NAME: MAHEK PAREKH (MPP83) [1 PT]

GEORGETOWN UNIVERSITY
McDonough SCHOOL of BUSINESS

OPIM 510: MANAGERIAL STATISTICS II
Fall 2022
Professor V. R. Jose

HOMEWORK 2

Due: December 1, 2022, Thursday, 11: 59 AM
Submission Method: Upload Answers on Canvas
75 POINTS

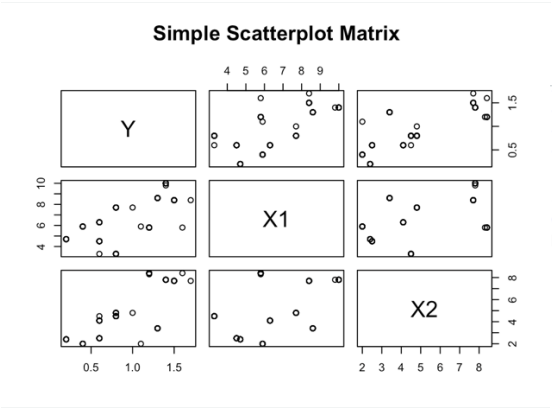
General Instructions: Please complete the following tasks. **Discussions with other MiM students are allowed but you must generate the printouts and write final answers/solutions by yourself.** Remember that copy and pasting someone else's work (using someone's plot or code without running it on your own) will be considered violations of the Honor Code. Paraphrasing someone else's answers is also not allowed. If you are unclear about any instruction, feel free to send me an email.

These questions are designed to be answered with the use of Excel and/or R Studio. When it is not stated, you can use either program to generate the output. I do not collect Excel spreadsheets or R files. ***The data sets are in the Excel sheet provided in different worksheets.***

Exercise 1. (22 pts) This *exercise requires parts (a), (b) and (e) must be done in R Studio. The rest can be done in Excel, R Studio or a mix of both*. This is a modified version of Problem 14 in the Chapter 7 of the textbook on Franchise. Read the story/context but you can ignore the original questions. Use the data provided to you and answer the following:

You must use R Studio for parts (a) and b).

a) Generate a matrix plot for the variables y, x_1, and x_2. No need to include Store and Location columns.
INSERT THE PLOT HERE




```
library(readxl)
library(dplyr)
library(readr)
library(ggplot2)

Store <- read_excel("HW2DataSet.xlsx", sheet = 1)
attach(Store)
pairs(~Y+X1+X2,data=Store,
      main="Simple Scatterplot Matrix")
cor(Store[,1:4])
```

b) Estimate the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where y is net profit, x_1 is counter sales, and x_2 is drive-through sales. All variables are measured in millions of dollars. Ignore the variable Location for parts (b) to (d)

	Estimated Value		Estimated Value		Estimated Value
b_0	-0.20739 million $	b_1	0.08523 (no unit)	b_2	0.11449 (no unit)

INCLUDE UNITS FOR THE TERMS ABOVE. THEN CUT AND PASTE THE R OUTPUT BELOW

```
Call:
lm(formula = Y ~ X1 + X2, data = Store)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26796 -0.19843 -0.04865  0.13763  0.57556

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.20739    0.07575  -2.738  0.00743 **
X1           0.08523    0.01259   6.772 1.17e-09 ***
X2           0.11449    0.01109  10.328  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2158 on 92 degrees of freedom
Multiple R-squared:  0.7613,    Adjusted R-squared:  0.7561
F-statistic: 146.7 on 2 and 92 DF,  p-value: < 2.2e-16

>
```

c) Interpret the coefficient attached to counter sales from the regression in part (b).
INSERT THE INTERPRETATION HERE. INCLUDE ALL THE ELEMENTS WE SAID WERE IMPORTANT IN INTERPRETING THIS COEFFICIENT

On Average, the net profit increases by 0.0852 million dollars or $85,200 for every unit increase in counter sales, holding the drive through sales constant and unchanged.

d) Use the regression in (b) to predict the net profit for the first three locations in dataset and then determine whether the model over or under-estimated the net profit for these three stores

Store	Actual Net Profit	Predicted Net Profit	Did the Model Over or Under-estimate this store's net profit?
1	$1.5 million	$1.38993 Million	**Under – estimated Net Profit**
2	$0.8 million	$0.58901 Million	**Under – estimated Net Profit**
3	$1.2 million	**$** 1.24856 Million	**Over-estimated Net Profit**

You must use R Studio for part (e).

e) Create a *new* regression model to predict net profit based on three predictors: counter sales, drive-through sales and location. ***Make Northeast the base/reference category for Location***.
CUT AND PASTE THE TABLE OF COEFFICIENTS FROM THIS NEW REGRESSION.

```
> Store$Location <- factor(Store$Location, levels=c("Northeast","MidAtlantic","Midwest","South"))
> Regression1 <- lm(Y~X1+X2+Location,data=Store)
> summary(Regression1)

Call:
lm(formula = Y ~ X1 + X2 + Location, data = Store)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31002 -0.05136  0.00634  0.04265  0.45682

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.096006   0.049444   1.942   0.0553 .
X1                  0.090496   0.007021  12.889  < 2e-16 ***
X2                  0.114382   0.006020  19.001  < 2e-16 ***
LocationMidAtlantic -0.548203   0.036643 -14.961  < 2e-16 ***
LocationMidwest     -0.231842   0.039068  -5.934 5.56e-08 ***
LocationSouth       -0.419992   0.034774 -12.078  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1098 on 89 degrees of freedom
Multiple R-squared:  0.9402,    Adjusted R-squared:  0.9368
F-statistic: 279.9 on 5 and 89 DF,  p-value: < 2.2e-16

>
```

f) Interpret the coefficient related to the dummy variable Location? Mid Atlantic from the regression in part (e). Recall Northeast is the base/reference category in the regression in part (e)

INSERT THE INTERPRETATION HERE. USE THE CONTEXT INSTEAD OF JUST SAYING X and Y. ALSO SHOW THE IMPORTANT ELEMENTS IN THE INTERPRETATION WE MENTIONED IN CLASS.

On average, the store in MidAtlantic location is $0.54820 million less profitable than store in Northeast keeping the drive through and counter sales as constants.

-

g) Holding all else constant, arrange the four geographic locations (MidAtlantic, Midwest, Northeast, South) from most profitable to least profitable for this fast-food chain.

Most Profitable Region/Location	**Northeast**
Second Most Profitable	**Midwest**
Third Most Profitable	**South**
Least Profitable Region/Location	**MidAtlantic**

Exercise 2. Creating new variables. Using the data provided, create new columns of data and then run regressions on the following using the following models and estimate the coefficients of these new models. The new variables you need to do are:

C = X^2 [Square of the X variable. If the value of X is 1.82 then C = 1.82^2 =3.3124]
D = LN(X) [Natural log of X. If the value of X is 1.82 then D = 0.598837. In Excel, this is "=LN(1.82)". In R, this is simply "log(1.82)"]
E = LN(Y) [Natural log of Y. If the first value of Y is 6003 then E = 8.700015. See comment on the variable D on how to generate this new value]

Once you have created these new variables. Estimate the coefficients for the following regression using these new variables.

Model 1: $B = \beta_0 + \beta_1 A + \varepsilon$
Model 2: $B = \beta_0 + \beta_1 A + \beta_2 C + \varepsilon$
Model 3: $B = \beta_0 + \beta_1 D + \varepsilon$
Model 4: $E = \beta_0 + \beta_1 D + \varepsilon$

	Model 1	Model 2	Model 3	Model 4
b_0	-2290.870555	-1658.7121	6404.03914	8.46314541
b_1	7915.778886	6280.83741	5995.31014	1.67704859
b_2		780.833048		
Coefficient of Determination*	0.82232985	0.82485519	0.71721758	0.92327177

*** For coefficient of determination, use the more appropriate measure, i.e., R-sq if there is only one predictor and adjusted R-sq if there is more than one predictor**

**** You can ignore units for this question. Do not include the outputs. Just type them in the boxes above. No need to provide interpretations as well.**

Exercise 3. (10 pts) ***must use Excel for Exercise 3.***

a) Data was collected on the salaries and how long people have been in a local firm.

Year – The number of years a person has been working in the firm
Salary – Their annual salary (without bonuses) expressed in 000s USD
Location – The office location of the employee (DC or Outside DC)

A regression model of the form:

$$Salary = \beta_0 + \beta_1 Years + \beta_2 Location + \varepsilon$$

is being used to study the relationship between how long a person has been in the company and their salary. Answer the following questions: (Units of salary is 000s $, Years is years in the firm)

a) Run the regression and cut-and paste the table of coefficients as well as the summary that contains (R-sq and S). Set DC as the base category, i.e., Location = 0 if Location is DC and 1, if outside DC.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.88043147
R Square	0.77515958
Adjusted R Sq	0.77210053
Standard Erro	11.1100529
Observations	150

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	62555.6183	31277.8091	253.398517	2.3055E-48
Residual	147	18144.6916	123.433276		
Total	149	80700.3099			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	90.3643695	2.21961178	40.7117904	6.1975E-82	85.9778988	94.7508403	85.9778988	94.7508403
Year	4.65539224	0.24407993	19.0732285	1.3558E-41	4.17303335	5.13775112	4.17303335	5.13775112
Location	-15.507972	1.84375204	-8.4110942	3.2675E-14	-19.151656	-11.864288	-19.151656	-11.864288

b) Test if the predictor *Years* is statistically significant at the 5% level of significance.
SHOW ALL THE RELEVANT STEPS IN THE HYPOTHESIS TEST.

H0: B1 = 0
HA: B1 not equal to 0
Alpha = 0.05
Using t-test, we see t= 19.073 and p-value is 1.3558E-41
Due to the fact that the p value is less than 0 and lesser than alpha which is 1.3558E-41<0.05, we reject the null hypothesis
Which implies that Years is now statistically significant at level of significance of 5%.

c) Compute for the residuals for each of these 150 points then compute the following summary statistics for the residuals:

Minimum	-33.6117898
Maximum	73.68023822
Mean	-3.0127E-14
Standard Deviation	11.03523692

How to proceed: First, compute the predicted salary for each employee from 1 to 150. Next, compute the residual for each employee: Residual = Actual – Predicted. You can also do this by just clicking on the button on the Regression option (as seen on the right). Finally, compute the summary statistics for the residuals you computed/got from the output: MIN, MAX, AVERAGE and STDEV.S. No need to show these steps for the homework. Just fill in the table. Do not include the intermediate computations.