Problem Set 2 Solutions
ECO3121 - Fall 2023

October 30, 2023

Due 12PM, 29/10/2023
Please remember to submit your Stata code and requested output as it will be graded.

Question 1

In this problem set, we'll continue to examine the effects of land rental activity in rural households on output and aggregate productivity. Please continue using the main dataset `"aghousehold.dta"` from your first assignment from the blackboard site. It is the National Fixed Point Survey (NPFS) in the year of 2010.

We continue with our inquiry on the causal effect of land rental behavior on agricultural productivity. The researcher plans to follow the specifications in Question 1-7 in Assignment 1 and regress the yield on two measures of land rentals.

$$yield_i = \beta_0 + \beta_1 rental\_in\_share_i + \beta_2 rental\_out\_share_i + \mu_i$$

1. Do you think this regression suffers from omitted variable bias? Explain why.

   Answers: Yes.

   [Open question] Explanation: There are other important determinants of yield which may also be correlated with land rental activities, including land quality, agricultural technologies, farmers' household characteristics, etc.

2. Using the expression for omitted variable bias,

   $$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X}$$

   assess whether the regression will likely over- or underestimate the effect of land rentals on agricultural yield. That is, do you think that $\hat{\beta}_1 > \beta_1$ or $\hat{\beta}_1 < \beta_1$? Illustrate with one example of omitted variable.

   [Open question] Answers: Suppose that the yield is positively affected by the agricultural technologies that used upon the land, and larger land is applicable with more advanced agriculture technologies. In this case, the rental-in land activity is likely to be positively correlated with agricultural technologies leading to a positive value for the omitted variable bias so that $\hat{\beta}_1 > \beta_1$.

Depend Var	Original yield	(1-3) yield	(1-5) yield	(1-6) yield
share100_rent_in	0.053	0.169	0.052	-0.182***
	(0.064)	(0.105)	(0.065)	(0.080)
share100_rent_out	0.235***	-0.236	0.232***	
	(0.048)	(0.302)	(0.049)	
education_lvl			0.940	
			(1.360)	
rental_summation				0.235***
				(0.0480)
Observations	14,171	14,171	14,171	14,171
R-squared	0.0017	0.0410	0.0017	0.0017

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

Table 1: All regression results in Question 1.

3. How would you solve address the concerns for omitted variable bias? Which variables would you add to the regression? Rerun the regression using your desired specification. Be sure to include all variables that you may find important. (*This is an open-ended question*)

   Answers: By including more covariates that might be correlated with land rental activity, here we choose ***f9, f10, f25, f26, f28, f29, f30, g1, g82, g114, g149*** *as additional variables. (We choose these variables by correlation matrix, and typically correlation coefficients under 0.1 can be regarded as negligible.)*

   See ***Tab.1 (1-3)*** *column for regression details, coefficients and standard errors for additional variables are omitted.*

4. Using your regression results in the previous question, what is the change in yield if rent-in land increases by 10 percentage points?

   Answers: Yield would increase by 1.69 kilograms.

5. Economist A argues that a higher education level is associated with a higher agriculture productivity. Test his hypothesis at the 5% significance level. Be clear about your testing procedure.

   Answers: Testing against one-sided alternatives (Not Equals Zero)

   - *Choose a null hypothesis and the alternative hypothesis:*

     $$H_0 : \beta_{educ} = 0, \quad H_1 : \beta_{educ} > 0$$

   - *Choose a* ***significance level*** *for the test, here we choose* 5%.

- *Run following regression:*

$$yield_i = \beta_0 + \beta_1 rental\_in\_share_i + \beta_2 rental\_out\_share_i + \beta_3 education\_lvl_i + \mu_i$$

see ***Tab.1 (1-5)*** *column for regression details.*

- *Calculate t-statistic:* $t_{\hat{\beta}_{educ}} = \frac{\hat{\beta}_{educ}}{se(\hat{\beta}_{educ})} = \frac{0.940}{1.360} = 0.691$*[The value for t-stat could be different by using different regression results.]*
- *Choose the corresponding critical value,* $c_{0.05} = 1.645$*, and the* ***rejection rule*** *is*

$$|t_{\hat{\beta}_{educ}}| > c_{0.05}$$

- *Since our* $|t_{\hat{\beta}_{educ}}| < c_{0.05}$*, we can't reject the null hypothesis. Under* 5% *significance level, a higher education level is* ***NOT*** *statistically associated with a higher agriculture productivity.*

6. Economist B argues that renting in and renting out have the same effects on agriculture productivity. Test her hypothesis at the 10% significance level. Be clear about your testing procedure.

   Answers: Testing against two-sided alternatives (Coefficient Equality)

   - *Choose a null hypothesis and the alternative hypothesis:*

$$H_0 : \beta_{rent-in} = \beta_{rent-out}, \quad H_1 : \beta_{rent-in} \neq \beta_{rent-out}$$

   - *Choose a* ***significance level*** *for the test, here we choose* 10%*.*
   - *Run following regression:*

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

   where $\theta = \beta_{rent-out} - \beta_{rent-in}$*. Testing* $\beta_{rent-in} = \beta_{rent-out}$ *equals to test* $\theta = 0$*, see* ***Tab.1 (1-6)*** *column for regression details.*

   - *Calculate t-statistic* $t_{\hat{\theta}} = \frac{\hat{\theta}}{se(\hat{\theta})} = \frac{-0.182}{0.08} = -2.275$
   - *Choose the corresponding critical value,* $c_{0.1} = 1.645$*, and that the* ***rejection rule*** *is*

$$|t_{\hat{\theta}}| > c_{0.1}$$

   - *Since our* $|t_{\hat{\theta}}| > c_{0.1}$*, we can reject the null hypothesis, therefore under* 10% *significance level, renting in and renting out have statistically different effects on agriculture productivity.*

7. Is R^2 in question 3 higher or lower than R^2 of the original regression? Explain. Is R^2 a good enough measure to tell us whether we need to include these additional variables in our regression? Why or why not?

   Answers: Higher, because more independent variables in the regression in Q3.

   R^2 *is not a good enough measure. Because* R^2 *increases whenever we add a regressor to the model, regardless of whether or not the coefficient of that regressor equals zero. And in the extreme case known as "kitchen sink" regression, we can even get a fairly high* R^2 *but may massively overfit the data.*

Question 2

In this question, we continue our inquiry on the treatment effect of the land rental contract law on agricultural yields. Suppose the Ministry of Agriculture and Rural Affairs has hired you as part of their impact evaluation team. Your first assignment is to evaluate a Randomized Control Trial (randomized experiment) that they have implemented before you arrived. Two years before you arrive, they have implement a land rental contract law to some trial villages, which they selected from a list of villages that all had expressed their interest in participating in the project.

The board asks you to estimate the Average Treatment Effect of their intervention (land rental contract law) on **village-level** yields. Suppose the selection of the treated villages is based on the following procedure: 50% of the villages in the coastal region were randomly assigned to the treated group to implement the land rental contract law and 50% do not implement. For villages in the inland region, 20% are randomly assigned to the treated group and 80% to untreated group. Let Y_i denote the average yield for the ith village, X_i denote a binary variable that equals 1 if the village is assigned to the treatment of land rental contract law, and W_i denote a binary variable that equals 1 if the village is in the coastal region, and 0 if in the inland region. Let β_1 denote the causal effect on yield of land rental contract law.

1. Consider the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$. Do you think that $E(u_i|X_i) = 0$? Is the OLS estimator of β_1 unbiased? Explain.

   Answers: Treatment (assignment with land rental contract law) was not randomly assigned in the population (the coastal and inland region) because of the difference in the proportion of treated region. Thus, the treatment indicator X is correlated with W. If farmers in coastal region perform systematically differently on standardized tests than those in inland region (perhaps because of advanced agricultural technology), then this becomes part of the error term u. This leads means that $\mathbb{E}(u_i|X_i) \neq 0$. Since $\mathbb{E}(u_i|X_i) \neq 0$, $\hat{\beta}_1$ is biased.

2. Consider the regression $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$.

   (a) Do you think that $E(u_i|X_i, W_i)$ depends on X_i? Is the OLS estimator of β_1 unbiased? Explain.

   Answers: Because treatment was randomly assigned conditional on village region (coastal or inland), $\mathbb{E}(u_i|X_i, W_i)$ will not depend on X_i. This means that the assumption of conditional mean independence is satisfied, and $\hat{\beta}_1$ is unbiased.

   (b) Do you think that $E(u_i|X_i, W_i)$ depends on W_i? Explain.

   Answers: Because W_i was not randomly assigned (coastal region may, on average, have other attributes that could affect average yield), $\mathbb{E}(u_i|X_i, W_i)$ may depend of W_i, hence that $\hat{\beta}_2$ may be a biased estimator for the causal effect of being in coastal/inland region.

Question 3

Consider the regression model

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

for $i = 1, \ldots, n$. (Notice that there is no constant term in the regression.).

1. Specify the least squares function that is minimized by OLS.

   Answers: $\hat{\beta}_1, \hat{\beta}_2 = arg\min_{\beta_1,\beta_2} \sum_{i=1}^{n}(Y_i - \beta_1 * X_{1i} - \beta_2 * X_{2i})^2$

2. Compute the partial derivatives of the objective function with respect to b_1 and b_2.

   Answers:
   F.O.C. to $\hat{\beta}_1$ *:* $\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 * X_{1i} - \hat{\beta}_2 * X_{2i}) * X_{1i} = 0$
   F.O.C. to $\hat{\beta}_2$ *:* $\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 * X_{1i} - \hat{\beta}_2 * X_{2i}) * X_{2i} = 0$

3. Suppose that $\sum_{i=1}^{n} X_{1i}X_{2i} = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^{n} X_{1i}Y_i / \sum_{i=1}^{n} X_{1i}^2$.

   Answers: From F.O.C. to $\hat{\beta}_1$ *:* $\hat{\beta}_1 = \sum_{i=1}^{n} X_{1i}Y_i / \sum_{i=1}^{n} X_{1i}^2$

4. Suppose that $\sum_{i=1}^{n} X_{1i}X_{2i} \neq 0$. Derive an expression for $\hat{\beta}_1$ as a function of the data $(Y_i, X_{1i}, X_{2i}), i = 1, \ldots, n$.

   Answers: From F.O.C. to $\hat{\beta}_1$ *:* $\hat{\beta}_1 = (\sum_{i=1}^{n} X_{1i}Y_i - \hat{\beta}_2 * \sum_{i=1}^{n} X_{1i}X_{2i}) / \sum_{i=1}^{n} X_{1i}^2$
   From F.O.C. to $\hat{\beta}_2$ *:* $\hat{\beta}_2 = (\sum_{i=1}^{n} X_{2i}Y_i - \hat{\beta}_1 * \sum_{i=1}^{n} X_{1i}X_{2i}) / \sum_{i=1}^{n} X_{2i}^2$
   Derive: $\hat{\beta}_1 =$

$$(\sum_{i=1}^{n} X_{1i}Y_i * \sum_{i=1}^{n} X_{2i}^2 - \sum_{i=1}^{n} X_{1i}X_{2i} * \sum_{i=1}^{n} X_{2i}Y_i) / (\sum_{i=1}^{n} X_{1i}^2 * \sum_{i=1}^{n} X_{2i}^2 - (\sum_{i=1}^{n} X_{1i}X_{2i})^2)$$

5. Suppose that the model includes an intercept: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Show that the least squares estimators satisfy $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2$

   Answers: F.O.C. to $\hat{\beta}_0$ *:* $\sum_{i=1}^{n} \hat{\beta}_0 = \sum_{i=1}^{n} Y_i - \hat{\beta}_1 * \sum_{i=1}^{n} X_{1i} - \hat{\beta}_2 * \sum_{i=1}^{n} X_{2i}$
   Thus, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2$

6. As in 5, suppose that the model contains an intercept. Also suppose that $\sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)\left(X_{2i} - \bar{X}_2\right) = 0$. Show that $\hat{\beta}_1 = \sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)\left(Y_i - \bar{Y}\right) / \sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)^2$. How does this compare to the OLS estimator of $\beta_1$from the regression that omits X_2 ?

   Answers:
   F.O.C. to $\hat{\beta}_1$ *:* $\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 * X_{1i} - \hat{\beta}_2 * X_{2i}) * X_{1i} = 0$
   Put $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2$ *into:* $\sum_{i=1}^{n}[(Y_i - \bar{Y}) - \hat{\beta}_1 * (X_{1i} - \bar{X}_1) - \hat{\beta}_2 * (X_{2i} - \bar{X}_2)] * X_{1i} = 0$
   $\sum_{i=1}^{n}[(Y_i - \bar{Y}) - \hat{\beta}_1 * (X_{1i} - \bar{X}_1) - \hat{\beta}_2 * (X_{2i} - \bar{X}_2)] * (X_{1i} - \bar{X}_1) = 0$
   Thus, $\hat{\beta}_1 = \sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)\left(Y_i - \bar{Y}\right) / \sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)^2$
   This equation is the same as the OLS estimator of β_1 *from the regression that omits* X_2.