University of Victoria
Department of Economics

Economics 565
Econometrics of Cross-Section Data

Practice problems
November 2012

1. Under what conditions is OLS BLUE?

2. Briefly explain what we mean by the "power" and "size" of a test statistic. Argue that selecting 1% size is not unambiguously better than selecting 5% size.

3. The DGP is $y_i = \beta + u_i$, where $u_i \sim N(0, \sigma^2)$.

   (a) Derive the OLS estimator of β and its variance.

   (b) Display the t–statistic against the null that $\beta = 0$.

   (c) Given that the true value is β, give an expression for the probability of rejecting the null that $\beta = 0$.

   (d) Show that the probability of rejecting the null converges to one as the sample grows without bound for all $\beta \neq 0$.

4. The DGP is $y_i = \beta x_i + u_i$, where $u_i \sim N(0, \gamma_0 + \gamma_1 x_i^2)$.

   (a) Is the OLS estimator from a regression of y on x unbiased?

   (b) Is the OLS estimator efficient? Briefly explain.

   (c) Give an expression for an efficient estimator of β if (γ_0, γ_1) are known.

   (d) Outline a two–step procedure using OLS in each step to obtain estimates of (γ_0, γ_1).

   (e) Outline a maximum likelihood procedure to obtain asymptotically efficient estimates of $(\beta, \gamma_0, \gamma_1)$.

   (f) Outline a procedure to obtain FGLS estimates.

5. The DGP takes the form $y_i = \beta x_i + u_i$.

   (a) Consider regressing x on y and using the reciprocal of the estimated coefficient on y as an estimate of β. Is this estimator consistent, assuming an OLS regression of y on x is?

   (b) Suppose you wish to estimate the model using z as an instrument for x. Show that the IV reciprocal of the IV estimator obtained by regressing x on y using z as an instrument for y is consistent so long as z is a valid instrument for x.

6. The model takes the form,

$$\begin{align} y &= \beta_1 x_1 + \beta 2 x_2 + u \tag{1}\\ x_1 &= \pi_{11} z_1 + \pi_{12} z_2 + \epsilon_1 \tag{2}\\ x_2 &= \pi_{21} z_1 + \pi_{22} z_2 + \epsilon_2 \tag{3}\\ & \tag{4} \end{align}$$

   (a) Under what conditions will OLS regressions of y on x_1 and x_2 produce consistent estimates?

   (b) Using Y, X1, X2, Z1, Z2 to denote variables, present Stata code which uses three OLS regressions to construct the IV estimate of the equation of interest.

   (c) What do we mean by a "weak instrument"?

   (d) How would you provide evidence on whether z_1 and z_2 are weak instruments for x_1 and x_2?

7. The DGP is of the class,

$$\begin{align} y_i &= \beta x_i + u_i \tag{5}\\ x_i &= \pi z_i + \epsilon_i \tag{6} \end{align}$$

where u and z are uncorrelated and (u, ϵ) are jointly normal with zero means and covariance Σ.

   (a) Under what restrictions on Σ is the OLS estimator from regressing y on x consistent?

   (b) Calculate the probability limits of the OLS and IV estimators of β.

   (c) Denote the OLS estimate from a regression of y on z $\hat{\gamma}$, and the estimate from a regression of x on z $\hat{\pi}$. Show that $(\hat{\gamma}/\hat{\pi})$ is a consistent estimator of β so long as $\pi \neq 0$.

   (d) Derive an expression for the relative asymptotic biases of the OLS and IV estimators in terms of the covariances between u and x, u and z, and the first-stage R^2.

8. You wish to estimate the causal effect of schooling S on log–earnings y. You have a randomly selected cross section of workers of size n. For each worker, you observe y, S, a vector of demographic variables denoted X, a dummy indicating the respondent grew up within a two–hour drive of a degree-granting institution denoted C, and a dummy indicating the respondent's parents were willing to financially support the respondent's college education denoted P.

   (a) You regress y on X and S. The coefficient on S is 0.10 with an associated standard error of 0.02. Carefully explain how you would interpret this estimate.

(b) Outline a two–step procedure using OLS in each step to obtain estimates of the effect of schooling on earnings using C as an instrument for S. Clearly display which variables appear in each regression you propose.

(c) Discuss the conditions under which the estimate produced in the previous procedure is consistent. Do you think these conditions hold? Can you test these conditions?

(d) Outline a procedure to test the null that the OLS and IV estimates of the model are equal.

(e) Suppose you instead use P as an instrument for S. Carefully explain how to interpret the IV resulting IV estimate.

(f) Briefly outline a two–step procedure using both P and C as instruments for S. Explain how to test the overidentifying restriction. What can you conclude if the data reject the restriction?

9. You wish to estimate the causal effect of income on health in a developing country. Your data consist of observations on rural families' incomes from farming y, a health measure H, a vector of demographic controls X, and the rainfall that the families' land received R.

   (a) Would regressing H on X and y produce compelling estimates of the causal effect of income on health?

   (b) Explain how to obtain estimates of the causal effect of income on health under the assumption that rainfall affects health only because rainfall affects farming income. Under this assumption, do you need to include X in your procedure? Should you include X?

   (c) Suppose families are not randomly distributed with respect to rainfall. Families with more resources or power are able to obtain farmland in areas with more rainfall. How would this phenomena affect your estimates?

10. "Missingness is *ignorable* if selection occurs on observables." Explain as clearly and technically as you are able what this statement means.

11. Consider estimating the model $y = X\beta + \delta C + u$, where y is log–wage, X is a vector of sociodemographic controls, and C is a dummy indicating the respondent has a college degree.

   (a) Explain how to interpret an OLS estimate of δ.

   (b) You observe T and E, local college tuition and the local employment rate. Explain the conditions under which each are valid instruments for C.

   (c) Explain how to construct a control function to correct for non–random selection into college specifying that C is linearly related to its determinants.

(d) Suppose C is determined by a probit. Explain how to construct a control function exploiting this distributional assumption. Suggest a test statistic against the null that there is no selection into college on unobservables.

12. Some provinces implement a \$5,000 per year scholarship to encourage students to attend university. Students are only eligible for the scholarship if their family income is less than \$40,000 per year. Later, you collect a cross-section on randomly sampled Canadians, observing $\{y_i, z_i, c_i\}$, where y is annual earnings, z is a dummy indicating the scholarship program was implemented in i's province, and c is a dummy indicating university graduation. Explain how to cleanly interpret an IV estimate of the effect of university on earnings using z as an instrument for university.

13. Suppose $y^* = X\beta + u$, where $u \sim F(u)$. You observe $y = 1(y^* > 0)$.

    (a) What are the advantages and disadvantages of estimating the relationship between y and X using OLS?

    (b) Derive the likelihood function for the data.

14. Consider the following Stata code,

```
program define E565, rclass
    args [ ,spam(integer 100) dude(real 0) ]
    set obs ‘spam’
    gen fred = invnorm(uniform())
    gen george = invnorm(uniform())
    gen harry = 1 + ‘dude’*fred + george
    regress harry fred
    return scalar bfred = _b[fred]
    return scalar sefred = _se[fred]
end
```

(a) Display a Stata command to run Monte Carlo simulations of this code 1000 times with a sample size of 100 and the default value of the slope coefficient, saving the values of "bfred" and "sefred" in variables of the same names from each replication.

(b) After running the Monte Carlo experiment:

i. How many observations will you have in memory?

ii. What will the mean of bfred be, roughly?

iii. What will the mean of sefred be, roughly?

(c) How would you estimate the realized size of the t–statistic against the null that the slope coefficient is zero at 5% nominal size?

(d) How would you estimate the power of the t–statistic against the false null that the slope coefficient is 1.0 at 5% nominal size?

15. Explain the difference between *economic significance* and *economic significance.*

16. The DGP takes the form $y^* = \beta_0 + \beta_1 x_i + u_i$, where $u_i \sim N(0, (\gamma + \alpha x_i)^2)$. You observe $y = min(0, y^*)$.

    (a) Will an OLS regression of y on a constant and x produce consistent estimates?

    (b) Under the assumption that $\alpha = 0$, suggest a consistent estimator for the β's.

    (c) Explain, including full Stata code, how to use simulation to estimate the bias of OLS and Tobit estimators for this model in the case of $n = 100$, $\beta_0 = \beta_1 = 0$, $\gamma = \alpha = 1$, and x is drawn from a standard normal distribution.

17. (10 marks) Consider the DGP,

$$y_i = \beta_0 + \beta_1 x_i + u_i;\ u_i \sim N(0, \gamma_0 + \gamma_1 x_i^2)$$

where x_i is a scalar. You observe a random sample of size n from this process.

    (a) If you regress y on x and a constant using OLS, will your estimates be efficient? unbiased? (Assert the answer and provide a brief explanation; you need not provide formal proofs.)

    (b) Propose a method to test the null hypothesis that $\gamma_1 = 0$.

18. (10 marks) You are modeling time to first job for graduates with an M.A. in Economics. You begin by assuming that time to first job for the i^{th} graduate, y_i, follows an exponential distribution

$$f(y_i, \theta) = \theta e^{-\theta y_i},\ y_i > 0,\ \theta > 0.$$

Assume that the y_i are statistically independent. You observe a random sample of size n from this process. Derive the likelihood function, and calculate the maximum likelihood estimator of θ and its asymptotic variance.

19. (15 marks) Consider OLS models of the form,

$$\begin{aligned}
(M1) \quad & y = \beta_0 + \beta_1 x_1 + u \\
(M2) \quad & y = \beta_0 + \beta_2 x_2 + u \\
(M3) \quad & y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u
\end{aligned}$$

and estimates thereof,

	M1	M2	M3
x_1	1.04		-28.78
	(33.55)		(-0.91)
x_2		1.04	29.83
		(33.55)	(0.95)
constant	.0384	.0382	.0372
	(1.23)	(1.23)	(1.21)
N	1000	1000	1000
R^2	.530	.530	.531
F	1125.8	1125.7	563.3

where the numbers in parentheses are t–ratios and F denotes the F-statistic against the null that all slope coefficients are zero (the critical value for an F–test with 2 and 998 degrees of freedom at 5% size is 3.005).

(a) Can you reject the hypothesis $\beta_1 = 0$ in model M1?

(b) Can you reject the hypothesis that $\beta_1 = 0$ in model M3?

(c) Can you reject the joint null that $\beta_1 = \beta_2 = 0$ in model M3?

(d) Briefly describe properties of the data that would yield results such as these.

20. (20 marks) Consider the following Stata code:

```
program define final, rclass
drop _all
set obs 1000
gen x = invnorm(uniform())
gen u = uniform() - 0.5
gen y = x + u
reg y x , noconstant
return scalar gilligan = _b[x]
return scalar maryanne = ( _b[x] - 1 ) / _se[x]
end
```

You issue the code above, then the following commands

```
simulate gill=r(gilligan) mary=r(maryanne), reps(10000) : final
summ
count if abs(mary) > 1.962
```

Defend each answer you give (undefended answers are worth no marks):

(a) After the **summ** command, how many observations will Stata report it has on the variables **gill** and **mary**?

(b) What would you expect the mean of **gill** to approximately equal?

(c) What would expect the standard deviation of **gill** to approximately equal?

(d) What would expect the mean and standard deviation of **mary** to approximately equal?

(e) Approximately what number will the **count** command return?

21. The table on the next page is from Acemoglu, Johnson, and Robinson's (2001) paper on institutions and growth. The footnote summarizes the models estimated. APE below is short for "average protection against appropriation risk, 1985–1995."

    (a) Interpret the coefficient on APE in model (1) panel C.

    (b) Interpret the coefficient on APE in model (1) panel A.

    (c) Interpret the coefficient on APE in model (2) panel C.

    (d) Interpret the coefficient on APE in model (2) panel A.

    (e) Clearly explain the intuition behind the instrumental variables strategy used in model 1.

    (f) Does the first–stage in model (1) indicate a weak instrument problem is present?

    (g) Does the sign on settler mortality in the first stage make sense, given the authors' hypotheses?

    (h) Can you reject the null that the coefficient on APE in model 1 is zero? Does your argument require that the error term is normally distributed?

    (i) Is the instrument weak in model (4)?

    (j) Does model 7 indicate that Africa has lower GDP than we would expect given its institutions?

    (k) The sample size ranges from 37 to 64 across models. Briefly discuss inferential issues these small samples might cause.

    (l) Construct an F–test against the null that latitude and the continent dummies are jointly irrelevant in the first stage of model (8).

22. Suppose laws banning workplace smoking are implemented over time in various provinces. Let $D_{jt} = 1$ if province j has implemented the policy in year t. You have annual cross-sections of randomly sampled Canadians spanning the period over which the bans are implemented. For each person i, you observe X_{ijt}, a vector of sociodemographic controls, and for each province you observe a vector of contextual controls, such as tobacco taxes, Z_{jt}.

    (a) Suppose the outcome in which you are interested is number of cigarettes smoked, s_{ijt}. Specify the model you would estimate to investigate the effect of D on s. Draw a diagram illustrating the variation your model uses to identify the effect of interest. Comment briefly on issues in interpretation or estimation you anticipate you may encounter.

(b) Suppose the outcome in which you are interested is wages, w_{ijt} (you may condition on employment). Explain how you would use an instrumental variable strategy to estimate the effect of smoking on wages using these data.