

## Study Questions for ECON 836 SET #1

1. Pendakur and Pendakur (1998) “Colour of Money” control for personal characteristics  $X$ , but do not control for job characteristics  $Z$  in some of their regressions. If  $X=[V\ W]$  where  $W$  represents variables of interest (in this case, ethnicity) and  $V$  represents variables that are controlled for, but are not of direct interest, (in this case, age, education and so on) does it matter if:
  - a.  $V$  and  $W$  are correlated? Can you give an example of this kind of correlation?
  - b.  $W$  and  $Z$  are correlated? Can you give an example of this kind of correlation?
  - c.  $E[ZZ']$  depends on  $W$ ? Can you give an example of this kind of dependence?
2. Pendakur and Pendakur estimate models of income from wages and salaries which control for education.
  - a. If there were unobserved quality variation for people with the same reported education level, how would this affect your interpretation of the estimates?
  - b. Assume that 'field-of-study' is available in the data (it is). Should it be included in the regression? Does excluding it induce bias? Why?
  - c. Does it matter that they drop all observations for which income from wages and salaries is zero?
  - d. Suppose these authors wanted to investigate the conditional median of log-earnings rather than the conditional mean. Would this give them a way to deal with zeroes?
3. Allen, Pendakur and Suen (2005) estimates a panel model with the standard deviation of the log of age at first marriage on the LHS and no-fault status and state and year dummies on the RHS.
  - a. They do not include any information about the population of the state in the model. Likewise, there is not information on education levels in the state. Does this matter? Under what conditions does it not matter? Are these plausible conditions?
  - b. Should lagged age at first marriage be a regressor? What would happen if there were over-time correlations in the error terms for a country? What could you do about this problem?
  - c. Why didn't they use the *random effects* FGLS estimator?
  - d. It could be that time affects every country differently. Why didn't they interact time dummies with country dummies?
  - e. These authors regress median age at first marriage in a state on legal characteristics. Could they have run a quantile regression to address their question? If so, what quantile regression?
4. OLS minimises the sum of squared vertical deviations from the regression line. What would it mean if we minimised the sum of squared *horizontal* deviations from the regression line? How might you interpret a regression line that was obtained by this minimisation?
5. Why do we say it doesn't matter if the disturbance term contains measurement error in  $Y$ ? In contrast, why does it matter if there is measurement error in  $X$ ?
6. Say the IV estimator for a coefficient is unbiased, but has a standard error of 1. If I tell you that the OLS estimator has a bias no greater than 0.5 and a standard error of only 0.5, how should you think about this trade-off? Is there an argument to use the biased OLS estimator in this case?
7. Why are residuals in OLS regressions mean-zero?
8. What is the correlation between a regressor and the OLS residual?
9. Say I estimate the effect of being a visible minority on the log of earnings with the public use sample of the 1991 Census, which has only 700,000 observations, and I get an estimate of -0.15 with an estimated standard error of 0.05.
  - a. How much extra precision should I expect to get if I go to Ottawa to use the main base which has 5,000,000 records in it?
10. If you have heteroskedasticity of unknown form, you cannot get a consistent estimator for the GLS weighting matrix, because you need one element for every observation.
  - a. How does the White hetero-robust covariance matrix estimator get around this problem? How does it affect the estimated coefficients, compared to the OLS estimates?
11. If observations are group-means for variables, like average incomes and education levels in cities,
  - a. how do you get the correct standard errors for OLS estimates?
  - b. can you get more efficient estimates than the OLS estimates?
12. If a variable is insignificant in a regression, should you throw it out and re-estimate?

- a. Why or why not?
  - b. If two variables are insignificant in a regression, should you throw them out and re-estimate?
13. Imagine that you estimate a model where earnings is a function of a bunch of covariates that are inter-related (they cause each other), and nothing in the model is statistically significant.
  - a. Does that mean that these covariates do not influence earnings?
  - b. If you believed that they really did influence earnings, what might you do?
14. Panels have 2 subscripts, not just 1, call them  $i$  and  $t$ . Can I estimate a panel model just with OLS regression of  $Y$  on covariates  $X$ ? Under what conditions is this okay? What about regressing  $Y$  on covariates  $X$  and dummies for each value of the subscript  $i$ ? What about regressing  $Y$  on  $X$  and dummies for each subscript  $t$ ? What about regressing  $Y$  on  $X$  and dummies for each value of the subscript  $i$  and dummies for each value of the subscript  $t$ ?
15. Assume that in a panel of individuals over time there are unit effects for each person that you want to control for. However, also assume that you are not really interested in the values of these unit effects.
  - a. How should you estimate the model if you believe that these unit effects are random and independent of the other RHS variables?
  - b. How should you estimate the model if you believe that these unit effects are non-stochastic?
16. Derive the following moment estimators:
  - a. the method-of-moments estimator for a linear model satisfying the classical conditions is the the OLS estimator.
  - b. the method of moments estimator for a linear model with exactly as many instruments as endogenous variables.
  - c. If you have more than 1 instrument for each endogenous variable, how could you modify the 'method-of-moments' estimator to accommodate? Show that this estimator is the 2SLS estimator when disturbances are homoskedastic.
17. Generalised least squares (GLS) says that if the vector of disturbance terms in a linear model has a variance of  $\Omega$ 
  - a. Define a matrix  $T$  such that I can premultiply  $Y$  and  $X$  by  $T$ , run OLS of  $TY$  on  $TX$ , and get efficient coefficient estimates. Would this regression give me the correct coefficient covariance matrix? What is the sampling distribution of an OLS coefficient estimated in this fashion (that is, of the GLS coefficient).
  - b. Feasible GLS (FGLS) says that if I can estimate  $\Omega$  with a consistent estimator  $G$ , then I can premultiply  $Y$  and  $X$  by a transformation matrix  $T$  as above, but based on  $G$  instead of  $\Omega$ , and use OLS. What is the sampling distribution of a GLS coefficient estimated in this fashion?
  - c. Imagine that I assume that  $\Omega$  is diagonal. How can I proceed with FGLS?
  - d. Imagine that I assume that  $\Omega = \text{diag}(a)$  where  $a$  is an  $N$ -vector, and  $a = Wb$ , where  $W$  is an  $N \times L$  matrix of exogenous variables, where  $L < N$  and  $b$  is an  $L$ -vector of constants. How can I proceed with FGLS?
18. Given a Stata data file data.dta with 1000 observations of  $y$  and  $x1$ - $x10$ , write Stata code to
  - a. Assume that  $x7$  is coded as 999 when missing. Change it so that is a . symbol when missing.
  - b. Get the mean of each variable for observations where  $x1$  is equal to 5 and  $x2$  is less than 10.
  - c. Run a regression of  $y$  on  $x1$ - $x10$  for observations where  $x1$  is equal to 5 and  $x2$  is less than 10.
  - d. Run a regression of  $y$  on  $x1$ - $x9$  for observations where  $x10$  is less than its median value.
  - e. Run regressions of  $y$  on  $x1$ - $x9$  separately for observations with each value of  $x10$ .
  - f. Assume that  $x1$ - $x10$  are the values of a single variable 'x' in each of 10 years, 1980 to 1989. Write code to run a regression of  $y$  on  $x$  with year dummies.