ECO3121 Problem Set 3

Xue Zhongkai (122090636)

November 20, 2023

Question 1.

1. Here are some possible reasons why it could be biased:

   - **Omitted variables bias.** There are variables that influence both the household's decision to rent out land and its yield, but are not included in the regression model.

     The bias direction depends on the property of the omitted variable itself. For example, if richer families tend to rent out the field, and the yield is higher because of better resources, then it will cause the **upwards** of the causality. Conversely, if the field has a lower yield thus being rent out, it will cause a **downwards** of the causality.
   - **Simultaneity or reverse causality.** There could be a reverse causality where higher yields influence the decision to rent out land. Households experiencing high yields might be more confident in renting out their land, knowing that their remaining land is productive enough to meet their needs.

     This could lead to an **upwards** bias, as the model would capture the effect of high yields on renting behavior, not the other way around.
   - **Measurement Error.** If there is a measurement error in the independent variable, it could bias the results. This is particularly relevant in agricultural studies where precise measurement can be challenging.

     Typically, single measurement error in the independent variable leads to attenuation bias, meaning the estimated effect would lead to a **downwards** bias to zero.

2. This IV is **not true** after justification. Here are assumptions I check for:

   - **Check relevance:** The instrumental variable (total land area) must be correlated with the endogenous explanatory variable (rental out share).

     It is plausible as households with more land might be more inclined or able to rent out a portion of it.
   - **Check exogeneity:** The total land area must not be correlated with the error term in the regression model, and only influence the agricultural yield through its effect on the rental out share.

     This assumption fail to be reasonable. The total land area could have a direct effect on agricultural yield, independent of land rental activities. Larger areas might lead to economies of scale or more efficient use of resources influencing the yield.

3. First we use the Rainfall.dta, transform its *vl_id* to string to guarantee unity, and save it to the working directory.

   ```
   use "/Users/kevinshuey/Github/coursework/eco3121/as3/Rainfall.dta", clear
   tostring vl_id, replace
   save Rainfall_temp.dta, replace
   ```

   After we prepare for the lyield:

```
use aghousehold.dta, clear
gen yield=d32/d31
gen lyield = ln(yield)
```

we merge the datasets and drop the missing values.

```
merge m:1 vl_id using Rainfall_temp.dta
drop if missing(d31) | missing(c13) | missing(av_rain) | missing(lyield)
```

Here we have the first stage regression model:

$$rental\_out\_share_i = \pi_1 av\_rain + \pi_0$$

then we perform the regression

```
gen rental_out_share = c13/d31*100
reg rental_out_share av_rain
```

and the output

Source	SS	df	MS	Number of obs	=	13,862
				F(1, 13860)	=	35.99
Model	173403.633	1	173403.633	Prob > F	=	0.0000
Residual	66786668.9	13,860	4818.66298	R-squared	=	0.0026
				Adj R-squared	=	0.0025
Total	66960072.6	13,861	4830.82552	Root MSE	=	69.417

rental_out~e	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
av_rain	.0079017	.0013172	6.00	0.000	.0053198	.0104836
_cons	1.271974	1.348389	0.94	0.346	-1.37105	3.914997

The result indicates that one unit of increase in *av_rain* will lead to 0.007902 increase in the percentage point of *rental_out_share*. We find the F-statistics to be 35.99 thus indicating a strong IV.

4. The 2nd stage IV point estimate is here as

$$lyield_i = \beta_0 + \beta_1 rental\_out\_share\_hat_i + \mu_i$$

with the code as

```
predict rental_out_share_hat
reg lyield rental_out_share_hat
```

and the output

Source	SS	df	MS	Number of obs	=	13,862
				F(1, 13860)	=	4.84
Model	1.37754549	1	1.37754549	Prob > F	=	0.0279
Residual	3946.39339	13,860	.284732568	R-squared	=	0.0003
				Adj R-squared	=	0.0003
Total	3947.77094	13,861	.284811409	Root MSE	=	.5336

lyield	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rental_out_share_hat	.0028185	.0012814	2.20	0.028	.0003068	.0053303
_cons	5.92472	.0118523	499.88	0.000	5.901488	5.947953

The result indicates that one unit of increase in *rental_out_share_hat* will lead to 0.2819% increase in the *yield*.

We have the t-statistics to be

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{0.0028185}{0.0012814} = 2.200 > 1.96$$

Thus we have 95% confidence to believe that there is causality between *rental_out_share_hat* and *rental_out_share*.

5. With the direct command

```
ivreg lyield (rental_out_share = av_rain)
```

we have

```
Instrumental variables (2SLS) regression

      Source |       SS           df       MS      Number of obs   =    13,862
-------------+----------------------------------   F(1, 13860)     =      4.38
       Model |  -414.149016         1 -414.149016   Prob > F        =    0.0364
    Residual |   4361.91995    13,860  .314712839   R-squared       =         .
-------------+----------------------------------   Adj R-squared   =         .
       Total |   3947.77094    13,861  .284811409   Root MSE        =    .56099

------------------------------------------------------------------------------
      lyield |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rental_out_s~e |   .0028185   .0013472     2.09   0.036     .0001779    .0054592
       _cons |    5.92472   .0124607   475.47   0.000     5.900296    5.949145
------------------------------------------------------------------------------
Instrumented:  rental_out_share
Instruments:   av_rain
------------------------------------------------------------------------------
```

We find β_{IV} to be the same, while the standard error seems a bit larger in (5).

6. After executing the results:

```
save aghousehold_temp.dta
use aghousehold_temp.dta, clear
drop _merge
gen vl_id2 = substr(vl_id, 1, 2)
merge m:1 vl_id2 using landlaw.dta
drop if missing(implemented) | missing(av_rain) | missing(lyield) /*
*/| missing(rental_out_share)
```

we begin to check for relevance and exogeneity:

- **Check relevance:** We perform the first-stage regression as

```
reg rental_out_share av_rain implemented
test av_rain implemented
```

and the result

Source	SS	df	MS	Number of obs	=	13,862
				F(2, 13859)	=	25.08
Model	241493.664	2	120746.832	Prob > F	=	0.0000
Residual	66718578.9	13,859	4814.09762	R-squared	=	0.0036
				Adj R-squared	=	0.0035
Total	66960072.6	13,861	4830.82552	Root MSE	=	69.384

rental_out~e	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
av_rain	.0075401	.0013201	5.71	0.000	.0049525	.0101276
implemented	4.774689	1.269582	3.76	0.000	2.286137	7.263242
_cons	-1.654896	1.55631	-1.06	0.288	-4.705475	1.395682

```
. test av_rain implemented

 ( 1)  av_rain = 0
 ( 2)  implemented = 0

       F(  2, 13859) =   25.08
            Prob > F =    0.0000
```

We find the F-statistics to be 25.08. Hence we have a strong IV indicating the significant relevance.

- **Check exogeneity:** As the single independent variable here is explained by 2 IVs, we apply J test for overidentification.

  a) First we perform the 2SLS estimator with 2 IVs, and predict $lyield\_hat$.

  ```
  ivreg lyield (rental_out_share = av_rain implemented)
  predict lyield_hat
  ```

  b) Then we compute the residual

  $$\hat{u}_i = lyield_i - lyield\_hat_i$$

  with the code

  ```
  gen resid = lyield  -  lyield_hat
  ```

  c) Finally we regress $\hat{u}_i$ on $rental\_out\_share_i$, av_rain and $implemented_i$,

  ```
  reg resid rental_out_share av_rain implemented
  test av_rain implemented
  ```

  with the result

  ```
  ( 1)  av_rain = 0
  ( 2)  implemented = 0

        F(  2, 13858) =    9.97
             Prob > F =    0.0000
  ```

  and compute the F-statistic with the corresponding J-statistic:

  $$J = 2F = 19.94 \sim \chi^2(1)$$

  We have the J-statistic to be 19.94, which is significantly large. Thus there is NO significant exogeneity.

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

```
use "/Users/kevinshuey/Github/coursework/eco3121/as3/Rainfall.dta", clear
tostring vl_id, replace
save Rainfall_temp.dta, replace

use aghousehold.dta, clear
gen yield=d32/d31
gen lyield = ln(yield)

merge m:1 vl_id using Rainfall_temp.dta
drop if missing(d31) | missing(c13) | missing(av_rain) | missing(lyield)

gen rental_out_share = c13/d31*100
reg rental_out_share av_rain

predict rental_out_share_hat
reg lyield rental_out_share_hat

ivreg lyield (rental_out_share = av_rain)

save aghousehold_temp.dta
use aghousehold_temp.dta, clear
drop _merge
gen vl_id2 = substr(vl_id, 1, 2)
merge m:1 vl_id2 using landlaw.dta
drop if missing(implemented) | missing(av_rain) | missing(lyield) /*
*/| missing(rental_out_share)


reg rental_out_share av_rain implemented
test av_rain implemented


ivreg lyield (rental_out_share = av_rain implemented)
predict lyield_hat

gen resid = lyield  -  lyield_hat

reg resid rental_out_share av_rain implemented
test av_rain implemented
```

********** This is the end of Problem Set 3. **********