ECO3121 Problem Set 2

Xue Zhongkai (122090636)

October 26, 2023

Question 1.

---

1. Yes. There are many other factors influencing the agricultural productivity, which could be highly correlated with *rental_in_share* and *rental_out_share*, thus causing the omitted variable bias of $\hat{\beta}_1$ and $\hat{\beta}_2$.

2. We have the bias to be
$$\hat{\beta}_1 - \beta_1 = \rho X u \frac{\sigma_u}{\sigma_X}$$
where $\rho X u$ indicates the correlation between the omitted variable and the variable X.

   Take the example of "the education level of the farmer". If the farmer with more education has a higher chance of renting in the lands, and education has a positive effect on *yield*, then we have $\rho X u$ to be positive, hence we may under-estimate β_1, i.e. $\hat{\beta}_1 < \beta_1$.

3. By adding other significant variables could the omitted variable bias be solved.

   According to the data description, *f9*, *f25*, *f29* represent quantity of fertilizer, diesel and agrochemicals respectively, which are key factors of agriculture production.

   Recall the definition of *yield* as the dependent variable:

```
gen yield=d32/d31
```

   and we had dependent variables *rental_in_share* and *rental_out_share*:

```
gen rental_in_share = c10/d31*100
gen rental_out_share = c13/d31*100
```

   There seems to be some missing value "." in the data for *f9*, *f25*, *f29*, and we replace them with the mean value:

```
egen mean_f9 = mean(f9)
replace f9 = mean_f9 if f9 == .
drop mean_f9

egen mean_f25 = mean(f25)
replace f25 = mean_f25 if f25 == .
drop mean_f25

egen mean_f29 = mean(f29)
replace f29 = mean_f29 if f29 == .
drop mean_f29
```

   and we add them into the regression

```
reg yield rental_in_share rental_out_share f9 f25 f29
```

   with the corresponding results:

Source	SS	df	MS			
Model	8007623.89	5	1601524.78	Number of obs	=	14,171
Residual	2.2051e+09	14,165	155671.452	F(5, 14165)	=	10.29
				Prob > F	=	0.0000
				R-squared	=	0.0036
				Adj R-squared	=	0.0033
Total	2.2131e+09	14,170	156181.633	Root MSE	=	394.55

yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rental_in_share	-.189134	.0859754	-2.20	0.028	-.357657	-.020611
rental_out_share	.2442282	.0479468	5.09	0.000	.1502462	.3382103
f9	.0153953	.0030828	4.99	0.000	.0093526	.021438
f25	-.0637091	.0284143	-2.24	0.025	-.1194049	-.0080132
f29	.0077934	.0157555	0.49	0.621	-.0230895	.0386764
_cons	427.2161	4.353218	98.14	0.000	418.6833	435.749

4. The unit for *rental_in_share* is percentage points, and we find β_1 to be -0.1891. Hence, a 10-percentage-points increase of rent-in land will result in 1.891 decrease in yield.

5. Assume $H_0 : \beta_3 = 0$ s.t. education has nothing to do with the yield.

   As a two-sided test, first we perform a regression between the yield and *huzhu*'s education:

```
reg yield huzhu_edu
```

   we have

Source	SS	df	MS			
Model	67063.3741	1	67063.3741	Number of obs	=	13,629
Residual	2.1805e+09	13,627	160013.614	F(1, 13627)	=	0.42
				Prob > F	=	0.5174
				R-squared	=	0.0000
				Adj R-squared	=	-0.0000
Total	2.1806e+09	13,628	160006.793	Root MSE	=	400.02

yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
huzhu_edu	.8810332	1.360906	0.65	0.517	-1.78653	3.548596
_cons	428.9546	9.819348	43.68	0.000	409.7074	448.2019

   From the diagram we know

$$\text{p-Value} = \text{P(P>|t|)} = 0.517$$

   which is much larger than the significance-level

$$\text{p-Value} > \alpha = 0.05$$

   thus we fail to reject H_0.

   As a result, there is no significance evidence that a higher education level is associated with a higher agriculture productivity.

6. Assume $H_0 : \beta_1 = \beta_2$ s.t. the renting in and renting out have the same effects. This statement is equivalent to $H_0' : \theta = \beta_1 - \beta_2 = 0$, thus we could re-write the model as

$$yield_i = \beta_0 + \theta \cdot rental\_in\_share_i + \beta_2(rental\_in\_share_i + rental\_out\_share_i) + u_i$$

As a two-sided test, first we perform a regression between the yield and the difference:

```
gen rental_total = rental_in_share  +  rental_out_share
reg yield rental_in_share rental_total
```

we have

Source	SS	df	MS
Model	3856365.43	2	1928182.71
Residual	2.2092e+09	14,168	155931.491
Total	2.2131e+09	14,170	156181.633

Number of obs	=	14,171
F(2, 14168)	=	12.37
Prob > F	=	0.0000
R-squared	=	0.0017
Adj R-squared	=	0.0016
Root MSE	=	394.88

yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rental_in_share	-.1815378	.0800299	-2.27	0.023	-.3384069	-.0246687
rental_total	.2349335	.0479369	4.90	0.000	.140971	.328896
_cons	431.3688	3.35498	128.58	0.000	424.7926	437.945

From the diagram we know

$$\text{p-Value} = \text{P(P>|t|)} = 0.023$$

which is smaller than the significance-level

$$\text{p-Value} < \alpha = 0.1$$

thus we could reject H_0.

As a result, there is significance evidence that the renting-in and renting-out have quite different effects.

7. The R^2 in question 3 is 0.0036, while the previous one is 0.0017. Thus R^2 in question 3 higher than that of the original regression.

   R^2 is not good enough as a guarantee to include the variables, for the following reasons:

   - There could be other confusing omitted variables;
   - The model could be over-fitting and less effective with new data;
   - The model is too complex to analyze and interpret.

Appendix: Here is the **.do File** for Problem 1.

```
use "/Users/kevinshuey/Github/Assignments/cuhksz_ECO3121/as2/aghousehold.dta"
gen yield=d32/d31
gen rental_in_share = c10/d31*100
gen rental_out_share = c13/d31*100

egen mean_f9 = mean(f9)
replace f9 = mean_f9 if f9 == .
drop mean_f9

egen mean_f25 = mean(f25)
replace f25 = mean_f25 if f25 == .
drop mean_f25

egen mean_f29 = mean(f29)
```

```
replace f29 = mean_f29 if f29 == .
drop mean_f29

reg yield rental_in_share rental_out_share f9 f25 f29

reg yield huzhu_edu

gen rental_total = rental_in_share + rental_out_share
reg yield rental_in_share rental_total
```

Question 2.

1. I don't think that $E(u_i|X_i) = 0$, as the u_i could be correlated with X_i, for the reason that an important independent variable W_i contains within u_i. Under this circumstance, u_i and X_i are not independent with each other, causing β_1 to be biased.

2. **(a)** Since X_i is randomly and equally assigned, we could still believe that $E(u_i|X_i)$ does not depend on X_i, thus β_1 is unbiased.

   (b) From the sample, 50% of the coastal regions are assigned treated group, while only 20% of inland ones. Under this circumstance, there could be some unobserved independent variables specific in some regions, hence $E(u_i|X_i)$ may depend on W_i.

Question 3.

1. We have the least squares function to be

$$LS(\beta_1, \beta_2) = \sum_{i=1}^{n}(Y_i - \beta_1 X_{1i} - \beta_2 X_{2i})^2$$

2. We have the partial derivatives as

$$\frac{\partial LS}{\partial \beta_1} = -2\sum_{i=1}^{n} X_{1i}(Y_i - \beta_1 X_{1i} - \beta_2 X_{2i})$$

$$\frac{\partial LS}{\partial \beta_2} = -2\sum_{i=1}^{n} X_{2i}(Y_i - \beta_1 X_{1i} - \beta_2 X_{2i})$$

3. Set the partial derivative equal to zero, we have

$$\sum_{i=1}^{n} X_{1i}Y_i = \beta_1 \sum_{i=1}^{n} X_{1i}^2 + \beta_2 \sum_{i=1}^{n} X_{1i}X_{2i}$$

   Since $\sum_{i=1}^{n} X_{1i}X_{2i} = 0$, we could see that

$$\hat{\beta}_1 = -\frac{\sum_{i=1}^{n} X_{1i}Y_i}{\sum_{i=1}^{n} X_{1i}^2}$$

4. Since $\sum_{i=1}^{n} X_{1i}X_{2i} \neq 0$, first we get

$$\hat{\beta}_2 = -\frac{\sum_{i=1}^{n} X_{2i}Y_i - \beta_1 \sum_{i=1}^{n} X_{1i}X_{2i}}{\sum_{i=1}^{n} X_{2i}^2}$$

Taking into another equation, we have the expression to be

$$\sum_{i=1}^{n} X_{1i}Y_i = \beta_1 \sum_{i=1}^{n} X_{1i}^2 - \frac{\sum_{i=1}^{n} X_{2i}Y_i - \beta_1 \sum_{i=1}^{n} X_{1i}X_{2i}}{\sum_{i=1}^{n} X_{2i}^2} \sum_{i=1}^{n} X_{1i}X_{2i}$$

5. By adding all equations and divide it by n, we have

$$\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n} \cdot n \cdot \hat{\beta}_0 - \frac{1}{n}\hat{\beta}_1 \sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} = 0$$

which is equivalent to

$$\bar{Y} = \hat{\beta}_0 0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 + \bar{u}$$

In OLS regression, we have the expected residual

$$E(u_i) = 0$$

thus we get

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

6. First we construct the error from the mean for each term

$$\begin{aligned} Y_i - \bar{Y} &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} - \bar{Y} + u_i \\ &= \beta_0 - \bar{Y} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\ &= \beta_0 - \bar{Y} + \beta_1 (X_{1i} - \bar{X}_1 + \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2 + \bar{X}_2) + u_i \\ &= (\beta_0 - \bar{Y} + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2) + \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + u_i \end{aligned}$$

As

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

we have

$$Y_i - \bar{Y} = \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + u_i$$

Then we get SSR as

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} [\beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + u_i]^2$$

To derive an appropriate $\hat{\beta}_1$, we need to partial out the β_1.

As the expectation of the residual is 0, all intersection terms concerning u_i will not take effect.

As

$$\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 0$$

all the intersection terms between X_{1i} and X_{2i} will not take effect as well.

As a result, the effect-equivalent equation for $\hat{\beta}_1$ could just be simplified as

$$\sum_{i=1}^{n}(Y_i - \bar{Y}) = \beta_1 \sum_{i=1}^{n}(X_{1i} - \bar{X}_1)$$

Multiplying $(X_{1i} - \bar{X}_1)$ on both sides,

$$\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(Y_i - \bar{Y}) = \beta_1 \sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2$$

Finally we derive

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$

Compared to the OLS estimator of β_1 from the regression that omits X_2, it seems identical between the two, under the implication that the two variable X_1, X_2 are not correlated (which is what $\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 0$ implies).

******************** **This is the end of Problem Set 2.** ********************