# Economics 361
# Problem Set #5

Jun Ishii *
Department of Economics
Amherst College

Fall 2022

This problem set will require you to use econometrics software, presumably STATA (but you may use other software, e.g. R). Amherst College does have a site license for STATA; instructions on how to download/install STATA has been emailed to you. Mateo Hoyos Lopez, our STATA TA, holds office hours in Converse 311 (Beals Computer Lab) on Tuesday/Wednesday/Thursday; for more details, see the announcement on our Moodle siate (under "Econometrics Software"). The data file required for this problem set, plants.csv, is available on the course website.

plants.csv contains a data set consisting of 5 "columns" and 100 "rows." Each column is a variable and each row an observation. The variables are:

1. plant: an indicator of whether the observation is from plant 1 (=1) or plant 2 (=2)
2. warm: an indicator of whether the observation is from a warm (=1) or cold (=0) season week
3. lnk: log(capital) for that observation
4. lnl: log(labor) for that observation
5. lny: log(output) for that observation

Copy the data file plants.csv into the main STATA data directory. This is usually C:\data

Start STATA. Use the following command to read-in the data

```
infile plant warm lnk lnl lny using plants.csv
```

If you stored plants.csv is a directory other than the STATA data directory, put the directory address in front of plants.csv. e.g. if the file is in C:\projects\data then use

```
infile plant warm lnk lnl lny using C:\projects\data\plants.csv
```

If you are using the Economics Department Computer Lab, remember to delete any STATA files you have created (including plants.csv) after you are done working on the assignment.

*Office: Converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

## Preliminaries

**NOTE:** Problems in this section are not graded. But you should do them as they set-up problems in the subsequent sections, which are graded.

**(a)** Use the STATA summarize command to calculate the sample mean and standard deviation for the five variables in plants.csv. Check to make sure that the sample means are

$\{$ 1.5, 0.5, 5.027027, 2.951686, 3.626923 $\}$

**(b)** Use the STATA summarize command with the if option to calculate the sample mean and standard deviation for the five variables for observations where plant == 1 & warm == 1. Check to make sure that the sample means are

$\{$ 1, 1, 4.858687, 3.201824, 3.420729 $\}$

**(c)** Use the STATA sort command to sort the data by plant warm. After you have sorted by plant warm, calculate the sample mean and standard deviation by plant warm. For each of the 4 possible plant warm combinations, there should be 25 observations.

**(d)** Go to the Menu Bar on top. Select Data and then Data Browser. Check to make sure that the first observation is

$\{$ 1, 0, 4.305153, 1.421877, 1.520981 $\}$

Close the Data Browser window after you are done.

**(e)** Use the STATA gen command to generate the following new variables

- gen ones = 1
- gen lnkXwarm = lnk × warm
- gen lnlXwarm = lnl × warm
- gen lnk0 = lnk - 5.027027
- gen lnl0 = lnl - 2.951686
- gen lny0 = lny - 3.626923

Note that ones is simply a variable with the value 1. STATA automatically includes a constant when running linear regressions. However, we will need ones for some other calculations later. The next two variables are **interaction terms**, interacting lnk or lnl with warm. For observations where warm == 1, lnkXwarm = lnk and lnlXwarm = lnl. For all other observations (warm == 0), lnkXwarm = lnlXwarm = 0. The other three variables, $\{\text{lnk0}, \text{lnl0}, \text{lny0}\}$, are the variables $\{\text{lnk}, \text{lnl}, \text{lny}\}$ after subtracting out their respective sample mean.

**NOTE:** In Stata, to assign a value, use a single equal sign. So "$A = B$" means assign the value of $B$ to $A$. To do a comparison, use double equal signs. So $A == B$ means $A$ is equal to $B$.

## Question 1: "Simple" OLS Model

You believe that the data in plants.csv come from manufacturing plants whose production function is characterized by the Cobb-Douglas function: $Y = AK^{\beta_k}L^{\beta_l}$ where $ln(A) \sim N(\mu_A, \sigma^2)$. You also believe that the data represents a size 100 *random* sample.

**(a)** Given the assumptions above, briefly explain why the OLS model is appropriate for estimating $\{\mu_A,\ \beta_k,\ \beta_l\}$.

**(b)** Use the STATA regress command to obtain the OLS estimates for $\{\mu_A,\ \beta_k,\ \beta_l\}$. Denote them $\{b_0,\ b_k,\ b_l\}$. Write down the values of $\{b_0,\ b_k,\ b_l\}$ ("Coef."), their standard errors ("Std. Err."), and their t-statistic ("t").

**(c)** Use the STATA vce command to obtain the estimated $\text{Var}(b \mid \mathsf{lnk}, \mathsf{lnl})$ where $b = \begin{pmatrix} b_0 \\ b_k \\ b_l \end{pmatrix}$

Compare the square root of the diagonal elements with what STATA calls the "Std. Err." Write down the estimated $\text{Var}(b \mid \mathsf{lnk}, \mathsf{lnl})$

**(d)** Show that what STATA calls "t" is simply the ratio "Coef." and "Std. Err." What null hypothesis does this t-statistic test?

**(e)** Conduct the two-sided hypothesis test of $H_0 : \beta_k + \beta_l = 1$ (Constant Returns to Scale) for the 5% ($\alpha = 0.05$) significance level. **NOTE:** Do *not* use the automated routine in STATA. Do it "by hand" using the information gathered above – show your work! (You can, of course, use a calculator)

**(f)** Now, use the STATA regress command with the option noconstant to regress lny on $\{\mathsf{lnk}, \mathsf{lnl}\}$ (no constant). Compare the coefficient estimates here to those in **(b)**.

**(g)** Now, use the STATA regress command with the option noconstant to regress lny0 on $\{\mathsf{lnk0}, \mathsf{lnl0}\}$. Compare the coefficient results here to those in **(b)** and **(f)**. Explain this result using the concept of best linear predictor.

## Question 2: Residual Regression

This section follows the discussion in Goldberger Chapter 17.3. Maintain the same assumptions about the production function and sample as in Question 1.

**NOTE:** Residuals are the difference between the real values (in the sample) and the predicted values using the OLS estimates. Recall that OLS estimates are the estimates that **minimize the sum of squared residuals**.

**(a)** Suppose instead of regressing lny on {lnk, lnl, constant}, you regressed lny on just {lnk, constant}; you omit lnl. You would get a different estimate for the coefficient before lnk. Explain why using the concept of best predictor / best linear predictor.

Goldberger discusses the common interpretation of an OLS coefficient, say the one before lnk, as the effect of lnk on lny after "controlling for the other variables" – in this case, lnl and the constant. He likens this concept to that of a "partial derivative" where we fix lnl and the constant.

**(b)** Do the following steps, re-creating Goldberger's residual regression

1. Use the STATA regress command to regress lnk on lnl and a constant. Recall that STATA automatically includes a constant, as default
2. Use the STATA predict command to save the residuals from the above regression into variable lnkres
3. Use the STATA regress command to regress lny on lnkres

Compare the estimated coefficient before lnkres (from the last regression in Step 3) to the estimated coefficient before lnk in Question 1 **(b)**.

**(c)** Do the following steps, re-creating Goldberger's residual regression

1. Use the STATA regress command to regress lnk on lnl (no constant)
2. Use the STATA predict command to save the residuals from the above regression into variable lnkres2
3. Use the STATA regress command to regress ones on lnl (no constant)
4. Use the STATA predict command to save the residuals from the above regression into variable oneres2
5. Use the STATA regress command to regress lny on lnkres2, oneres2 (no constant)

Compare the estimated coefficients before lnkres2 and before oneres2 (from the last regression in Step 3) to the estimated coefficient before lnk and before constant in Question 1 **(b)**.

**(d)** Relate your results in **(b)** and **(c)** to Goldberger's discussion in Chapter 17.3

## Question 3: Interaction Terms

Suppose you are now told that the production process differs depending on whether production occurs during a warm or cold season week. If warm $==1$ then $Y = AK^{\beta_{k1}}L^{\beta_{l1}}$ and if warm $==0$ then $Y = AK^{\beta_{k0}}L^{\beta_{l0}}$. $\beta_{k1} \neq \beta_{k0}$ and $\beta_{l1} \neq \beta_{l0}$. $ln(A) \sim N(\mu_A, \sigma^2)$ and the sample being *random* are still assumed to be true.

**(a)** Explain why the above implies that

$$E[\ \mathsf{lny} \mid \mathsf{lnk}, \mathsf{lnl}, \mathsf{warm}\ ] \quad = \quad \gamma_0 + \gamma_1 \mathsf{lnk} + \gamma_2 \mathsf{lnl} + \gamma_3 \mathsf{lnkXwarm} + \gamma_4 \mathsf{lnlXwarm}$$

Express $\{\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4\}$ in terms of $\{\mu_A, \beta_{k1}, \beta_{l1}, \beta_{k0}, \beta_{l0}\}$.

**(b)** Use the STATA regress command to obtain the OLS estimate of $\{\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4\}$. Use the STATA vce command to obtain the estimated variance-covariance matrix associated with those OLS estimate.

**(c)** You are told that capital is not as productive during the cold season. Test the null hypothesis $H_0 : \beta_{k1} = \beta_{k0}$ against the alternative $H_a : \beta_{k1} > \beta_{k0}$ using a 10% significance level ($\alpha = 0.1$).

**(d)** Suppose that you have doubts that $\mu_A$ is the same for both warm and cold season weeks. Explain the regression you would run and the associated hypothesis test you would use to test your suspicion. You do not have to conduct this test (but you may)

## Question 4: Weighted Least Squares

Maintain the same assumptions as Question 3 except assume that $ln(A) \sim N(\mu_A, \sigma_i^2)$ where $\sigma_i^2 = \sigma^2$ if observation $i$ is from plant 1 (plant == 1) and $\sigma_i^2 = 4\sigma^2$ if observation $i$ is from plant 2 (plant == 2). The observations are still statistically independent of each other (but, now, not identically distributed).

Use the following STATA commands to generate new variables wt and wt2

1. gen wt = 1
2. replace wt = 0.5 if plant == 2
3. gen wt2 = wt*wt

Use the Data Browser to make sure that wt = 1 for plant == 1 and wt = 0.5 for plant == 2.

Now, use wt and the STATA gen command to generate the following transformed variables

- gen lnywt = lny * wt
- gen lnkwt = lnk * wt
- gen lnlwt = lnl * wt
- gen lnkXwt = lnkXwarm * wt
- gen lnlXwt = lnlXwarm * wt

**(a)** Briefly explain why regressing lnywt on {wt, lnkwt, lnlwt, lnkXwt, lnlXwt} **without** a constant can provide you with estimates of $\{\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4\}$ (and thus $\{\mu_A, \beta_{k1}, \beta_{l1}, \beta_{k0}, \beta_{l0}\}$) with lower variance.

**(b)** Use the STATA regress command to conduct the estimation described in **(a)**. Use the STATA command vce to get the estimated variance-covariance matrix. Why is your answer here different from that in Question 3 **(b)**?

**(c)** Use the STATA regress command with the option [w = wt2] to regress lny on {lnk, lnl, lnkXwarm, lnlXwarm} **with** constant. Note: this option comes at the end of the command without a comma (but in brackets). Compare your result with **(b)**. What does the [w = wt2] option do?

**(d)** Using your result in **(b)** re-do the hypothesis test in Question 3 **(c)**.