Ashish Kumar Gupta,Niraj Kumar,Vishesh Bhagat

GREAT LAKES INSTITUTE OF MANAGEMENT  PGPBABI – MARCH 2017

advanced statistics group assignment GROUP

Table of Contents

[Assignment 1 3](#_Toc487575785)

[Advanced Statistics (one-way ANOVA, Two-way ANOVA, ANCOVA and Regression Analysis) 3](#_Toc487575786)

[Question 1: 3](#_Toc487575787)

[Solution 1 3](#_Toc487575788)

[Significance of experimental treatments on the outcome variables: 3](#_Toc487575789)

[One-Way ANNOVA Tests 4](#_Toc487575790)

# Assignment 1

# Advanced Statistics (one-way ANOVA, Two-way ANOVA, ANCOVA and Regression Analysis)

## Question 1:Design of Experiment – study the impact of superiority and uniqueness on customer response to Ads

A marketing Researcher has conducted an experiment to test the effects of different print ad copy elements on attitude toward the product and intention to buy. The design of the experiment was a simple 2×2 balanced design (with 25 subjects in each cell). The first factor involved either the inclusion (or exclusion) of a competitive claim (i.e. our product is superior to other available products on a particular dimension). The second factor involved either the inclusion (or exclusion) of a uniqueness claim (i.e., our product is unlike any other available product). Thus, 25 subjects saw the basic print ad (without either claim), 25 subjects saw basic ad plus competitive claim, 25 subjects saw the basic ad plus uniqueness claim and 25 subjects saw the basic ad plus both claims. The following four response measures were taken:

Y­1 How much do you like this product? (from “Not at all” to “Very Much”)

Y2 I like this product (from “Strongly Disagree” to “Strongly agree”)

Y­3 I would buy this product (from “Strongly Disagree” to “Strongly agree”)

Y4 What is the likelihood you would buy this product? (stated probability)

All four items above are measured on a scale of 1 to 100. The data are available in the file AD\_CLAIM\_TEST. Analyze the data from the experiment and answer the following questions.

1. Which of the experimental treatments if any is significant at the 0.05 level for each of the four outcome variables Y­1, Y­2, Y­3, Y­4. Conduct Individual one –way Analysis of variance. State the hypothesis (null and alternate), test for assumptions, analyze and report your results
2. Test for interaction effects for each of the four outcome variables.
3. How would you test by taking into account simultaneously all these variables?
4. Now do the above analysis using regression approach. Report your results.

## Solution

### Significance of experimental treatments on the outcome variables:

### One-Way ANNOVA Tests

#### One-Way ANOVA Test for Y1

‘Y1 - How much do you like this product?’ is dependent variable. ‘superior’ is independent variable.

> aov(adclaimtest\_data$Y1~adclaimtest\_data$superior, data=adclaimtest\_data) ->A11

> A11

Call:

aov(formula = adclaimtest\_data$Y1 ~ adclaimtest\_data$superior,

data = adclaimtest\_data)

Terms:

adclaimtest\_data$superior Residuals

Sum of Squares 0.16 7789.84

Deg. of Freedom 1 98

Residual standard error: 8.915613

Estimated effects may be unbalanced

> summary(A11)

Df Sum Sq Mean Sq F value Pr(>F)

adclaimtest\_data$superior 1 0 0.16 0.002 0.964

Residuals 98 7790 79.49

Since P value is above 0.05, we do not reject the null hypothesis – there exists an evidence that a customer is likely to buy a product when the a competitive claim was included in the Ad.

#### One-Way ANOVA Test for Y2

‘Y2 - I like this product?’ is dependent variable. ‘superior’ is independent variable.

> aov(adclaimtest\_data$Y2~adclaimtest\_data$superior, data=adclaimtest\_data) ->A12

> summary(A12)

Df Sum Sq Mean Sq F value Pr(>F)

adclaimtest\_data$superior 1 1 1.44 0.019 0.89

Residuals 98 7277 74.25

> A12

Call:

aov(formula = adclaimtest\_data$Y2 ~ adclaimtest\_data$superior,

data = adclaimtest\_data)

Terms:

adclaimtest\_data$superior Residuals

Sum of Squares 1.44 7276.92

Deg. of Freedom 1 98

Residual standard error: 8.617093

Estimated effects may be unbalanced

Since P value is above 0.05, we do not reject the null hypothesis – there exists an evidence that a customer likes the product when a competitive claim was included in the Ad.

#### One-Way ANOVA Test for Y3

‘Y3 - I would buy this product?’ is dependent variable. ‘superior’ is independent variable.

> aov(adclaimtest\_data$Y3~adclaimtest\_data$superior, data=adclaimtest\_data) ->A13

> A13

Call:

aov(formula = adclaimtest\_data$Y3 ~ adclaimtest\_data$superior,

data = adclaimtest\_data)

Terms:

adclaimtest\_data$superior Residuals

Sum of Squares 141.61 6883.14

Deg. of Freedom 1 98

Residual standard error: 8.380699

Estimated effects may be unbalanced

> summary(A13)

Df Sum Sq Mean Sq F value Pr(>F)

adclaimtest\_data$superior 1 142 141.61 2.016 0.159

Residuals 98 6883 70.24

Since P value is above 0.05, we do not reject the null hypothesis – there exists an evidence that a customer buys the product when a competitive claim was included in the Ad.

#### One-Way ANOVA Test for Y4

‘Y4 – What is the likelihood you would buy this product?’ is dependent variable. ‘superior’ is independent variable.

|  |
| --- |
| > aov(adclaimtest\_data$Y4~adclaimtest\_data$superior, data=adclaimtest\_data) ->A14  > summary(A14)  Df Sum Sq Mean Sq F value Pr(>F)  adclaimtest\_data$superior 1 676 676.0 11.06 0.00124 \*\*  Residuals 98 5991 61.1  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > A14  Call:  aov(formula = adclaimtest\_data$Y4 ~ adclaimtest\_data$superior,  data = adclaimtest\_data)  Terms:  adclaimtest\_data$superior Residuals  Sum of Squares 676.00 5990.84  Deg. of Freedom 1 98  Residual standard error: 7.818633  Estimated effects may be unbalanced |
|  |
|  |

Since P value is less than 0.05, we reject the null hypothesis – there is not enough evidence to prove that there is a likelihood that the customer would by the product when a competitive claim was included in the Ad.

#### One-Way ANOVA Test for Y1

‘Y1 - How much do you like this product?’ is dependent variable. ‘unique’ is independent variable.

> aov(adclaimtest\_data$Y1~adclaimtest\_data$unique, data=adclaimtest\_data) ->A21

> A21

Call:

aov(formula = adclaimtest\_data$Y1 ~ adclaimtest\_data$unique,

data = adclaimtest\_data)

Terms:

adclaimtest\_data$unique Residuals

Sum of Squares 432.64 7357.36

Deg. of Freedom 1 98

Residual standard error: 8.664589

Estimated effects may be unbalanced

> summary(A21)

Df Sum Sq Mean Sq F value Pr(>F)

adclaimtest\_data$unique 1 433 432.6 5.763 0.0183 \*

Residuals 98 7357 75.1

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Since P value is less than 0.05, we **reject** the null hypothesis – there is not enough evidence to prove that there is a likelihood that the customer would by the product when a uniqueness claim was included in the Ad.

#### One-Way ANOVA Test for Y2

‘Y2 - I like this product?’ is dependent variable. ‘unique’ is independent variable.

> aov(adclaimtest\_data$Y2~adclaimtest\_data$unique, data=adclaimtest\_data) ->A22

> A22

Call:

aov(formula = adclaimtest\_data$Y2 ~ adclaimtest\_data$unique,

data = adclaimtest\_data)

Terms:

adclaimtest\_data$unique Residuals

Sum of Squares 324.00 6954.36

Deg. of Freedom 1 98

Residual standard error: 8.423945

Estimated effects may be unbalanced

> summary(A22)

Df Sum Sq Mean Sq F value Pr(>F)

adclaimtest\_data$unique 1 324 324 4.566 0.0351 \*

Residuals 98 6954 71

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Since P value is less than 0.05, we **reject** the null hypothesis – there is not enough evidence that a customer likes the product when a uniqueness claim was included in the Ad.

#### One-Way ANOVA Test for Y3

‘Y3 - I would buy this product?’ is dependent variable. ‘unique’ is independent variable.

> aov(adclaimtest\_data$Y3~adclaimtest\_data$unique, data=adclaimtest\_data) ->A23

> A23

Call:

aov(formula = adclaimtest\_data$Y3 ~ adclaimtest\_data$unique,

data = adclaimtest\_data)

Terms:

adclaimtest\_data$unique Residuals

Sum of Squares 110.25 6914.50

Deg. of Freedom 1 98

Residual standard error: 8.399769

Estimated effects may be unbalanced

> summary(A23)

Df Sum Sq Mean Sq F value Pr(>F)

adclaimtest\_data$unique 1 110 110.25 1.563 0.214

Residuals 98 6914 70.56

Since P value is above 0.05, we do not reject the null hypothesis – there exists an evidence that a customer buys the product when a uniqueness claim was included in the Ad.

#### One-Way ANOVA Test for Y4

‘Y4 – What is the likelihood you would buy this product?’ is dependent variable. ‘unique’ is independent variable.

|  |
| --- |
| > aov(adclaimtest\_data$Y4~adclaimtest\_data$unique, data=adclaimtest\_data) ->A24  > A24  Call:  aov(formula = adclaimtest\_data$Y4 ~ adclaimtest\_data$unique,  data = adclaimtest\_data)  Terms:  adclaimtest\_data$unique Residuals  Sum of Squares 1.44 6665.40  Deg. of Freedom 1 98  Residual standard error: 8.247077  Estimated effects may be unbalanced  > summary(A24)  Df Sum Sq Mean Sq F value Pr(>F)  adclaimtest\_data$unique 1 1 1.44 0.021 0.885  Residuals 98 6665 68.01  Since P value is greater than 0.05, we do not reject the null hypothesis – there is evidence to prove that there is a likelihood that the customer would by the product when a uniqueness claim was included in the Ad. |

### Interaction between two variables

#### Two-Way ANOVA Test for Y1

‘Y1 - How much do you like this product? is a dependent variable. ‘superior’ and ‘unique’ are independent variables.

> aov(adclaimtest\_data$Y1 ~ adclaimtest\_data$superior+adclaimtest\_data$unique, data=adclaimtest\_data)-> A31

> A31

Call:

aov(formula = adclaimtest\_data$Y1 ~ adclaimtest\_data$superior +

adclaimtest\_data$unique, data = adclaimtest\_data)

Terms:

adclaimtest\_data$superior adclaimtest\_data$unique Residuals

Sum of Squares 0.16 432.64 7357.20

Deg. of Freedom 1 1 97

Residual standard error: 8.709043

Estimated effects may be unbalanced

> summary(A31)

Df Sum Sq Mean Sq F value Pr(>F)

adclaimtest\_data$superior 1 0 0.2 0.002 0.9635

adclaimtest\_data$unique 1 433 432.6 5.704 0.0189 \*

Residuals 97 7357 75.8

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Since P value is greater than 0.05, we do not reject the null hypothesis – there is evidence to prove that there is a customer liked the product more when a superior and uniqueness claim was included in the Ad.

**Question 2**

Ofir and Simonson (2001)conducted several experiments investigating the effect of “expecting to evaluate” on ratings of product quality. They found that in a variety of situations, consumers are more critical (and provide ore negative ratings) when they are told in advance that they will be asked to provide an evaluation after they have experienced a product or service.

We now consider a subset of the data collected by Ofir and Simonson in one of their studies. A group of 201 subjects was asked to read an article that “appeared in the first issue of a new magazine” and all subjects were told that “we will ask for your evaluation of the writing quality of the magazine based on the article”. Thus all subjects expect to evaluate the article after reading it. For this group there were two experimental treatments: the quality of the article (high, low) and expectations about the quality of the article (hight, low). To accomplish the first manipulation, Ofir and Simonson took an article from “The New York Times”. The “high-quality” treatment was the original article; the “low-quality” version was created by using slang and poor grammar without changing the content. They manipulated expectations of quality by telling the subjects the new magazine was either “started by experienced journalists” (high expectations) or “started by freshman students at a local high school” (low expectations). For the dependent variable, Ofir and Simoson asked, “What is your evaluation of the magazine based on the professionalism and quality of writing, grammar, language, and editing of the article?” Subjects responded on a seven point scale from “Very favorable” (1) to “Very unfavorable” (7).

Data from the experiment are available in the file EXPECT\_EVAL. There are six columns of data in the file, described below:

Col 1: Subject ID

Col 2: Evaluation manipulation (a=expecting to evaluate)

Col 3: Quality manipulation (b=bad quality, g-good quality)

Col 4: Expectations manipulation ( h=expect high quality, 1 = expect low quality)

Col 5: Y1 (evaluation of magazine, 7-point scale)

Col 6: Y2 (agreement with issue in the article, 7-point scale)

1. Using ANOVA, test the null hypothesis that there are no differences in the evaluation of the magazine across the experimental treatment groups. Can you reject the hypothesis at the 0.05 level? If so, which of the treatments is significant?
2. Describe the nature of interaction between quality and expected quality.

Please use both techniques (AOV) and LM command for the analysis.

**Null hypothesis: there is no significance difference between the quality of manipulation and expectations manipulation on Evaluation of a magazine**

**#The null hypothesis for ANOVA is that the mean**

**#(average value of the dependent variable) is the same for all groups.**

**# The alternative #or research hypothesis is that the average is not the**

**# same for all groups.**

**set.seed(123)**

1. **setwd("G:/Grreat lakes/Advance stat/Group Assignment")**

**mydata<- read.csv("Expectations Evaluation.csv")**

**mydata$Y2<- as.numeric(mydata$Y2)**

**mydata$Y1<- as.numeric(mydata$Y1)**

**#glm(Y1~.,data=mydata)**

**# Outlier Analysis - Varaiable**

**outlier\_upper=function(x){**

**q = quantile(x)**

**names(q) = NULL**

**q1 = q[2]**

**q3 = q[4]**

**QR = q3-q1**

**return(q3+1.5\*QR);**

**}**

**outlier\_lower=function(x){**

**q = quantile(x)**

**names(q) = NULL**

**q1 = q[2]**

**q3 = q[4]**

**QR = q3-q1**

**return(q1-1.5\*QR);**

**}**

**# outlier limits validation ------------------**

**Y2\_upper = outlier\_upper(mydata$Y2)**

**Y2\_lower = outlier\_lower(mydata$Y2)**

**Y1\_upper = outlier\_upper(mydata$Y1)**

**Y1\_lower = outlier\_lower(mydata$Y1)**

**# Outlier data**

**mydata[mydata$Y2>Y2\_upper | mydata$Y2<Y2\_lower , ]**

**mydata[mydata$Y1>Y1\_upper | mydata$Y1<Y1\_lower , ]**

**mydata = subset( mydata, mydata$Y2<=Y2\_upper & mydata$Y2>=Y2\_lower)**

**mydata = subset( mydata, mydata$Y1<=Y1\_upper & mydata$Y1>=Y1\_lower)**

**nrow(mydata)**

**#mydata$Expectatations.Manipulation=ifelse(mydata$Expectatations.Manipulation=="h",7,1)**

**HI<-table(mydata$Expectatations.Manipulation,mydata$Quality.Manipulation)**

**#mean(HI[,1])**

**#mean(HI[,2])**

**aov(Y1~Quality.Manipulation,mydata)->ft**

**summary(ft)**

**TukeyHSD(ft)**

**tab<-table(mydata$Y1,mydata$Quality.Manipulation)**

**mean(tab[,1])**

**mean(tab[,2])**

**aov(Y1~Expectatations.Manipulation,mydata)->ft**

**summary(ft)**

**TukeyHSD(ft)**

**tab<-table(mydata$Y1,mydata$Expectatations.Manipulation)**

**mean(tab[,1])**

**mean(tab[,2])**

**aov(Y1~Expectatations.Manipulation+Quality.Manipulation,mydata)->ft**

**summary(ft)**

**second question. Interaction term**

**aov(Y1~Expectatations.Manipulation\*Quality.Manipulation,mydata)->ft**

**summary(ft)**

**TukeyHSD(ft)**

**To check if there is difference in between Expecation Manipulation and Quality Manipulation**

**aov(Expectatations.Manipulation~Quality.Manipulation,mydata)->ft**

**summary(ft)**

**TukeyHSD(ft)**

**# Using lm function**

**mydata$random <- runif(nrow(mydata), 0, 1);**

**mydata <- mydata[order(mydata$random),]**

**mydata.dev <- mydata[which(mydata$random <= 0.75),]**

**mydata.val <- mydata[which(mydata$random > 0.75),]**

**c(nrow(mydata.dev), nrow(mydata.val))**

**ft1<- lm(Y1~Quality.Manipulation,data=mydata.dev)**

**#ft1<- lm(Y1~.,data=mydata.dev)**

**str(mydata.val$Quality.Manipulation)**

**summary(ft1)**

**Quality.Manipulation<-data.frame(Quality.Manipulation= mydata.val$Quality.Manipulation)**

**str(Quality.Manipulation)**

**predict(ft1 , c(Quality.Manipulation))**

**mydata.val$predicted\_score<-predict(ft1 , c(Quality.Manipulation))**

**View(mydata.val)**

**plot(ft1)**

**#some correlation**

**cor(mydata$Y1,mydata$Y2)**

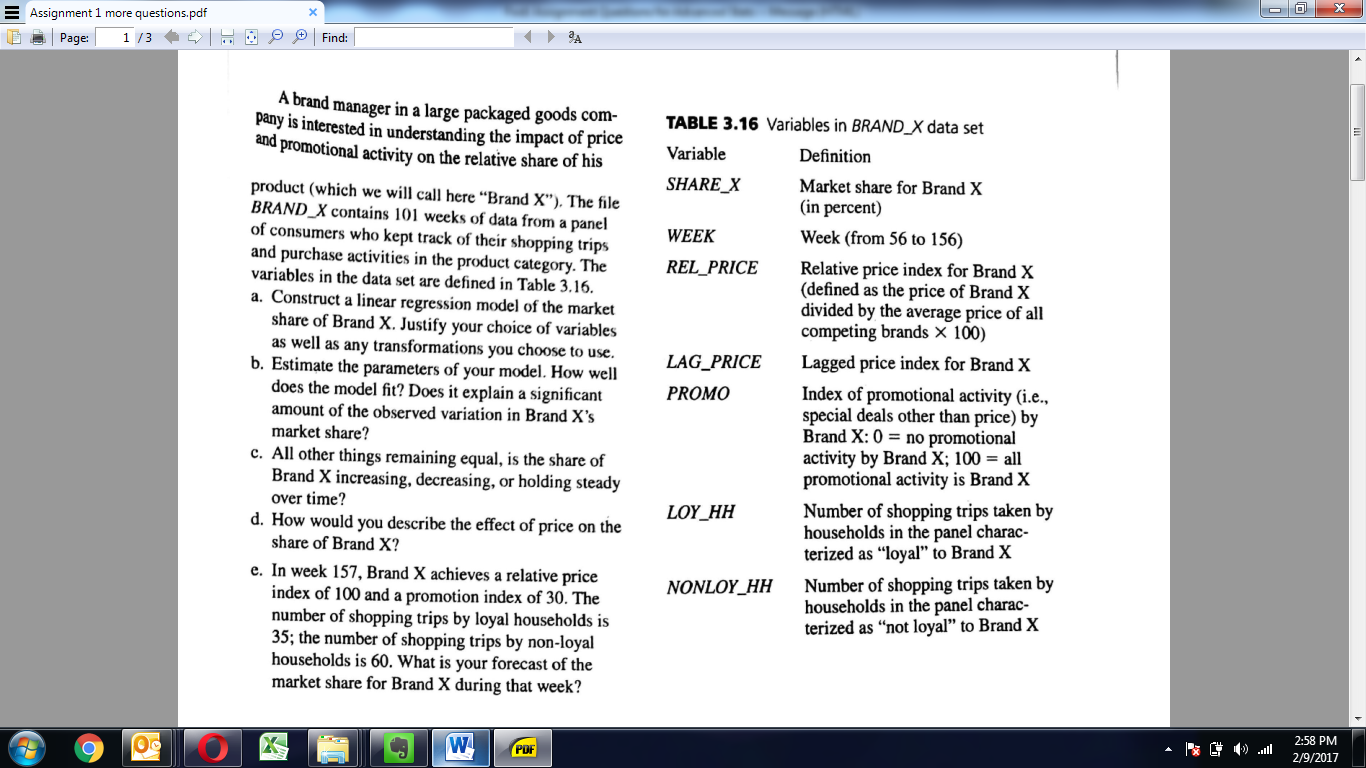
**plot(mydata$Y1,mydata$Y2)**

**table(mydata$Y1,mydata$Y2)**

**aov(Expectatations.Manipulation~Quality.Manipulation,mydata)->ft2**

**summary(ft2)**

**Questions 3**



#read data

#set working directory

setwd("G:/Grreat lakes/Advance stat/Group Assignment")

getwd()

library(readxl)

Brand\_X <- read\_excel("Brand\_X.xlsx")

View(Brand\_X)

str(Brand\_X)

#Variables in Dataset

#Col1: Rel\_price\_lag5

#correlation between variables

cor(Brand\_X)

# Replacing missing obs for missing promo values

library(Hmisc)

Brand\_X$Promo <- impute(Brand\_X$Promo,mean)

View(Brand\_X)

cor(Brand\_X)

#a. linear regression model of the market share of Brand\_X

lm(Brand\_X$Share\_X ~ Brand\_X$Rel\_Price+Brand\_X$Promo, data = Brand\_X )->B1

B1

summary(B1)

lm(Brand\_X$Share\_X ~ Brand\_X$Rel\_Price+Brand\_X$Promo+Brand\_X$Loy\_HH+Brand\_X$Nonloy\_HH, data = Brand\_X )->B2

B2

summary(B2)

lm(Brand\_X$Share\_X ~ Brand\_X$Rel\_Price+Brand\_X$Promo+Brand\_X$Loy\_HH+Brand\_X$Nonloy\_HH+Brand\_X$Rel\_price\_lag1+Brand\_X$Rel\_price\_lag5, data = Brand\_X )->B3

B3

summary(B3)

> lm(Brand\_X$Share\_X ~ Brand\_X$Rel\_Price+Brand\_X$Promo+Brand\_X$Loy\_HH+Brand\_X$Nonloy\_HH+Brand\_X$Rel\_price\_lag1+Brand\_X$Rel\_price\_lag5, data = Brand\_X )->B3

> B3

Call:

lm(formula = Brand\_X$Share\_X ~ Brand\_X$Rel\_Price + Brand\_X$Promo +

Brand\_X$Loy\_HH + Brand\_X$Nonloy\_HH + Brand\_X$Rel\_price\_lag1 +

Brand\_X$Rel\_price\_lag5, data = Brand\_X)

Coefficients:

(Intercept) Brand\_X$Rel\_Price Brand\_X$Promo Brand\_X$Loy\_HH Brand\_X$Nonloy\_HH

63.363819 0.002037 0.067844 0.521706 -0.007809

Brand\_X$Rel\_price\_lag1 Brand\_X$Rel\_price\_lag5

0.003099 -0.498909

> summary(B3)

Call:

lm(formula = Brand\_X$Share\_X ~ Brand\_X$Rel\_Price + Brand\_X$Promo +

Brand\_X$Loy\_HH + Brand\_X$Nonloy\_HH + Brand\_X$Rel\_price\_lag1 +

Brand\_X$Rel\_price\_lag5, data = Brand\_X)

Residuals:

Min 1Q Median 3Q Max

-14.5518 -4.0815 0.4392 3.0378 18.1259

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 63.363819 12.846927 4.932 3.50e-06 \*\*\*

Brand\_X$Rel\_Price 0.002037 0.002016 1.010 0.315

Brand\_X$Promo 0.067844 0.053218 1.275 0.206

Brand\_X$Loy\_HH 0.521706 0.111775 4.667 1.01e-05 \*\*\*

Brand\_X$Nonloy\_HH -0.007809 0.011617 -0.672 0.503

Brand\_X$Rel\_price\_lag1 0.003099 0.002142 1.447 0.151

Brand\_X$Rel\_price\_lag5 -0.498909 0.107791 -4.628 1.18e-05 \*\*\*

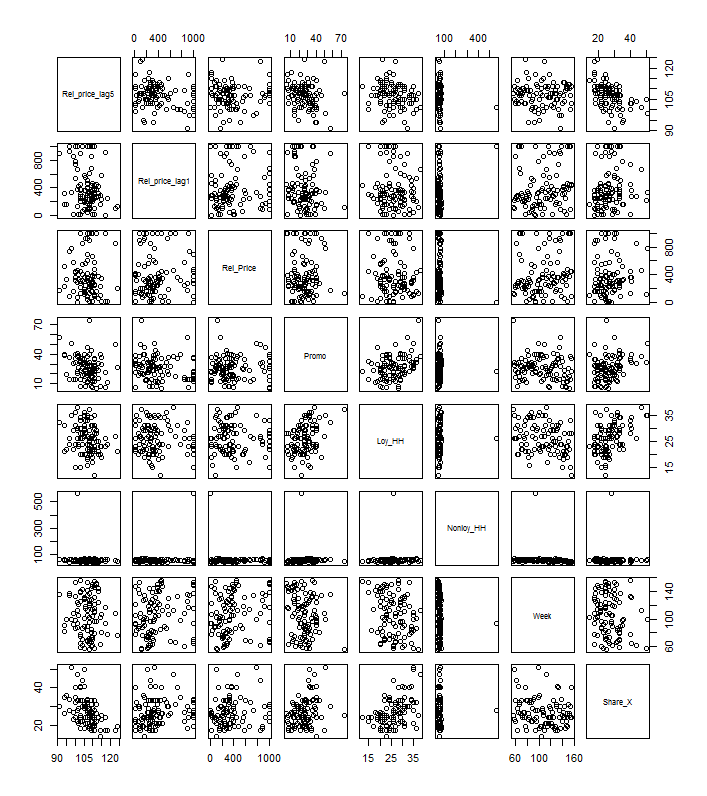
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

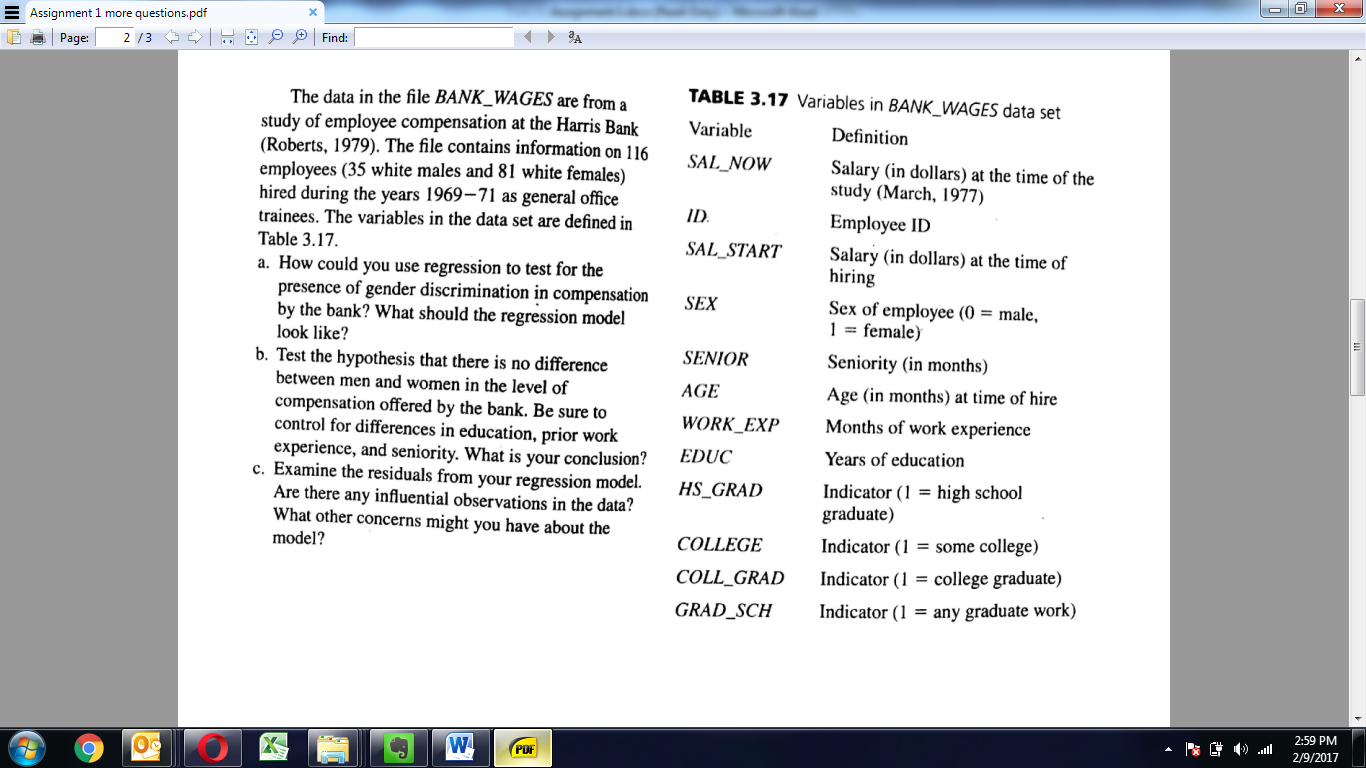
Residual standard error: 5.746 on 94 degrees of freedom

Multiple R-squared: 0.4405, Adjusted R-squared: 0.4048

F-statistic: 12.33 on 6 and 94 DF, p-value: 3.371e-10



**Question 4**



**Code: Not sure abt Lm function what wll be the dependent variable and independent variable.**

**setwd("C:/Users/Ashishkumar\_GUpta02/Downloads/Assign")**

**mydata<- read.csv("Excel Data for bank Wages.csv")**

**female<- subset(mydata,mydata$Sex==1)**

**female\_newdata<- data.frame(female[1:35,])**

**head(female\_newdata)**

**nrow(female\_newdata)**

**male<- subset(mydata,mydata$Sex==0)**

**nrow(male)**

**mean(female\_newdata$Salary\_Now)**

**mean(male$Salary\_Now)**

**lm((mydata$Salary\_Now)~.,data=mydata)->ft**

**summary(ft)**

**table(mydata$Salary\_Now,mydata$Sex)**

**lm((mydata$Work\_Exp)~.,data=mydata)->ft**

**summary(ft)**