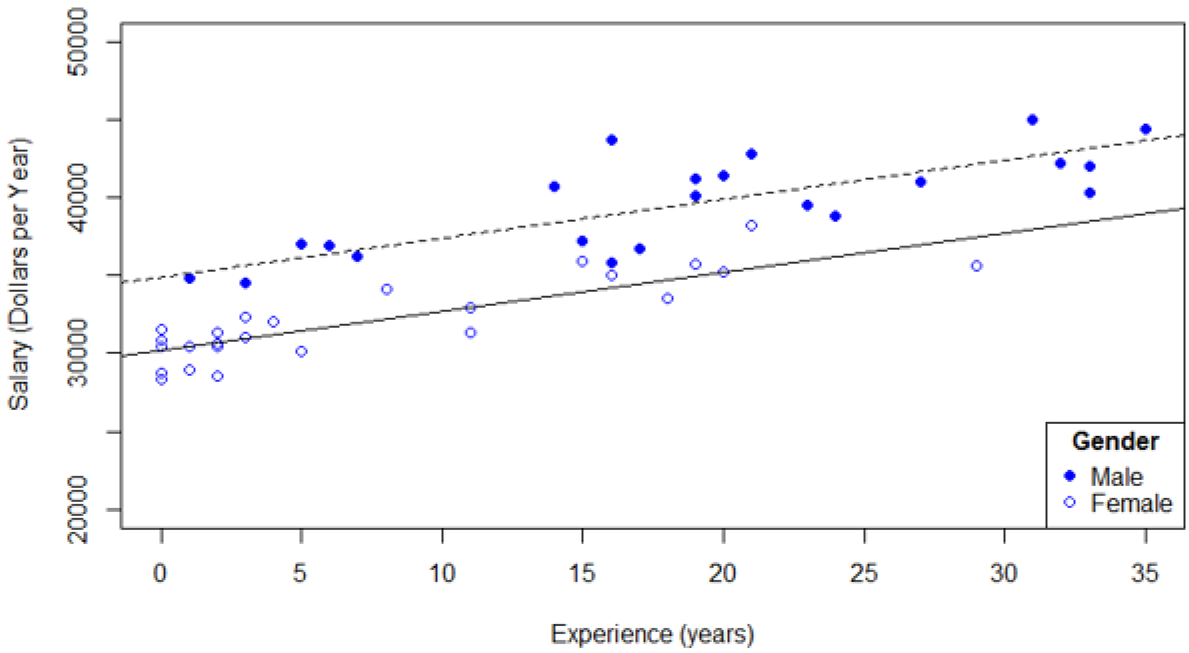
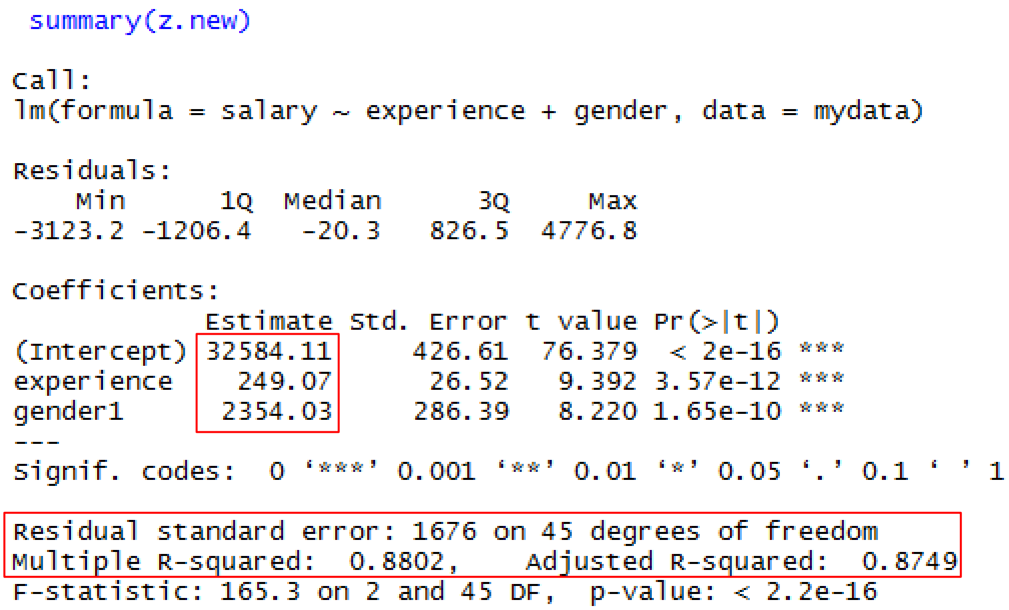


FIGURE 13: Salary as a function of experience applied to the Final Model. The data for both males and females appears linear, clustered for ages 0 - 5 and 15 - 20 years, with no obvious outliers, and in the positive direction.

In R, I verified these values were correct using the summary command as shown below in Figure 14.



```
 summary(z.new)

Call:
lm(formula = salary ~ experience + gender, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3123.2 -1206.4   -20.3   826.5  4776.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 32584.11     426.61  76.379  < 2e-16 ***
experience    249.07      26.52   9.392 3.57e-12 ***
gender1      2354.03     286.39   8.220 1.65e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1676 on 45 degrees of freedom
Multiple R-squared:  0.8802,	Adjusted R-squared:  0.8749 
F-statistic: 165.3 on 2 and 45 DF,  p-value: < 2.2e-16
```

FIGURE 14: Output of summary command applied to the Final Model. Red boxes encircle estimates of the measure of goodness of fit and predicted intercept and slope values.

3.2 Recommendations to Equalize Salary for Men and Women

Recommended Adjustments for Salary

In order to equalize salary for men and women, I would increase the salary of female employees by $4708.06. According to the model, this would help eliminate the gender wage gap at Spectra Technologies. In contrast, I would not recommend decreasing the salary of male employees by $4708.06. This may be against the law, and psychologically demoralizing, and likely result in male employees leaving Spectra Technologies. At the very least, it would be likely the male employee's morale would be lowered possibly leading the lowered productivity.

Recommended Changes in Hiring Process

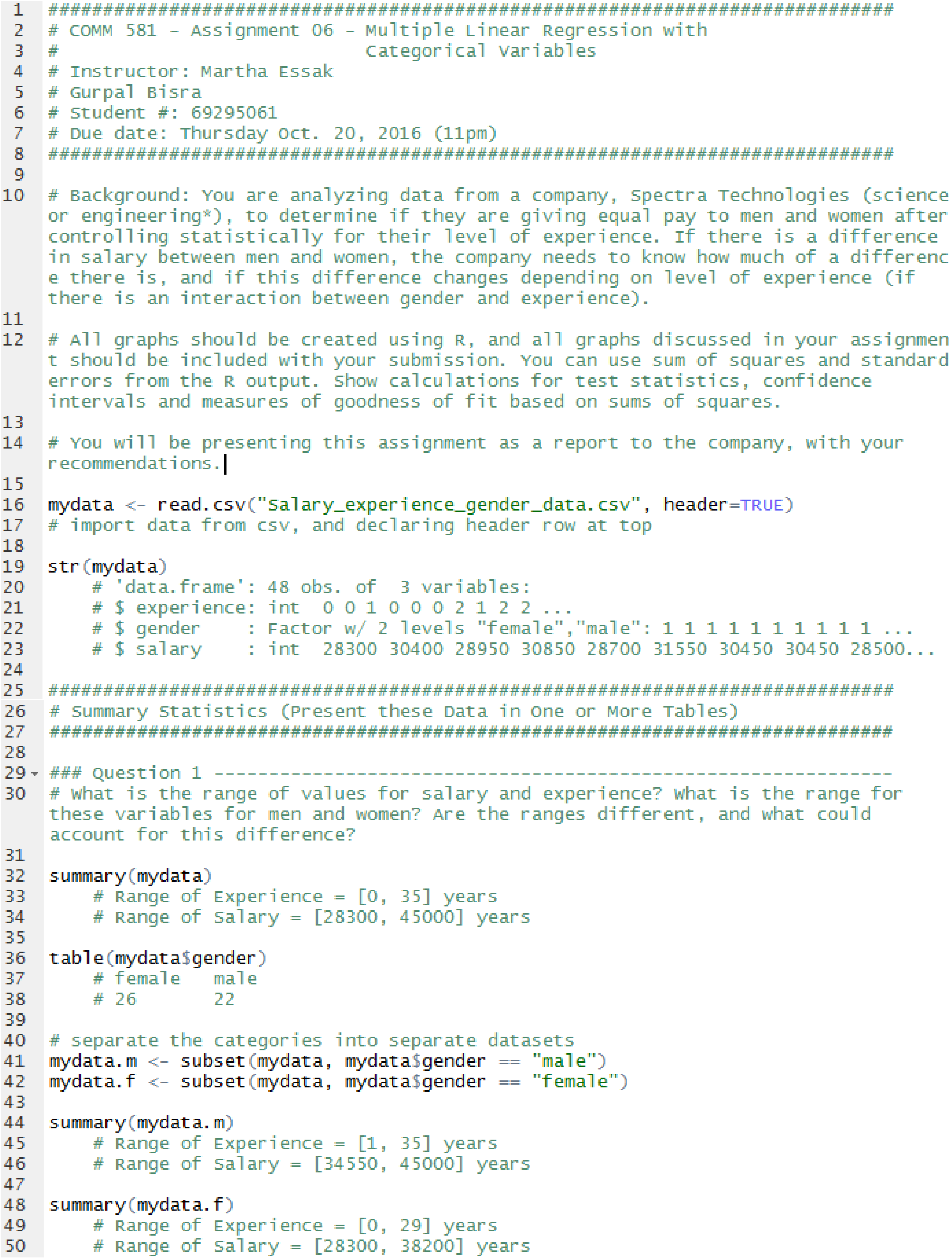
My aforementioned recommendations were based on the dataset provided by Spectra Technologies. In order for me to state specific changes in the hiring process, I would need to be provided further information. For instance, it may be the case the entire dataset only includes observations from a single department such as engineering. If this were the case, then a quota of the male-dominated profession should be established such that it is mandatory to hire a certain number of female engineers to avoid occupational segregation. Likewise, it may be the case future human resources and marketing hires will be male to balance the male to female ratio in those departments. In general, one may argue that women are paid less than men because they tend to be more submissive during salary negotiations. In this instance, I would suggest a pre-set salary is established per role such that gender bias cannot influence salary negotiations. For higher positions, a merit system should be enforced to determine which employees, regardless of gender, earn raises. In particular, an internal position should be dedicated to ensuring employee evaluations are screened for objectivity prior to promotions.

3.3 Potential Improvements to Model

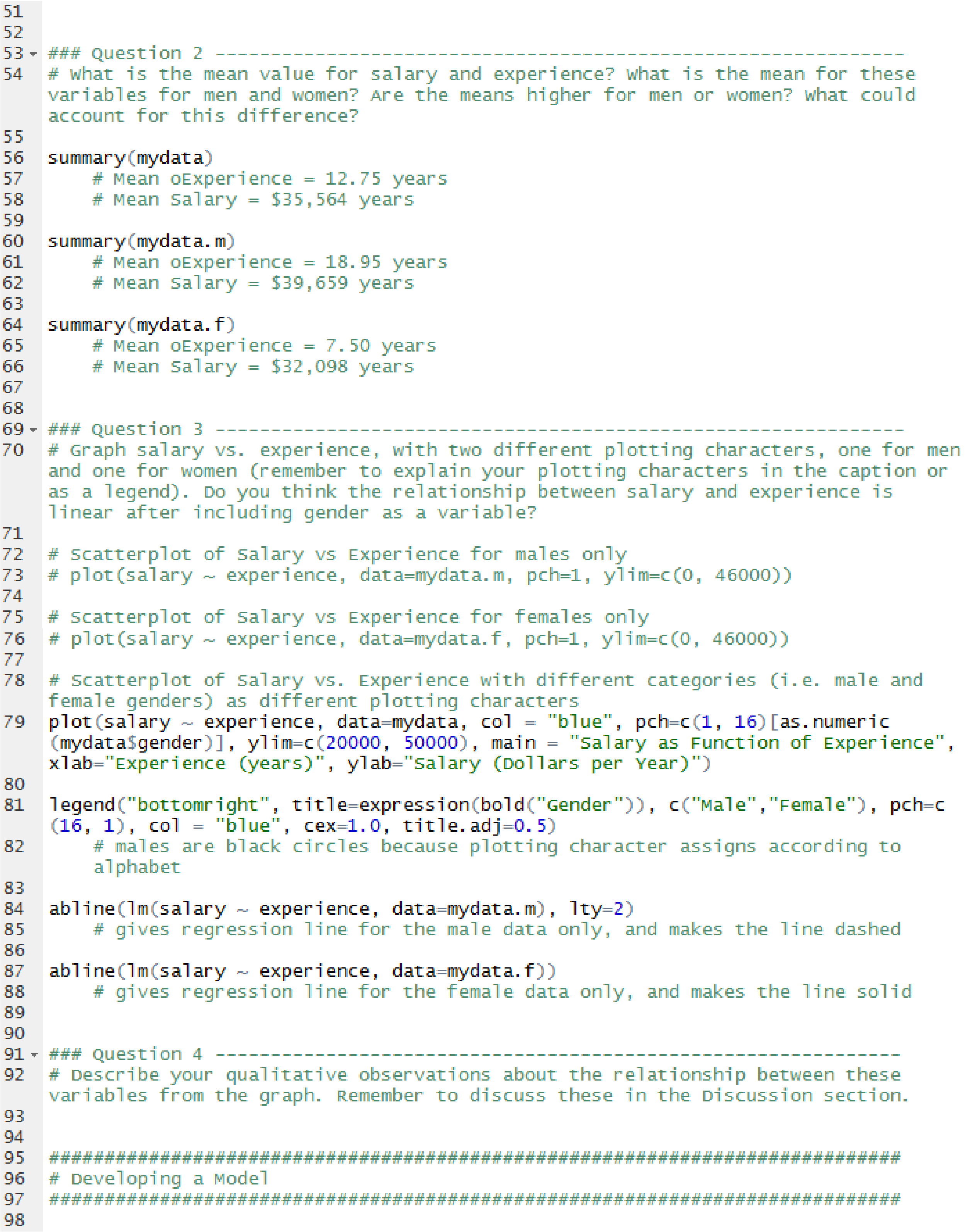
The model can be improved by: (1) incorporating another continuous variable to predict the response variable (salary [$/year]) better; (2) replacing the explanatory variable with another continuous variable such as generated profit; (3) adding another categorical variable with levels for which department they are situated in Spectra Technologies; (4) adding another categorical variable with levels of whether an employee works full- or part-time; (5) adding another categorical variable with levels of whether an employee has an undergraduate or graduate degree; and (6) obtaining more data points across a wider range of explanatory variable values. A wider range of data points would keep the slope and intercept values more stable and less sensitive to change upon receiving additional future data. In general, it is preferred to use continuous variables instead of more categorical variables because they use up fewer degrees of freedom. If more degrees of freedom are used, then the significance of the regression may be compromised if enough data points are not obtained.

4 Appendix

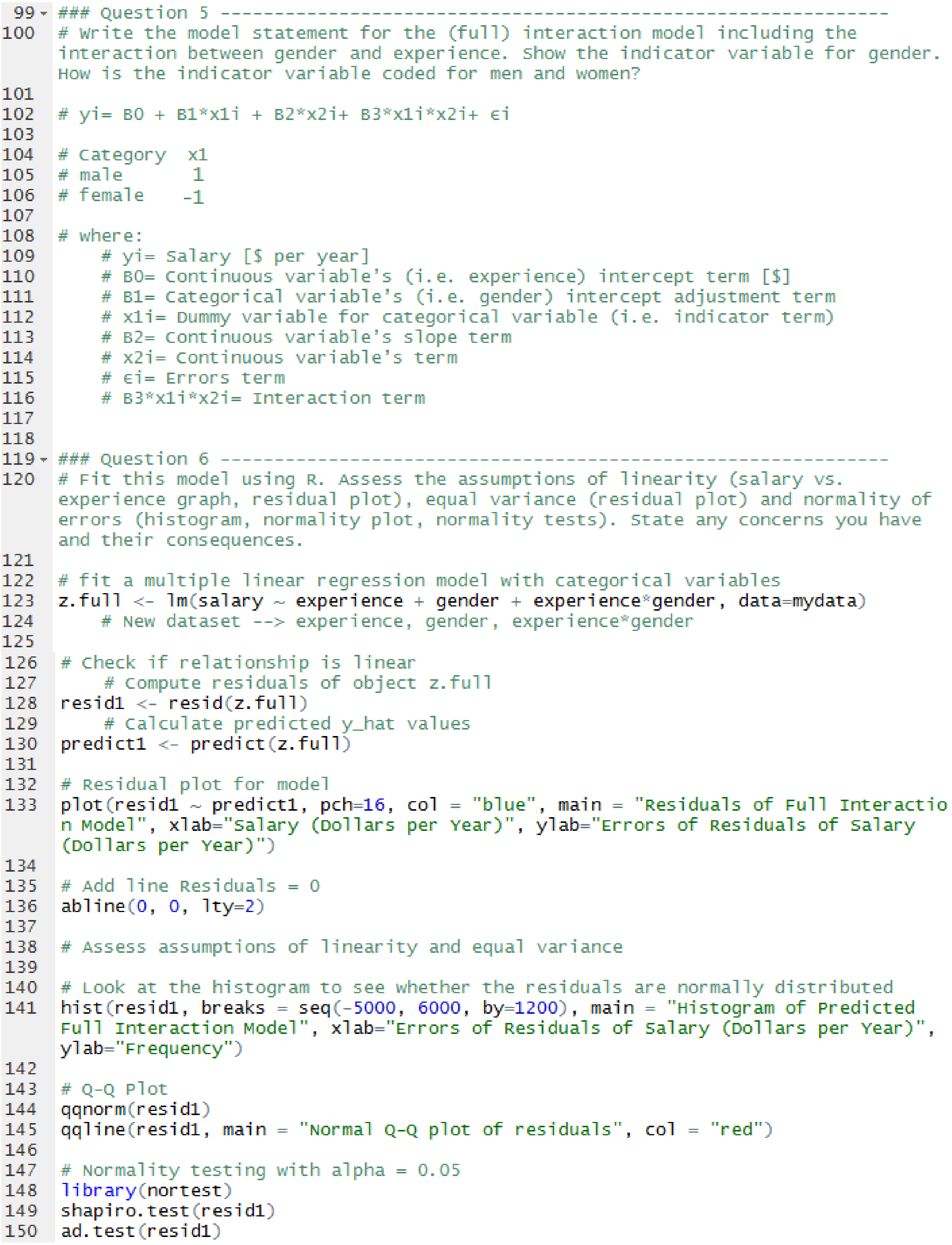
R Code for Analysis



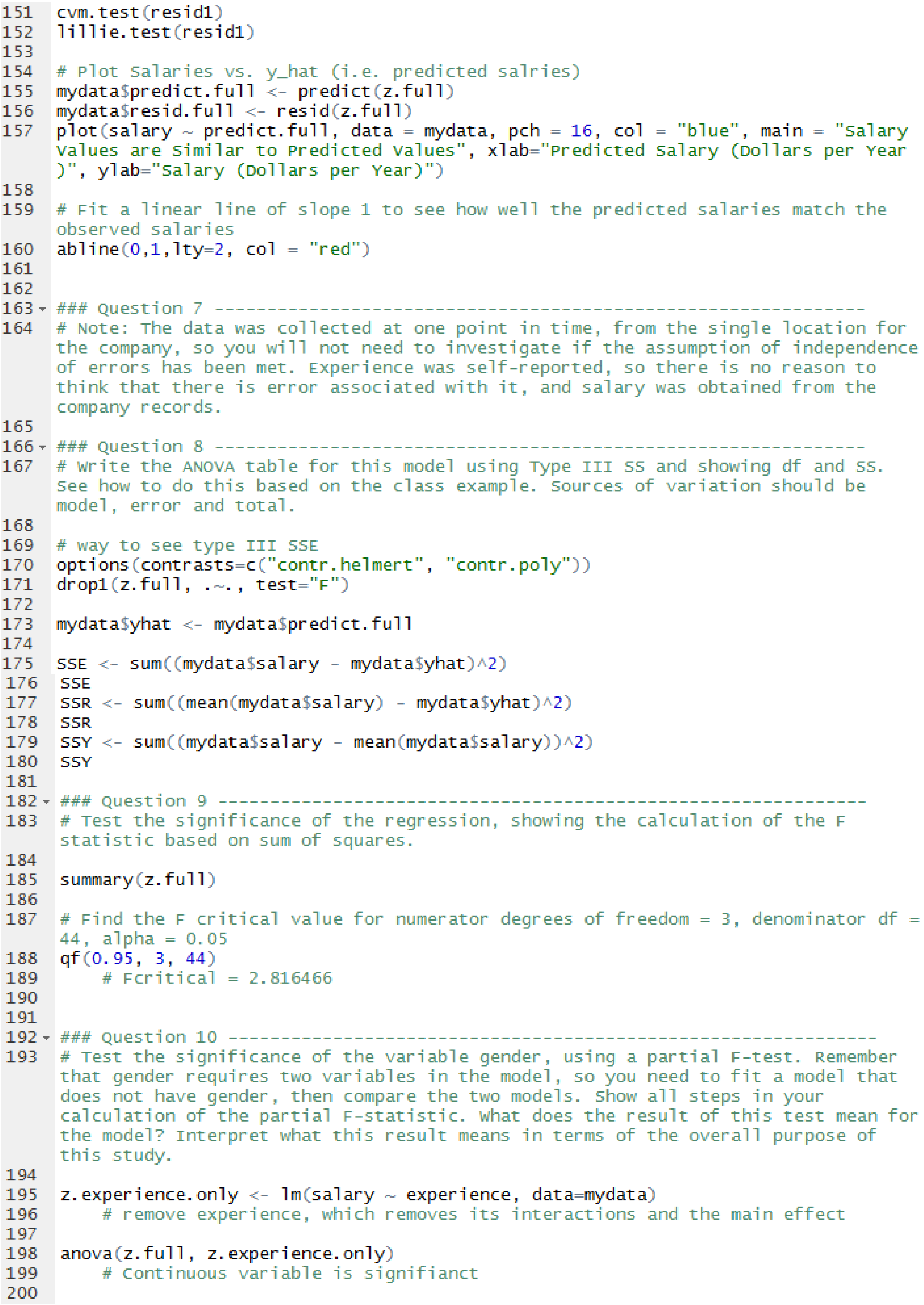
```
#############################################################################
# COMM 581 - Assignment 06 - Multiple Linear Regression with
#                            Categorical Variables
# Instructor: Martha Essak
# Gurpal Bisra
# Student #: 69295061
# Due date: Thursday Oct. 20, 2016 (11pm)
#############################################################################

# Background: You are analyzing data from a company, Spectra Technologies (science or engineering*), to determine if they are giving equal pay to men and women after controlling statistically for their level of experience. If there is a difference in salary between men and women, the company needs to know how much of a difference there is, and if this difference changes depending on level of experience (if there is an interaction between gender and experience).

# All graphs should be created using R, and all graphs discussed in your assignment should be included with your submission. You can use sum of squares and standard errors from the R output. Show calculations for test statistics, confidence intervals and measures of goodness of fit based on sums of squares.

# You will be presenting this assignment as a report to the company, with your recommendations.

mydata <- read.csv("Salary_experience_gender_data.csv", header=TRUE)
# import data from csv, and declaring header row at top

str(mydata)
    # 'data.frame': 48 obs. of  3 variables:
    # $ experience: int  0 0 1 0 0 0 2 1 2 2 ...
    # $ gender     : Factor w/ 2 levels "female","male": 1 1 1 1 1 1 1 1 1 1 ...
    # $ salary     : int  28300 30400 28950 30850 28700 31550 30450 30450 28500...

#############################################################################
# Summary Statistics (Present these Data in One or More Tables)
#############################################################################

### Question 1 --------------------------------------------------------------
# What is the range of values for salary and experience? What is the range for these variables for men and women? Are the ranges different, and what could account for this difference?

summary(mydata)
    # Range of Experience = [0, 35] years
    # Range of Salary = [28300, 45000] years

table(mydata$gender)
    # female   male
    # 26       22

# separate the categories into separate datasets
mydata.m <- subset(mydata, mydata$gender == "male")
mydata.f <- subset(mydata, mydata$gender == "female")

summary(mydata.m)
    # Range of Experience = [1, 35] years
    # Range of Salary = [34550, 45000] years

summary(mydata.f)
    # Range of Experience = [0, 29] years
    # Range of Salary = [28300, 38200] years
```



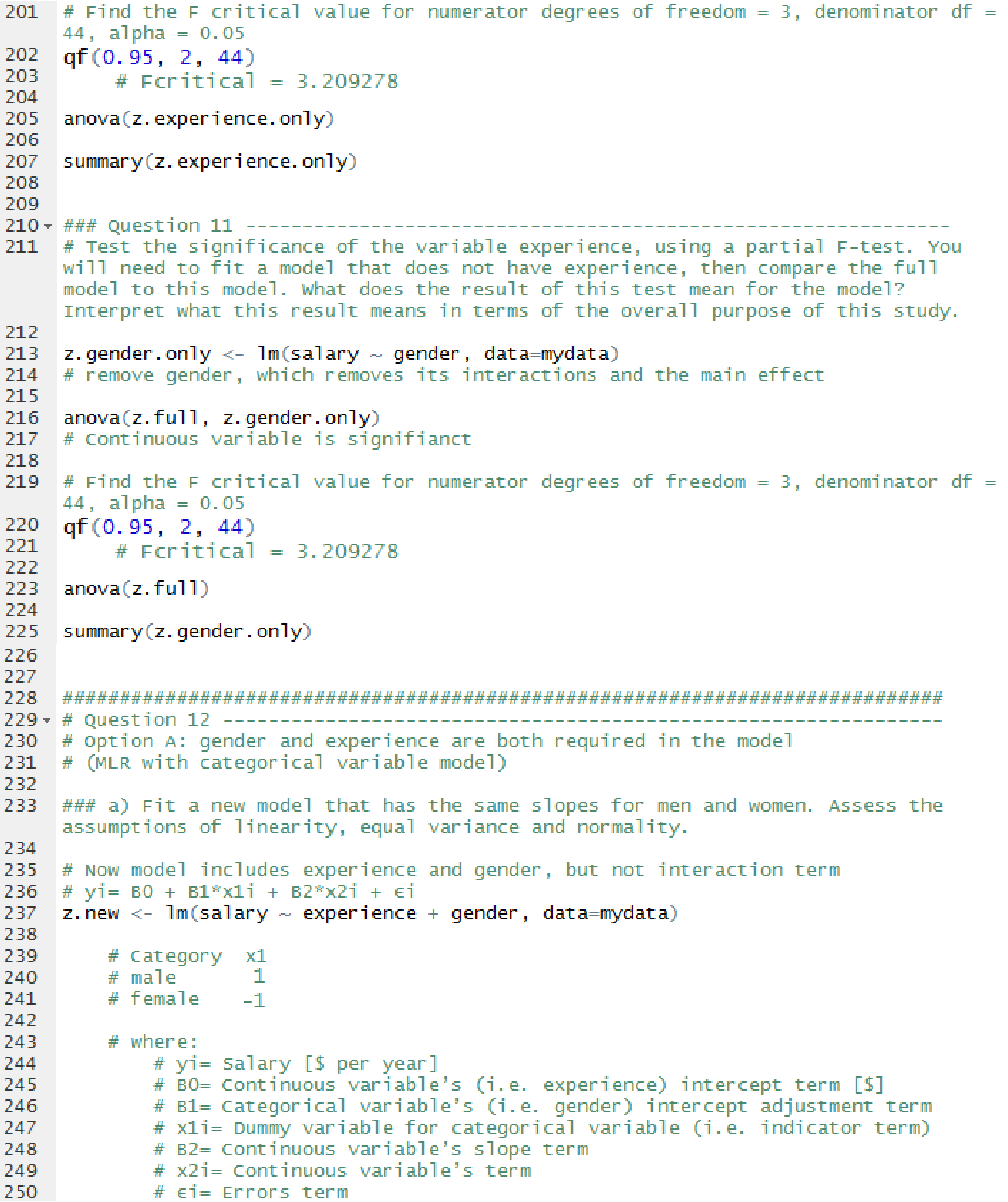
```r


### Question 2 --------------------------------------------------------------
# What is the mean value for salary and experience? What is the mean for these variables for men and women? Are the means higher for men or women? What could account for this difference?

summary(mydata)
    # Mean oExperience = 12.75 years
    # Mean Salary = $35,564 years

summary(mydata.m)
    # Mean oExperience = 18.95 years
    # Mean Salary = $39,659 years

summary(mydata.f)
    # Mean oExperience = 7.50 years
    # Mean Salary = $32,098 years


### Question 3 --------------------------------------------------------------
# Graph salary vs. experience, with two different plotting characters, one for men and one for women (remember to explain your plotting characters in the caption or as a legend). Do you think the relationship between salary and experience is linear after including gender as a variable?

# Scatterplot of Salary vs Experience for males only
# plot(salary ~ experience, data=mydata.m, pch=1, ylim=c(0, 46000))

# Scatterplot of Salary vs Experience for females only
# plot(salary ~ experience, data=mydata.f, pch=1, ylim=c(0, 46000))

# Scatterplot of Salary vs. Experience with different categories (i.e. male and female genders) as different plotting characters
plot(salary ~ experience, data=mydata, col = "blue", pch=c(1, 16)[as.numeric(mydata$gender)], ylim=c(20000, 50000), main = "Salary as Function of Experience", xlab="Experience (years)", ylab="Salary (Dollars per Year)")

legend("bottomright", title=expression(bold("Gender")), c("Male","Female"), pch=c(16, 1), col = "blue", cex=1.0, title.adj=0.5)
    # males are black circles because plotting character assigns according to alphabet

abline(lm(salary ~ experience, data=mydata.m), lty=2)
    # gives regression line for the male data only, and makes the line dashed

abline(lm(salary ~ experience, data=mydata.f))
    # gives regression line for the female data only, and makes the line solid


### Question 4 --------------------------------------------------------------
# Describe your qualitative observations about the relationship between these variables from the graph. Remember to discuss these in the Discussion section.


#############################################################################
# Developing a Model
#############################################################################

```



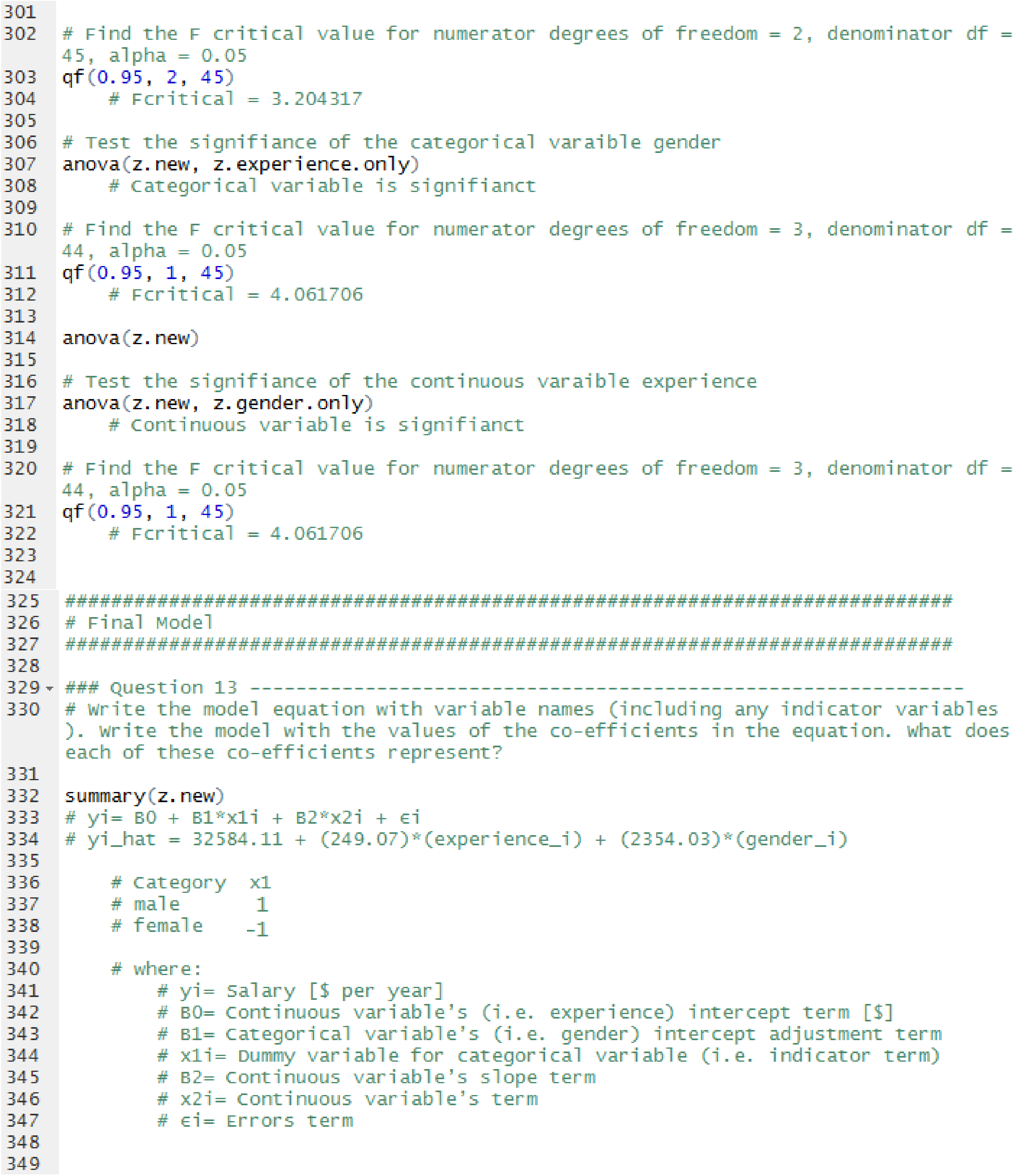
```
### Question 5 ----------------------------------------------------------------
# Write the model statement for the (full) interaction model including the interaction between gender and experience. Show the indicator variable for gender. How is the indicator variable coded for men and women?

# yi= B0 + B1*x1i + B2*x2i+ B3*x1i*x2i+ ϵi

# Category  x1
# male       1
# female    -1

# where:
    # yi= Salary [$ per year]
    # B0= Continuous variable’s (i.e. experience) intercept term [$]
    # B1= Categorical variable’s (i.e. gender) intercept adjustment term
    # x1i= Dummy variable for categorical variable (i.e. indicator term)
    # B2= Continuous variable’s slope term
    # x2i= Continuous variable’s term
    # ϵi= Errors term
    # B3*x1i*x2i= Interaction term


### Question 6 ----------------------------------------------------------------
# Fit this model using R. Assess the assumptions of linearity (salary vs. experience graph, residual plot), equal variance (residual plot) and normality of errors (histogram, normality plot, normality tests). State any concerns you have and their consequences.

# fit a multiple linear regression model with categorical variables
z.full <- lm(salary ~ experience + gender + experience*gender, data=mydata)
    # New dataset --> experience, gender, experience*gender

# Check if relationship is linear
    # Compute residuals of object z.full
resid1 <- resid(z.full)
    # Calculate predicted y_hat values
predict1 <- predict(z.full)

# Residual plot for model
plot(resid1 ~ predict1, pch=16, col = "blue", main = "Residuals of Full Interaction Model", xlab="Salary (Dollars per Year)", ylab="Errors of Residuals of Salary (Dollars per Year)")

# Add line Residuals = 0
abline(0, 0, lty=2)

# Assess assumptions of linearity and equal variance

# Look at the histogram to see whether the residuals are normally distributed
hist(resid1, breaks = seq(-5000, 6000, by=1200), main = "Histogram of Predicted Full Interaction Model", xlab="Errors of Residuals of Salary (Dollars per Year)", ylab="Frequency")

# Q-Q Plot
qqnorm(resid1)
qqline(resid1, main = "Normal Q-Q plot of residuals", col = "red")

# Normality testing with alpha = 0.05
library(nortest)
shapiro.test(resid1)
ad.test(resid1)
```



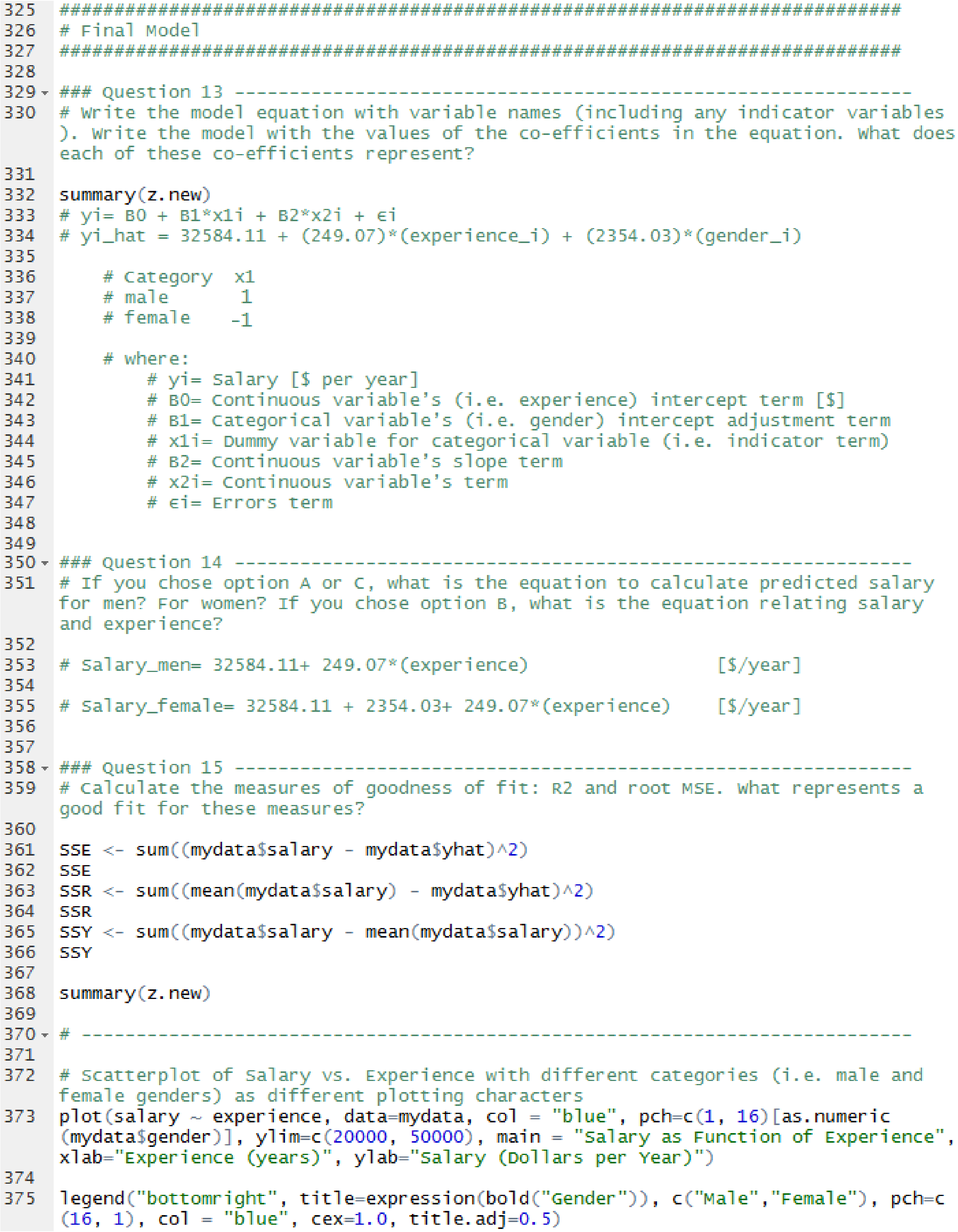
```
cvm.test(resid1)
lillie.test(resid1)

# Plot Salaries vs. y_hat (i.e. predicted salries)
mydata$predict.full <- predict(z.full)
mydata$resid.full <- resid(z.full)
plot(salary ~ predict.full, data = mydata, pch = 16, col = "blue", main = "Salary Values are Similar to Predicted Values", xlab="Predicted Salary (Dollars per Year)", ylab="Salary (Dollars per Year)")

# Fit a linear line of slope 1 to see how well the predicted salaries match the observed salaries
abline(0,1,lty=2, col = "red")


### Question 7 --------------------------------------------------------------
# Note: The data was collected at one point in time, from the single location for the company, so you will not need to investigate if the assumption of independence of errors has been met. Experience was self-reported, so there is no reason to think that there is error associated with it, and salary was obtained from the company records.

### Question 8 --------------------------------------------------------------
# Write the ANOVA table for this model using Type III SS and showing df and SS. See how to do this based on the class example. Sources of variation should be model, error and total.

# way to see type III SSE
options(contrasts=c("contr.helmert", "contr.poly"))
drop1(z.full, .~., test="F")

mydata$yhat <- mydata$predict.full

SSE <- sum((mydata$salary - mydata$yhat)^2)
SSE
SSR <- sum((mean(mydata$salary) - mydata$yhat)^2)
SSR
SSY <- sum((mydata$salary - mean(mydata$salary))^2)
SSY

### Question 9 --------------------------------------------------------------
# Test the significance of the regression, showing the calculation of the F statistic based on sum of squares.

summary(z.full)

# Find the F critical value for numerator degrees of freedom = 3, denominator df = 44, alpha = 0.05
qf(0.95, 3, 44)
    # Fcritical = 2.816466


### Question 10 -------------------------------------------------------------
# Test the significance of the variable gender, using a partial F-test. Remember that gender requires two variables in the model, so you need to fit a model that does not have gender, then compare the two models. Show all steps in your calculation of the partial F-statistic. What does the result of this test mean for the model? Interpret what this result means in terms of the overall purpose of this study.

z.experience.only <- lm(salary ~ experience, data=mydata)
    # remove experience, which removes its interactions and the main effect

anova(z.full, z.experience.only)
    # Continuous variable is signifiianct

```



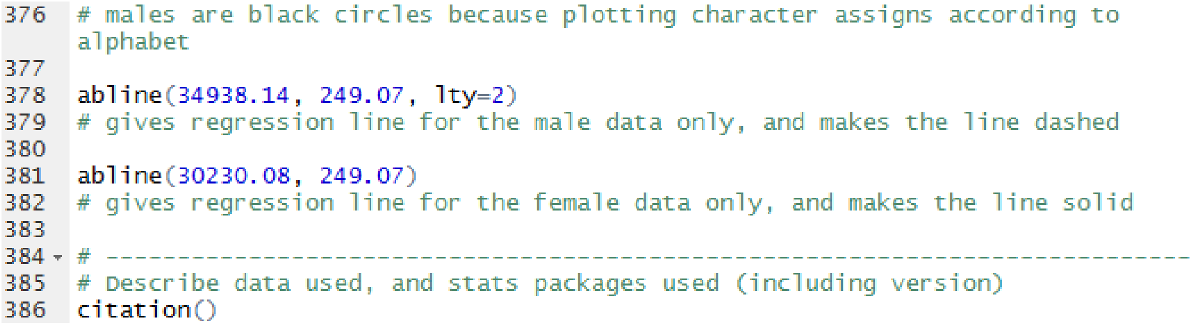
```r
# Find the F critical value for numerator degrees of freedom = 3, denominator df = 44, alpha = 0.05
qf(0.95, 2, 44)
    # Fcritical = 3.209278

anova(z.experience.only)

summary(z.experience.only)


### Question 11 ---------------------------------------------------------------
# Test the significance of the variable experience, using a partial F-test. You will need to fit a model that does not have experience, then compare the full model to this model. What does the result of this test mean for the model? Interpret what this result means in terms of the overall purpose of this study.

z.gender.only <- lm(salary ~ gender, data=mydata)
# remove gender, which removes its interactions and the main effect

anova(z.full, z.gender.only)
# Continuous variable is signifianct

# Find the F critical value for numerator degrees of freedom = 3, denominator df = 44, alpha = 0.05
qf(0.95, 2, 44)
    # Fcritical = 3.209278

anova(z.full)

summary(z.gender.only)


##############################################################################
# Question 12 -----------------------------------------------------------------
# Option A: gender and experience are both required in the model
# (MLR with categorical variable model)

### a) Fit a new model that has the same slopes for men and women. Assess the assumptions of linearity, equal variance and normality.

# Now model includes experience and gender, but not interaction term
# yi= B0 + B1*x1i + B2*x2i + ϵi
z.new <- lm(salary ~ experience + gender, data=mydata)

    # Category  x1
    # male       1
    # female    -1

    # where:
        # yi= Salary [$ per year]
        # B0= Continuous variable’s (i.e. experience) intercept term [$]
        # B1= Categorical variable’s (i.e. gender) intercept adjustment term
        # x1i= Dummy variable for categorical variable (i.e. indicator term)
        # B2= Continuous variable’s slope term
        # x2i= Continuous variable’s term
        # ϵi= Errors term

# Check if relationship is linear
    # Compute residuals of object z.new
resid2 <- resid(z.new)
    # Calculate predicted y_hat values
predict2 <- predict(z.new)

# Residual plot for model
plot(resid2 ~ predict2, pch=16, col = "blue", main = "Residuals of Non-Interaction Model", xlab="Salary (Dollars per Year)", ylab="Errors of Residuals of Salary (Dollars per Year)")

# Add line Residuals = 0
abline(0, 0, lty=2)

# Assess assumptions of linearity and equal variance

# Look at the histogram to see whether the residuals are normally distributed
hist(resid2, breaks = seq(-5000, 6000, by=1200), main = "Histogram of Predicted Non-Interaction Model", xlab="Errors of Residuals of Salary (Dollars per Year)", ylab="Frequency")

# Q-Q Plot
qqnorm(resid2)
qqline(resid2, main = "Normal Q-Q plot of Residuals", col = "red")

# Normality testing with alpha = 0.05
library(nortest)
shapiro.test(resid2)
ad.test(resid2)
cvm.test(resid2)
lillie.test(resid2)

# Plot Salaries vs. y_hat (i.e. predicted salries)
mydata$predict.new <- predict(z.new)
mydata$resid.new <- resid(z.new)
plot(salary ~ predict.new, data = mydata, pch = 16, col = "blue", main = "Salary Values are Similar to Non-Interaction Predicted Values", xlab="Predicted Salary (Dollars per Year)", ylab="Salary (Dollars per Year)")

# Fit a linear line of slope 1 to see how well the predicted salaries match the observed salaries
abline(0,1,lty=2, col = "red")

### b) Use a partial F-test to determine if the slopes are the same or different. What does the result of this test mean for the model? Interpret what this result means in terms of the overall purpose of this study.

# Test the signficiance of the regression when compared to z.new
anova(z.full, z.new)
    # Interaction is not significant

# Find the F critical value for numerator degrees of freedom = 1, denominator df = 46, alpha = 0.05
qf(0.95, 1, 46)
    # Fcritical = 4.061706

# Test the significance of the regression, showing the calculation of the F statistic based on sum of squares.

summary(z.new)
```



```

# Find the F critical value for numerator degrees of freedom = 2, denominator df = 45, alpha = 0.05
qf(0.95, 2, 45)
    # Fcritical = 3.204317

# Test the signifiance of the categorical varaible gender
anova(z.new, z.experience.only)
    # Categorical variable is signifianct

# Find the F critical value for numerator degrees of freedom = 3, denominator df = 44, alpha = 0.05
qf(0.95, 1, 45)
    # Fcritical = 4.061706

anova(z.new)

# Test the signifiance of the continuous varaible experience
anova(z.new, z.gender.only)
    # Continuous variable is signifianct

# Find the F critical value for numerator degrees of freedom = 3, denominator df = 44, alpha = 0.05
qf(0.95, 1, 45)
    # Fcritical = 4.061706


###########################################################################
# Final Model
###########################################################################

### Question 13 ------------------------------------------------------------
# Write the model equation with variable names (including any indicator variables). Write the model with the values of the co-efficients in the equation. What does each of these co-efficients represent?

summary(z.new)
# yi= B0 + B1*x1i + B2*x2i + ϵi
# yi_hat = 32584.11 + (249.07)*(experience_i) + (2354.03)*(gender_i)

    # Category  x1
    # male       1
    # female    -1

    # where:
        # yi= Salary [$ per year]
        # B0= Continuous variable's (i.e. experience) intercept term [$]
        # B1= Categorical variable's (i.e. gender) intercept adjustment term
        # x1i= Dummy variable for categorical variable (i.e. indicator term)
        # B2= Continuous variable's slope term
        # x2i= Continuous variable's term
        # ϵi= Errors term


```



```
########################################################################
# Final Model
########################################################################

### Question 13 ----------------------------------------------------------------
# Write the model equation with variable names (including any indicator variables). Write the model with the values of the co-efficients in the equation. What does each of these co-efficients represent?

summary(z.new)
# yi= B0 + B1*x1i + B2*x2i + ϵi
# yi_hat = 32584.11 + (249.07)*(experience_i) + (2354.03)*(gender_i)

    # Category  x1
    # male       1
    # female    -1

    # where:
        # yi= Salary [$ per year]
        # B0= Continuous variable's (i.e. experience) intercept term [$]
        # B1= Categorical variable's (i.e. gender) intercept adjustment term
        # x1i= Dummy variable for categorical variable (i.e. indicator term)
        # B2= Continuous variable's slope term
        # x2i= Continuous variable's term
        # ϵi= Errors term


### Question 14 ----------------------------------------------------------------
# If you chose option A or C, what is the equation to calculate predicted salary for men? For women? If you chose option B, what is the equation relating salary and experience?

# Salary_men= 32584.11+ 249.07*(experience)                [$/year]

# Salary_female= 32584.11 + 2354.03+ 249.07*(experience)    [$/year]


### Question 15 ----------------------------------------------------------------
# Calculate the measures of goodness of fit: R2 and root MSE. What represents a good fit for these measures?

SSE <- sum((mydata$salary - mydata$yhat)^2)
SSE
SSR <- sum((mean(mydata$salary) - mydata$yhat)^2)
SSR
SSY <- sum((mydata$salary - mean(mydata$salary))^2)
SSY

summary(z.new)

# ------------------------------------------------------------------------------

# Scatterplot of Salary vs. Experience with different categories (i.e. male and female genders) as different plotting characters
plot(salary ~ experience, data=mydata, col = "blue", pch=c(1, 16)[as.numeric(mydata$gender)], ylim=c(20000, 50000), main = "Salary as Function of Experience", xlab="Experience (years)", ylab="Salary (Dollars per Year)")

legend("bottomright", title=expression(bold("Gender")), c("Male","Female"), pch=c(16, 1), col = "blue", cex=1.0, title.adj=0.5)
```



```r
# males are black circles because plotting character assigns according to alphabet

abline(34938.14, 249.07, lty=2)
# gives regression line for the male data only, and makes the line dashed

abline(30230.08, 249.07)
# gives regression line for the female data only, and makes the line solid

# ---------------------------------------------------------------------------
# Describe data used, and stats packages used (including version)
citation()
```