Theoretical Problems

Question 1

1A

$x \sim N(2,3)\ E(X) = 2\ \ var(x) = 3;\ \ y \sim N(0,1)\ E(y) = 0\ var(y) = 1$

$E(z) = E(2x - y) = E(2x) - E(y) = 2E(x) - E(y) = 2 * 2 - 0 = 4$

1B

Since x, y is independent,

$corr(x, y) = 0$

$$corr(x, y) = \frac{cov(x, y)}{\sqrt{var(x)var(y)}} = \frac{0}{\sqrt{3 * 1}} = 0$$

1C

$var(z) = var(2x - y) = var(2x) + var(y) - 2 * 2cov(x, y)$

$= 4var(x) + var(y) - 4cov(x, y)$

$= 4 * 3 + 1 - 4 * 0 = 13$

1D

$cov(x, z) = cov(x, 2x - y) = 2cov(x, x) - cov(x, y) = 2var(x) - cov(x, y) = 2 * 3 - 0 = 6$

$$corr(x, z) = \frac{cov(x, z)}{\sqrt{var(x)var(z)}} = \frac{6}{\sqrt{3 * 13}} = 0.961$$

1E

$x_i \sim N(2,3) \qquad y_i \sim N(0,1)$

Since $z_i = 2x_i - y_i$ *and* $Z = \beta_0 + \beta_1 X + u$

X, Y are independent, $Z = \beta_0 + \beta_1 X + u$ will not affected by Y.

We use this model to estimate the linear relationship between x and z.

As x increase by 1 unit and holding y constant, z will expect to increase by 2 unit.

We expect $\beta_1 = 2$.

Question 2

2A

Bias refers to the tendency of a measurement process to over or under estimate the value of a population parameter.

$$E\left(\frac{y_j - y_k}{x_j - x_k}\right) = E\left(\frac{\beta_0 + \beta_1 x_j + u_j - (\beta_0 + \beta_1 x_k + u_k)}{x_j - x_k}\right) = E\left(\beta_1 + \frac{u_j - u_k}{x_j - x_k}\right)$$

$$= E(\beta_1) + E\left(\frac{u_j - u_k}{x_j - x_k}\right) = \beta_1$$

Thus, $bias = E(\widehat{\beta_1}) - \beta_1 = 0\ \left(note: u_j - u_k = u_i - u_i = 0\right)\ \widehat{\beta_1}\ is\ unbiased.$

2B

$$\widehat{\beta_1} = \frac{y_j - y_k}{x_j - x_k} = \frac{y_j}{x_j - x_k} - \frac{y_k}{x_j - x_k}$$

$\widehat{\beta_1}$ is a linear estimator, it can be expressed as a linear function of x and y.

2C

By Gauss Markov Theorem, the least squares estimator is appealing because

1. Under the assumptions of linear regression model, it is the best linear unbiased estimator
2. Under the assumptions of linear regression model, of all linear and unbiased estimates of β, $\hat{\beta}$ has the minimum variance and it is unique

2D

True. OLS is used to find the best fit line. If we only know observations j and k, the most fitted line is the line that cross two points. The slope of this line is $\hat{\beta}_1 = \frac{y_j - y_k}{x_j - x_k}$ which is as same as your friend's estimation.

Question 3

3A

β_1 is the SLOPE of the regression line. β_1 represents the difference in the predicted value of Y for each one-unit difference in X1.

In this case, if tutorial hours(X) differed by one-unit the GPA (Y) will differ by -0.02 units on average.

3B

No. There may have other variables in "u" that influences the GPA score. For example, the innate ability is not observed in u term. Noticed that the higher the innate ability, the lower the hours spend and more likely have a higher GPA. So, cov (innate ability, hours) < 0 and $E(u|x) \neq 0$. The omitted variable bias which causes assumption 2 :$E(u|x)=0$ not hold. Regarding whether the school decides to cancel the tutoring program, it cannot only depend on the tutoring hour.

Computer based problems

Problem 1 weight loss

A. The weight loss data is a time series data since the series of data points listed in time order.

B. BMI=(WeightPounds/2.2)/1.73^2

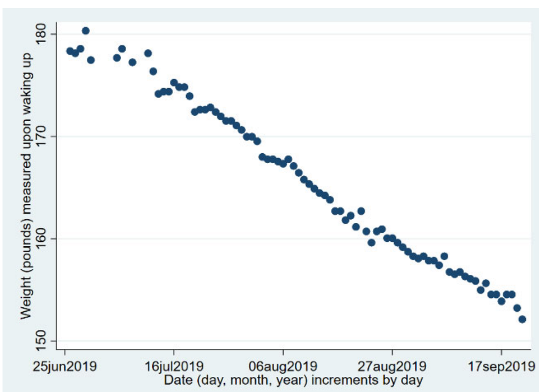
The mean food consumption is 2.913. It tells that the average of plates of food consumption of TA is 2.9 per day.

```
. tabstat WeightPounds WaistInches PlatesFoodCons BMI, s(mean v sd n)

   stats |  Weight~s  WaistI~s  Plates~s       BMI
---------+----------------------------------------
    mean |  165.4989  34.93957   2.91358  25.13508
variance |  63.48545  1.806475  .9299383  1.464348
      sd |  7.967776  1.344052  .9643331  1.210103
       N |        81        46        81        81
--------------------------------------------------
```

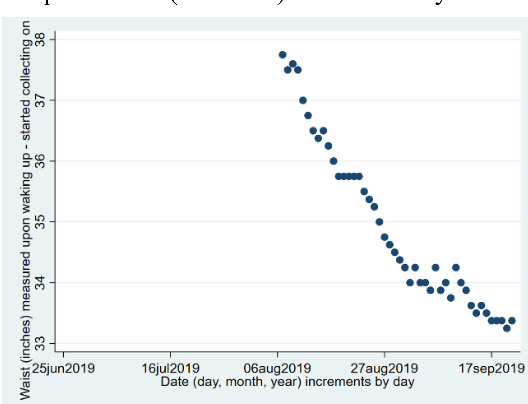
C.

Graph 1: Weight(in pounds) vs time in days



This graph shows that the TA's weight in pounds decreases over time (i.e. There is a negative relationship). The relationship appears linear. The points line in a tight band indicating that the relationship is quite strong. The variability appears reasonably constant.

Graph 2: Waist(in inches) vs time in days



The TA's waist in inches decreases as date increases. There appears to be a moderate negative linear relationship. There is quite a lot of variability.
In general, there is a significant weight loss for TA. Two graphs cover the same time periods, from 07 AUG 2019 to 17 SEPT 2019.

D.

i. WeightPounds = 7221.293 - 0.3240695 * TimeUnitDay

```
. reg 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
------------------------------------------------------------------------------
```

ii. In general, the higher the R-squared, the better the model fits the data. In this case, R^2 is 0.98. It indicates that the model explains 98% of the variability of the response data around its mean. Also, when we look at the scatter plots (Weight (in pounds) vs. time in days), it shows the strong linear relationship between two variables. So, the simple linear regression fits the data well.

iii. β_1 represents the difference in the predicted value of Y for each one-unit difference in X1. In this case, the weight is expected to decrease by 0.3240695 pounds on average when each day passes.

iv.

WeightPounds = 7221.293 - 0.3240695 * TimeUnitDay

7221.293 - 0.3240695 * TimeUnitDay = 145

TimeUnitDay =21835.7266

on 13 oct 2019, he will be expected to achieve his goal.

v. $\frac{1}{3500} = \frac{0.324}{deficit}$

daily deficit = 1134.24

The average daily calorie deficit from the mean rate of daily weight loss is 1134.24.

E.

It violates the 3rd assumption: random sample-iid data. In time series data, the observations are not independent. In this question, x and y are related to the particular person (the TA).

Problem 2 Exports and Employment

A.

Graph 1: sample mean, standard deviation, the median, the 25^{th} and 75^{th} percentiles for Exports and total employment

```
tabstat exports total_employment, s(mean sd median p25 p75 )
```

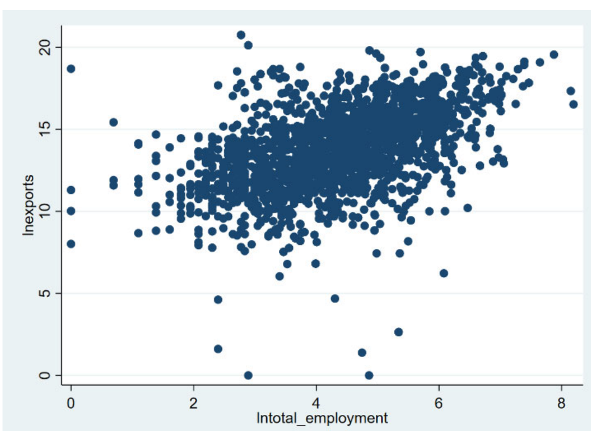
stats	exports	total_~t
mean	1.42e+07	172.4819
sd	5.44e+07	243.153
p50	1385333	94
p25	233964	39
p75	7752355	215

Graph 2: sample mean, standard deviation, the median, the 25^{th} and 75^{th} percentiles for ln(exports) and ln (total employment)

```
.  tabstat lnexports lntotal_employment, s(mean sd median p25 p75 )
```

stats	lnexpo~s	lntota~t
mean	14.03814	4.481492
sd	2.553711	1.214396
p50	14.14145	4.543295
p25	12.36292	3.663562
p75	15.86351	5.370638

B.



The scatter plot appears to be a positive linearly relationship between In(exports) and ln(employment). There is quite a lot of variability and hence this relationship appears moderate to week.

The graph support what the entrepreneurs thought. As total_employment increases, the exports' value increases.

C.

$\ln(\text{exports}) = \beta_0 + \beta_1 \ln(total\_employment) + u_i$

$\ln(\text{exports}) = 9.567017 + 0.9976858\ \ln(total\_employment)$

The estimate of slope is $\widehat{\beta_1} = 0.9977$. As In (total_employment) changes by 1 unit, the ln(exports) will differ by 0.9977 units. A 1% change in total_employment is associated with a 0.998% change in export on average.

```
. reg lnexports lntotal_employment

      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]
-------------+----------------------------------------------------------------
lntotal_em~t |   .9976858   .0386238    25.83   0.000     .9219446    1.073427
       _cons |   9.567017   .1793324    53.35   0.000     9.215346    9.918687
------------------------------------------------------------------------------
```

D.

median of total_employment is 94

median of lntotal_employment is 4.543295

median of exports is 1385333

The predicted exports for a plant with the median employment:

lnest_exports = 9.567 + 0.9976 * 4.543295 = 14.0998

est_exports = exp(14.0998) = 1328818

est_exports =1328818 < median of exports = 1385333

The predict exports is smaller than the median of exports. The prediction is not accurate since u contains other unobservable variables. The prediction is biased.

E.

```
. reg  lnexports lntotal_employment lnmaterials lncapital

      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]
-------------+----------------------------------------------------------------
lntotal_em~t |  -.0281651   .0516769    -0.55   0.586    -.1295034    .0731731
 lnmaterials |   .8166265   .0333824    24.46   0.000     .7511637    .8820893
   lncapital |   .0634026   .0330915     1.92   0.055    -.0014897    .1282949
       _cons |   .5112672   .3612492     1.42   0.157    -.1971418    1.219676
------------------------------------------------------------------------------
```

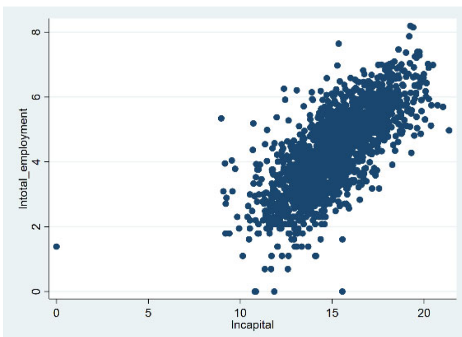
$\ln EX_i = \beta_0 + \beta_1 \ln(total_{employment}) + \beta_2 \ln(Matrials) + \beta_3 \ln(Captial) + Ui$

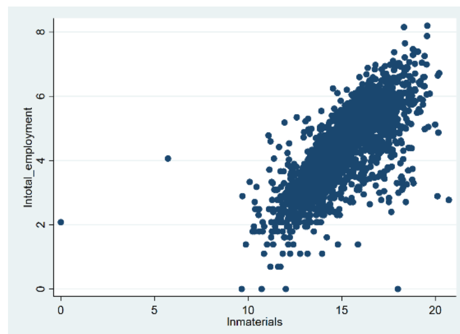
In general, the results change much with respect to the previous model's results.

The omitted variable bias caused misspecification analysis. The capital and materials are omitted from first model, so the estimators of β_1 is biased. In second model, $\widehat{\beta_2}$ *and* $\widehat{\beta_3}$ both positive and measured as 0.816 and 0.634 separately. Thus, we noticed that the materials and capital are positively correlated to total_employment.

In first model β_1 has an upward bias.

In second model, holding materials and capital constant, a 1% change in total_employment is associated with a -0.28 % change in export on average.





F.

```
. reg lntotal_employment lnmaterials lncapital

      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

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

. reg lnexports est_error

      Source |       SS           df       MS      Number of obs   =     2,299
-------------+----------------------------------   F(1, 2297)      =      0.17
       Model |  1.08317372         1  1.08317372   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]
-------------+----------------------------------------------------------------
   est_error |  -.0281651   .0691215    -0.41   0.684    -.1637121    .1073819
       _cons |   14.03814   .0532698   263.53   0.000     13.93368     14.1426
------------------------------------------------------------------------------
```

$\widehat{\beta_1} = -0.028$ in model $\ln EX_i = \beta_0 + \beta_1\,\widehat{\epsilon_i} + \widehat{u_i}$ is the same in model $\ln EX_i = \beta_0 + \beta_1 \ln(total_{employment}) + \beta_2 \ln(Matrials) + \beta_3 \ln(Captial) + Ui$

This example is a partialling out Approach Example. In model $\ln L_i = \alpha_0 + \alpha_1 lnM_i + \alpha_2 lnK_i + \epsilon_i$, ϵ_i can be interpreted as the variation in $\ln L_i$ that cannot be explained by $\ln M_i$ *and* lnK_i. In other words, ϵ_i is the part of $\ln L_i$ that is uncorrelated with $\ln M_i$ *and* lnK_i. When we use this unique variation ϵ_i to do linear regression, the influences that ϵ_i have on y are the same as x1 on y when we are holding other factors constant. From $\ln EX_i = \beta_0 + \beta_1\,\widehat{\epsilon_i} + \widehat{u_i}$, as ϵ_i increase by 1 unit, we expect ln(export) decrease by -0.028 unit on average.

Problem 3 Monte Carlo Simulation

A.

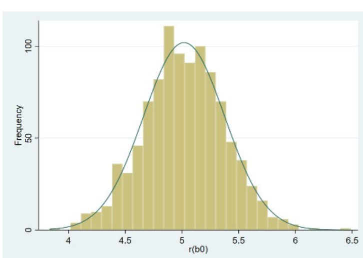
The average of $\widehat{\beta_0} = 5.018, \widehat{\beta_1} = -2.00721, \widehat{\beta_2} = -2.984$.

The average of $\hat{\beta}$ is close to the true β

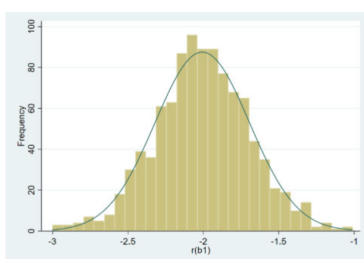
The average and std.Dev of $\hat{\beta}$:

Variable	Obs	Mean	Std. Dev.	Min	Max
b0	1,000	5.018323	.3576911	3.834016	6.486395
b1	1,000	-2.00721	.3125933	-2.999369	-1.010637
b2	1,000	-2.984774	.1537827	-3.468918	-2.520949

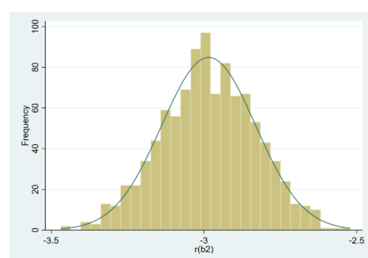
Histogram Graph:



Histogram of β_0



Histogram of β_1



Histogram of β_2

B

Variable	Obs	Mean	Std. Dev.	Min	Max
b0	**1,000**	**8.000995**	**.6729103**	**5.68796**	**10.17895**
b1	**1,000**	**-3.192236**	**.7180337**	**-5.499341**	**-.4678245**

True model : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$ $\widehat{\beta_0} = 5.018, \widehat{\beta_1} = -2.00721$

Observed model (omit x2): $Y = \beta_0 + \beta_1 X_1 + V$ $\widehat{\beta_0} = 8.0$ $\widehat{\beta_1} = -3.19$

The average of $\widehat{\beta_0}$ *and* $\widehat{\beta_1}$ are not close to the values I would expect because of omitted variable bias. The observed model excludes a relevant variable x2.

PROOF:

Since x2 $= 0.4\, x_1 + V$

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$

$Y = \beta_0 + \beta_1 X_1 + \beta_2 (0.4\, x_1 + V) + U$

$Y = \beta_0 + (\beta_1 + 0.4\beta_2) x_1 + (\beta_2 V + U)$

Let $\beta_0 = \delta_0$; $(\beta_1 + 0.4\beta_2) = \delta_1$; $(\beta_2 V + U) = \xi$

$E(\widehat{\delta_1}) = \delta_1 = \beta_1 + 0.4\beta_2 < \beta_1$ $since\ \widehat{\beta_2} = -2.984$

β_1 $is\ biased$

bias $(\widehat{\delta_1})$= $0.4\beta_2$= 0.4 * -2.984= -1.19

$note$: $\beta_{1\widehat{omit\ x2}} - \beta_{1\,\widehat{true\ model}} = -3.19 - (-2.0) = -1.19$