NAME: MAHEK PAREKH (MPP83) [1 PT]

GEORGETOWN UNIVERSITY
McDonough SCHOOL of BUSINESS

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

HOMEWORK 1

Due: November 15, 2022, Tuesday Before the Start of Class
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. This assignment is worth 50 points.

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.*** If you are having a hard time finding symbols, you can just type words instead of using symbols. For example, you can just type "beta1" instead of β_1 or "rho" instead of "ρ".

Exercise 1. (36 pts) ***Parts (a) to (f) can be done in Excel, R Studio or a mix of both. Parts (g) and (h) must be done in R Studio.*** The dataset contains the monthly price data for Johnson and Johnson (JNJ) and Merck (MRK) as well as the SP500 Index. Returns for SP500 and JNJ have been computed as well. Use the dataset to answer the following questions and treat the data as a sample not as a population.

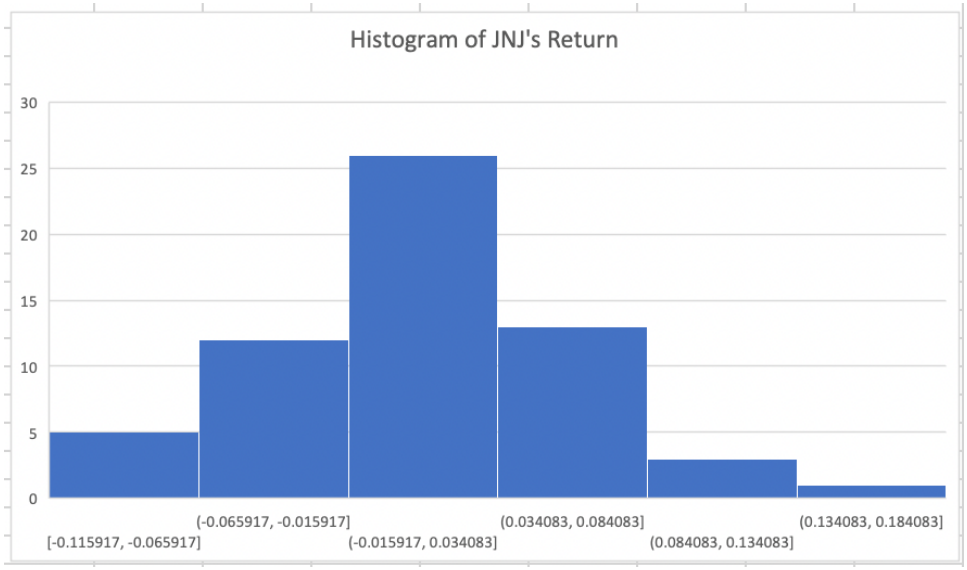
a) Create the monthly return series (R_t) for MRK using Excel. (Reminder: The formula is provided in our slides. And assume no dividends). To help me check if you did this correctly, compute for the following returns R_2, R_{20} and R_{50} and write the returns in the table below as a decimal with at least 4 decimal places.

Period (t)	Date	JNJ Return
2	1-Jan-18	**0.061914**
20	1-Jul-19	**-0.003634**
50	1-Jan-22	**0.073333**

*No need to include the other returns.

b) Generate a histogram for the *returns* of JNJ. Label your plots properly. Describe briefly the shape that you see.

INSERT YOUR PLOT HERE AS WELL AS YOUR SHORT DESCRIPTION. REMEMBER TO LABEL YOUR PLOTS PROPERLY.



The JNJ Return histogram is unimodal, it is almost symmetrical and is normally bell-shaped.

c) Similar to Slide Set 1, I would like you to create a portfolio by dividing your money (75-25) between JNJ and a fund that mimics the SP500. This will create a portfolio return that will be equal to 0.75 JNJReturns + 0.25SP500Returns. After creating the returns for this new series. No need to show me the intermediate calculations, just complete the following table:

	JNJ Returns	SP500 Returns	Portfolio Returns
Mean/Average Return	**0.0069679**	**0.0073059**	**0.0070524**
Standard Deviation of Returns	**0.051390288**	**0.0530652**	**0.0474638**

*Include at least 4-5 decimal places in your answer.

d) Looking at your answers in part (c), does the portfolio provide diversification (i.e., lowering risk or standard deviation than the two individual stock's standard deviations)? If yes, what statistics/numbers did you look at to come up with this conclusion? If not, try to think of one possible reason why this portfolio did not lower the standard deviation.

INSERT YOUR ANSWER HERE

Yes, this portfolio is diversified because mean of portfolio returns is higher than JNJ returns but lower than SP500 returns, which means that it is more profitable than individual JNJ returns but less profitable than individual SP500 returns.

Standard deviation of portfolio returns is less than both individual stocks, therefore portfolio returns are comparatively less risky.

e) Compute for the correlation and covariance between the following: Include appropriate units for these quantities

Cov(JNJ Price, SP500 Value)	**12612.69926 dollars squared**
Corr(JNJ Price, SP500 Value)	**0.898614**
Cov(JNJ Returns, SP500 Returns)	**0.001550466 dollars squared**
Corr(JNJ Returns, SP500 Returns)	**0.578189**

*SP500Value is in dollars, SP500Returns is unitless

f) Consider the mean and standard deviations you computed for JNJ Returns in part (c) and assume that Company XYZ's returns have a mean of 0.015 and a standard deviation of 0.12 over the same period. If the correlation between JNJ and Company XYZ's returns is 0, compute for

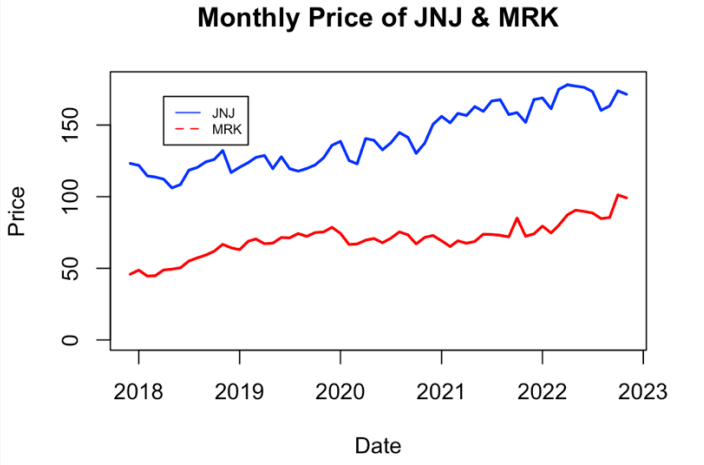
E[0.3 JNJ Returns + 0.7 Company XYZ Returns]	**0.0125903601**
Var[0.3 JNJ Returns + 0.7 Company XYZ Returns]	**0.007293687**

(Note there is no raw data set for Company XYZ. Use the formula for the mean and variance of linear combinations of random variables)

Parts (g) and (h) in Exercise 1 must be done in R Studio.

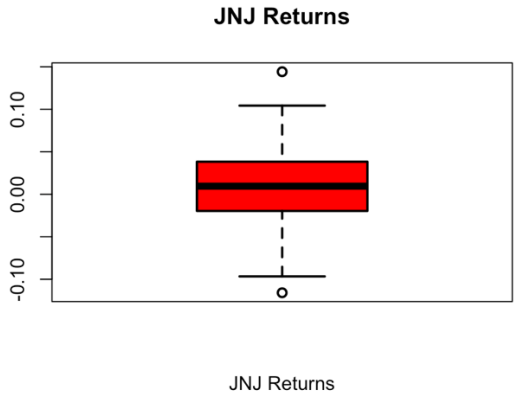
g) Create a single plot that contains both time series data for JNJ (Price) and MRK (Price). Label your plots properly (i.e., title, axes label, legend, etc). For the x-axis, you can use either Period or Date.

INSERT YOUR PLOT HERE. INCLUDE YOUR R CODE AS WELL FOR THIS PART




```
setwd("~/Desktop/STATS HOMEWORK1 /")
library(readxl)
Price <- read_excel("PRICES.xlsx", sheet = 1)
Price$Date <- as.Date(Price$Date, "%Y/%m/%d")
summary(cbind(Price$JNJPrice,Price$MRKPrice))
plot(Price$JNJPrice ~ Price$Date,ylab="Price", xlab="Date", type="l", col="blue", lwd=2,
     ylim=c(0,180), main="Monthly Price of JNJ & MRK")
points(Price$MRKPrice ~ Price$Date, type="l", col="red", lwd=2)
legend(Price$Date[5],170, legend=c("JNJ", "MRK"),col=c("blue", "red"), lty=1:2, cex=0.6)
```

h) Create a boxplot for JNJ's returns. Then answer whether there are outliers in this return series based on the IQR/boxplot approach.
INSERT YOUR PLOT AND DESCRIPTION HERE. INCLUDE YOUR R CODE AS WELL FOR THIS PART



.

```
boxplot(Price$JNJReturns, xlab="JNJ Returns", col="red", lwd=2, main="JNJ Returns")
```

Yes, since there are data points outside of the plot, there are outliers in this given boxplot.

Exercise 2. (16 pts) Transforming Data by Linear Combinations. Using the *sample* data set provided, create four new variables C, D, E, and F in your Excel worksheet as follows C = 3*B, D = B + 500, E = 500 + 3*B, and F = 500 – 3B. This means that if the first value for B is 20, then the first entry for C is 60, D is 520, E is 560, and F is 440. Remember to add appropriate units. C, D, and E are only changed by constants and should have the same units as B. Make sure to include the units for these in the table.

a) Compute for the following: Include appropriate units

	ANSWER		**ANSWER**
Cov(A,B)	**409.3473 seconds miles**	Corr(A,B)	**0.669538**
Cov(A,C)	**1228.042 seconds miles**	Corr(A,C)	**0.669538**
Cov(A,D)	**409.3473 seconds miles**	Corr(A,D)	**0.669538**
Cov(A,E)	**1228.042 seconds miles**	Corr(A,E)	**0.669538**
Cov(A,F)	**-1228.04 seconds miles**	Corr(A,F)	**-0.66954**

b) From your answer in (b), what happens to the covariance and correlation when you add a constant to it? How about when you multiply it by a constant?

Adding or multiplying a constant to the data does not change the correlation but it will change the covariance.

c) Suppose you take two of these variables and estimate a regression line. Complete the following table

Dependent Var. or Response ("Y" variable)	Independent Var. or Predictor ("X" variable)	**ESTIMATES**	
		INTERCEPT (b_0)	SLOPE (b_1)
A	B	**36.7755**	**0.8433**
A	C	**36.7755**	**0.281101**
A	D	**-384.876**	**0.843304**
A	E	**-103.775**	**0.281101**
A	F	**177.326**	**-0.2811**

(Note you can easily estimate the slope and intercept using =SLOPE () & =INTERCEPT() function in Excel)

Exercise 3. (22 pts) College basketball is big business with coaches' salaries, program's revenues and program's expenses in millions of dollars. Using old data in 2017, you plan to develop a regression model to possibly understand a coach's next year salary based on one of the three variables: winning percentage, team's revenues, or team's expenses.

a) Create a matrix of correlations for the four variables (salary, winning percentage, revenues, expenses). Then answer: which variable is most correlated with salary? Least correlated?

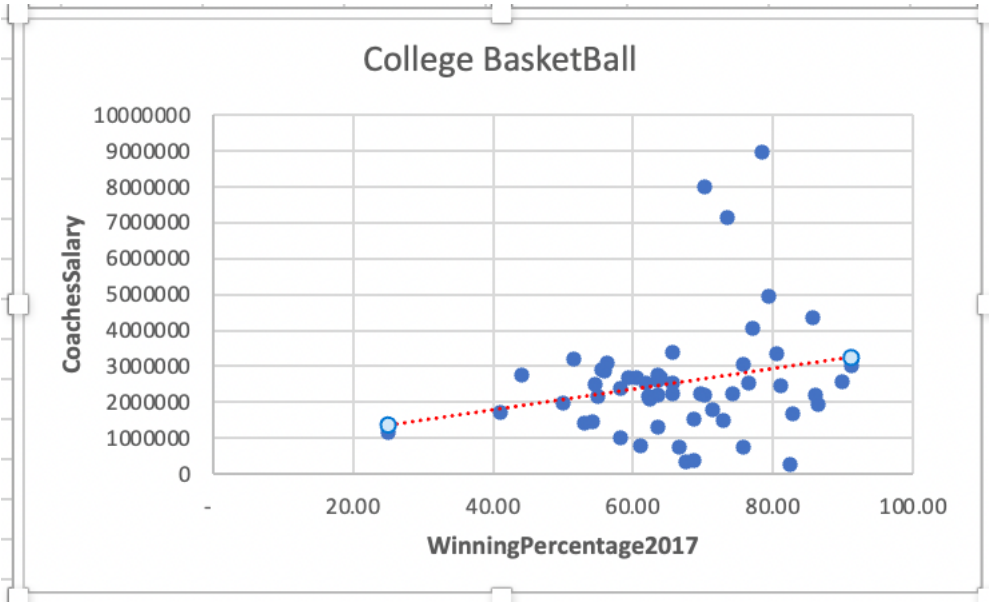
CUT AND PASTE MATRIX OUT HERE. TYPE YOUR ANSWER HERE

A	B	C	D	E
	CoachesSalary2018	*WinningPercentage2017*	*TotalRevenue2017*	*TotalProgramExpenses2017*
CoachesSalary2018	1			
WinningPercentage2017	0.221378286	1		
TotalRevenue2017	0.269248152	-0.125936699	1	
TotalProgramExpenses20	0.27115673	-0.141759251	0.985723971	1

The variable which is most correlated with salary is TotalProgramExpense and the variable least correlated with salary is WinningPercentage2017.

b) Create a scatterplot where Winning Percentage is on the x-axis and Salary is on the y-axis. Label the axes and chart properly.

CUT AND PASTE YOUR PLOT HERE.



Part (c) in Exercise 3 must be done both in Excel and R Studio.

c) Run a regression using the winning percentage as the only predictor, i.e.,

$$Salary = \beta_0 + \beta_1 Winning\ Percentage + \varepsilon$$

Cut and paste the table of coefficients of your regression. This exercise checks your ability to run a regression both in Excel and R.

CUT AND PASTE THE TABLE OF COEFFICIENTS HERE FOR EXCEL AND R.

```
> Regression1 <- lm(CoachesSalary2018 ~ WinningPercentage2017, data = Book3)
> summary(Regression1)

Call:
lm(formula = CoachesSalary2018 ~ WinningPercentage2017, data = Book3)

Residuals:
     Min       1Q   Median       3Q      Max
-2745990  -846340  -234961   349057  6099879

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           656987    1177023   0.558    0.579
WinningPercentage2017  28386      17340   1.637    0.108

Residual standard error: 1629000 on 52 degrees of freedom
Multiple R-squared:  0.04901,   Adjusted R-squared:  0.03072
F-statistic:  2.68 on 1 and 52 DF,  p-value: 0.1077
```

	A	B	C	D	E	F	G	H	I	J	K
1	SUMMARY OUTPUT										
2											
3	*Regression Statistics*										
4	Multiple R	0.22137829									
5	R Square	0.04900835									
6	Adjusted R Sq	0.03072004									
7	Standard Erro	1629269.42									
8	Observations	54									
9											
10	ANOVA										
11		*df*	*SS*	*MS*	*F*	*Significance F*					
12	Regression	1	7.1135E+12	7.1135E+12	2.67976482	0.10766931					
13	Residual	52	1.3803E+14	2.6545E+12							
14	Total	53	1.4515E+14								
15											
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*		
17	Intercept	656987.13	1177022.89	0.55817702	0.57911901	-1704882.1	3018856.35	-1704882.1	3018856.35		
18	WinningPerce	28385.9573	17340.244	1.63699872	0.10766931	-6409.7879	63181.7025	-6409.7879	63181.7025		
19											
20											
21											

d) Provide a 1-sentence interpretation of the slope of the regression in part (c). Use the variables in the regression and do not simply refer to them as X and Y.

INSERT YOUR ANSWER HERE

On average, an increase in one percentage point for 2017 winning percentage of the college basketball team increases the 2018 salary received to the coaches by $28,386.

e) Regardless of the quality of fit, using the regression in part (c), predict the salary for a coach where the winning percentage was 70%.

INSERT YOUR ANSWER HERE

If the winning percentage was 70%
Predicted salary = 656987+ (28,386*70) + ε
= $2,644,007

f) Consider one of the points in this regression: Sean Miller from Arizona. Based on your answer in (e), does the regression model over- or under-estimate this point? What is the residual for this data point?

INSERT YOUR ANSWER HERE

Sean Miller from Arizona has a winning percentage of 77.10 and has a salary of $4,054,853
Therefore, his expected salary = 656,987+(28,386*77.10)+ $\varepsilon = \$2{,}845{,}547.60$
This model underestimates the salary for Sean Miller from Arizona. The residual is
=4,054,853-2,845,547.60
=$1,209,305.40