Financial Econometrics
Problem Set 1: review of probability and statistics, linear regression, and inferences
Due date: Sunday 24, 2023, before 11:59pm

1. Consider each of these statements whether it is true, false, or uncertain (not enough information to conclude). Let assume that we have an iid sample.
   i. let $E[U|X]= 3$. U is mean dependent of X
   ii. let $E[U|X]= 3$. $E[U]E[X] = E[UX]$
   iii. let $E[U|X]= 3$. $Var[U - X] = Var[U] + Var[X]$
   iv. let $E[U|X] = 0$ and $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$. The model is homoskedastic.
   v. OLS estimator of the model $Y = \beta_0 + U$ always has $R^2 = 0$.
   vi. Let $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$ and $cov[X_1, X_2] \neq 0$. The OLS estimator is not consistent.
   vii. let $\hat{\theta}_n$ be an unbiased estimator of θ. $plim(\hat{\theta}_n) = \theta$
   viii. Multicollinearity makes OLS estimator biased and inconsistent
   ix. $\bar{R}^2$ must have a value falling within [0,1]
   x. Significance level is always less than 1

2. Suppose $Z = 2^X - 1$, $Y = X^2$, and X is a discrete random variable with the p.m.f

$$P\{X = -2\} = \frac{1}{6}, P\{X = -1\} = \frac{1}{3}, P\{X = 1\} = \frac{1}{3}, P\{X = 2\} = \frac{1}{6}$$

   a) Is Z a random variable? Why?
   b) Find $E[Z]$
   c) Find $E[Z^2]$
   d) What is the variance of Z?
   e) What is $Cov[X, Y]$?
   f) Is Y mean independent of X?
   g) Is Y independent of X?

3. Let GPA denote a random variable for the grade point average of a student enrolling in the Master of Finance program in 2020 and GMAT denote a random variable for the student's GMAT score. Suppose $E[GPA|GMAT] = 0.007GMAT - 1.73$
   a) Is $E[GPA|GMAT]$ a random variable? Why?
   b) What is the expected value of GPA when $GMAT$ score is 650? What is the expected GPA when $GMAT$ score is 790?
   c) If $E[GMAT] = 700$, what is $E[GPA]$?

4. Let X denote the annual salary of bankers in Thailand measured in thousand Baht. Suppose that $E[X] = 27.6$ and the standard deviation of X = 11.2. Let Y denote the monthly salary of bankers in Thailand measured in Baht. What is $E[Y]$ and $Var[Y]$?

5. Let (X, Y) be a random vector, and let $(X_1, Y_1), \ldots, (X_n, Y_n) \sim iid(X, Y)$. If $Var[X] < \infty$ and $Var[Y] < \infty$. Consider estimating $E[X]E[Y]$ using the estimators

$$\hat{\theta}_n = \frac{1}{2n}(X_1 + X_n)\left(\sum_{j=1}^{n} Y_j\right)$$

$$\bar{X}_n \bar{Y}_n = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)$$

a) Is $\hat{\theta}_n$ an unbiased estimator of $E[X]E[Y]$? Explain.
b) Is $\hat{\theta}_n$ a consistent estimator of $E[X]E[Y]$? Explain.
c) Is $\bar{X}_n\bar{Y}_n$ a consistent estimator of $E[X]E[Y]$? Explain.

6. Suppose a researcher wants to study how the distance between household's residence and the nearest bank branch is related to rate of returns on investment of household enterprise. He uses the following model:

$$R = \alpha_0 + \alpha_1 dist + \alpha_2 dist^2 + U$$

where $dist$ is the distance between household's residence and the nearest bank branch measured in kilometers. Suppose he interprets this model as the *ceteris paribus* causation from distance to rate of returns.

a) What is the effect of the distance on rate of returns? Does it depend on the distance?
b) The threshold effect is defined as the distance that has zero impact on the rate of returns. What is this threshold in terms of $\alpha_0, \alpha_1, \alpha_2$?
c) Do you think the OLS gives a consistent estimator of the threshold impact?
d) Now, if he wants to study correlation (instead of causation) between the distance and rate of returns by using this model. The threshold will simply mean the point that the correlation between distance and relationship changes from negative to positive. Will OLS gives a consistent estimate of the threshold in this case? Explain.

7. Starting with the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$, how would you transform the regression to

$$\tilde{Y} = \gamma_0 + \gamma_1 X_1 + \gamma_2 \tilde{X}_2 + U$$

so that you can test the null hypothesis $H_0: \gamma_1 = 0$ vs $H_1: \gamma_1 \neq 0$, using the following t-statistics.

$$t = \frac{\hat{\gamma}_1}{SE(\hat{\gamma}_1)} \rightarrow N(0,1)$$

In other words, find $\tilde{Y}$ and $\tilde{X}_2$ after transforming if you want to do the following tests:

a) $H_0: 2\beta_1 - \beta_2 = 0 \; vs \; H_1: 2\beta_1 - \beta_2 \neq 0$
b) $H_0: \beta_1 = 3\beta_2 \; vs \; H_1: \beta_1 \neq 3\beta_2$
c) $H_0: \beta_1 + \beta_2 = 1 \; vs \; \beta_1 + \beta_2 \neq 1$

8. (Previous Midterm) Suppose that you as a researcher would like to study the effect of household income on its debt level by using the linear regression model:

$$\ln Y = \beta_0 + \beta_1 X_1 + U$$

where Y = Household debt in unit of thousand Baht
X_1 = Household monthly income in unit of Baht

a) Suppose that you run the OLS regression and get $\hat{\beta}_1 = -0.03$, $SE(\hat{\beta}_1) = 0.01$, and $R^2 = 0.287$. If you change to use household annual income in unit of thousand Baht, what are the values of your new $\hat{\beta}_1, SE(\hat{\beta}_1)$ and R^2?
b) If you change to measure Y in unit of Baht instead, how would this affect your OLS estimates $\hat{\beta}_0, \hat{\beta}_1$?

From now on, suppose that you adjust your model to

$$\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U$$

where X_2 = dummy variable, taking value 1 if the household's main source of income is from agriculture, and X_3 = dummy variable, taking value 1 if the household's an ordinary wage earner.

c) how would you interpret $\beta_2 - \beta_3$ in plain English?

d) If you want to test the hypothesis $H_0: \beta_2 - \beta_3 = 0; H_1: \beta_2 - \beta_3 \neq 0$ but do not have an access to a program to estimate $cov(\hat{\beta}_3, \hat{\beta}_2)$, what should be your new regressand $\tilde{Y}$ and new regressors $\tilde{X}_1, \tilde{X}_3$ if you transform the model to

$$\tilde{Y} = \gamma_0 + \gamma_1 \tilde{X}_1 + (\beta_2 - \beta_3) X_2 + \gamma_3 \tilde{X}_3 + U$$

so that you can run the regression and get $Se(\hat{\beta}_2 - \hat{\beta}_3)$ automatically?

9. (STATA exercise) Use the data file wage.dta from blackboard, run OLS regression to estimate the following linear regression model:

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 female \cdot educ + \beta_3 grad \cdot educ + \beta_4 grad \cdot female \cdot educ + U$$

where the variables $wage$, $educ$, and $female$ are given in the dataset and denote monthly salary in unit of thousand Baht, years of education, and dummy variable for being female respectively. Let the variable $grad$ be the dummy variable taking value 1 if the observation has at least 16 years of education. In other words, $grad$ is a dummy variable for college graduate. Suppose that you want to interpret this linear model as a causal relationship from education to wage.

a) Estimate the model and report the results in an outreg format table.

b) Report the OLS estimates of $\beta_3 + \beta_4$ along with its standard errors.

c) Test the hypothesis that, for college graduates, the impact of education on wage is different between male and female. Can you reject the null hypothesis at 10%, 5%, and 1% significance levels?

d) Test the hypothesis that, for college graduates, the impact of education on wage for male is higher than that for female. Can you reject the null hypothesis at 10%, 5%, and 1% significance levels respectively?