Study Questions For ECON 836 Midterm

1) Pendakur and 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?
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?
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) Why are residuals in regressions mean-zero?
7) Say I estimate the effect of being a visible minority on the log of earnings with the public use sample of the 2006 Census, which has only 700,000 observations, and I get an estimate of -0.15 with an estimated standard error of 0.05.
      i) How much extra precision should I expect to get if I go to the RDC to use the main base which has 5,000,000 records in it?
8) 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?

9) 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?

10) Panels have 2 subscripts, not just 1. Can I estimate a panel model just with OLS? Under what conditions is this okay?

11) 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 and the model error term?
   b) How should you estimate the model if you believe that these unit effects are non-stochastic?

12) Generalised least squares (GLS) says that if the vector of disturbance terms in a linear model has a variance of Ω.
   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 Ω 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 Ω, and use OLS. What is the sampling distribution of a GLS coefficient estimated in this fashion?
   c) Imagine that I assume that Ω is diagonal. How can I proceed with FGLS?
   d. Imagine that I assume that Ω=diag(a) where a is an N-vector, and $a=Wb$, where W is an NxL matrix of exogenous variables, where $L<N$ and b is an L-vector of constants. How can I proceed with FGLS?

13) 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.
   g) Write code to run a regression of y on x with year dummies.