Assignment 1

Tahir Muhammad

1002537613

1 Theoretical Problems

Question 1

a) The Expected value of Z:

$$\begin{aligned} E(Z) &= E(2X - Y) \\ &= E(2X) + E(-Y) \\ &= 2E(X) - E(Y) \\ &= 2(2) - (0) \\ &= 4 \end{aligned}$$

b) The Co-Variance of X and Y:

$$\begin{aligned} COV(X,Y) &= E[(X - u_x)(Y - u_y)] \\ &= E[(X-2)(Y-0)] \\ &= E[(X-2)(Y)] \\ &= E[(XY - 2Y)] \\ &= E(XY) + E(-2Y) \\ &= E(X)E(Y) - 2E(Y) \\ &= (2)(0) - 2(0) \\ &= 0\ . \end{aligned}$$

The Co-relation of X and Y is:

$$\begin{aligned} Corr(X,Y) &= \frac{COV(X,Y)}{\sigma_x \sigma_y} \\ &= \frac{(0)}{(\sqrt{3})(1)} \\ &= 0 \end{aligned}$$

Which is expected, as X and Y are independent.

c) The Variance of Z is:

$$\begin{aligned} Var(Z) &= Var(2X - Y) \\ &= Var(2X) + Var(Y) \\ &= 4Var(X) + Var(Y) \\ &= 4(3) + (1) \\ &= 13 \end{aligned}$$

d) The Co variance of X and Z:

$$\begin{aligned} Cov(X,Z) &= Cov(X, 2X - Y) \\ &= 2Cov(X,X) + (-1)(1)Cov(X,Y) \\ &= 2Var(X) - Cov(X,Y) \\ &= 2(3) - 0 \qquad \text{From Part B} \end{aligned}$$

$= 6$

f) The Expected Value of $\hat{\beta_1}$ is 2, due to the Gauss Markov assumptions. Below is a description on why they hold:

i) Random Sample: Holds due to how the sample data was drawn out from the distributions. **ii)** Linear in Parameters: The assumptions made by the researcher led him to use a linear model. Therefore, by default, the model is linear in parameters. **iii)** No multi Co-linearity: Drawn by random generating computer, hence the probability of drawing the same value 100 times is extremely unlikely (pretty much impossible, even though computers are finite). **iv)** Exogenous: $E(U|X) = 0$ holds due to the fact that the true data consists of two variables, $z_i = 2x_i - y_i$ but the model consists of only one. This means that our omitted variable Y is contained inside of U, the error term. Furthermore, since X and Y are independent $\implies E(U|X) \implies E(Y|X) \implies E(Y) = 0$. Hence, the exogenous assumption also holds, and we have an unbiased estimator of $\hat{\beta}$ which illustrates that $\hat{\beta} = \beta = 2$.

Question 2

a) We can find the bias of an estimator by the following equation:

$$Bias(\hat{\beta_1}) = E(\hat{\beta_1}) - E(\beta)$$

Consequently, if an estimator is unbiased, then

$$E(\hat{\beta_1}|X) = \beta_1$$

Hence we use the second condition to show the bias of

$$\hat{\beta_1} = \frac{y_i - yk}{x_j - x_k}$$

$$E(\hat{\beta_1}|X) \implies E(\frac{y_i - yk}{x_j - x_k}|X) \implies E(\frac{(\beta_0 + \beta_1 x_i + u_i) - (\beta_0 + \beta_1 x_j + u_j)}{x_j - x_k}|X)$$

$$\implies E(\frac{\beta_1 x_i + u_i - \beta_1 x_j - u_j}{x_j - x_k}|X) \implies E(\frac{\beta_1(x_i - x_k) + u_i - u_j}{x_j - x_k}|X)$$

$$\implies E(\beta_1 + \frac{u_i - u_j}{x_j - x_k}|X)$$

$$\implies \beta_1$$

Hence, we can conclude that the estimator is unbiased which leads to the bias being 0.

b) We know that $\hat{\beta_1}$ is linear in parameters if it can be expressed as:

$$\hat{\beta}_j = \sum_{i=1} w_{ij} y_i$$

or as a sum of constants. In our case we have: $\hat{\beta}_1 = \frac{y_i - yk}{x_j - x_k}$, and since the values y_i, y_k, x_j, x_k are two points (i.e. constants), we can Let $w_{ij} = y_i - y_k$ and $y_i = \frac{1}{x_j - x_k}$.

$$\begin{aligned} \implies \hat{\beta}_j &= \textstyle\sum_{i=1}(y_i - y_k)(\frac{1}{x_j - x_k}) \\ &= \textstyle\sum_{i=1} w_{ij} y_i \\ &= \hat{\beta}_1 \ . \end{aligned}$$

Hence, the estimator $\hat{\beta}_1$ is an linear estimator.

c) Well we can just say that OLS is BLUE, and hence it is the best out there. However, intuitively, we can also tell him that if the data consists of more than just two points (i.e which he drew a line through with his estimator, and got the slope), his estimator would not consider any other data points. Hence, it is not accurate/bad estimator.

d) Yes, this is true. However, as mentioned above, if there is more than 2 data points, the friend's estimator will not be consider those data points and will give us a bad estimate.

Question 3

a) The value of $\hat{\beta}_1$ being -0.02 means that one more hour of studying would result in a decrease of GPA by 0.02 points on average.

b)I do not agree this was the correct move, because even though the regression suggests that tutoring has not been effective on the average GPA, the model in this scenario is not correct. For example, the model excludes other key factors such as special education requirements, family-related issues effecting the grades, mental health problems, incompetent tutors, etc. Furthermore, the Simple Linear Regression model also does not hold all of its assumptions. Assuming that the error term U (gpa going down) and the variable X1 (time spent studying) are co-related is not true. Hence, since the assumptions does not hold, we conclude that this model isnt the most accurate representation, and that the school should not shut down the tutoring program.

2 Computer Based Problems

Question1

a) The data set seems to be of the Time-Series data structure. This is because of the following two reasons: i) The variables do not seem independent of each other, and ii) There are several variables which consists of observations over a period of time.

b) Refer to Appendix, figure 1. The mean number of food consumption is 2.91358, and this tells us that on average, the TA Hammad eats approximately 3 plates of food per day.

c) Refer to Appendix A, figures 2 and 3. The overall pattern of both the plot shows a negative co-relation, which means that as time increases, the weight in pounds and the waist in inches of Hammad decreases. Both plots cover the same time-period, but we have missing data for the first half until august $7th$ for waist in inches variable which is important to note. After that, both variables are measured in same time.

di) The results of the regression are shown in Appendix A, Figure 4.

ii) The simple linear regression fits the data well as we only have one variable, and the relation between the two variables WeightPounds TimeUnitDays is linear, as it can be seen from the scatter plot.

iii) The estimated slope is -.3240695, and it means that on average, Hammad loses about 0.324 pounds per day.

iv) Approximately 103 days

v)

e) In part a we assumed that the data set seems to come from a time-series model, and that the variables are dependent. This model violates that condition as it assumes that the time is independent of the error term, which is not true as the error term u can consist of factors that contribute towards Hammads weight.

Question 2

a) Refer to *Appendix A*, Figure 5

b) Refer to *Appendix A, Figure 6.* Yes, the graph does indeed support what the entrepreneurs thought as the scatter plot illustrates that as total employment increases, the amount of exports also increase.

c) Refer to *Appendix A, Figure 7.* The estimated slope is .9976858 and it represents that a 1% change in ln of total employment is associated with a .9976858 increase in ln of exports.

d) Refer to *Appendix A, Figure 8.* We can see that the true exported value is 1385333. However, the estimated value we got for export is $e^{14.09978} = 1328814.9$ from which we can conclude that since the values are close, but not too accurate.

e) Refer to *Appendix A, Figure 9.* It can be seen that from the previous model, we had the $\hat{\beta}_1$ as .9976858. However, with the new model we get the coefficient to be -0.0281651. This is a huge difference. The results show a negative relationship between the variables and total exports ,so much so that a percentage increase in total employments you get a decrease in exports by ≈ 0.0281651. This might be the case because in the first model, we have a case of omitted variables. The coefficients of the omitted variables could have impacted the coefficient to be bias, and hence was positively co-related with total employment.

f) Refer to *Appendix A, Figure 10.*

As we partial out the other regressers, we can see think of the remaining as pure Ln exports without the cor relational effect of other variables. This can be seen as the Total Employment comes to become completely the error term.

Question3

1a) The averages and standard deviations of the parameters are as follows:

$$\hat{\beta}_0 = 4.996074 \qquad \text{standard deviation} = 0.2104072$$

$$\hat{\beta}_1 = -1.992957 \qquad \text{standard deviation} = 0.1845446$$

$$\hat{\beta}_2 = -3.00047 \qquad \text{standard deviation} = 0.1251213$$

It is clear that the average of the estimated Betas is extremely close to their true means. The histograms of each beta is in *Appendix A* Figures $11, 12, 13$.

b) The averages and standard deviations of the parameters are as follows:

$$\hat{\beta}_0 = 7.980198 \qquad \text{standard deviation} = 0.454295$$

$$\hat{\beta}_1 = -3.184726 \qquad \text{standard deviation} = 0..4661129$$

No, the averages are not close to what we expected. This is because $X2 = 0.4X1 + V$ and they are co-related, which violates the assumptions of OLS and leads to different values than the true means.

3 *APPENDIX A*

The following section contains all pictures for the above report.

```
. tabstat WeightPounds WaistInches PlatesFoodCons BMI, stat(mean var sd count)
```

stats	Weigh~ds	WaistI~s	Plates~s	BMI
mean	165.4989	34.93957	2.91358	25.08241
variance	63.48545	1.806475	.9299383	1.458217
sd	7.967776	1.344052	.9643331	1.207567
N	81	46	81	81

Figure 1: Summary of required variables for 1b

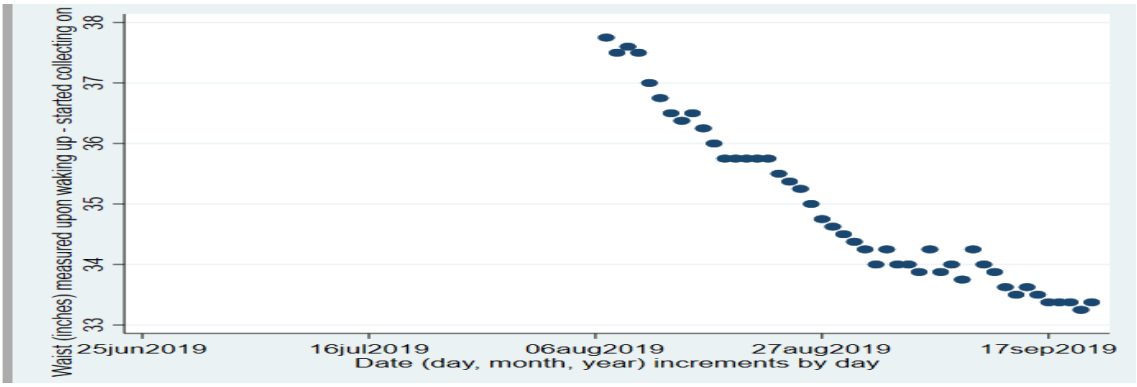

Figure 2: Waist in inches vs Time in Days

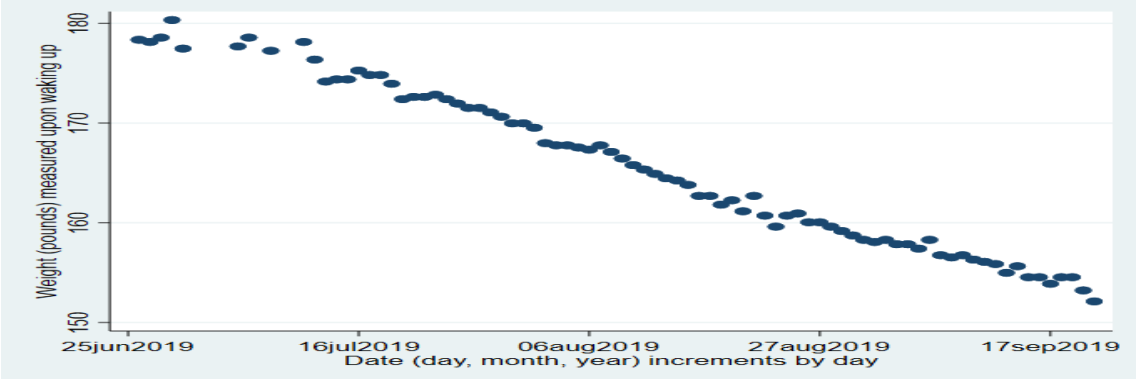

Figure 3: Weight in Pounds vs Time in Days.

```
. * Part D; Regression
. regress WeightPounds TimeUnitDay
```

Source	SS	df	MS	Number of obs	=	81
				F(1, 79)	=	5985.26
Model	5012.67309	1	5012.67309	Prob > F	=	0.0000
Residual	66.1627609	79	.837503302	R-squared	=	0.9870
				Adj R-squared	=	0.9868
Total	5078.83585	80	63.4854481	Root MSE	=	.91515

WeightPounds	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
TimeUnitDay	-.3240695	.0041889	-77.36	0.000	-.3324073	-.3157318
_cons	7221.293	91.20207	79.18	0.000	7039.759	7402.826

Figure 4: Result for the Regression Output.

```
. *Showing the mean, standard deviation, median, 25% and 50% quartiles for required variables
. tabstat exports TotalEmployment LnTotalEmployment LnExports, stat(mean sd median p25 p75)
```

stats	exports	TotalE~t	LnTota~t	LnExpo~s
mean	1.42e+07	172.4819	4.481492	14.03814
sd	5.44e+07	243.153	1.214396	2.553711
p50	1385333	94	4.543295	14.14145
p25	233964	39	3.663562	12.36292
p75	7752355	215	5.370638	15.86351

Figure 5: Result from tabstat for required variables

Figure 6: Scatter plot of Total Employments effect on Exports

```
. reg LnExports LnTotalEmployment
```

Source	SS	df	MS	Number of obs	=	2,299
				F(1, 2297)	=	667.23
Model	3373.32341	1	3373.32341	Prob > F	=	0.0000
Residual	11612.9476	2,297	5.05570205	R-squared	=	0.2251
				Adj R-squared	=	0.2248
Total	14986.271	2,298	6.52144083	Root MSE	=	2.2485

LnExports	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LnTotalEmploy~t	.9976858	.0386238	25.83	0.000	.9219446	1.073427
_cons	9.567017	.1793324	53.35	0.000	9.215346	9.918687

Figure 7: Regression of Total Employments effect on Exports

```
. * Part D
. * Getting the true median for Total Employment and Exports
. tabstat TotalEmployment exports, stat(median)

   stats |  TotalE~t   exports
---------+--------------------
     p50 |         94   1385333
---------+--------------------

.
. * Storing the ln of median total employees
. generate LnMedianTotalEmployemees = ln(94)

.
. * Predicting the median exports by the regression model
. generate EstimatedExports =  9.567017 +  0.9976858*LnMedianTotalEmployemees

.
. * Show the results of the predicted exports
. display EstimatedExports
14.099798
```

Figure 8: Part D: Estimated exports vs True Exports

```
. regress LnExports LnTotalEmployment LnMaterial LnCaptial
```

Source	SS	df	MS	Number of obs	=	2,299
				F(3, 2295)	=	604.95
Model	6617.71837	3	2205.90612	Prob > F	=	0.0000
Residual	8368.55265	2,295	3.64642817	R-squared	=	0.4416
				Adj R-squared	=	0.4409
Total	14986.271	2,298	6.52144083	Root MSE	=	1.9096

LnExports	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LnTotalEmployment	-.0281651	.0516769	-0.55	0.586	-.1295034	.0731731
LnMaterial	.8166265	.0333824	24.46	0.000	.7511637	.8820893
LnCaptial	.0634026	.0330915	1.92	0.055	-.0014897	.1282949
_cons	.5112672	.3612492	1.42	0.157	-.1971418	1.219676

Figure 9: Part e: Regression with Capital and Materials included

```
. *Part F
. regress LnTotalEmployment LnMaterial LnCaptial
```

Source	SS	df	MS	Number of obs	=	2,299
				F(2, 2296)	=	1701.29
Model	2023.54234	2	1011.77117	Prob > F	=	0.0000
Residual	1365.44832	2,296	.594707455	R-squared	=	0.5971
				Adj R-squared	=	0.5967
Total	3388.99066	2,298	1.4747566	Root MSE	=	.77117

LnTotalEmp~t	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LnMaterial	.2074299	.0127675	16.25	0.000	.182393	.2324669
LnCaptial	.2783011	.0120358	23.12	0.000	.2546988	.3019033
_cons	-3.039001	.1313826	-23.13	0.000	-3.296642	-2.78136

```
.
. *Estimating the error term, u
. predict Error_rate, resid

.
. *running the last regression for part F as required
. regress LnExports Error_rate
```

Source	SS	df	MS	Number of obs	=	2,299
				F(1, 2297)	=	0.17
Model	1.0831756	1	1.0831756	Prob > F	=	0.6837
Residual	14985.1878	2,297	6.52380838	R-squared	=	0.0001
				Adj R-squared	=	-0.0004
Total	14986.271	2,298	6.52144083	Root MSE	=	2.5542

LnExports	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Error_rate	-.0281651	.0691215	-0.41	0.684	-.1637121	.1073819
_cons	14.03814	.0532698	263.53	0.000	13.93368	14.1426

Figure 10:

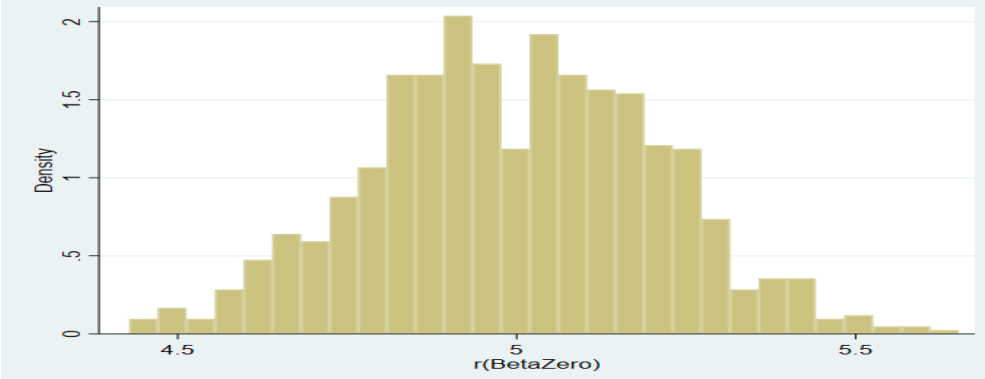

Figure 11: histogram of Beta not

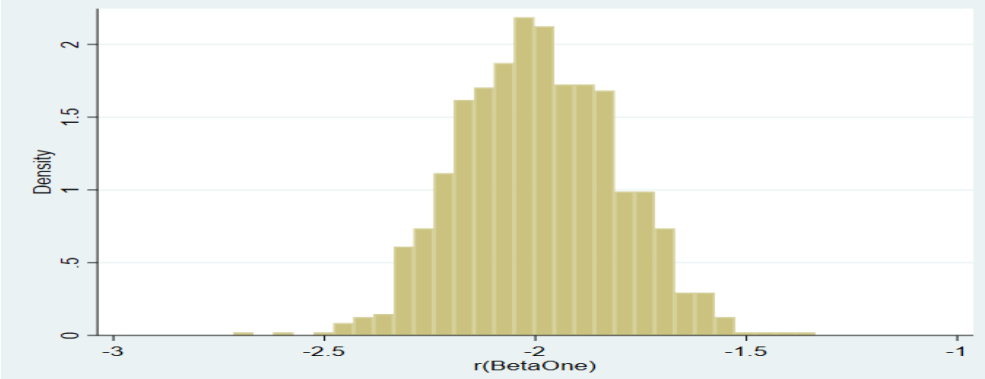

Figure 12: histogram of Beta one

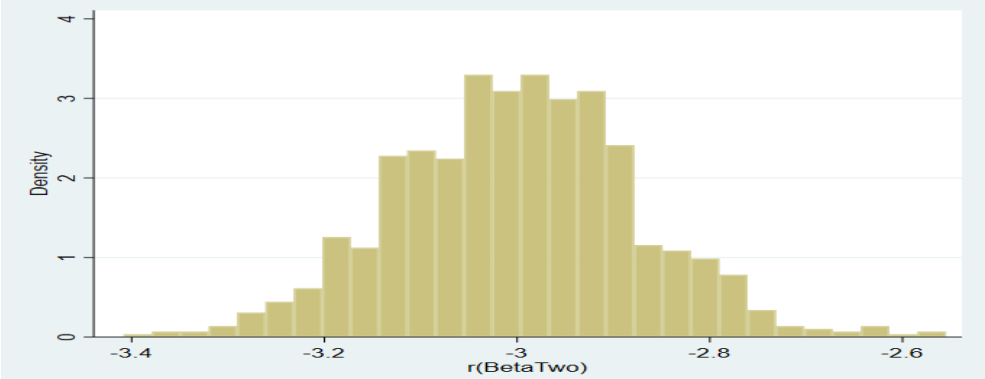

Figure 13: histogram of Beta two