Problem Set 3
ECO3121 - Fall 2024

Due 3 PM, November 28, 2024
No late submission is allowed
Please combine your answer, Stata code and requested output in **one** pdf file and upload it to Blackboard

Question 1

Download *aghousehold.dta* and *village_rainfall.dta* datasets from the blackboard site and load into Stata. The main data we use is the National Fixed Point Survey (NPFS), which is a nationally representative panel dataset (unbalanced) of roughly 5000 households in 88 villages between 1995 and 2002. It is collected by the Ministry of Agriculture and Rural Affairs of China.

On the blackboard, we also uploaded a second dataset regarding precipitation records in each village in 2001. See the variable list below. Write up the answers to 1) - 6) below. In addition, also attach the do file that you used to answer the following questions.

variable name	type	format	label
vl_id	int	%8.0g	village ID, consistent with household data
lat_o	float	%8.0g	latitude of each village
lon_o	float	%8.0g	longitude of each village
provname	str24	%24s	province name in Chinese
cityname	str33	%33s	city name in Chinese
countyname	str33	%33s	county name in Chinese
year	float	%9.0g	year of the record
av_rain	float	%9.0g	average precipitation in 2001, measured by mm
sd_rain	float	%9.0g	standard deviation of precipitation across months in 2001
z_rain	float	%9.0g	zero score of precipitation in 2001

First, limit the sample to households **in the year 2001** by running code "keep if year==2001" in Stata. You can generate variable *yield* (output per unit of land) via $\frac{hx95\_nh125}{hx95\_nh112}$, and variable *fertilizer* (fertilizer application per unit of land) through $\frac{nh6\_3}{hx95\_nh112}$.

1. The variable *h95_nh269* indicates how many days for household members in each household had been working as temporal migrants (measured by days) in 2001.

   Generate a binary variable regarding household migration decision. It takes the value of 1 if *h95_nh269* is greater than zero, otherwise it is 0. Generate the natural log of yield and fertilizer application intensity (using **log(fertilizer + 1)** and **log(yield + 1)** to smooth zero values.). (2 points)

   Ans:
   See Stata code.

2. Run a linear probability model of household migration decision in question (1) on the natural log of yield.
   Interpret the result and comment on its statistical significance. (2 points)

   Ans:
   The result is (1 point)

$$\widehat{migration}_i = \underset{(0.0743338)}{0.1773844} + \underset{(0.0128412)}{0.0495398}\ log(yield)$$

   Interpretation: holding other factors fixed, a 1% increase in the yield increases the probability of migration by 0.000495398.
   Statistical significance: the p-value of this regression is 0.000, so it is significant at the 1% significance level. (1 point)

3. List three plausible arguments why the point estimate in question (2) could be biased, and the corresponding bias directions (**upwards** or **downwards**) relative to the true causal effect of household's agricultural production on household migration decision. (3 points)

   Ans:

   (a) Omitted variable bias. For example, *soil quality* can be a potential omitted variable. First, *soil quality* is positively correlated with *yield*. Second, *soil quality* is negatively correlated with migration decisions, since bad soil quality is detrimental to health. Thus, there is a downward bias. (1 point)

   (b) Measurement error in yield. Downwards bias (attenuation bias). (1 point)

   (c) Simultaneity (reverse causality). Similar to question 2.2. (1 point)

4. Now your professor suggests that the rainfall (precipitation) could be a valid instrumental variable (IV) for your measure of household's yield. Try to merge the household's production dataset and precipitation dataset via *vl_id*, the specific village identifier, using stata command *merge* (Many-to-one merge, type *"help merge"* in Stata for assistance).

   You decide to use the natural log of average rainfall in 2001 (*log(av_rain)*) as the IV for the natural log of yield. Verify if the assumption of instrument relevance is satisfied using the first stage regression, and obtain the results in Stata. **Write down the first stage regression model** and interpret the result and statistical significance of your result. (2 points)

   Ans:
   The specification is

$$\log(yield) = \gamma_0 + \gamma_1 \log(av\_rain) + \mu$$

   The result is (1 point)

$$\widehat{\log(yield)} = \underset{(0.1365663)}{2.526537} + \underset{(0.0196173)}{0.4644991} \log(av\_rain)$$

   The F value = 560.65. So we don't think $\log(av\_rain)$ is a weak instrument. (1 point)

5. Now use Stata to estimate the 2nd stage IV point estimate (using linear probability model) as suggested by the professor, and export your result. **Write down the second stage regression model** and interpret the result and statistical significance of your result. (2 points)

   Ans:
   The specification is

$$migreation = \beta_0 + \beta_1 \widehat{\log(yield)} + v$$

   The result is (1 point)

$$\widehat{migreation} = \underset{(0.208899)}{-2.012801} + \underset{(0.0362857)}{0.4295868} \widehat{\log(yield)}$$

   yield significantly (at the 1% significance level) increases the probability of migration decision. (1 point)

6. Now your professor tells you that you can use *ivregress 2sls* command directly to replicate the results in question (5).

   (a) Do you find any difference in the IV estimations (β_{IV}) using *ivregress 2sls* command **regarding the coefficients and standard errors** relative to (5). (2 points)

   (b) In reference to your answer to questions (2) and (3), is the difference between the linear probability model and IV point estimates as you expected or rather not? (2 points)

Ans:
(a) The IV estimation result is (1 point)

$$\widehat{migreation} = \underset{(0.234359)}{-2.012801} + \underset{(0.0407081)}{0.4295868}\ \log(yield)$$

The coefficients are the same, but the standard errors are different. (1 point)
(b) Yes. There is a downward bias in the OLS estimation. (2 points)

Question 2

Consider the two-way relationship between crop yield and fertilizer usage

$$Crop = \alpha_0 + \beta_0 Fertilizer + u$$

$$Fertilizer = \alpha_1 + \beta_1 Crop + v$$

The first equation models the determinant of crop yield given the amount of fertilizer usage. The second equation models the amount of fertilizer the farmer chooses to apply given the crop yield in the area.

1. What do you expect the signs of β_0 and β_1 are? Explain. (2 points)

   Ans:
   $\beta_0 > 0$. The more fertilizer used, the more nutrition the crops have, so more crop yield. (1 point)
   $\beta_1 > 0$. The more crops the farmer plants, the more fertilizer she needs to buy. (1 point)
   Note: $\beta_0,\ \beta_1 < 0$ can also be an acceptable answer if you give reasonable justification.

2. Explain why the OLS estimator for β_0 and β_1 are biased. If we use the OLS estimator to estimate β_0 and β_1, what directions are the biases? Explain. (4 points)

   Ans:
   Because of the simultaneity (reverse causality), $\hat{\beta}_0$ and $\hat{\beta}_1$ are biased (2 points).
   We now denote $Crop$ as y, and $Fertilizer$ as x, then we have (2 points)

   $$\hat{\beta}_0 = \frac{\hat{c}ov(x,y)}{\hat{v}ar(x)} = \frac{\hat{c}ov(x, \alpha_0 + \beta_0 x + u)}{\hat{v}ar(x)} = \beta_0 + \frac{\hat{c}ov(x,u)}{\hat{v}ar(x)} = \beta_0 + \frac{\frac{\beta_1}{1-\beta_0\beta_1}\hat{v}ar(u)}{\hat{v}ar(x)}$$

   $$\hat{\beta}_1 = \frac{\hat{c}ov(y,x)}{\hat{v}ar(y)} = \frac{\hat{c}ov(y, \alpha_1 + \beta_1 y + v)}{\hat{v}ar(y)} = \beta_1 + \frac{\hat{c}ov(y,v)}{\hat{v}ar(y)} = \beta_1 + \frac{\frac{\beta_0}{1-\beta_0\beta_1}\hat{v}ar(v)}{\hat{v}ar(y)}$$

   Therefore, there is an upward bias for $\hat{\beta}_0$ if $\frac{\beta_1}{1-\beta_0\beta_1} > 0$. There is an upward bias for $\hat{\beta}_1$ if $\frac{\beta_0}{1-\beta_0\beta_1} > 0$.

3. Suppose the only variables available are *Crop*, *Fertilizer*, *Sunshine* (the sunshine of the area), and *Budget* (the budget constraint of the farmer). To estimate β_0 and β_1 by the two-stage least squares estimator, which variables among the data you have should be used as instruments? Be specific, what IV is for *Fertilizer* and which IV is for *Crop*. Explain. (4 points)

Ans:

The IV for *Fertilizer* is the *Budget*.

(a) Relevance: the more money the farmer has in hand, the more fertilizer she can apply. (1 point)

(b) Exogeneity: the budget is not correlated with other factors that affect *Crop*. (1 point)

The IV for *Crop* is the *Sunshine*.

(a) Relevance: the better sunshine conditions, the more crop output. (1 point)

(b) Exogeneity: the sunshine condition is determined by local climate, so it is not correlated with other factors besides *Crop* that affect *Fertilizer*. (1 point)