ECO3121 Problem Set 1

Xue Zhongkai (122090636)

October 7, 2023

Question 1.

All float numbers in this problem are restricted to at most 4 decimal places, if not mentioned.

1. First we import the data with the corresponding Stata code:

```
use "/Users/kevinshuey/Github/Assignments/cuhksz_ECO3121/as1/aghousehold.dta"
```

   Then we generated the variable **yield** with:

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

   For summarization information, we have

```
sum yield c10 c13
```

   with corresponding output:

Variable	Obs	Mean	Std. Dev.	Min	Max
yield	14,171	433.6169	395.1982	0	36461.98
c10	14,171	.5031543	4.466769	0	200
c13	14,171	.2164561	1.489389	0	60

   Thus we could conclude from the table that:

   There are 14171 observations as a data scale.

   The mean for **yield** per unit is 433.6169 kg/mu, with the standard deviation 395.1982 kg/mu. We have the minimum value for yield to be 0, and the maximum value to be 36461.98 kg/mu.

   As to **arable area subcontracted into**, we have mean value to be 0.5032 mu, with the standard deviation to be 4.4668 mu. We have the minimum value to be 0, and the maximum value to be 200 mu.

   Also, we have the mean for **arable area subcontracted out** to be 0.2165 mu, with the standard deviation to be 1.4894 mu. We have the minimum value to be 0, and the maximum value to be 60 mu.

2. I predict β_1 to be **negative**, β_2 to be **positive**. Here is the reason:

   For a considerable period of time, farmers in rural China tends to employ primitive and straightforward farming methods. In order to develop agriculture in a modern approach, aggregate agricultural methods are introduced and recommended by local governments. Once the fields are collected, there could be potential tenants with advanced technology and aggregate productivity, while individual farmers, most of the time, still remain original ones.

   Pay attention that we invest on individual farming households. Under these circumstances, the **more** one rent-in, the **lower** efficiency, the lower yields there will be; the **more** rent-out, the higher efficiency, the **higher** yields there will be. Thus I predict β_1 to be negative, β_2 to be positive.

3. Here we use regression operation in Stata respectively. First we predict β_1:

```
reg yield c10
```

with the output:

Source	SS	df	MS	Number of obs	=	14,171
				F(1, 14169)	=	1.10
Model	172005.961	1	172005.961	Prob > F	=	0.2940
Residual	2.2129e+09	14,169	156180.516	R-squared	=	0.0001
				Adj R-squared	=	0.0000
Total	2.2131e+09	14,170	156181.633	Root MSE	=	395.2

yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
c10	-.7799978	.7432502	-1.05	0.294	-2.236866	.6768703
_cons	434.0094	3.340807	129.91	0.000	427.461	440.5578

Then we predict β_2:

```
reg yield c13
```

with the output:

Source	SS	df	MS	Number of obs	=	14,171
				F(1, 14169)	=	0.09
Model	13342.3764	1	13342.3764	Prob > F	=	0.7701
Residual	2.2131e+09	14,169	156191.714	R-squared	=	0.0000
				Adj R-squared	=	-0.0001
Total	2.2131e+09	14,170	156181.633	Root MSE	=	395.21

yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
c13	.6515137	2.229133	0.29	0.770	-3.717879	5.020907
_cons	433.4759	3.354809	129.21	0.000	426.9001	440.0518

Thus we conclude from the regression that:

The estimated $\hat{\beta}_1$ is -0.7800, with the standard deviation 0.7433, with $R^2 = 0.0001$. The $\hat{\beta}_2$ is 0.6515, with the standard deviation 2.2291, with $R^2 = 0.0000$.

It agrees with my predictions, though it seems very insignificant concerning R^2.

4. For the equation $yield_i = \alpha_1 + \beta_1 rental\_in_i + \mu_i$, $\hat{\beta}_1 = 0.7800$ indicates that:

Holding other constants identical, one unit change in rent-in area will cause the yield to decrease by 0.7800 units.

For the equation $yield_i = \alpha_2 + \beta_2 rental\_out_i + \mu_i$, $\hat{\beta}_2 = 0.6515$ indicates that:

Holding other constants identical, one unit change in rent-out area will cause the yield to increase by 0.6515 units.

5. First we use Stata codes respectively with:

```
reg yield c10
predict yield_hat1, xb
gen residual1=yield - yield_hat1
```

and

```
reg yield c13
predict yield_hat2, xb
gen residual2=yield - yield_hat2
```

then we could see the corresponding variables with data in the data browser.

Finally for the sum of residuals:

```
egen res1_total=total(residual1)
egen res2_total=total(residual2)
```

we could access from the data browser that

$$\textbf{residual1} = 0.1147, \quad \textbf{residual2} = 0.3537$$

which is approximately 0, respectively.

6. For $\hat{\beta}_1$, we have $R^2 = 0.0001$; for $\hat{\beta}_2$, we have $R^2 = 0.0000$.

   These two R^2 values could be really insignificant, which indicates that a significant portion of the variation in yields cannot be explained by the variables **arable area subcontracted into/out** themselves.

7. For the requirements to generate new variables **rent_in_prop** and **rent_out_prop** with 100 multiplied, we have:

   ```
   gen rent_in_prop=c10/d31*100
   gen rent_out_prop=c13/d31*100
   ```

   We re-run the regression models like:

   ```
   reg yield rent_in_prop
   reg yield rent_out_prop
   ```

   with corresponding outputs

Source	SS	df	MS	Number of obs	=	14,171
				F(1, 14169)	=	0.71
Model	111081.633	1	111081.633	Prob > F	=	0.3991
Residual	2.2130e+09	14,169	156184.816	R-squared	=	0.0001
				Adj R-squared	=	-0.0000
Total	2.2131e+09	14,170	156181.633	Root MSE	=	395.2

yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rent_in_prop	.0540097	.0640427	0.84	0.399	-.0715225	.1795418
_cons	433.37	3.332743	130.03	0.000	426.8374	439.9026

and

Source	SS	df	MS	Number of obs	=	14,171
				F(1, 14169)	=	24.04
Model	3747795.28	1	3747795.28	Prob > F	=	0.0000
Residual	2.2093e+09	14,169	155928.149	R-squared	=	0.0017
				Adj R-squared	=	0.0016
Total	2.2131e+09	14,170	156181.633	Root MSE	=	394.88

yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rent_out_prop	.2350118	.0479362	4.90	0.000	.1410505	.3289732
_cons	431.6123	3.342233	129.14	0.000	425.0611	438.1635

Thus we conclude from the regression that:

The estimated $\hat{\beta}_1$ is 0.0540, with the standard deviation 0.0640, with $R^2 = 0.0001$. The $\hat{\beta}_2$ is 0.2350, with the standard deviation 0.0479, with $R^2 = 0.0017$.

Here is the interpretation:

a) For the equation $yield_i = \alpha_1 + \beta_1 rental\_in\_share_i + \mu_i$, $\hat{\beta}_1 = 0.0540$ indicates that:

Holding other constants identical, one unit change in the proportion of rent-in area will cause the yield to increase 0.0540 units.

b) For the equation $yield_i = \alpha_2 + \beta_2 rental\_out\_share_i + \mu_i$, $\hat{\beta}_2 = 0.2350$ indicates that:

Holding other constants identical, one unit change in the proportion of rent-out area will cause the yield to increase by 0.2350 units.

8. For the *log-level* regression, we have

```
gen log_yield=log(yield)
reg yield rent_in_prop
reg yield rent_out_prop
```

with the outputs

Source	SS	df	MS	Number of obs	=	14,135
				F(1, 14133)	=	1.66
Model	.465258697	1	.465258697	Prob > F	=	0.1982
Residual	3970.8567	14,133	.280963469	R-squared	=	0.0001
				Adj R-squared	=	0.0000
Total	3971.32196	14,134	.280976508	Root MSE	=	.53006

log_yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rent_in_prop	.0001105	.0000859	1.29	0.198	-.0000578	.0002789
_cons	5.95214	.0044757	1329.87	0.000	5.943367	5.960913

and

Source	SS	df	MS	Number of obs	=	14,135
				F(1, 14133)	=	24.13
Model	6.76799638	1	6.76799638	Prob > F	=	0.0000
Residual	3964.55396	14,133	.28051751	R-squared	=	0.0017
				Adj R-squared	=	0.0016
Total	3971.32196	14,134	.280976508	Root MSE	=	.52964

log_yield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rent_out_prop	.0003159	.0000643	4.91	0.000	.0001898	.0004419
_cons	5.94995	.0044885	1325.59	0.000	5.941152	5.958749

Here is the interpretation:

a) For the equation $log(yield_i) = \alpha_1 + \beta_1 rental\_in\_share_i + \mu_i$, $\hat{\beta}_1 = 0.0001105$ indicates:

Holding other constants identical, one unit change in the proportion of rent-in area will cause the change in percentage of the yield to increase by 0.0001105 units.

b) For the equation $log(yield_i) = \alpha_2 + \beta_2 rental\_out\_share_i + \mu_i$, $\hat{\beta}_2 = 0.0003159$ indicates that:

Holding other constants identical, one unit change in the proportion of rent-out area will cause the change in percentage of the yield to increase by 0.0003159 units.

9. I prefer model (3) in q8, as it provides a smaller standard error, which could better handle skewed data and reduce heteroscedasticity.

10. **SLR4 Assumption:** For any households with their rent-in share and rent-out share, we have the expectation of the residual to be 0.

    Preposition: For the cause of biased estimation, one might be because **some variables are omitted**. There are other factors as important as well, e.g. farming technics. Another might be for **endogeneity**, when one or more independent variables are correlated with the error term in the regression model.

11. The model indicates that the land rental activities can **increase** yields. Here are two possible mechanisms:

    a) Efficient aggregate farming: Land rental activities can lead to a more efficient allocation of resources in agriculture, which allows for better utilization of available resources like labor, machinery, and fertilizers.

    b) Specialization: By renting land from others, some potential tenants can focus on one specific produce within their specification, leading to increased productivity and yields in those areas.

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

```
use "/Users/kevinshuey/Github/Assignments/cuhksz_ECO3121/as1/aghousehold.dta"
gen yield=d32/d31
sum yield c10 c13

reg yield c10
predict yield_hat1, xb
gen residual1=yield - yield_hat1
```

```
reg yield c13
predict yield_hat2, xb
gen residual2=yield - yield_hat2

egen res1_total=total(residual1)
egen res2_total=total(residual2)

gen rent_in_prop=c10/d31*100
gen rent_out_prop=c13/d31*100

reg yield rent_in_prop
reg yield rent_out_prop

gen log_yield=log(yield)
reg log_yield rent_in_prop
reg log_yield rent_out_prop
```

Question 2.

1. Here are variables I need:

Symbol	Meaning
D_i	A dummy indicating variable, $D_i = 1$ if the village is selected else $D_i = 0$.
X_i	Characteristic of the village, e.g. population, climate, soil, etc.
Y_i	Crop yield per unit in the scale of a whole village.

2. Here is the regression model as

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

Here is the index and the interpretation:

a) Y_i is **the Crop yield per unit in the scale of a whole village**. It is the major dependent variable, from which we observe the changes corresponding to independent variables.

b) β_0 is **the intercept**, indicating the yields when all independent variables are 0.

c) β_1 is **the coefficient indicating Average Treatment Effect**. When all variables except D_i remain unchanged, applying the law (increasing D_i from 0 to 1) would bring about increase in yield for β_1 units.

d) β_2 is the coefficient of other influencing factors, indicating **the characteristic of the village**. When all variables except X_i remain unchanged, an 1-unit increase in X_i would bring about increase in yield for β_2 units.

e) u_i is **the residual** of the regression model, indicating unobserved influencing factors or purely random changes.

3. From the text, it indicates that selection bias may have occurred. This endogeneity might lead to insignificant or even incorrect indications on Average Treatment Effect.

   We could check the fairness of the random selection process; or add a term like $\beta_3 S_i$, which evaluates the selection bias in the process with a score.

********** **This is the end of Problem Set 1.** **********