Problem Set 1 Solutiuons
ECO3121 - Fall 2023

October 12, 2023

Due 5PM, 08/10/2022
Please remember to submit your Stata code and requested output as it will be graded.

Question 1

In this and later problem sets, we'll try to replicate a research paper titled "Property Rights, Land Misallocation and Agricultural Efficiency in China" (Chari, Liu, Wang, and Wang, 2021) step by step. The paper examines the impact of a property rights reform in rural China that allowed farmers to lease out their land. They examine the effects of land rental activity in rural households on output and aggregate productivity.

To replicate this paper, please download main dataset `"aghousehold.dta"` 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 20,000 households in 360 villages between 1986 and 2013. It is collected by the Ministry of Agriculture and Rural Affairs of China. Since we are learning simple linear regression and multiple linear regression at the current stage, we provide a cross sectional dataset this time which only includes households in the year of 2010.

It will be a univariate model, we'll be estimating regressions of the form:

(1)

$$yield_i = \alpha_1 + \beta_1 rental\_in_i + \mu_i$$

$$yield_i = \alpha_2 + \beta_2 rental\_out_i + \mu_i$$

First, let's analyze the effect of land renetal activity (rent in and rent out) on agricultural yield.

1. You can generate variable yield (output per unit of land) via $\frac{d32}{d31}$, and variable rent in through variable $c10$ and rent out via variable $c13$. Visualize and export a table that lists the number of observations, the mean, the standard deviation, the minimum value and the maximum value for these three variables in the dataset (edit and include the table in your written up answer). Briefly describe what we learn from the table about these households. (5 points)

Answers: Figure 1 and 2 show the distributions of these three variables, and Table 1 presents the summary statistics. Apparently, all box plots indicate there exist relatively large amount of anomaly data among three variables.

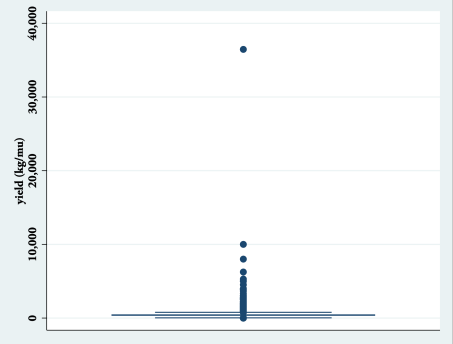

Figure 1: Box plot of *yield*

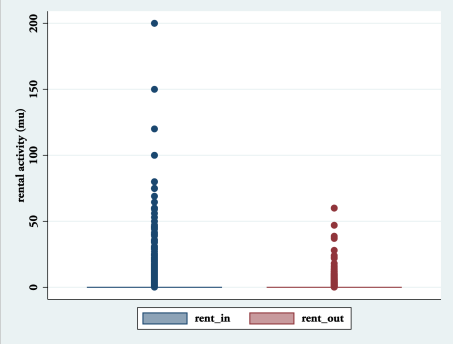

Figure 2: Box plot of rental activity

Variables	yield	rent_in	rent_out
Observations	14,171	14,171	14,171
Mean	433.617	0.503	0.216
Std. Dev.	395.198	4.467	1.489
Minimum	0	0	0
Maximum	36,461.98	200	60

Table 1: Summary Statistics for the three variables

Overall, Tab. 1 provides summary statistics that give us a basic understanding of the distribution, range, and variability of yield, rent_in and rent_out. All three variables contains 14,171 observations, and have the same minimum value as 0. yield has a larger mean, standard deviation and maximum value. Average rent-in land is larger than land that rent out. ...

2. What sign do you predict β_1 and β_2 will have? Why? (3 points)

   [Open Question] *E.g. Negative for β_1, spending more on renting in land might cause less expenditure on factors that can contribute to yield growth. Positive for β_2, renting out land could make extra income that can be used on factors that can contribute to yield growth.*

   OR *positive for β_1, larger land is applicable with advanced irrigation systems, precision agriculture technologies, or improved crop varieties, households can benefit from these resources and adopt innovative techniques that enhance productivity and yield. Negative for β_2, if land is rented out without proper land management practices, it may lead to soil degradation and nutrient depletion, which can reduce productivity and negatively impact crop yields.*

3. Run the regression and report the estimated coefficients, their standard errors, and R^2. Does the estimated value of β_1 and β_2 agree with your predictions? (3 point)

	(1)	(2)
Depend Var	yield	yield
rent_in	-0.780	
	(0.743)	
rent_out		0.652
		(2.229)
constant	434.009***	433.476***
	(3.341)	(3.355)
Observations	14,171	14,171
R-squared	0.0001	0.0000

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

Answers: See above table for detailed regression results. The estimated value of β_1 and β_2 both agree with the predictions.

$\hat{\beta}_1 = -0.780,\ se(\hat{\beta}_1) = 0.743,\ R^2 = 0.0001$

$\hat{\beta}_2 = 0.652,\ se(\hat{\beta}_2) = 2.229,\ R^2 = 0.0000$

4. Interpret the value of the estimated parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ from the models. (3 points)

Answers: An additional area (in unit mu) increases in rental-in land is associated with -0.78 kilogram increase in yield. And an additional area (in unit mu) increases in rental-out land is associated with 0.652 kilogram increase in yield.

5. Compute the fitted value $\widehat{yield}_i$ and the residual $\hat{\mu}_i$ for each observation for both regressions, and verify that the residuals (approximately) sum to 0. (2 points)

Answers: See code for details. For β_1, the sum of residuals is 0.115 (-0.0088) and for β_2, it's 0.035 (0.0156). Both approximately equal to 0.

6. How much of the variation in yields for these farmers is explained by their rent-in and rent-out activities? (2 points)

Answers: 0.01%. And if we want a more reliable result, we could conduct a bivariate regression as

$$yield_i = \alpha + \beta_1 rental\_in_i + \beta_2 rental\_out_i + \mu_i$$

7. A professor asks you to create two new variables which measure the proportion of rent-in and rent-out land to total land area ($d31$) **(remember multiply these fractions by 100)**. She also asks you to use these two measures as independent variables to re-run the two regressions,

(2)

$$yield_i = \alpha_1 + \beta_1 rental\_in\_share_i + \mu_i$$

$$yield_i = \alpha_2 + \beta_2 rental\_out\_share_i + \mu_i$$

Please report and interpreter the new estimated parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ from the models. (4 points)

	(1)	(2)
Depend Var	yield	yield
share100_rent_in	0.054	
	(0.064)	
share100_rent_out		0.235***
		(0.048)
constant	433.370***	431.612***
	(3.333)	(3.342)
Observations	14,171	14,171
R-squared	0.0001	0.0017

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

Answers: See above table for detailed regression results.

1) An additional unit increases in the proportion of rent-in to total land is associated with 0.054 kilogram increase in yield.

2) An additional unit increases in the proportion of rent-out to total land is associated with 0.235 kilogram increase in yield.

8. The professor sees your analysis. She asks you to explore alternative functional forms for this relationship. Using the provided data, you estimate the following equations:
* A log-linear relationship:

(3)

$$log(yield_i) = \alpha_1 + \beta_1 rental\_in\_share_i + \mu_i$$

$$log(yield_i) = \alpha_2 + \beta_2 rental\_out\_share_i + \mu_i$$

`We take natural log here. In Stata, the coding is "gen newvar = log(variable)"`

Run the regressions recommended by the professor. Interpret the parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ in each of these equations. (4 points)

	(1)	(2)
Depend Var	log_yield	log_yield
share100_rent_in	0.0001	
	(0.0001)	
share100_rent_out		0.0003***
		(0.0001)
constant	5.9408***	5.9387***
	(0.0050)	(0.0050)
Observations	14,171	14,171
R-squared	0.0001	0.0014

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

Answers: Since the minimum value is 0, we use `log(var+1)` *to generate variables in logarithm form. See above table for detailed regression results.*

1) An additional unit increases in the proportion of rent-in to total land is associated with 0.01% increase in yield.

2) An additional unit increases in the proportion of rent-out to total land is associated with 0.03% increase in yield.

9. Compare model (2) in q7 and model (3) in q8, which do you prefer? And why? (2 points)

   Answers: I prefer model (3) in q8. Since the dependent variable yield is positive skewness, natural log transformation would be an incremental improvement yielding the variable log_yield roughly approximates a normal distribution.

10. What are the SLR4 assumptions under this context. Propose two reasons that might cause biased estimation of β_1 and β_2 (4 points)

    Answers: The independent variables land rental activities (rent_in and rent_out) must not contain information about the mean of the unobserved factors μ.

    [Open question] *E.g. 1) Some households' characteristics would simultaneously affect the land rental activities and yield.*

    2) Land rental activities would affect certain households' characteristics that have influence on dependent variable yield.

3) Measurement error in the explanatory variables, rent_in and rent_out or the in share form.

11. Think about two mechanisms that why land rental activities can increase yields or lower yields (depending on the sign of your β_1 and β_2) (4 points)

    Answers: In our regression analysis, both rent-in and rent-out activities can increase yields, though the beta associated with rent-in are all not statistically significant.

    ***[Open question]** E.g. 1) Land rental activities could help household to obtain land beneficial to the crops and discard barren land.*

    2) Land rental activity can influence the allocation of resources, such as labor, capital, and technology.

Question 2

[Potential Outcome Framework] Background knowledge: the land rental contract law grants farmers the legal right to rent out and rent in land, outlining rules for leasing, transferring leases, and how to address land leasing disputes. Prior to this law reform, land rental activities are very rare since they were not protected by a formal law.

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 **100 rural villages**, which they randomly selected from a list of **200 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.

1. What data (which variables) do you require from their project team to answer this question? (2 points)

   Answers: We need: one dummy variable whether a village has implemented the land rental contract law and a variable that describes the total yield in a village. (You can also use yield difference as the explained variable, defined by yield this year minus yield two years ago)

2. Which regression equation would you want to use these data for? Describe each variable and subscripts that indexes basic research entity and explain the interpretation of the regression coefficients. (3 points)

   Answers: Simple linear regression: $yield_i = \beta_0 + \beta_1 contract_i + u_i$
   $yield_i$: the number of yield in the village i now
   $contract_i$: a dummy variable whether the village i have implemented the land rental contract two years ago
   β_0 : without the land rental contract law (contract = 0), the number of yield in the village i β_0
   β_1 : the land rental contract law is associated with an increase by β_1 the number of yield in the village i

3. After you have started working on this, the former leader of their project implementation team tells you that she is concerned that the program may not actually have randomly allocated treatments across the 200 villages, and that some selection may have gone on (she heard reports that the richest villages were more likely to be put into the treatment group). What concern would this bring to your estimation? (3 points)

   Answers: We'll have selection bias. If land plots were selected based on their initial land quality ("richest lands"). we would expect that two year later the average number of yield in the law-implemented group is higher than the average number of yield in the group without such contract to protect the legal right because of the fact that land were richer two year prior and not because of the law. Therefore, our estimate of the treatments is biased upwards