Financial Econometrics
Suggested Solution to Problem Set 1:

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**

   False: $E[U|X]$ is a constant, meaning that U is mean independent of X

   ii. **let $E[U|X]$= 3. $E[U]E[X] = E[UX]$**

   True: Since U is mean independent of X, we know that

   $$cov(U, X) = E[U]E[X] - E[UX] = 0$$

   iii. **let $E[U|X]$= 3. $Var[U - X] = Var[U] + Var[X]$**

   True: Since the covariance between U and X is zero, then

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

   iv. **let $E[U|X] = 0$ and $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U$. The model is homoscedastic.**

   Uncertain: scedasticity is about $Var[U|X]$ which isn't given herein

   v. **OLS estimator of the model $Y = \beta_0 + U$ always has $R^2 = 0$.**

   True: This is a special case in which there is only the constant term without any regressor. Thus, the OLS estimator is $\hat{\beta}_0 = \bar{Y}_n$, and thereby $\hat{Y}_i = \hat{\beta}_0 = \bar{Y}_n$. So, we have

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

   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.**

   Uncertain: The OLS estimator is consistent if the error term is not correlated with the regressors, regardless of whether the regressors themselves are correlated with each other. Without the information about $cov[U, X_1]$ and $cov[U, X_2]$, we cannot tell whether the OLS estimator is consistent.

   vii. **let $\widehat{\theta}_n$ be an unbiased estimator of θ. $plim(\widehat{\theta}_n) = \theta$**

   Uncertain: For $\hat{\theta}_n$ to be an unbiased estimator of θ, we have $E\left[\hat{\theta}_n\right] = \theta$. This unbiased condition provides no information regarding the probability limit of $\hat{\theta}_n$. Some unbiased estimators are also consistent, e.g. $\bar{X}_n$ is both an unbiased and consistent estimator for $E[X]$. In contrast, some unbiased estimators are not consistent, e.g. if we use X_n from an iid sample from the population of X as an estimator for $E[X]$, then it is unbiased but not consistent, as $plim(X_n)$ does not exists because the distribution of X_n is the same as the distribution of X rather than collapsing to a single point.

viii. Multicollinearity makes OLS estimator biased and inconsistent

False: with multicollinearity, we cannot even use the OLS to estimate the model. Multicollinearity has nothing to do with biasedness and consistency.

ix. $\bar{R}^2$ must have a value falling within $[0, 1]$

False: $\bar{R}^2$ can be negative.

x. Significance level is always less than 1

True: Significance level is the probability, below which we can reject the null hypothesis. Since probability cannot exceed one, and we never want to always reject the null. Then, significance level must be set below one.

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?

Yes. It is a function of a random variable X; so, the value of Z is random as well

b) Find $E[Z]$

$$\begin{aligned} E[Z] &= E[2^X - 1] \\ &= (2^{-2} - 1)P\{X = -2\} + (2^{-1} - 1)P\{X = -1\} + (2^1 - 1)P\{X = 1\} \\ &\quad + (2^2 - 1)P\{X = 2\} \\ &= (2^{-2} - 1) \cdot \frac{1}{6} + (2^{-1} - 1) \cdot \frac{1}{3} + (2^1 - 1) \cdot \frac{1}{3} + (2^2 - 1) \cdot \frac{1}{6} \\ &= \frac{13}{24} \end{aligned}$$

c) Find $E[Z^2]$

$$\begin{aligned} E[Z^2] &= E[(2^X - 1)^2] \\ &= \left(\frac{1}{4} - 1\right)^2 \cdot \frac{1}{6} + \left(\frac{1}{2} - 1\right)^2 \cdot \frac{1}{3} + (2 - 1)^2 \cdot \frac{1}{3} + (4 - 1)^2 \cdot \frac{1}{6} = \frac{193}{96} \end{aligned}$$

d) What is the variance of Z?

$$Var[Z] = E[Z^2] - (E[Z])^2 = \frac{193}{96} - \left(\frac{13}{24}\right)^2$$

e) What is $Cov[X, Y]$?

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] = E[X(X^2)], \text{ because } E[X] = 0 \\ &= (-2)^3 \cdot \tfrac{1}{6} + (-1)^3 \cdot \tfrac{1}{3} + (1)^3 \cdot \tfrac{1}{3} + (2)^3 \cdot \tfrac{1}{6} = 0 \end{aligned}$$

Note that we can actually conclude that $E[X^3] = 0$ without plugging in the value and computing out, because X has a symmetric distribution which implies that its skewness is zero; thus, $E[X^3] = 0$ when $E[X] = 0$.

f) Is Y mean independent of X?
No. Since $Y = X^2$, therefore $E[Y|X] = E[X^2|X] = X^2$, which is obviously not constant, as X^2 can either be 4 or 1

g) Is Y independent of X?
No, because Y is not mean 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?
Yes, it is a function of the random variable $GMAT$; so, its outcome value is uncertain.

b) What is the expected value of GPA when $GMAT$ score is 650? What is the expected GPA when $GMAT$ score is 790?

$$E[GPA|GMAT = 650] = 0.007(650) - 1.73 = 2.82$$
$$E[GPA|GMAT = 790] = 0.007(790) - 1.73 = 3.80$$

c) If $E[GMAT] = 700$, what is $E[GPA]$?
By Law of Iterated Expectation,

$$\begin{aligned} E[GPA] &= E(E[GPA|GMAT]) \\ &= E(0.007GMAT - 1.73) = 0.007E[GMAT] - 1.73 \\ &= 0.007(700) - 1.73 \quad = 3.17 \end{aligned}$$

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 $Var[Y]$?

X is annual salary in the unit of 1000 Baht, and we are given $E[X]$ = 27.6 and (X) = 11.2. Then, we let Y be monthly salary in 1 Baht unit, so that $Y = \frac{1000X}{12}$.W e are asked to find $[Y]$ and $\sigma^2(Y)$.

- First we find $[Y]$ (using the linearity of expectations)

$$E[Y] = E\left[\frac{1000X}{12}\right] = \frac{1000}{12}E[X] = \frac{1000}{12} * 27.6 = 2{,}300$$

- Next we find (Y)

$$Var[Y] = Var\left[\frac{1000X}{12}\right] = \left(\frac{1000}{12}\right)^2 Var[X] = \left(\frac{1000}{12}\right)^2 * (11.2)^2 = \left(\frac{11{,}200}{12}\right)^2$$

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

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

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

a) Is $\widehat{\theta}_n$ an unbiased estimator of $E[X]E[Y]$? Explain.

To prove unbiasedness, we need $E\left[\hat{\theta}_n\right] = E[X]E[Y]$

$$\begin{aligned} E\left[\hat{\theta}_n\right] &= E\left[\frac{1}{2n}(X_1 + X_n)\left(\sum_{j=1}^{n} Y_j\right)\right] \\ &= \frac{1}{2n}\left\{E\left[X_1\left(\sum_{j=1}^{n} Y_j\right)\right] + E\left[X_n\left(\sum_{j=1}^{n} Y_j\right)\right]\right\} \\ &= \frac{1}{2n}\{E[X_1Y_1] + \cdots + E[X_1]E[Y_n] + E[X_nY_1] + \cdots + E[X_n]E[Y_n]\} \end{aligned}$$

Note that

$$E\left[X_iY_j\right] = E[X_i]E\left[Y_j\right] + cov\left(X_i, Y_j\right) = E[X]E[Y] + cov(X, Y)$$

Since $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent, we know that X_i and Y_j are independent for $i \neq j$. Therefore, for $i \neq j, E\left[X_iY_j\right] = E[X_i]E\left[Y_j\right] = E[X]E[Y]$. However, we don't know if $cov(X_1, Y_1)$ and $cov(X_n, Y_n) = 0$. So,

$$\begin{aligned} E\left[\hat{\theta}_n\right] &= \tfrac{1}{2n}\left\{E[X_1Y_1] + \underbrace{\ldots + E[X_1]E[Y_n] + E[X_n]E[Y_1] + \cdots}_{2n-2\ terms\ of\ E[X]E[Y]} + E[X_nY_n]\right\} \\ &= \tfrac{1}{2n}\{(2n-2)E[X]E[Y] + 2[E[X]E[Y] + cov(X, Y)]\} \\ &= \frac{1}{2n}\{2nE[X]E[Y] + 2cov(X, Y)\} \\ &= E[X]E[Y] + \frac{cov(X, Y)}{n} \end{aligned}$$

Therefore, $\hat{\theta}_n$ is unbiased only if X and Y are uncorrelated.

b) Is $\widehat{\theta}_n$ a consistent estimator of $E[X]E[Y]$? Explain.

To prove consistency, we need $plim\left(\hat{\theta}_n\right) = E[X]E[Y]$

$$\begin{aligned} plim\left[\hat{\theta}_n\right] &= plim\left[\frac{1}{2n}(X_1 + X_n)\left(\sum_{j=1}^{n} Y_j\right)\right] \\ &= plim\left(\frac{1}{2}(X_1 + X_n)\right) plim\left[\frac{1}{n}\left(\sum_{j=1}^{n} Y_j\right)\right] \\ &= \frac{1}{2}\left(plim(X_1 + X_n)\right) \cdot E[Y] \end{aligned}$$

where the last equality is implied by WLLN. However, $plim(X_1 + X_n)$ does not exist, because it is not converging to any single point even when n is large. To see this, note that $Var(X_1 + X_n) = 2Var(X)$, instead of approaching to zero even when n is large. So, $\hat{\theta}_n$ is not a consistent estimator of $E[X]E[Y]$.

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

Yes, by applying the WLLN:

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

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:

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

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?

The effect depends on distance and is given by $\frac{\partial R}{\partial dist} = \alpha_1 + 2\alpha_2 dist$

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 $\boldsymbol{\alpha_0, \alpha_1, \alpha_2}$?

$$\alpha_1 + 2\alpha_2 dist = 0 \;\rightarrow\; dist = -\frac{\alpha_1}{2\alpha_2}$$

Therefore, the threshold is $-\frac{\alpha_1}{2\alpha_2}$

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

Note that if $plim(\hat{\alpha}_1) = \alpha_1; plim(\hat{\alpha}_2) = \alpha_2$, we will then have

$$plim\left(\frac{-\hat{\alpha}_1}{2\hat{\alpha}_2}\right) = -\frac{\alpha_1}{2\alpha_2}$$

So, the question is whether $plim\left(\hat{\alpha}_1^{OLS}\right) = \alpha_1; plim\left(\hat{\alpha}_2^{OLS}\right) = \alpha_2$, or whether the OLS estimator is consistent in this case. In other words, do we have that

$$Cov(dist, U) = Cov(dist^2, U) = 0?$$

Because we interpret the linear regression model as causal relationship in this case, there is no econometric/statistic theory to justify the zero covariance conditions. So, we need to think if there is any factor that can contribute to rate of returns on investment of these households and is correlated with the distance. If so, and since it's included in the error term U, then the OLS will be inconsistent. For this particular example, type of business enterprise (e.g. agricultural/livestock businesses vs. retailers) or size of an enterprise may explain rate of returns (larger size gains from economy of scale; so, may be more profitable). And these variables may be correlated with the distance because agricultural/livestock business and relatively larger size enterprises are likely to require more space and, thus, located further away from city-center area where bank branches are usually present.

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.

Yes, because under this interpretation, we always get the condition that the error term is uncorrelated with each of the regressors. Therefore, the OLS is consistent, and we have

$$plim\left(\frac{-\hat{\alpha}_1^{OLS}}{2\hat{\alpha}_2^{OLS}}\right) = -\frac{plim\left(\hat{\alpha}_1^{OLS}\right)}{2plim(\hat{\alpha}_2^{OLS})} = -\frac{\alpha_1}{2\alpha_2}$$

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:

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

According to the question, we should make the coefficient of X_1 to be what we want to test, which is $2\beta_1 - \beta_2$ in this case. However, it's more convenient to change the hypothesis to

$$H_0: \beta_1 - \frac{\beta_2}{2} = 0 \ \ vs \ H_1: \beta_1 - \frac{\beta_2}{2} \neq 0,$$

so that we can just add the $-\frac{\beta_2}{2}$ to the coefficient of X_1:

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

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

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

Re-write the hypothesis as $H_0: \beta_1 - 3\beta_2 = 0 \ vs \ H_1: \beta_1 - 3\beta_2 \neq 0$

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

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

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

For this problem, we can transform the model to just have a coefficient equal to $\beta_1 + \beta_2$, so that we can get the standard error for $\beta_1 + \beta_2$ and use it to calculate the t-stats:

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

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

and calculate t-stat as $\frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{se(\hat{\beta}_1 + \hat{\beta}_2)}$

Or we can set the coefficient of X_1 as $\beta_1 + \beta_2 - 1$:

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

Note that we cannot have the last term $(+ X_1)$, becuase we cannot have the same variables be a regressor twice. (this is like a multicollinearity problem). So, a method is to move the last term X_1 to the left-hand side of the regression:

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

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:

$$\boldsymbol{lnY = \beta_0 + \beta_1 X_1 + U}$$

where $\boldsymbol{Y}$ **= Household debt in unit of thousand Baht**
$\boldsymbol{X_1}$ **= Household monthly income in unit of Baht**

a) Suppose that you run the OLS regression and get $\boldsymbol{\widehat{\beta}_1 = -0.03}$, $\boldsymbol{SE(\widehat{\beta}_1) = 0.01}$, **and** $\boldsymbol{R^2 = 0.287}$. **If you change to use household annual income in unit of thousand Baht, what are the values of your new** $\boldsymbol{\widehat{\beta}_1, SE(\widehat{\beta}_1)}$ **and** $\boldsymbol{R^2}$**?**

Assume you have a monthly salary of 2,000 Baht. Then, $X_1 = 2{,}000$. 2,000 a month is 24,000 Baht a year; therefore, the new regressor $\tilde{X}_1$, which measure yearly income in unit of thousand Baht will take value 24. Hence,

$$\tilde{X}_1 = 24 = \left(\frac{12}{1000}\right)2000 = \left(\frac{12}{1000}\right)X_1$$

New $\hat{\beta}_1$:

$$\ln Y = \beta_0 + \frac{\beta_1}{\left(\frac{12}{1000}\right)} \underbrace{\left(\frac{12}{1000}\right)X_1}_{\substack{\text{the new regressor} \\ \text{which is } \tilde{X}_1}} + U$$

Then, the new $\hat{\beta}_1$ is

$$\frac{1000\hat{\beta}_1}{12} = \frac{-0.03 * 1000}{12} = -\frac{30}{12} = -2.5$$

Standard error of the new $\hat{\beta}_1$:

$$se\left(\frac{1000\hat{\beta}_1}{12}\right) = \left(\frac{1000}{12}\right)se(\hat{\beta}_1) = \frac{10}{12}$$

R^2 is not affected by re-scaling regressand or regressors.

b) If you change to measure $\boldsymbol{Y}$ **in unit of Baht instead, how would this affect your OLS estimates** $\boldsymbol{\widehat{\beta}_0, \widehat{\beta}_1}$**?**

Assume you have 1,000 Baht in debt, then $Y = 1$ thousand Baht and the new regressand $\tilde{Y}$ will take value 1,000. Therefore, $\tilde{Y} = 1000Y$

Transform model:

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

Therefore, the new intercept is $\ln 1000 + \beta_0$, and the coefficient doesn't change. Consequently, the new $\hat{\beta}_0$ is $\hat{\beta}_0 + \ln 1000$, and $\hat{\beta}_1$ is unchanged.

From now on, suppose that you adjust your model to

$$\boldsymbol{ln\, Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + U}$$

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?

If the household's main source of income is from agriculture, then $X_2 = 1; X_3 = 0$:

$$\ln Y = (\beta_0 + \beta_2) + \beta_1 X_1$$

If the household's main source of income is from working for wage, then $X_2 = 0; X_3 = 1$

$$\ln Y = (\beta_0 + \beta_3) + \beta_1 X_1$$

So, for the agricultural households and wage-earner households with the same level of income,

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

Note that a difference in natural log is approximately $d(lnY) = \frac{dY}{Y}$, which is rate of change of Y, or if multiplied by 100 represents % change in Y. So, roughly speaking, debt of agricultural households is about $100(\beta_2 - \beta_3)$ percent higher than that of the wage earners if they have the same level of income.

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 $cov(\widehat{\beta}_3, \widehat{\beta}_2)$, 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 $Se(\widehat{\beta}_2 - \widehat{\beta}_3)$ automatically?

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

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

So, $\tilde{Y} = \ln Y$; $\tilde{X}_1 = X_1; \tilde{X}_3 = X_2 + X_3; \gamma_0 = \beta_0; \gamma_1 = \beta_1; \gamma_3 = \beta_3$

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.

```
. gen lwage = ln(wage)
. gen grad = educ>=16
. gen fedu = educ*female
. gen gedu = grad*educ
```

```
. gen fgedu = grad*female*edu

. reg lwage educ fedu gedu fgedu
. outreg2 using "ps2.xls", replace label adjr2 ctitle(ln(wage))
```

VARIABLES	(1) ln(wage)
educ	0.0714***
	(0.00987)
educ*female	-0.0302***
	(0.00362)
educ*grad	0.00621
	(0.00492)
educ*female*grad	0.0142**
	(0.00685)
Constant	0.871***
	(0.115)
Observations	526
Adjusted R-squared	0.306

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

b) Report the OLS Estimate of $\beta_3 + \beta_4$ along with its standard errors.
consider the effect of education on wage is

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

	$female = 0$	$female = 1$
$grad = 0$	Male without college degree: β_1	Female without college degree: $\beta_1 + \beta_2$
$grad = 1$	Male college graduate: $\beta_1 + \beta_3$	Female college graduate: $\beta_1 + \beta_2 + \beta_3 + \beta_4$

To interpret $\beta_3 + \beta_4$, consider the effect or return of an additional year of education on wage of female is

$$\beta_1 + \beta_2 + (\beta_3 + \beta_4) grad$$

Hence, $\beta_3 + \beta_4$ is the additional return of education on wage that female college graduates receive compared with female non-college graduates.

To get the OLS estimates, we can use STATA as follows:

```
. reg lwage educ fedu gedu fgedu
```

Source	SS	df	MS	Number of obs	=	526
Model	46.1096544	4	11.5274136	F(4, 521)	=	58.75
Residual	102.220096	521	.1961998	Prob > F	=	0.0000
				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
fedu	-.0302179	.0036203	-8.35	0.000	-.03733	-.0231057
gedu	.0062146	.0049153	1.26	0.207	-.0034415	.0158708
fgedu	.0142159	.0068539	2.07	0.039	.0007512	.0276805
_cons	.8710988	.1146053	7.60	0.000	.6459535	1.096244

According to the table above, $\hat{\beta}_3 + \hat{\beta}_4 = 0.0062146 + 0.0142159 = 0.0204305$. Now, to find standard error of this linear combination of the coefficient, we can either transform the model or manually calculate it:

<u>Manual Calculation:</u>

```
. vce
```

```
Covariance matrix of coefficients of regress model
```

e(V)	educ	fedu	gedu	fgedu	_cons
educ	.00009749				
fedu	-7.229e-06	.00001311			
gedu	-.00003163	6.952e-06	.00002416		
fgedu	7.539e-06	-.00001311	-.00001757	.00004698	
_cons	-.00109109	4.590e-06	.00029832	-8.312e-06	.01313438

$$SE(\hat{\beta}_3 + \hat{\beta}_4) = \sqrt{[SE(\hat{\beta}_3)]^2 + [SE(\hat{\beta}_4)]^2 + 2cov(\hat{\beta}_3, \hat{\beta}_4)}$$
$$= \sqrt{(.0049153)^2 + (.0068539)^2 + 2(-.00001757)} = 0.005999$$

<u>Transformation:</u> the newly transform model is

$$\ln(wage) = \beta_0 + \beta_1 educ + \boldsymbol{\beta_2 female \cdot educ} + (\beta_3 + \beta_4) grad \cdot educ$$
$$+\beta_4 \underbrace{(grad \cdot female \cdot educ - grad \cdot educ)}_{\text{let define this as a new variable called "ww"}} + U$$

The new regression result is in the table below. We can see that the standard error (highlighted) is the same as what we calculated manually.

```
. gen ww = fgedu - gedu
. reg lwage educ fedu gedu ww, nohead
```

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

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?

For college graduates, the effect of education on wage is

$$(\beta_1 + \beta_3) + (\beta_2 + \beta_4) female$$

So, the difference of the impact between male and female is $\beta_2 + \beta_4$. Then, the hypothesis testing that we need to do is

$$H_0: \beta_2 + \beta_4 = 0 \quad vs \quad H_1: \beta_2 + \beta_4 \neq 0$$

Here, you can either 1. transform the regression model, 2. conduct the t-test by using the post-estimation command "vce" to find the covariance between $\hat{\beta}_2$ and $\hat{\beta}_4$ and calculate $SE(\hat{\beta}_2 + \hat{\beta}_4)$, or 3. use STATA postestimation command "test".

For the first method, the new transform model is

$$\ln(wage) = \beta_0 + \beta_1 educ + (\boldsymbol{\beta_2 + \beta_4}) \boldsymbol{female \cdot educ} + \beta_3 grad \cdot educ$$
$$+\beta_4 \underbrace{(grad \cdot female \cdot educ - female \cdot educ)}_{let\ define\ this\ as\ a\ new\ variable\ called\ "xx"} + U$$

The new regression result is

```
. gen xx = fgedu - fedu
. reg lwage educ fedu gedu xx, nohead
```

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

So, based on the homoscedastic standard error, we can reject the null hypothesis at as low as 1% significance levels because the p-value is 0.6%. However, based on the heteroscedasticity-robust standard error, we cannot reject the null hypothesis at 1% but at 5% significance level because the p-value is 1.1%. Notice that the 95% confidence intervals don't cover zero, implying that we can reject the null at 5% significance level.

For the third method, we have the result:

```
. quietly reg lwage educ fedu gedu fgedu
. test fedu+fgedu=0

  (1)  fedu + fgedu = 0

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

Notice that the test-statistic for the command "test" is distributed $F_{1,521}$ and is equal to the square of the t-statistic:

$$F_{stat} = 7.56 = (-2.749)^2 = (t_{stat})^2$$

because, if $t \sim t_{n-k-1}$, then $t^2 \sim F_{1,n-k-1}$. Also notice that the p-values from the two methods are exactly equal; hence, the conclusions of the hypothesis tests are the

same. Note that if you don't do the first and second method above, but just starting with this STATA test result, you may conclude that the t-stat is -2.749, not 2.749, because based on the non-transformed regression result you can see that $\hat{\beta}_2 + \hat{\beta}_4 =$ -.0302179 + .0142159 < 0

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?

$$H_0: \beta_2 + \beta_4 = 0 \quad vs \quad H_1: \beta_2 + \beta_4 < 0$$

Now, we know from b). that the t-stat is -2.75 and, for two-sided test, the p-value is 0.006. So, the p-value for this question is just $0.006/2 = 0.003$. Hence, we can reject the null at all the given levels 10%, 5%, and 1%. So, there is statistical evidence to conclude that, for graduated people, the impact is higher for male