Econometrics Preliminary Exam

Agricultural and Resource Economics, UC Davis

June 30, 2016

There are **FOUR** questions. Answer each part of each question. All questions are weighted equally. Within each question, each part will receive equal weight in grading.

I. There are two parts to this question

(a) We observe random variables $Y_1, Y_2, \ldots, Y_n$ and fixed variables $X_1, \ldots, X_n$ such that $E[Y_i|X_i] = X_i'\beta$, where β is an unknown parameter. Conditional on $X_1, \ldots, X_n$, the elements of $Y_1, Y_2, \ldots, Y_n$ are mutually independent, each with a normal distribution and variance σ^2.

(i) Find the maximum likelihood estimator for β. Compute its mean and variance.

(ii) Write down the Cramer-Rao lower bound for the variance of an unbiased estimator for β.

(iii) Find the best linear unbiased estimator for β.

(b) Consider the following model:

$$y_i = x_i'\beta + \delta T_i + u_i,$$

where y_i is log wages, x_i includes a constant, grades completed, square of grades completed and job tenure, and T_i equals 1 if individual i receives a job training, and 0 otherwise.

(i) Suppose a researcher is interested in the effectiveness of a job training program on wages. OLS estimation of the model produces a δ coefficient that is positive and highly significant ($p - value = 0.0002$). Would you conclude that the job training program is extremely effective? Why or why not? In your response, comment on which, if any, of the classical linear regression assumptions may be violated.

(ii) Now assume that $u_i = a_i + \epsilon_i$, where a_i is unobserved ability. Derive the bias in the OLS estimator of δ due to unobserved ability. State any assumptions you require for your derivation. Does it lead to an over-estimation or under-estimation of the effect of job training?

II. For $i = 1, \ldots, n$, consider the following production function $Y_i = (\alpha K_i^\rho + (1-\alpha)L_i^\rho)^{1/\rho} + \epsilon_i$, where Y is output, K is capital, and L is labor.

(a) Assume that $E[\epsilon_i|K_i, L_i] = 0$, propose a least squares, method of moments and generalized method of moments estimator for the parameter vector $\mu = (\alpha, \rho)'$.

(b) For each of the estimators in (a), give sufficient conditions for identification of μ.

(c) Is any of the estimators in (a) efficient? If yes, explain your answer. If not, propose an efficient estimator.

(d) Using the efficient estimator, propose a Wald test for the null hypothesis: $H_0 : \alpha = \rho = 1$.

(e) Using the least squares estimator, write down the score that is used to construct the score test for the same null hypothesis in (d).

III. For $i = 1, \ldots, n$ and $t = 1, \ldots, T$, let x_{it} and β be $k \times 1$ column vectors, y_{it} and u_{it} are scalar, where $y_{it} = x_{it}'\beta + a_i + u_{it}$. Let $X_i = (x_{i1}, ..., x_{iT})$ and $u_i = (u_{i1}, ..., u_{iT})'$. Assume $E[u_{it}|X_i, a_i] = 0$ for all t.

(a) Assume $E[a_i|X_i] = 0$, $E[a_i^2|X_i] = 0$, $E[u_{it}^2|X_i, a_i] = \sigma_u^2$ and $E[u_{it}u_{i,t-\tau}|X_i] = \rho^\tau \sigma_u^2$.
Under the above scenario:

  (i) Write down the conditional variance-covariance matrix of $v_i = a_i + u_i$.

  (ii) Would the pooled OLS estimator yield a consistent estimator for β? To answer this question, show the probability limit of the estimator and give any conditions required for the result.

  (iii) Is the pooled OLS estimator asymptotically efficient? Provide a formal argument for your answer.

  (iv) If not, propose an asymptotically efficient estimator and show how you would estimate all components required.

(b) Now assume $E[a_i|X_i] \neq 0$, $E[a_i^2|X_i] > 0$, $E[u_{it}^2|X_i, a_i] = \sigma_u^2$ and $E[u_{it}u_{i,t-\tau}|X_i] = \rho^\tau \sigma_u^2$.
Under the above scenario:

  (i) Show whether or not the pooled OLS estimator is consistent for β. If not, propose an estimator that would be consistent.

  (ii) Is the consistent estimator (according to your answer in (i)) asymptotically efficient? Give a formal argument for your answer.

(c) If you would like to test whether scenario (a) or (b) are true in practice, how would you proceed? Describe the test statistic you would use and any specific issues that would arise in its application under the above assumptions.

IV. (a) State a Law of Large Numbers (LLN) and explain in words what it means.

(b) State a Central Limit Theorem (CLT)and explain in words what it means.

(c) Suppose that 40% of all auto accidents are partly caused by alcohol consumption and 30% of all auto accidents involve bodily injury. Further, of those accidents that involve bodily injury, 50% are partly caused by alcohol consumption.

  (i) What is the probability that a randomly chosen accident is partly caused by alcohol consumption and involves bodily injury?

(ii) If a randomly chosen accident was partly caused by alcohol consumption, what is the probability that it involved bodily injury?

(iii) Are the events partly caused by alcohol consumption and involved bodily injury independent? Why or why not?

(iv) Suppose that you have data on 100 randomly sampled auto accidents, including whether the accident involved bodily injury and the amount of alcohol in the blood of each driver involved in the accident. In your sample, 20% of the accidents involve bodily injury. Form a 95% asymptotic confidence interval for the proportion of accidents involving bodily injury. If you invoke the LLN or CLT to obtain the confidence interval, state precisely why you do so.

(v) Based on your answer in (iv), do you think your data are a true random sample of the population of auto accidents? Why or why not?

(vi) Suppose that you have data on 100 randomly sampled auto accidents, including whether the accident involved bodily injury and the amount of alcohol in the blood of each driver involved in the accident. How would you test the hypothesis that the events "partly caused by alcohol consumption" and "involved bodily injury" are independent? State precisely the test statistic you would use and justify your choice.

(vii) Derive the asymptotic null distribution of your test statistic in (vi). Highlight all points in your derivation where you invoke the LLN or CLT.

Econometrics Preliminary Exam

Agricultural and Resource Economics, UC Davis

August 8, 2016

There are **FOUR** questions. Answer each part of each question. All questions are weighted equally. Within each question, each part will receive equal weight in grading.

1. Given the following linear regression model

$$
\begin{aligned}
Y &= X\beta + u \\
E[u|X] &= 0 \\
E[uu'|X] &= \sigma^2\Omega = \Sigma
\end{aligned}
$$

   where $Y = (y_1, \ldots, y_T)'$ and $u = (u_1, \ldots, u_T)'$ are $T \times 1$ vectors, $X = (x_1, \ldots, x_T)'$ is an $T \times k$ matrix and β is a $k \times 1$ vector.

   (a) Show that the OLS estimator of β is unbiased.

   (b) Find the variance-covariance matrix of the OLS estimator of β.

   (c) Show that the OLS estimator of β is consistent.

   (d) Is the OLS estimator efficient? If not, propose an efficient estimator of β.

   (e) Now assuming that there is no heteroskedasticity. Suppose that the disturbance term in the above model is an AR(1) process, i.e., $u_t = \rho u_{t-1} + \epsilon_t$.

      i. Consider the case where $|\rho| < 1$. Is the OLS estimator consistent? Is yes, give sufficient conditions for its consistency.

      ii. Now what happens to your answer in (i) if $\rho = 1$?

2. Consider the following objective function, $Q(\theta) = E[m(y_i, x_i; \theta)]$, where θ and x_i are $k \times 1$ vectors and y_i is a scalar.

   (a) Given n observations of $\{y_i, x_i\}$, propose an estimator of $\theta_0 = \arg\max_{\theta\in\Theta} Q(\theta)$ and give conditions for its consistency.

   (b) Derive an expression for the sampling error, $\sqrt{n}(\hat{\theta} - \theta_0)$, and give conditions for the asymptotic normality of the estimator in (a). Make sure to give the expression for the asymptotic variance, $Avar(\sqrt{n}(\hat{\theta} - \theta_0))$. Explain briefly how the conditions you propose imply the asymptotic normality of the sampling error.

(c) Now consider the following generalizations of ordinary least squares (OLS) estimation, ridge regression ($\hat{\theta}_R$) and LASSO ($\hat{\theta}_L$). Let θ^j denote the j^{th} element of θ,

$$\hat{\theta}_R = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (y_i - x_i'\theta) + \lambda \sum_{j=1}^{k} (\theta^j)^2$$

$$\hat{\theta}_L = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (y_i - x_i'\theta) + \lambda \sum_{j=1}^{k} |\theta^j|$$

where λ is specified by the empirical researcher.

i. Show that the OLS estimator is a special case of $\hat{\theta}_R$ and $\hat{\theta}_L$.

ii. Define the population objective function for $\hat{\theta}_R$ and $\hat{\theta}_L$.

iii. Now consider the conditions you specified in (a) and (b). Do $\hat{\theta}_R$ and $\hat{\theta}_L$ fulfill these conditions? If not, which conditions do they violate? Be precise and clear in your answer.

3. Consider the following estimation problems.

(a) Let W_1 and W_2 be dependent random variables, where $W_2 = \alpha W_1 + \eta$, and η and W_1 are i.i.d. N(0,1) random variables. Suppose you observe an i.i.d. random sample of $\{W_{1i}, W_{2i}\}_{i=1}^{n}$, propose an estimator of α and give conditions for its consistency and asymptotic normality. Is the estimator you proposed efficient?

(b) Now consider the two-period dynamic panel data model, where

$$Y_{i2} = \rho Y_{i1} + X_{i2}'\beta + \alpha_i + u_{it},$$

Given an i.i.d. random sample of $\{Y_{i1}, Y_{i2}, X_{i1}, X_{i2}\}_{i=1}^{n}$, can you use insights from the estimation strategy you proposed in (a) to estimate β? State any additional assumptions you require. Define the estimator you propose and its objective function.

4. This question is based on a 1996 paper by Steve Levitt in the Quarterly Journal of Economics, which addresses the effect of imprisonment on violent crime. Levitt's data are measured annually at the state level, i.e., one observation for each U.S. state in each year from 1980-1993. Consider the equation:

$$gcriv = \beta_0 + \beta_1 gpris + \beta_2 gincpc + \varepsilon$$

where *gcriv* denotes the annual growth rate in violent crime, *gpris* denotes the annual growth rate in the number of prison inmates per resident, and *gincpc* denotes per capita income.

The parameter of interest is β_1, which measures the marginal effect of an increase in imprisonment on the crime rate.

(a) Would ordinary least squares (OLS) produce a consistent estimate of β_1? Justify your answer clearly in words.

(b) Write the model in the notation from question 1. Clearly define each term and show which property of the model determines whether OLS is consistent for β_1.

(c) Levitt observes two additional variables: (i) *final1* is a dummy variable denoting a final decision in the current year on legislation to reduce prison overcrowding, and (ii) *final2* is a dummy variable denoting a final decision in the last two years on legislation to reduce prison overcrowding. Levitt uses these variables to instrument for *gpris*. What would need to be true for this instrumental variables (IV) estimator to produce a consistent estimate of β_1? Justify your answer clearly in words.

Using Levitt's data, we estimated the model by OLS and IV. The STATA output is shown on the last page of the exam.

(d) Both the OLS and IV estimation uses the robust command to correct the standard errors for heteroscedasticity. How might the results differ if this correction were not done?

(e) The IV estimate of β_1 is a larger negative number than the OLS estimate. Explain in words whether this result makes sense.

(f) Describe how you would check the strength of *final1* and *final2* as instruments for imprisonment. Explain the implications for the results if the instruments are weak.

```
. regress gcriv gpris gincpc, robust

Linear regression                                      Number of obs =     714
                                                       F(  2,   711) =   15.62
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0461
                                                       Root MSE      =  .08661

------------------------------------------------------------------------------
             |               Robust
       gcriv |     Coef.  Std. Err.     t   P>|t|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
       gpris |-.1954158  .0531582  -3.68  0.000    -.2997817   -.09105
      gincpc | .4667109  .1468838   3.18  0.002     .1783331    .7550887
       _cons | .0043254  .0106637   0.41  0.685    -.0166106    .0252615
------------------------------------------------------------------------------


. ivregress gcriv (gpris=final1 final2) gincpc, robust

Instrumental variables (2SLS) regression               Number of obs =     714
                                                       F(  2,   711) =    8.82
                                                       Prob > F      =  0.0002
                                                       R-squared     =       .
                                                       Root MSE      =  .10484

------------------------------------------------------------------------------
             |               Robust
       gcriv |     Coef.  Std. Err.     t   P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       gpris |-1.082207  .3154422  -3.43  0.001    -1.701517   -.4628977
      gincpc | .3798519  .2007459   1.89  0.059    -.0142738    .7739776
       _cons | .0684567  .0255884   2.68  0.008     .0182188    .1186946
------------------------------------------------------------------------------
Instrumented:  gpris
Instruments:   gincpc final1 final2
------------------------------------------------------------------------------
```