ECO 375–Homework 1
University of Toronto
Due: 13 October, 2019
Late assignments will not be accepted
For full credit, please show your work

1 Theoretical Problems

1. Suppose $X \sim N(2,3)$ and $Y \sim N(0,1)$, and that X and Y are independent. Define $Z = 2X - Y$.

   (a) What is expected value of Z?

   (b) What is the covariance between X and Y? What is their correlation?

   (c) What is the variance of Z?

   (d) What is the covariance between X and Z? What is their correlation? *(Hint: use the fact that, for any random variables A, B, and C, and constants d, e, and f, $Cov[dA, eB + fC] = deCov[A,B] + dfCov[A,C]$.)*

   (e) Suppose we draw 100 independent observations of $x_i \sim N(2,3)$ and $y_i \sim N(0,1)$ and define $z_i = 2x_i - y_i$. Now, suppose a researcher estimates the model $Z = \beta_0 + \beta_1 X + u$ using this data. What is the expected value of β_1? *(Hint: you don't need to use any statistical formulas, as you did in the previous parts of this problem. Instead, think carefully about what the researcher is trying to do.)*

2. A friend of yours is writing an research paper. The friend knows that $Y = \beta_0 + \beta_1 X + u$, where u has mean zero and is independent of X. The friend gathered 100 independent and identically distributed observations x_i and y_i, all of which have different values of x_i. (Note that this means that, for example, x_1 is independent and identically distributed with x_2; however, x_1 is *not* necessarily independent and identically distributed with y_1.)

   Your friend wants to estimate β_1 but doesn't want to use Stata. Therefore, instead of running OLS to estimate β_1, they randomly choose two observations j and k and estimate β_1 using the slope of the line connecting them, $\tilde{\beta}_1 = \frac{y_j - y_k}{x_j - x_k}$.

   (a) What is the bias of $\tilde{\beta}_1$?

   (b) Is $\tilde{\beta}_1$ a linear estimator?

   (c) Your friend thinks that their estimator will have a lower variance than the usual OLS estimator. Without using any formulas, how would you try to convince your classmate that using the usual OLS estimator to estimate the linear regression model is the better choice?

(d) To counter your argument, your classmate argues that $\tilde{\beta}_1$ would be exactly the OLS estimator if they only knew observations j and k. Is that true? *(Hint: you don't need to use any statistical formulas. Instead, think carefully about what $\tilde{\beta}_1$ is estimating.)*

3. A high school wants to help its struggling students, so it offers a tutoring program: anyone who received a D or F in the 2018-2019 school year may receive free tutoring in 2019-2020. To evaluate that program, they collect data from the whole school on GPA in 2019-2020 and hours of tutoring received (where hours is zero for anyone who did not receive any tutoring). They then estimate the model parameters in $GPA = \beta_0 + \beta_1 hours + u$ using OLS, and interpret $\hat{\beta}_1$ (the OLS estimate of β_1) as the causal effect of tutoring hours on GPA. They estimate $\hat{\beta}_1 = -0.02$.

   (a) For now, assume all of the usual OLS assumptions apply and the estimate is statistically significantly different from zero. In words that a non-economist could understand, what does "$\hat{\beta}_1 = -0.02$" mean?

   (b) Based on these results, the school decides to cancel the tutoring program. Do you agree with this decision? Why or why not? Use the assumptions of the simple linear regression model to explain.

2 Computer Based Problems

1. **Weight Loss.** Over the PhD one of the course TAs at UTM (Hammad) gained some weight. To get back in shape he has been exercising and dieting since June, 2019. Given his interest in data and analytics, he collected data on his weight, waist size, and food consumption almost on a daily basis; he has made significant progress thus far. You will be using "WeightFood-Days.dta" for this problem. Below is a table describing the variables in this data:

Variable	Description
TimeUnitDay	Date (day, month, year) increments by day
WeightPounds	Weight (pounds) measured upon waking up
WaistInches	Waist (inches) measured upon waking up - started collecting on aug 7
PlatesFoodCons	Plates of food consumed at end of each day

   (a) Econometrics involves analyzing different types of data sets including cross sectional, repeated cross sectional, panel, and time series data. What kind of data set is this? Please justify your answer.

   (b) Create a table with the mean, variance, standard deviation, and number of observations for weight (pounds), waist (inches), food consumption (# of plates), and Body Mass Index (BMI). *(Hint: you can use the Stata command* `tabstat`*.)* You will need to use $BMI = \frac{\text{weight (kilograms)}}{(\text{height in metres})^2}$, noting that Hammad's height is 1.73 metres, and using the fact that 1 kg $\approx$ 2.2 pounds. What is the mean of food consumption, and what does that tell us?

   (c) Create two scatter plots: one with weight (in pounds) on the y-axis, and the other with waist (in inches) on the y-axis. For both, time in days should be on the x-axis. Briefly describe the overall patterns observed in the resulting plot. Do both graphs cover the same time periods?

(d) Regress the weight (pounds) on time (days) and use the estimated slope to answer the following:

i. Show the results of the regression.

ii. Do you think a simple linear regression fits the data well? Justify your answer.

iii. Interpret the estimated slope.

iv. Hammad's goal is to achieve a weight of 145 lbs. Assuming a linear rate of weight loss, on what calendar date is he expected to achieve his goal?

v. The primary method to losing weight is to create a calorie deficit. This involves consuming fewer calories than you burn through your metabolism. Estimate the average daily calorie deficit from the mean rate of daily weight loss; assume that Hammad loses one pound for every 3500 calories of calorie deficit that he creates.

(e) The previous question involved estimating a simple linear regression of weight (pounds) on time (days). Does this violate any of the regression assumptions? *(Hint: refer to your answer in part (a) of this problem.)*

2. **Exports and Employment.** The Colombian Annual Manufacturing Survey (AMS) is a Census of all the Manufacturing Establishments with more than 9 employees. The dataset "AMS_exporters.dta" contains information on the exporting establishments from the AMS for the year 2012. Below is a table describing the variables in this data:

Variable	Description
nordest	Establishment identificator
employment_w	Number of women employees
employment_m	Number of men employees
capital	Value of fixed capital (in Colombian currency)
materials	Value of the materials used in the production (in Colombian currency)
exports	Value of exports (in Colombian currency)
revenue	Value of revenue (in Colombian currency)

Colombian entrepreneurs believe that the exports' value depends positively on the productive plants' size, in terms of the number of employees. To test if they are right, you were asked to estimate the following model:

$$\ln EX_i = \beta_0 + \beta_1 \ln L_i + u_i$$

Where EX_i is the exports' value, L_i the number of employees and u_i the error of the model.

(a) Present a table that shows the sample mean, standard deviation, the median, the 25th and 75th percentiles for exports, total employment, ln(exports) and ln(total employment).

(b) Construct a scatterplot to explore the relationship between the export value and the total number of employees, both in logs. Visually, does the graph support what the entrepreneurs thought?

(c) Estimate the model using OLS and show the results. What is the estimated slope? Interpret the estimated β_1.

(d) Using the estimates, predict the exports for a plant with the median employment and compare this prediction with the median value of exports.

(e) Now you realized that the model you are using is misspecified, and the right model is:

$$\ln EX_i = \beta_0 + \beta_1 \ln L_i + \beta_2 \ln M_i + \beta_3 \ln K_i + u_i$$

Where M_i and K_i are the value of materials and capital used by the plant. Estimate this new model. Do the results change much with respect to the previous model's results? Why do you think this happens?

(f) Run the following regression:

$$\ln L_i = \alpha_0 + \alpha_1 \ln M_i + \alpha_2 \ln K_i + \varepsilon_i$$

Predict $\hat{\varepsilon}_i$, then run the following regression and show the results:

$$\ln EX_i = \beta_0 + \beta_1 \hat{\varepsilon}_i + \tilde{u}_i$$

Compare the results with the results from the previous question. Explain your findings.

3. **Monte Carlo Simulation**. Simulate the following model in STATA:

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

where

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 5 \\ -2 \\ -3 \end{pmatrix}$$

$$X_1 \sim N(0,1),$$

$$\begin{aligned} X_2 &= 0.4X_1 + V, \\ V &\sim N(-1,2). \end{aligned}$$

and

$$U \sim N(0,3).$$

For each simulation, generate a data set $\{y_i, x_{i1}, x_{i2} : i = 1, ..., n\}$ with $n = 100$ observations. Then, for each sample, estimate β using OLS and save the result $\widehat{\beta}$. Run $m = 1000$ simulations.

(a) What is the average and the standard deviation of $\widehat{\beta} = \left(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2\right)$ across the simulations? Is the average of $\widehat{\beta}$ close to the true β? Plot the histograms of $\widehat{\beta}$.

(b) Using the same specification as before, now omit X_2 in the regressions. I.e., generate the same model but for each simulation regress Y on X_1 only. Report the averages and the standard deviation of both $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Are the averages of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ close to the values you would expect? Explain why.
(Hint: If you want to, you can refer to the OLS formulas to explain the results.)

Provide your do file and log file as part of your submission.