Suparat Patarasupanit (FT27)
6684069326

Financial Econometrics : Problem set 1

I

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.

i) False, if $E[U|X]=3$, then U is mean independent of X because 3 is a constant.

ii) If $E[U|X]=3$, then U is mean independent of X as 3 is a constant.
Also, because U is mean independent of X, $cov(U,X)=0$. Thus, the statement $E[UX] = E[U]\cdot E[X]$ is true.

iii) If $E[U|X]=3$, then U is mean independent of X and $cov(U,X)=0$. Then,

$$
\begin{aligned}
Var[U-X] &= Var[U+(-X)] \\
&= Var[U] + Var[-X] + 2\,\underbrace{cov(U,X)}_{0} \\
&= Var[U] + (-1)^2 Var[X] \\
&= Var[U] + Var[X]
\end{aligned}
$$

Thus, the statement $Var[U-X] = Var[U] + Var[X]$ is true.

iv) Homoskedasticity is an assumption of linear regression that the variance of the residual term is constant for all observations.
↳ Homoskedasticity implies that $Var[U|X] = Var[U] = \sigma^2$
residual error น้อยมาก R^2 เข้าใกล้ 1

However, $E[U|X]=0$ and $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$ do not imply anything regarding $Var[U|X]$
Thus, the statement is uncertain.

I

✓ 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

v) OLS estimator of the model $Y = \beta_0 + U$;

OLS estimator : $\hat{Y}_i = \hat{\beta}_0$

and

$$\hat{\beta}_0 = \bar{Y}_n - \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} \bar{X}_n \quad (\to 0)$$

$$\hat{\beta}_0 = \bar{Y}_n$$

Then, $\hat{\beta}_0 = \hat{Y}_i = \bar{Y}_n$

As,
$$R^2 = \frac{\sum_{i=1}^{n} [(\hat{Y}_i - \bar{Y}_n)^2]}{\sum_{i=1}^{n} [(Y_i - \bar{Y}_n)^2]} \quad (\to 0)$$

$$= 0$$

Thus, the statement OLS estimator of the model $Y = \beta_0 + U$ always has $R^2 = 0$ is true.

vi) Uncertain, we cannot conclude that OLS estimator is not consistant unless $cov(U, X_1) = 0$ and $cov(U, X_2) = 0$.

vii) Uncertain, we cannot conclude that $plim(\hat{\theta}_n) = 0$ or $\hat{\theta}_n$ is a consistent estimator because unbiasedness and consistency of an estimator do not imply to each other.

viii) False, if the linear regression model has multicollinearity problem, then we cannot use OLS estimator to estimate the model as β is not unique.

ix) False, $\bar{R}^2$ can be negative if we keep adding non-meaningful regressor (x) to the model

x) True, Significance level is the probability of rejecting the null hypothesis and probability cannot exceed 1. Thus, Significance level is always less than 1.

II

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?

a) Z is a random variable because a function of random variable X.

b) $E[Z] = E[2^X - 1]$

$= E[2^X] - 1$

$= \{(2^{-2} \cdot \frac{1}{6}) + (2^{-1} \cdot \frac{1}{3}) + (2^1 \cdot \frac{1}{3}) + (2^2 \cdot \frac{1}{6})\} - 1$

$= (\frac{1}{24} + \frac{1}{6} + \frac{2}{3} + \frac{2}{3}) - 1$

$= \frac{13}{24}$

c) $E[Z^2] = E[(2^X - 1)^2]$

$= (2^{-2} - 1)^2 \cdot \frac{1}{6} + (2^{-1} - 1)^2 \cdot \frac{1}{3} + (2^1 - 1)^2 \cdot \frac{1}{3} + (2^2 - 1)^2 \cdot \frac{1}{6}$

$= \frac{3}{32} + \frac{1}{12} + \frac{1}{3} + \frac{3}{2}$

$= \frac{193}{96}$

d) $Var[Z] = E[Z - E[Z]]$

$= E[Z^2] - (E[Z])^2$

$= \frac{193}{96} - \left(\frac{13}{24}\right)^2$

$= \frac{193}{96} - \frac{169}{576}$

$= \frac{989}{576}$

II

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?

e) $Cov(X,Y) = E[XY] - E[X]E[Y]$

$= E[X^3] - E[X]E[X^2]$

$= \left\{-8\cdot\frac{1}{6} + (-1)\cdot\frac{1}{3} + 1\cdot\frac{1}{3} + 8\cdot\frac{1}{6}\right\} - \left(\cancel{-2\cdot\frac{1}{6} - 1\cdot\frac{1}{3} + 1\cdot\frac{1}{3} + 2\cdot\frac{1}{6}}\right)\left(E[X^2]\right)$ $\to 0$ $\to E[X] = 0$

$= -\frac{4}{3} - \frac{1}{3} + \frac{1}{3} + \frac{4}{3} = 0$

f) Is Y mean independent of X?

↳ $E[Y|X]$ must equal to $E[Y]$ or a constant value

$E[Y|X] = E[\overset{\text{treat as constant}}{X^2}|X] = X^2 \neq$ constant $\to X^2 \in \{1,4\}$

∴ Y is not mean independent of X

g) Is Y independent of X?

No, since Y is not mean independent of X, Y cannot be independent (strongest condition) of X.

Let p represents Y is independent of X and q represents Y is mean independent of X. As independence is stronger than mean independence, $\underbrace{p \to q \equiv T}$. However, from this question, Y is not mean independent of X. Thus, from $\sim q \to \sim p \equiv T$, Y is also not independent of X.

$* \; p \to q \equiv \sim q \to \sim p$

III

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 $E[GPA]$?

a) $E[GPA \mid GMAT]$ is a random variable because it is a function of random variable "GMAT".

b) $E[GPA \mid GMAT = 650] = 0.007(650) - 1.73 = 2.82$

$E[GPA \mid GMAT = 790] = 0.007(790) - 1.73 = 3.80$

c) $E[GPA] = E[E[GPA \mid GMAT]]$

$= E[0.007\,GMAT - 1.73]$

$= 0.007\,E[GMAT] - 1.73$

$= 0.007(700) - 1.73$

$= 3.17$

IV

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 $Var[Y]$?

$Y = \frac{1000}{12} \cdot X$

- $E[Y] = E\left[\frac{1000}{12} \cdot X\right] = \frac{1000}{12} \cdot E[X] = \frac{1000}{12} \cdot (27.6) = 2{,}300$

- $Var[Y] = Var\left[\frac{1000}{12} \cdot X\right]$

$= \left(\frac{1000}{12}\right)^2 \cdot Var[X]$

$= \left(\frac{1000}{12}\right)^2 \cdot (11.2)^2$

$= 871{,}111.11$

V

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.

a) $\hat{\theta}_n$ is an unbiased estimator of $E[X]E[Y]$ if and only if $E[\hat{\theta}_n] = E[X]E[Y]$

$$E[\hat{\theta}_n] = E\left[\overset{\text{constant}}{\frac{1}{2n}}(X_1 + X_n)\left(\sum_{j=1}^{n} Y_j\right)\right]$$

$$= \frac{1}{2n} \cdot E\left[X_1 \cdot \sum_{j=1}^{n} Y_j + X_n \sum_{j=1}^{n} Y_j\right]$$

$$= \frac{1}{2n} \cdot E\left[X_1Y_1 + X_1Y_2 + \ldots + X_1Y_n + X_nY_1 + X_nY_2 + \ldots + X_nY_n\right]$$

$$= \frac{1}{2n} \cdot \left\{ \overbrace{\begin{array}{l} E[X_1]E[Y_1] + COV(X_1, Y_1) \\ + E[X_1]E[Y_2] + \cancel{COV(X_1, Y_2)}^{\,0\ \text{iid}} \\ + \ldots \\ + E[X_1]E[Y_n] + \cancel{COV(X_1, Y_n)} \end{array}}^{n \text{ terms}} \Bigg\}_{(n-1)} + \overbrace{\begin{array}{l} E[X_n]E[Y_1] + \cancel{COV(X_n, Y_1)}^{\,0\ \text{iid}} \\ + E[X_n]E[Y_2] + COV(X_n, Y_2) \\ + \ldots \\ + E[X_n]E[Y_n] + COV(X_n, Y_n) \end{array}}^{n \text{ terms}} \Bigg\}_{(n-1)} \right\}$$

$$= \frac{1}{2n} \cdot \left\{ \begin{array}{l} E[X_1]E[Y_1] + COV(X_1, Y_1) + E[X_n]E[Y_n] + COV(X_n, Y_n) \\ + (2n-2)E[X]E[Y] \end{array} \right\}$$

$$= \frac{1}{2n} \cdot \left\{ 2\left[\cancel{E[X]E[Y]} + COV(X, Y)\right] + \cancel{(2n-2)}E[X]E[Y] \right\}$$

$$= \frac{1}{2n} \cdot \left\{ 2nE[X]E[Y] + 2COV(X, Y) \right\}$$

$$= E[X]E[Y] + \frac{1}{n}COV(X, Y)$$

$\therefore$ Thus, $\hat{\theta}_n$ is not an unbiased estimator of $E[X]E[Y]$ as $E[\hat{\theta}_n] \neq E[X]E[Y]$.

↳ If $COV(X, Y) = 0$, then $\hat{\theta}_n$ is an unbiased estimator of $E[X]E[Y]$.

$\boxed{V}$

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.

b) $\hat{\theta}_n$ is a consistent estimator of $E[X]E[Y]$ if and only if $\text{plim}[\hat{\theta}_n] = E[X]E[Y]$

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

$$= \text{plim}\left(\frac{1}{2}(X_1 + X_n)\right) \cdot \text{plim}\left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)$$

weak law of large number (WLLN)

$$= \underbrace{\text{plim}\left(\frac{1}{2}(X_1 + X_n)\right)}_{} \cdot E[Y]$$

$\text{plim}\left(\frac{1}{2}(X_1 + X_n)\right)$ does not approach $E[X]$

$\therefore$ Thus, $\hat{\theta}_n$ is not a consistent estimator as $\text{plim}[\hat{\theta}_n] \neq E[X]E[Y]$

c) $\bar{X}_n\bar{Y}_n$ is a consistent estimator of $E[X]E[Y]$ if and only if $\text{plim}[\bar{X}_n\bar{Y}_n] = E[X]E[Y]$

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

$$= \underbrace{\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}_{} \cdot \underbrace{\text{plim}\left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)}_{}$$

WLLN WLLN

$$= E[X]E[Y]$$

$\therefore$ Thus, $\bar{X}_n\bar{Y}_n$ is a consistent estimator of $E[X]E[Y]$ as $\text{plim}[\bar{X}_n\bar{Y}_n] = E[X]E[Y]$

VI

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?

a) Effect of the distance on rate of return :

$$\frac{dR}{d\,dist} = \alpha_1 + 2\alpha_2 dist^2$$

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 $\alpha_0, \alpha_1, \alpha_2$?

b) Distance has zero impact on the rate of returns :

$$\frac{dR}{d\,dist} = \alpha_1 + 2\alpha_2 dist = 0$$

$$dist = \frac{-\alpha_1}{2\alpha_2}$$

$\therefore$ Threshold in terms of $\alpha_0, \alpha_1, \alpha_2$ is $\frac{-\alpha_1}{2\alpha_2}$

c) Do you think the OLS gives a consistent estimator of the threshold impact?

c) OLS estimators : $\hat{\alpha}_1$, $\hat{\alpha}_2$

- OLS gives a consistent estimator of the threshold impact if $plim(\hat{\alpha}_1) = \hat{\alpha}_1$ and $plim(\hat{\alpha}_2) = \hat{\alpha}_2$ or $cov(dist, U) = 0$ and $cov(dist^2, U) = 0$
- As the researcher interprets this model as the ceteris paribus causation, then we cannot conclude that the OLS gives a consistent estimator. This is because causal model does not imply that regressors is uncorrelated with the error term. Thus, we need to justify by some theories that the regressors are uncorrelated with the error term.
- In this case, if there is a factor in error term that can have impact on the rate of returns and is correlated to distance, then OLS gives unconsistent estimator.

VI

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.

d)

- OLS gives a consistent estimator of the threshold impact if $\text{plim}(\hat{\alpha}_1) = \hat{\alpha}_1$ and $\text{plim}(\hat{\alpha}_2) = \hat{\alpha}_2$ or $\text{cov}(dist, u) = 0$ and $\text{cov}(dist^2, u) = 0$
- As researcher interprets this model as correlation, he can conclude that OLS gives a consistent estimator. This is because, under this interpretation, regressors are always uncorrelated with the error term.
- Then, $\text{plim}\left(-\dfrac{\hat{\alpha}_1}{2\hat{\alpha}_2}\right) = -\dfrac{\alpha_1}{2\alpha_2}$

VII

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) $H_0: \beta_1 + \beta_2 = 1 \; vs \; \beta_1 + \beta_2 \neq 1$

a) $H_0: 2\beta_1 - \beta_2 = 0$ vs. $H_1: 2\beta_1 - \beta_2 \neq 0$

$\downarrow$

$H_0: \beta_1 - \frac{1}{2}\beta_2 = 0$

$$Y = \beta_0 + \left(\beta_1 - \tfrac{1}{2}\beta_2\right) X_1 + \tfrac{1}{2}\beta_2 X_1 + \beta_2 X_2 + U$$

$$\underbrace{Y}_{\tilde{Y}} = \beta_0 + \underbrace{\left(\beta_1 - \tfrac{1}{2}\beta_2\right)}_{\gamma_1} X_1 + \beta_2 \underbrace{\left(\tfrac{1}{2}X_1 + X_2\right)}_{\tilde{X}_2} + U$$

b) $H_0: \beta_1 = 3\beta_2$ vs. $H_1: \beta_1 \neq 3\beta_2$

$\downarrow$

$H_0: \beta_1 - 3\beta_2 = 0$

$$Y = \beta_0 + (\beta_1 - 3\beta_2) X_1 + (3\beta_2) X_1 + \beta_2 X_2 + U$$

$$\underbrace{Y}_{\tilde{Y}} = \beta_0 + \underbrace{(\beta_1 - 3\beta_2)}_{\gamma_1} X_1 + \beta_2 \underbrace{(3X_1 + X_2)}_{\tilde{X}_2} + U$$

VII

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) $H_0: \beta_1 + \beta_2 = 1 \; vs \; \beta_1 + \beta_2 \neq 1$

c) $H_0: \beta_1 + \beta_2 = 1$ vs. $\beta_1 + \beta_2 \neq 1$

$\downarrow$ $H_0: \beta_1 + \beta_2 - 1 = 0$

$$Y = \beta_0 + (\beta_1 + \beta_2 - 1) X_1 - (\beta_2 - 1) X_1 + \beta_2 X_2 + U$$

$$Y = \beta_0 + (\beta_1 + \beta_2 - 1) X_1 - \beta_2 X_1 + X_1 + \beta_2 X_2 + U$$

$$\underbrace{Y - X_1}_{\tilde{Y}} = \beta_0 + (\underbrace{\beta_1 + \beta_2 - 1}_{\gamma_1}) X_1 + \beta_2 (\underbrace{X_2 - X_1}_{\tilde{X}_2}) + U$$

VIII

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?

a) monthly & Baht $\rightarrow$ annual & thousand baht

multiply $\frac{12}{1000}$ to regressor

▸ New $\hat{\beta}_1$:

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

new regressor with annual & thousand Baht unit

$$\ln Y = \beta_0 + \frac{\beta_1}{\frac{12}{1000}} \cdot \frac{12}{1000} X_1 + U$$

$\therefore$ Thus, new $\hat{\beta}_1 = \frac{-0.03 \times 1000}{12} = -2.5$

▸ New $SE(\hat{\beta}_1)$:

$$SE(\hat{\beta}_1) = SE\left(\frac{1000\,\hat{\beta}_1}{12}\right) = \left|\frac{1000}{12}\right| \cdot SE(\hat{\beta}_1) = \left|\frac{\cancel{1000}^{10}}{12}\right| \cdot \cancel{0.01} = 0.83$$

▸ New R^2 :

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2}$$

} scaling $\frac{12}{1000}$ to regressor X_1 doesn't change this numerator

} is the same for all models from the same population

$\therefore$ R^2 doesn't change.

VIII

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 $\hat{\beta}_0, \hat{\beta}_1$?

b) Thousand Baht → Baht
multiply 1000 to Y

$$\ln(1000Y) = \beta_0 + \beta_1 X_1 + U$$

$$\ln(1000) + \ln Y = \beta_0 + \beta_1 X_1 + U$$

$$\ln Y = (\beta_0 + \ln(1000)) + \beta_1 X_1 + U$$

∴ New $\hat{\beta}_0$ increases by $\ln(1000)$ but $\hat{\beta}_1$ doesn't change.

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?

c)

	$X_2 = 1$	$X_2 = 0$
$X_3 = 1$	$\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3$	Household's main source of income is from ordinary wage $\ln Y = \beta_0 + \beta_1 X_1 + \beta_3$
$X_3 = 0$	Household's main source of income is from agriculture $\ln Y = \beta_0 + \beta_1 X_1 + \beta_2$	$\ln Y = \beta_0 + \beta_1 X_1$

$$\ln Y^{agriculture} - \ln Y^{wage} = \beta_2 - \beta_3$$

∴ This means that, at the same level of income, the debt level of household with agriculture as its main source of income is $(\beta_2 - \beta_3)\%$ higher than the debt level of household with ordinary wage as its main source of income.

VIII

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?

d)

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

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

$$\underbrace{\ln Y}_{\tilde{Y}} = \beta_0 + \beta_1 \underbrace{X_1}_{\tilde{X}_1} + (\beta_2 - \beta_3) X_2 + \beta_3 \underbrace{(X_2 + X_3)}_{\tilde{X}_3} + U$$

$\therefore$ New regressand $\tilde{Y} = \ln Y$, new regressors $\tilde{X}_1 = X_1$ & $\tilde{X}_3 = X_2 + X_3$

IX

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?

a) Firstly, we need to create new variables for ln(wage), dummy variable grad, and interaction terms.

- Command in STATA
  - gen lwage = ln(wage)
  - gen grad = educ >= 16

or gen grad = 0
replace grad = 1 if educ >= 16

Interaction terms:

- gen femeduc = female * educ
  gen gradeduc = grad * educ
  gen gradfemeduc = grad * female * educ

IX ▸ Result from STATA

command: reg lwage educ femeduc gradeduc gradfemeduc
outreg2 using " HW1Q9A. doc",

```
. reg lwage educ femeduc gradeduc gradfemeduc

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(4, 521)       =     58.75
       Model |  46.1096544         4  11.5274136   Prob > F        =    0.0000
    Residual |  102.220096       521    .1961998   R-squared       =    0.3109
-------------+----------------------------------   Adj R-squared   =    0.3056
       Total |   148.32975       525  .282532857   Root MSE        =    .44294

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ | β1  .0714335   .0098735     7.23   0.000     .0520367    .0908304
     femeduc | β2 -.0302179   .0036203    -8.35   0.000      -.03733   -.0231057
    gradeduc | β3  .0062146   .0049153     1.26   0.207    -.0034415    .0158708
  gradfemeduc | β4  .0142159   .0068539     2.07   0.039     .0007512    .0276805
       _cons |   .8710988   .1146053     7.60   0.000     .6459535    1.096244
------------------------------------------------------------------------------

. outreg2 using "HW1Q9A.doc"
HW1Q9A.doc
```

VARIABLES	(1) lwage
educ	0.0714***
	(0.00987)
femeduc	-0.0302***
	(0.00362)
gradeduc	0.00621
	(0.00492)
gradfemeduc	0.0142**
	(0.00685)
Constant	0.871***
	(0.115)
Observations	526
R-squared	0.311

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

b) OLS estimates of $\beta_3 + \beta_4$ along with its standard errors:

$$\frac{d(\ln(wage))}{d\,educ} = \beta_1 + \beta_2 \text{female} + \beta_3 \text{grad} + \beta_4 \text{grad} \cdot \text{female}$$

	female = 1	female = 0
grad = 1	$\beta_1 + \beta_2 + \beta_3 + \beta_4$	$\beta_1 + \beta_3$
grad = 0	$\beta_1 + \beta_2$	β_1

$(\beta_3 + \beta_4)\%$ tells the % difference of impact of educ on wage between female college graduates $\%(\beta_1 + \beta_2 + \beta_3 + \beta_4)$ and female non-college graduates $\%(\beta_1 + \beta_2)$

▸ From STATA, $\%(\beta_3 + \beta_4)$ is equal to $0.0062146 + 0.0142159 = 0.0204305 = 2.04\%$

∴ For female, the impact of educ on wage for college graduates is 2.04% different from the impact of educ on wage for non-college graduates.

IX

- Find standard error of $\beta_3 + \beta_4$

* use transformation

↓

new model: $\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 female \cdot educ + (\beta_3 + \beta_4) grad \cdot educ - \beta_4 grad \cdot educ + \beta_4 grad \cdot female \cdot educ + u$

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

- Then, we need to create new variable for $grad \cdot female \cdot educ - grad \cdot educ$

  ↓ Command in STATA: gen gfeminusge = gradfemeduc - gradeduc

- Run regression

  ↳ reg lwage educ femeduc gradeduc gradfemeduc

```
. reg lwage educ femeduc gradeduc gfeminusge

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(4, 521)       =     58.75
       Model |  46.1096544         4  11.5274136   Prob > F        =    0.0000
    Residual |  102.220096       521    .1961998   R-squared       =    0.3109
-------------+----------------------------------   Adj R-squared   =    0.3056
       Total |   148.32975       525  .282532857   Root MSE        =    .44294

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0714335   .0098735     7.23   0.000     .0520367    .0908304
     femeduc |  -.0302179   .0036203    -8.35   0.000      -.03733   -.0231057
    gradeduc |   .0204305   .0059991     3.41   0.001      .008645    .0322159
  gfeminusge |   .0142159   .0068539     2.07   0.039     .0007512    .0276805
       _cons |   .8710988   .1146053     7.60   0.000     .6459535    1.096244
------------------------------------------------------------------------------

. outreg2 using "HW1Q9B2.doc",
HW1Q9B2.doc
dir : seeout
```

VARIABLES	(1) lwage
educ β_1	0.0714***
	(0.00987)
femeduc β_2	-0.0302***
	(0.00362)
gradeduc $(\beta_3 + \beta_4)$	0.0204***
	(0.00600)
gfeminusge β_4	0.0142**
	(0.00685)
Constant	0.871***
	(0.115)
Observations	526
R-squared	0.311

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

∴ standard error of $\beta_3 + \beta_4$ is 0.0059991

$\overset{\text{female}}{\beta_1+\beta_2+\beta_3+\beta_4} \neq \overset{\text{male}}{\beta_1+\beta_3}$

$H_0: \beta_2+\beta_4=0$ $\quad$ $\beta_2+\beta_4 \neq 0$

IX

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?

ⓒ $H_0: \beta_2+\beta_4=0$

$H_1: \beta_2+\beta_4 \neq 0$

Transformation :

new model : $\ln(wage) = \beta_0 + \beta_1 educ + (\beta_2+\beta_4)\, female \cdot edu + \beta_3\, grad \cdot educ$

$+ \beta_4 \underbrace{(grad \cdot female \cdot educ - female \cdot edu)}_{\text{gen this new variable as "gfeminusfe"}} + u$

```
. reg lwage educ femeduc gradeduc gfeminusfe

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(4, 521)       =     58.75
       Model |  46.1096544         4  11.5274136   Prob > F        =    0.0000
    Residual |  102.220096       521    .1961998   R-squared       =    0.3109
-------------+----------------------------------   Adj R-squared   =    0.3056
       Total |   148.32975       525  .282532857   Root MSE        =    .44294

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0714335   .0098735     7.23   0.000     .0520367    .0908304
     femeduc |   -.016002   .0058195    -2.75   0.006    -.0274345   -.0045695
    gradeduc |   .0062146   .0049153     1.26   0.207    -.0034415    .0158708
  gfeminusfe |   .0142159   .0068539     2.07   0.039     .0007512    .0276805
       _cons |   .8710988   .1146053     7.60   0.000     .6459535    1.096244
------------------------------------------------------------------------------

. outreg2 using "HW1Q9C1.doc",
HW1Q9C1.doc
dir : seeout
```

VARIABLES		(1) lwage
educ	β_1	0.0714***
		(0.00987)
femeduc	$(\beta_2+\beta_4)$	-0.0160***
		(0.00582)
gradeduc	β_3	0.00621
		(0.00492)
gfeminusfe	β_4	0.0142**
		(0.00685)
Constant		0.871***
		(0.115)
Observations		526
R-squared		0.311

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

▶ Test hypothesis:

test (female·edu = 0)

```
. test (femeduc = 0)

 ( 1)  femeduc = 0

       F(  1,   521) =    7.56
            Prob > F =    0.0062
```

* For two tailed test, p-value is equal to 0.62%.

∴ At 10%, 5%, and 1% significant levels, we reject the null hypothesis $H_0: \beta_2+\beta_4=0$ because p-value of 0.62% is less than significant levels of 10%, 5%, and 1%. Thus, for college graduates, the impact of education on wage is different between male and female.

IX

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?

$H_1: \beta_2 + \beta_4 < 0$

$\llcorner\ \beta_1 + \beta_2 + \beta_3 + \beta_4 < \beta_1 + \beta_3$

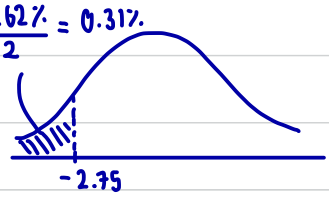
$\beta_2 + \beta_4 < 0$

ⓓ $H_0: \beta_2 + \beta_4 \geq 0$

$H_1: \beta_2 + \beta_4 < 0$

▸ $t\text{-stat} = \dfrac{\hat{\beta}_2 + \hat{\beta}_4 - 0}{se(\hat{\beta}_2 + \hat{\beta}_4)} = \dfrac{-0.0160 - 0}{0.00582} = -2.75$

$\dfrac{0.62\%}{2} = 0.31\%$



* For one-tailed test, p-value is equal to $\dfrac{0.62\%}{2} = 0.31\%$.

∴ At 10%, 5%, and 1% significant levels, we reject the null hypothesis $H_0: \beta_2 + \beta_4 \geq 0$ because p-value of 0.31% is less than significant levels of 10%, 5%, and 1%. Thus, for college graduates, there is enough evidence to conclude that the impact of education on wage for male is higher than that of female.