Economics 361
Problem Set #5 (Suggested Solutions)

Jun Ishii *
Department of Economics
Amherst College

Fall 2022

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 $\{\mu_A,\ \beta_k,\ \beta_l\}$.

ANS: Log-linearize the production function (take log-on both sides)

$$lnY \quad = \quad lnA + \beta_k lnK + \beta_L lnL$$

From above, we can show that $E[lnY \mid lnK, lnL] = \mu + \beta_k lnK + \beta_l lnL$ and $\text{Var}[lnY \mid lnK, lnL] = \sigma^2$ as $E[ln(A) \mid lnK, lnL] = \mu$ and $\text{Var}[lnA \mid lnK, lnL] = \sigma^2]$. Additionally, you have a random sample. So linearity and spherical errors will be satisfied. (We assume the data is such that full rank of [constant, lnK, lnL] is assured)

(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").

ANS:

	Coef.	Std. Err.	t
lnK	0.5569253	0.1541305	3.61
lnL	0.7238385	0.0791245	9.15
Constant	-1.3093	0.8133547	-1.61

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

(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 $\text{Var}(b \mid \mathsf{lnk}, \mathsf{lnl})$

	lnK	lnL	Constant
lnK	0.02375621		
lnL	-0.0006511	0.00626069	
Constant	-0.11923095	-0.01815228	0.66154587

The square roots of the diagonal elements should be the same as the "Std Err" in the earlier regression table.

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

ANS: The "t" is the test statistic for testing the null hypothesis that the coefficient is zero. So, for the "t" for lnK, the null hypothesis is $H_0 : \beta_k = 0$

(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)

ANS: First, calculate the estimated variance of $b_k + b_l$

$$\begin{aligned} \hat{\text{Var}}[b_k + b_l \mid lnK, lnL] &= \hat{\text{Var}}[b_k \mid lnK, lnL] + \hat{\text{Var}}[b_l \mid lnK, lnL] + 2\ \hat{\text{Cov}}[b_k, b_l \mid lnK, lnL] \\ &= 0.02375621 + 0.00626069 + 2 \times -0.0006511 \quad = \quad 0.02988668 \end{aligned}$$

Then, calculate the appropriate "t" test statistic

$$\begin{aligned} \text{t-stat} &= \frac{b_k + b_l - 1}{\sqrt{\hat{\text{Var}}[b_k + b_l \mid lnK, lnL]}} \\ &= \frac{0.5569253 + 0.77238385 - 1}{\sqrt{0.02988668}} \quad = \quad 1.62406077 \end{aligned}$$

Now, compare to the proper critiical region: the value on the t-distribution (with 100-3 = 97 degrees of freedom) for which the cumulative distribution is valued 0.975 (5% significance, two-sided).

The standard Normal can be used to approximate t-distributions with high degrees of freedom. In which case, the critical region is less than -1.96 and greater than +1.96. Our calculated t-statistic does not fall in either region. So, we fail to reject $H_0 : \beta_k + \beta_l = 1$.

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

ANS:

	Coef.	Std. Err.	t
lnK	0.3209491	0.0479993	6.69
lnL	0.6879124	0.0765256	8.99

The coefficients before lnK and lnL have both changed. This makes sense as this regression estimates a different relationship. We are no longer estimating the best (linear) predictor of lnY given lnK and lnL given that we are forcing the intercept term to be zero. We are estimating the best (linear but without an intercept) predictor.

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

	Coef.	Std. Err.	t
lnK	0.5569253	0.1533421	3.63
lnL	0.7238385	0.0787198	9.20
Constant	-1.3093	0.8133547	-1.61

The coefficient estimates are the same as in **(b)**. The standard errors are slightly different (and thus the reported "t") only because of rounding errors: when constructing (lny0, lnk0, lnl0), we subtracted a rounded version of the sample means rather than the actual sample mean. (If we had subtracted out the full mean, there would be no difference).

Recall from our derivation of the best linear predictor (under MSE), that the intercept term was essentially $E[Y] - E[X]\beta$ where β was the slope term for the BLP. Th intercept simply reflects the aspect of the mean value of Y (the variable being predicted) that cannot fully be explained by the mean values of the explanatory variables X. If $E[Y] = 0 = E[X]$ then the intercept $= 0$.

Similarly, in our derivation of the OLS estimator, we saw that the OLS estimate of the intercept was simply $\bar{Y} - \bar{X}^*b$ where X^* are the explanatory variables excluding the intercept and b the OLS estimate of the slope coefficients for those variables. If $\bar{Y} = 0 = \bar{X}$ then out estimated intercept will be zero. Excluding the intercept in that case does not make a difference. There is no "unexplained mean value" that needs to be reflected by the intercept as the means have been, effectively, normalized to zero.

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.

ANS: In this case, you are no longer estimating the BLP of lnY given (lnK, lnL); you are estimating the BLP pf lnY given just lnK.

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)**.

ANS: They should be the same (0.5569253).

(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)**.

ANS: They should be the same (0.5569253 and -1.3093).

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

ANS: Read the last two paragraphs of 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 $\{\mu_A, \beta_{k1}, \beta_{l1}, \beta_{k0}, \beta_{l0}\}$.

ANS:

$$E[\ \mathsf{lny} \mid \mathsf{lnk}, \mathsf{lnl}, \mathsf{warm}\ ] \quad = \quad \begin{cases} \gamma_0 + \gamma_1 lnK + \gamma_2 lnL & \text{if } \mathsf{warm} = 0 \\ \gamma_0 + (\gamma_1 + \gamma_3) lnK + (\gamma_2 + \gamma_4) lnL & \text{if } \mathsf{warm} = 1 \end{cases}$$

$$\gamma_0 = \mu \quad \gamma_1 = \beta_{k0} \quad \gamma_2 = \beta_{l0} \quad \gamma_3 = \beta_{k1} - \beta_{k0} \quad \gamma_4 = \beta_{l1} - \beta_{l0}$$

(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.

ANS:

	Coef.	Std. Err.	t
lnK	0.610032	0.1582006	3.86
lnL	0.7421699	0.1047326	7.09
lnkXwarm	-0.105356	0.094541	-1.11
lnlXwarm	0.0224858	0.1502265	0.15
Constant	-1.400954	0.7964369	-1.76

	lnk	lnl	lnkXwarm	lnlXwarm	constant
lnk	.02502742				
lnl	-.00297627	.01096892			
lnkXwarm	-.00456557	.00568248	.00893801		
lnlXwarm	.00672079	-.01047368	-.01312258	.02256801	
constant	-.11604729	-.01534798	.00420329	-.00511947	.6343117

(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 ($\alpha = 0.1$).

ANS: Note that $\beta_{k1} - \beta_{k0}$ is essentially the coefficient before lnkXwarm (i.e. γ_3). So, the hypothesis test can be re-expressed as

- $H_0 : \gamma_3 = 0$
- $H_a : \gamma_3 > 0$

STATA actually already calculates the appropriate t-statistic for this hypothesis test; it is the "t" for lnKXwarm, namely 1.11. (This is the ratio of the estimated γ_3 and the estimated standard error for the estimated γ_3.) Again using the standard normal as an approximation for t_{100-3} distribution, we find that the appropriate critical value (this time, the value at which the cumulative distribution function is valued at 0.9) is 1.28. We fail to reject the null hypothesis.

(d) Suppose that you have doubts that μ_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).

ANS: You should run the same regression as above except add another explanatory variable: warm. The coefficient before warm would represent the difference in μ_A between warm and cold seasons. The hypothesis test would be on whether the coefficient before warm is equal to zero.

	Coef.	Std. Err.	t
lnK	0.6503477	0.2255546	2.88
lnL	0.7475019	0.1073587	6.96
lnkXwarm	-0.1782186	0.3043679	-0.59
lnlXwarm	0.011977	0.1574786	0.07
warm	0.4049511	1.607068	0.25
Constant	-1.621319	1.185506	-1.37

FYI, you would fail to reject this hypothesis too (t-stat is a measly 0.25).

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.

ANS: You are essentially running Generalized Least Squares (GLS). [GLS, which we will cover later, is an extended version of linear regression that we use when the Spherical Errors assumption is violated in a known way. Intuitively, GLS is OLS applied to a judiciously transformed data.] You do not need the constant as wt is essentially the properly transformed constant.

(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)**?

	Coef.	Std. Err.	t
wt	-.9985026	.539215	-1.85
lnkwt	.4687586	.1143323	4.10
lnlwt	.780808	.0776487	10.06
lnkXwt	-.0140833	.0693019	-0.20
lnlXwt	-.081048	.1086748	-0.75

	wt	lnkwt	lnlwt	lnkXwt	lnlXwt
wt	.29075286				
lnkwt	-.05633835	.01307188			
lnlwt	-.00220677	-.00289582	.00602932		
lnkXwt	.00424191	-.00297731	.00329123	.00480276	
lnlXwt	-.0071963	.00471783	-.00595795	-.00699222	.01181021

The data has been transformed. Therefore, the regression is no longer on the same sample, per se. The estimated coefficients and variance/covariance can be different.

(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?

ANS: The [w = wt2] does the data transformation for you. OLS applied to this type of data transformation, where the observations are weighted, is known as "Weighted Least Squares." It is a specific example of GLS used when the only violation of the spherical errors assumption is heteroskedasticity.

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

ANS: Again, STATA provides the appropriate t-statistic. The magnitude is now even lower, -0.20. As the hypothesis test is one-sided, the appropriate critical region is the set of values greater than 1.28. The result is the same as before; we fail to reject the null hypothesis.