2007 Ph.D. Econometrics Prelim Examination

Department of Agricultural and Resource Economics
University of California, Davis
July 5, 2007

Please answer all FIVE questions. Each question will be weighted equally. Within each question, each part will be weighted equally.

Question One

Consider the linear regression model $y = X\beta + u$ and the estimator $\hat{\beta} = (X'X)^{-1}X'y$, where *X* is an *N*×*k* full rank matrix of constants (i.e. are nonstochastic). Suppose $E[u] = 0$ and $V[u] = \Omega$ where $\Omega = Diag[\sigma_i^2]$. State clearly any additional assumptions you need to make. There is no need to justify the use of any laws of large numbers or central limit theorems.

(a) Show that $\hat{\beta}$ is consistent for β.

(b) Obtain the limit distribution of $\sqrt{N}(\hat{\beta} - \beta)$.

(c) Provide a consistent estimate for the variance matrix given in part (b) when σ_i^2 is unknown and a functional form for σ_i^2 is unknown.

(d) Would there be any advantage to actually knowing the functional form for σ_i^2? A brief answer will do.

(e) Show using matrix calculus that $\hat{\beta} = (X'X)^{-1}X'y$ minimizes $u'u$.

Question Two

The STATA output on the next page provides summary statistics and regression results using 1976 data on young adult males with variables:

wage = hourly wage
lnwage = natural logarithm of hourly wage
schooling = highest grade of completed schooling
experience = years of work experience

You are to answer the following questions given the attached output.
You can use the output in a way that gets the answer as quickly as possible.

(a) Give a 95% confidence interval for the mean hourly wage of young adult males in 1976.

(b) Comment on the individual and joint statistical significance of the regression slope coefficients.

(c) Comment on the magnitude of the regression slope coefficients. Are these large or small effects? Provide some explanation.

(d) What is the R^2 for this regression?

(e) Provide a 95% confidence interval for the population mean log hourly wage of a young adult male in 1976 with 12 years of schooling and 10 years of experience.

```
. sum wage lnwage schooling experience

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |        60    5.917719    2.883179   2.080067   18.00051
      lnwage |        60    1.673003    .4631765      .7324     2.8904
   schooling |        60       12.95    2.770456          6         18
  experience |        60    9.866667    3.553307          2         17

. reg lnwage schooling experience

      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  2,    57) =   19.68
       Model |  5.16971452     2  2.58485726           Prob > F      =  0.0000
    Residual |  7.48770287    57  .131363208           R-squared     =
-------------+------------------------------           Adj R-squared =
       Total |  12.6574174    59  .214532498           Root MSE      =  .36244

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   schooling |   .1146237   .0185691     6.17   0.000     .0774396    .1518077
  experience |   .0504573   .0144781     3.49   0.001     .0214655    .0794491
       _cons |   -.309219   .3283185    -0.94   0.350    -.9666654    .3482274
------------------------------------------------------------------------------

. matrix varols = e(V)

. matrix list varols

symmetric varols[3,3]
             schooling  experience       _cons
 schooling   .00034481
experience   .00010711   .00020961
     _cons  -.00552214  -.00345526   .10779303

. lincom _cons + 12*schooling + 10*experience

 ( 1)  12 schooling + 10 experience + _cons = 0

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.570838   .0497712    31.56   0.000     1.471173    1.670504
------------------------------------------------------------------------------
```

Question Three

Consider the model in Question Two. Suppose you want to use this model to estimate the effect on wages of a change in education policy that would increase schooling levels.

(a) Suppose some individuals in your sample have higher IQ than others, which causes them to have higher schooling levels and higher income than others. Given this information, what will be the properties of the OLS estimate of the coefficient on schooling presented in Question Two?

(b) Continuing from (a), suppose you do not observe IQ but you do observe an instrumental variable *z*. What properties must the instrumental variable possess for it to be useful in producing a consistent estimate of the coefficient on schooling with a small finite-sample bias?

(c) Suppose you learn that the individuals in your sample did *not* choose their schooling levels. As part of a government experiment, each individual was assigned a schooling level based on the day of the week on which they were born (i.e., those born on a Monday received schooling up to 10th grade, those born on a Tuesday received schooling up to 11th grade, etc). Given this information, how would you estimate the effect of schooling on wages? Justify your answer.

(d) Suppose you learn of the government experiment in (c), but you note that many of the individuals in your sample received a schooling level different from that to which they were assigned. Explain how you would test whether these deviations from the assigned education level cause bias in your estimate. Write down the test statistic you would use and state its asymptotic null distribution.

Question Four

Annual data for the years 1959-1994 on consumption expenditures (Ct), wage income (Wt), and profit income (Pt), all in billions of 1992 dollars, were used to estimate the model

$$C_t = \beta_1 W_t + \beta_2 P_t + \beta_3 C_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim N(0,\sigma^2)$$

with the following results:

```
Model 1: OLS estimates using the 35 observations 1960-1994
Dependent variable: Ct
  VARIABLE           COEFFICIENT        STDERROR      T STAT    P-VALUE
  Wt                    0.519            0.0768        6.763  <0.00001 ***
  Pt                    0.458            0.0645        7.104  <0.00001 ***
  Ct_1                  0.279            0.1079        2.590   0.01433 **

  Mean of dependent variable = 2851.68
  Standard deviation of dep. var. = 927.12
  Sum of squared residuals = 33268.8
  Standard error of residuals = 32.24
  Unadjusted R-squared = 0.999894
  Adjusted R-squared = 0.999887
  F-statistic (3, 32) = 100615 (p-value < 0.00001)
  Durbin-Watson statistic = 0.867
  First-order autocorrelation coeff. = 0.565


LM (Breusch-Godfrey) test for first-order autocorrelation
OLS estimates using the 34 observations 1961-1994
Dependent variable: Ehat
  VARIABLE           COEFFICIENT        STDERROR      T STAT   P-VALUE
  Wt                   -0.009             0.065         -0.139   0.89063
  Pt                   -0.007             0.055         -0.124   0.90218
  Ct_1                  0.012             0.092          0.130   0.89771
  Ehat_1                0.566             0.151          3.756   0.00074 ***

Unadjusted R-squared = 0.32
Test statistic: TR^2 = 10.88,
with p-value = P(Chi-square(1) > 10.88) = 0.000974



Model 2: Cochrane-Orcutt estimates using the 34 observations 1961-1994
Dependent variable: Ct

Iterative calculation of rho:
                 ITER       RHO        ESS
                   1      0.565     21845.1
                   2      0.600     21800.6
                   3      0.602     21800.5
                final     0.602

  VARIABLE           COEFFICIENT        STDERROR      T STAT   P-VALUE
  Wt                    0.586             0.0680         8.622  <0.00001 ***
  Pt                    0.518             0.0570         9.084  <0.00001 ***
  Ct_1                  0.181             0.0950         1.914   0.06487 *
```

(Note that Ct_1 denotes C_{t-1} and Ehat_1 denotes the lagged residual.) Answer the questions below in a logical way; algebraic proof is not necessary.

(a) Since wages, profits, and lagged consumption might be expected to move together over time, multicollinearity is possible. Verbally, if multicollinearity were the only problem in a hypothetical regression and the model reflected all information available, what effect(s) would it have on the calculated estimates and their standard errors? (Specifically, would the coefficient estimates still be efficient? Would the hypothesis tests be valid? Just assert the answers with whatever informal justification you can muster -- do not attempt to prove anything.)

(b) Do you see evidence that multicollinearity has had an effect on the estimates reported? What is the basis of your conclusion?

(c) If you could calculate anything, what would you examine to more formally determine whether or not multicollinearity is a problem in this regression (briefly)?

(d) As an aside, for this part only, suppose the errors in the OLS regression were not serially correlated but instead were (only) heteroskedastic. Briefly, what are the consequences of this heteroskedasticity for hypothesis tests using the conventionally calculated OLS estimates?

(e) We suspect serial correlation of the classic AR(1) form

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \;\; u_t \sim iid \;\; N(0,\sigma^2)$$

Formally test for this serial correlation, being explicit about how you tested and why. Report the results of your test in a precise, complete, and unambiguous way.

(f) Formally test the hypothesis that the short-run marginal propensity to consume from wage income is equal to .65 using the regression results supplied above. Be explicit about whether you are using the OLS, LM test, or Cochrane-Orcutt estimates in your test, and why. Write the null and alternate hypotheses, the test statistic used and its distribution, and write your conclusion in plain but specific language understandable by non-econometricians.

Question Five

Suppose *y_i* is a random variable measuring a count (e.g., the number of children in a household). Let *x_i* denote a vector of explanatory variables including a constant. A common model for $y_i \mid x_i$ uses the Poisson distribution, i.e.,

$$y_i \mid x_i \sim \text{Poisson}(\lambda_i)$$

where $\lambda_i = \exp(x_i'\beta_0)$. The probability density function for the Poisson distribution is

$$f(y \mid x_i, \beta) = \frac{\lambda_i^y \exp(-\lambda_i)}{y!}.$$

(a) Write down the log-likelihood function for the unknown parameters β_0. State any assumptions that you make.

(b) Write down the moment conditions for an asymptotically efficient GMM estimator of β_0. Justify your answer.

(c) Consider a GMM estimator based on the moment condition $E_0\left(z_i(y_i - \exp(x_i'\beta_0))\right) = 0$, where *$z_i$* is different from *$x_i$*. Why might an applied econometrician choose this GMM estimator instead of the one in (b)?

2007 Ph.D. Econometrics Prelim Examination

Department of Agricultural and Resource Economics
University of California, Davis
August 20, 2007

Please answer all FOUR questions. Each question will be weighted equally. Within each question, each part will be weighted equally.

Question One

Consider the simple regression model $y_i = x_i\beta + \varepsilon_i$, where the scalar x_i is nonstochastic and $\varepsilon_i \sim iidN(0,\sigma^2)$. Suppose β is estimated using the formula $\tilde{\beta} = \bar{y}/\bar{x}$.

(a) Show that the estimator $\tilde{\beta}$ is unbiased.

(b) Derive the variance of the estimator $\tilde{\beta}$.

(c) Show that the estimator $\tilde{\beta}$ is consistent.

(d) Write down the probability distribution of the estimator $\tilde{\beta}$.

(e) Would you recommend using the estimator $\tilde{\beta}$ instead of OLS? Why or why not? (A logical answer in words is sufficient.)

Question Two

Consider the linear regression model $y = X\beta + u$, where X is an $N\times k$ full rank matrix of constants (i.e. are nonstochastic), and $u \sim N(0,\sigma^2 I)$. We consider estimation and testing given linear restrictions $R\beta - r = 0$.

(a) Suppose that k = 5 and we wish to test $H_0 : \beta_2 + \beta_3 - 2 = 0, \quad \beta_4 = 1$. Give R and r in this case.

[Note: The remainder of this question is to be answered for general R and r.]

(b) Suppose we wish to test $H_0 : R\beta - r = 0$ based on the estimator $\hat{\beta} = (X'X)^{-1}X'y$. Derive an appropriate test statistic, assuming σ^2 is known. State how you would use this test statistic to reject H_0 at 5%.

(c) Now suppose we wish to obtain the restricted least squares estimator that imposes H_0. Give the objective function for the restricted least squares estimator. [Note: There is no need to derive the estimator. Just give the objective function].

(d) The restricted least squares estimator $\tilde{\beta}$ can be shown to be

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$$

where $\hat{\beta} = (X'X)^{-1}X'y$. Suppose now that X is potentially stochastic and that sufficient assumptions have been made to ensure that $\hat{\beta}$ is consistent for β. Obtain the probability limit of $\tilde{\beta}$, stating any extra assumptions needed.

(e) Given your answer in (d), under what conditions will $\tilde{\beta}$ be consistent for β?

Question Three

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 system of equations:

$$gcriv = \beta_{11} + \gamma_{12}\, gpris + \beta_{12}\, gincpc + \varepsilon_1$$

$$gpris = \beta_{21} + \gamma_{21}\, gcriv + \beta_{23}\, final1 + \beta_{24}\, final2 + \varepsilon_2$$

where *gcriv* denotes the annual growth rate in violent crime

gpris denotes the annual growth rate in the number of prison inmates per resident

gincpc denotes per capita income

*final*1 is a dummy variable denoting a final decision in the current year on legislation to reduce prison overcrowding

*final*2 is a dummy variable denoting a final decision in the last two years on legislation to reduce prison overcrowding.

Define y = (*gcriv* *gpris*) and X = (1 *gincpc* *final1* *final2*) and specify the moment condition

$$E(y \mid X) = -XB\Gamma^{-1}$$

where

$$\Gamma = \begin{bmatrix} 1 & -\gamma_{21} \\ -\gamma_{12} & 1 \end{bmatrix}, \qquad B' = \begin{bmatrix} \beta_{11} & \beta_{12} & 0 & 0 \\ \beta_{21} & 0 & \beta_{23} & \beta_{24} \end{bmatrix}.$$

(a) Are the parameters of the model identified? Justify your answer.

(b) The model specifies that prison overcrowding legislation affects violent crime in a particular way. In words, explain why this specification implies that the parameter γ_{12} is or is not identified.

Using Levitt's data, I estimated the first equation in this system by OLS and IV. The STATA output follows.

```
. 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
----------------------------------------------------------------------

. est store ols1
. ivreg 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
---------------------------------------------------------------------

. est store iv1
. hausman iv1 ols1

                 ---- Coefficients ----
             |    (b)        (B)          (b-B)    sqrt(diag(V_b-V_B))
             |    iv1           .        Difference         S.E.
-------------+---------------------------------------------------------
       gpris |  -1.082207   -.1954158      -.8867915      .3109309
      gincpc |   .3798519    .4667109      -.086859       .136836
-----------------------------------------------------------------------
   b = consistent under Ho and Ha; obtained from ivreg
   B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic
                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        8.15
                Prob>chi2 =      0.0170
```

Question 3 continued

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

(d) The IV estimate of γ_{12} is a larger negative number than the OLS estimate. Explain in words whether this result makes sense.

(e) Compare the OLS and IV estimates of γ_{12}. Do the results suggest that the variable *gpris* is endogenous in the first equation? Explain.

(f) Levitt uses the variables *final*1 and *final*2 as instruments for imprisonment. Describe how you would check the strength of these instruments, and explain the implications for the results if the instruments are weak.

Question Four

Suppose y_i is a random variable measuring the duration of an event (e.g., the length of time until a customer walks into a store). Let x_i denote a vector of explanatory variables including a constant. A common model for $y_i \mid x_i$ uses the exponential distribution, i.e.,

$$y_i \mid x_i \sim Exp(\lambda_i)$$

where $\lambda_i = e^{x_i'\beta_0}$. The probability density function for the exponential distribution is

$$f(y \mid x_i, \beta) = \frac{1}{\lambda_i} e^{-y_i/\lambda_i},$$

and the distribution has the property that $E(y_i \mid x_i) = \lambda_i$.

(a) Write down the log-likelihood function for the unknown parameters β_0. State any assumptions that you make.

(b) Write down the moment conditions for an asymptotically efficient GMM estimator of β_0. Justify your answer.

(c) Consider a GMM estimator based on the moment condition $E_0\left(z_i(y_i - e^{x_i'\beta_0})\right) = 0$, where z_i is different from x_i. Why might an applied econometrician choose this GMM estimator instead of the one in (b)?

Econometrics Preliminary Examination

July 3, 2008

Answer each part of each question.

1. X and Y are independent normal random variables. Recall that the probability density function of a normal random variable is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

for $-\infty<x<\infty$, and that the moment-generating function of a normal random variable is given by

$$M(t) = \exp\left(\mu t + (1/2)t^2\sigma^2\right).$$

What is the probability distribution of the new random variable formed by

$$W = 6X - 4Y?$$

Give a complete proof of each result you claim for *W.*

2. A random variable X is distributed as Uniform(*a,b*). Both *a* and *b* are unknown parameters.

   a. With two observations on X, how would you estimate these two parameters? Defend your answer.

   b. How would your answer to part (a) change if you had n observations?

   c. You wish to test the hypothesis that *b-a>1*. Your strategy is to reject the hypothesis if the difference in your estimates exceeds 1.

   Suppose the true value of *b-a* is exactly 1. What is the probability of a type-1 error for your test?

3. Two equations of interest to you are given below:

$$y_{1t} = \beta_1 + \beta_2 X_t + \beta_3 W_t + e_{1t}$$
$$y_{2t} = \delta_1 + \delta_2 R_t + \delta_3 M_t + e_{2t}$$

a. Explain why Seemingly Unrelated Regressions (SUR) estimation might be preferred to Ordinary Least Squares (OLS) for obtaining estimates of the six parameters. Under what circumstances is there no gain from using SUR?

b. Assume that you estimated the parameters of these two equations using SUR. Explain in detail how you would test the hypothesis that the vector of parameters in the first equation is equal to the vector of parameters in the second equation, i.e., how you would test the joint hypothesis that β1=δ1, β2=δ2, and β3=δ3. Give the relevant mathematical expressions and a complete justification in words (i.e., not a complete proof or derivation) for your approach; do not base your answer on Stata or other computer commands.

c. A colleague of yours viewed the two-equation system as a single equation, "stacking" the system as follows:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} \iota & X & W & 0 & 0 & 0 \\ 0 & 0 & 0 & \iota & R & M \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$$

where ι denotes a column vector of 1s. Your colleague estimated the six-parameter model above using OLS.

Are your colleague's estimates unbiased? Are they consistent? Are they efficient? In each case, if you answer yes, then state any assumptions you make OR if you answer no, then state conditions needed to answer yes.

d. Suppose that your colleague wanted to test the same joint hypothesis as you, but within the OLS framework. Give complete details for how such a test might be performed.

Is there anything wrong with this procedure? Express your answer in terms of the size and power of the test.

e. You are interested in only the first equation from the two-equation system in question 1. Explain why it is possible to claim that OLS estimates are efficient, if this is the only equation under study, yet they are inefficient, if you are studying two equations.

f. Suppose the error terms e_{1t} and e_{2t} are *iid* and jointly normally distributed with known covariance matrix. In words, compare the properties of the SUR and maximum likelihood estimators of the six parameters. Suppose instead the errors have an unknown covariance matrix. In words, compare the properties of the SUR and maximum likelihood estimators of the six parameters.

4. Consider the following model:

$$y = \beta_1 + \beta_2 x_{t2} + ... + \beta_K x_{tK} + u_t$$

$$u_t = \sqrt{h_t} v_t \ ,$$

$$v_t = \rho v_{t-1} + \varepsilon_t \ , |\rho| < 1 ,$$

where the explanatory variables (each *x*) are independent of ε_t, for all t, and h_t, is a function of the explanatory variables. The process $\{\varepsilon_t\}$ has a zero mean and a constant variance σ_ε^2 and is serially uncorrelated.

a. Suppose $\rho = 0$; that is, there does not exist autocorrelation in the disturbance terms. Prove that the ordinary least squares estimator of the coefficients is consistent.

b. Now suppose that $\rho \neq 0$, but that h_t is known. Obtain a feasible generalized least squares estimator of the coefficients.

c. What are the properties of the estimator in part (b)? Explain.

d. What are the consequences of assuming that there is no autocorrelation, given the part (b) setup? Be specific.

e. Now suppose that $\rho \neq 0$ and h_t is unknown. Obtain a feasible generalized least squares estimator of the coefficients. Also explain how you can obtain an estimator of h_t.

f. What are the properties of the estimators in part (e)? Explain.

g. Suppose ρ=1. Describe in words the implications of this fact for the OLS estimator of β.

5. The four equations below represent a dynamic model of housing prices:

$$
\begin{aligned}
&P_t \equiv 0.7H_t + 0.2S_t + 0.1F_t \\
&H_t = \alpha P_t + \varepsilon_{1t} \\
&S_t = \beta \Delta P_t + \varepsilon_{2t} \\
&\Delta P_t \equiv P_t - P_{t-1}
\end{aligned}
$$

where P_t denotes an aggregate housing price, H_t denotes prices paid by households, S_t denotes prices paid by speculators, and F_t denotes prices paid by foreigners. All prices are indices. For your convenience, the share weights in the first equation are fixed and known. ε_{1t} and ε_{2t} denote stochastic errors.

(a) What are the endogenous variables? What are the exogenous variables? What is the identification status of each of the stochastic equations?

(b) Comment on the identification status of the identities (the non-stochastic equations).

(c) Without framing the test statistic explicitly, is it possible to test whether or not P_{t-1} should be a regressor in the second equation, i.e. to test whether $\gamma=0$ in the equation

$$H_t = \alpha P_t + \gamma P_{t-1} + \varepsilon_{1t} \text{ ?}$$

(d) Without framing the test statistic explicitly, is it possible to test whether or not both P_{t-1} and F_t should be included as regressors in the second equation? (Briefly comment on the relevant issues.)

(e) How many reduced-form equations will there be for this system?

(f) What variables will appear on the left-hand and right-hand sides of each of the reduced-form equations? (Do not solve for the reduced-form, just tell what variables will be in each equation.)

(g) What instruments(s) would you use to consistently estimate the parameter α, and how would you calculate the instrumental variables estimator?

(h) What are the two conditions required for acceptable instruments?

(i) Verbally but precisely, what additional condition is required for asymptotic efficiency in the class of estimators using the *same* information set?

(j) Instrumental variables is often thought of as regression on a "hat" variable, in which a troublesome right-hand side variable is replaced by a predicted value using instruments. In this sense, the method is often confused with regression on a "proxy variable", in which the right-hand side variable in a regression is replaced by some other variable. In general (not particular to this model), what is the difference in computational formulas and properties between two stage least squares and a regression model that uses a proxy variable?

(k) Why would we ever use three-stage least squares for this (or any) model?

Econometrics Preliminary Examination

August 18, 2008

Answer each part of each question.

1. You observe characteristics Y_1 and Y_2 for *n* individuals. Hence, your data set consists of *2n* random variables,

$$\mathrm{Y}_{\mathrm{ij}} = \mathrm{Y}_{11},\ \mathrm{Y}_{12},\ \ldots\ ,\ \mathrm{Y}_{1n}, \mathrm{Y}_{21},\ \mathrm{Y}_{22},\ \ldots .,\ \mathrm{Y}_{2n},$$

where i=1,2 and j=1,2,3,…,n.

You believe that all *2n* variables are independent, that

$$Y_{1j} \sim N(\mu_1, \sigma_1^2)\ ,$$

and that

$$Y_{2j} \sim N(\mu_2, \sigma_2^2).$$

Recall that the probability density function of a normal random variable is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

for $-\infty < x < \infty$, and that the moment-generating function of a normal random variable is given by

$$M(t) = \exp\left(\mu t + (1/2)t^2\sigma^2\right).$$

a. Give the likelihood function for your sample of *2n* observations.

b. Derive maximum-likelihood estimates for the four parameters.

c. Is your estimate of $E(Y_{1i})$ unbiased? Give a complete proof.

d. Is your estimate of $E(Y_{1i})$ consistent? Give a complete proof.

e. Is your estimate of $V(Y_{1i})$ unbiased? Give a complete proof.

f. Is your estimate of $V(Y_{1i})$ consistent? Give a complete proof.

g. Explain how your answers for (a), (b), and (c)-(f) change when you make the assumption that $V(Y_{1i})=V(Y_{2j})$ for every *i,j*. You do not need to repeat every detail of each particular proof, but your answers should be sufficiently complete to illustrate how those proofs are affected.

h. Continuing with the setup in part (g), what is the maximum-likelihood estimate of the difference between E(Y_{1i}) and E(Y_{2i})? What is the probability distribution of your estimate of the difference? Be specific about the family of distributions you expect, as well as the mean and variance of the random variable representing your estimate.

i. Explain in complete detail how you would use the result in (h) to test the hypothesis that the difference is +2, against the alternative hypothesis that the difference is not equal to +2.

j. Suppose the true difference is zero. Give an expression for the power of your test.

k. What would you do if you did not make the assumption about equal variances in part (g), but wanted to test the same hypothesis as in part (i)? Your answer should be sufficiently complete to indicate all of the steps you would undertake.

l. A final concern with these data relates to your assumption that the variables are independent. As a check, you calculate the sample covariance between Y_{1i} and Y_{2i}, using the formula

$$\hat{\sigma}_{12} = \frac{\sum_{i=1}^{n}(Y_{1i} - \overline{Y}_1)(Y_{2i} - \overline{Y}_2)}{n-1}.$$

Does this provide a consistent estimator? Is it unbiased? Give proofs for each of your answers.

2. Consider the model $Y_t = \beta X_t + u_t$ in which there is no intercept and X_t is a scalar.

a. Two alternative optimality conditions, $\sum \hat{u}_t = 0$ and $\sum X_t \hat{u}_t = 0$, yield two different estimators for β in this model. Derive these two estimators for β.

b. Show that the two different estimators derived in part (a) are unbiased. State all the assumptions you make to prove your result.

c. Show that one of the fitted regression lines goes through the average point $(\overline{X}, \overline{Y})$, but the other one does not.

d. Prove that one of the above estimators is the least squares estimator of β.

e. Show that the least squares estimator of β is BLUE (the best linear unbiased estimator).

f. Explain why the Gauss-Markov theorem is applicable here, and prove that the least squares estimator is superior to the other derived estimator.

g. It is often the case that a linear combination of two estimators is preferred to either separate estimator. Comment on whether or not a linear combination of the two estimators from part (a) would produce an improved estimate of β.

3. The following two equation supply and demand model has been proposed by a distinguished agricultural economist and so is taken as absolute truth:

$$(1) \quad q_t = \alpha_1 + \alpha_2 p_t + u_t, \quad u_t \sim iid(0, \sigma_u^2)$$

$$(2) \quad p_t = \beta_1 + \beta_2 q_t + v_t, \quad v_t \sim iid(0, \sigma_v^2)$$

where q_t is the equilibrium quantity, p_t is the equilibrium price, the $\alpha's$ and $\beta's$ are parameters to be estimated, and u_t and v_t are random errors that are contemporaneously and serially uncorrelated.

a. The equation errors u_t and v_t are uncorrelated across equations (i.e., they are contemporaneously uncorrelated). In this case, would Ordinary Least Squares estimates of the parameters of equations (1) and (2) be consistent? Formally, why or why not?

b. Solve algebraically for the two reduced form equations. Do not use matrix algebra, use the scalar symbols in the equations above.

c. Does the fact that the structural form equation errors u_t and v_t are uncorrelated across equations imply that the reduced form errors are uncorrelated across equations? Show explicitly.

d. Regardless of your answer to part (c) above, assume for this part that the reduced form errors are correlated across equations. How would you estimate the parameters of the reduced form, and why?

Econometrics Preliminary Examination

July 2, 2009

Answer each part of each question. Each part of every question is weighted equally.

1. Y is a random variable with unknown probability density function f(y). It is known that Y is a continuous random variable, and that its range is from a to c, both finite constants.

    a. While you don't know f(y), you will be provided with a random sample from this probability distribution. Prove that the sample mean from such a random sample provides an unbiased estimate of the population mean.

    b. Prove that the sample mean is consistent.

    c. Prove that the sample mean is BLUE.

    d. You are also interested in estimating the probability that Y is less than b, for a<b<c. Explain how you can use your random sample to give an estimate of this probability.

    e. Prove that your estimate is unbiased and consistent.

    f. Explain in words how you could use random sampling to estimate the probability that the sample mean will be less than b, and how you could use asymptotic theory to estimate this same probability.

2. X is a normal random variable. It is known that either X ~ N(0,1) or X ~ N(1,4). You wish to test the null hypothesis that X ~ N(0,1), and to do so, you will use one of two critical regions, based on a single observation for X. One choice is to reject the null hypothesis if X > 1.645. The other choice is to reject the null hypothesis if |X|>1.96.

    a. Evaluate the two alternative choices for your critical region. Which would you prefer to use, and why? Be specific. If you make use of any normal probabilities, use the normal probability table provided with the exam to make sure your probability statements are accurate.

    b. Indicate whether there is a third possible critical region you could specify, that would be preferred to either. Give a justification for your answer.

3. Suppose a researcher estimates the model

$$y = X_1\beta_1 + u \qquad (1)$$

but the true model is

$$y = X_1\beta_1 + X_2\beta_2 + v \qquad (2)$$

where y is $nx1$, X_1 is nxk_1, β_1 is k_1x1, X_2 is nxk_2, β_2 is k_2x1, and u and v are disturbance terms. Assume that the assumptions of the Gauss-Markov Theorem hold for the true model, and furthermore assume that the explanatory variables are exogenous (or else treat all expectations as conditional on the X's).

a. Show that the ordinary least squares estimator of β_1 from model (1) is biased and provide an interpretation of the bias term in terms of least squares regression. Also prove whether or not the OLS estimator of β_1 is consistent. Is it efficient? Explain.

b. Suppose that a researcher obtained the following estimates by applying ordinary least squares to model (2), the true model

$$\hat{y}_t = 0.5X_{1t} + 1.2X_{2t}$$

where y represents average worker productivity at manufacturing firms and depends on average hours of training (X_{1t}) and average worker ability (X_{2t}). Assume that X_1 and X_2 are negatively correlated. The researcher estimates model (1) omitting X_2 and obtains

$$\hat{y}_t = 1.4X_{1t}.$$

On average do we expect the bias to be positive or negative? Is this result possible?

c. A typical textbook reason for autocorrelation is that slowly changing relevant explanatory variables have been omitted; however, the texts also state that the ordinary least squares estimator remains unbiased in the presence of autocorrelation. Rationalize this apparent contradiction.

d. Since all models have omitted variables, ordinary least squares estimators are always biased. True or false. Explain.

e. How would you determine if your model is valid?

4. Consider the consumer demand model for a single good

   $$q_t = \beta_0 + \beta_1 p_t + \beta_2 x_t + \varepsilon_t ,$$

   where q_t denotes the log of quantity, p_t denotes the log of price, and x_t denotes the log of income. You have data on q_t, p_t, and x_t. The parameter β_1 can be interpreted as the price elasticity of demand and β_2 as the income elasticity of demand. Suppose that the variation in price across the sample stems from both supply and demand differences across observations. Suppose also that you observe an instrumental variable z_t.

   a. Would OLS consistently estimate β_1? If your answer is no, would OLS over- or under-estimate β_1 on average? Justify your answer.

   b. What properties must the instrumental variable z_t possess for it to be useful in producing consistent estimates of β_1 with a small finite-sample bias?

   c. Write down a set of moment conditions that could be used to consistently estimate β_0, β_1, and β_2 using GMM.

   d. Write down the statistic you would use to test the null hypothesis that the variation in price across the sample stems from supply sources only. State the asymptotic null distribution of your statistic.

   e. Suppose you were unable to reject the null hypothesis in (d). How would you estimate β_0, β_1, and β_2? Justify your answer.

   f. Suppose that variation in price across the sample stems from demand sources only. Describe the implications of this supposition for demand curve estimation?

5. Suppose a fellow student has time series data on some exogenous variable *x* and another variable *y* and wants to understand how *x* affects *y*. The student specifies the model

   $$y_t = x_t \beta + \varepsilon_t$$

   a. Suppose you know that $x_t = x_{t-1} + v_t$, where v_t is *iid* with mean zero. Explain the implications of this fact for the model specified by your fellow student. Describe how you would specify the model and how you would interpret the estimates. In words, justify your chosen estimation approach.

   b. Suppose you know that x_t follows the AR(1) process $x_t = \rho x_{t-1} + v_t$, where v_t is *iid*. Derive the autocorrelation function for x_t.

Cumulative Normal Probability Tables (Z-Values)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

Econometrics Preliminary Examination

August 17, 2009

Answer each part of each question. Each part of every question is weighted equally.

1. A single fair die is tossed. Let X denote the random variable corresponding to the number showing. The outcome for X, denoted by x, is observed, and then a fair coin is flipped x times. Let Y denote the number of times the coin shows Heads.

   a. Suppose this experiment is performed, and Y=4. Find the conditional probability distribution for X, given that Y=4.

   b. Find the expected value of X, given that Y=4.

   c. Find the expected value of Y.

   d. Are X and Y independent random variables?

2. Y is a random variable with unknown probability density function f(y). It is known that Y is a continuous random variable, and that its range is from a to c, both finite constants. *These two constants (a and c) are unknown.*

   a. While you don't know f(y), you will be provided with a random sample from this probability distribution. How would you estimate a and c? Be specific.

   b. Are your estimates unbiased? Give a justification for your answer.

   c. Are your estimates consistent? Give a justification for your answer.

3. Consider the model $S_i = \beta_1 + \beta_2 A_i + \varepsilon_i$, in which S_i is the average sales in industry i and A_i denotes the average advertising budget for that industry i=1,2,…,n. The industry average is computed for all the firms in the industry. Assuming independence across firms in an industry, the variance of ε_i equals σ^2 / N_i, where N_i is the known number of firms in the industry. You have data on S_i, A_i and N_i.

   a. Are OLS estimators of β_1 and β_2 unbiased? Are OLS estimators BLUE? Provide proofs.

   b. Carefully describe step-by-step how you would obtain weighted least squares (WLS) estimates of β_1 and β_2 . In particular, explain how the OLS method can be used to obtain WLS estimates.

c. Are the WLS estimates biased or unbiased? Provide a proof in your response.

d. Show that WLS estimators are consistent?

e. Are they BLUE? Justify your answer.

f. Suppose you know that the heteroscedasticity is a function of N_i , but you do not know the functional form. Describe the regression(s) you should run. What are the properties of your estimators?

g. How would you test the hypothesis of no heteroscedasticity? Write down the test statistic, state the null and alternative hypotheses, and state the null distribution of the test statistic.

4. Suppose you have a sample of size n on the random variables y_i and x_i, and you model the distribution of $y_i \mid x_i$ as

$$y_i \mid x_i \sim N\left(x_i'\beta_0, (x_i'\beta_0)^2\right).$$

Note: the probability density function for a random variable $z \sim N(\mu, \sigma^2)$ is $f(z) = (2\pi\sigma^2)^{-1/2} \exp(-0.5\sigma^{-2}(z-\mu)^2)$.

a. Write down the log likelihood function for the unknown parameters β_0. State any assumptions that you make.

b. Suppose the true distribution of $y_i|x_i$ is nonnormal. How would you interpret the parameter value β_0, which is the probability limit of your maximum likelihood estimate?

c. Would OLS produce a consistent estimate of β_0? Justify your answer.

d. Explain why maximum likelihood provides an asymptotically efficient estimate of β_0 relative to a GMM estimator based on the moment condition $E_0\left(x_i(y_i - x_i'\beta_0)\right) = 0$

5. Spending on medical care varies widely across counties in the United States, as does life expectancy. Suppose you are interested in estimating the causal effect of medical-care spending on life expectancy.

You begin by specifying the model:

$$LE_i = \beta_0 + \beta_1 M_i + \varepsilon_i$$

where LE_i denotes life expectancy in county i and M_i denotes spending per person on medical care in county i.

a. Suppose some counties in your sample have healthier populations than other counties, which causes them to have higher life expectancies. Given this information, what will be the properties of the OLS estimate of β_1 in the above regression model?

b. Continuing from (a), suppose you observe for each county a variable *X* that measures the health of the population. How would you incorporate *X* into your regression model? Justify your answer.

c. Continuing from (a), suppose you observe an instrumental variable *Z*. What properties must the instrumental variable possess for it to be useful in producing a consistent estimate of β_1 with a small finite-sample bias?

d. Suppose you learn that your data do not come from the United States as we know it. Instead, they come from an imaginary country that assigns medical spending to counties based on the letter that starts the county name (i.e., counties that start with A-M are assigned high spending levels and the rest are assigned low spending). Given this information, how would you estimate β_1? Justify your answer.

e. Continuing from (d), suppose that in spite of the government assignment of spending levels, many counties spend a different amount on medical care than they were assigned. Explain how you would test whether these deviations cause bias in your estimate of β_1. Write down the test statistic you would use and state its asymptotic null distribution. You may assume that you observe the variables *Z* and *X* from parts (b) and (c).

Econometrics Preliminary Examination

July 1, 2010

Answer each part of each question. Each question receives equal weight. Within each question, each part receives equal weight.

QUESTION ONE

A person leaves for work between 8 am and 8:30 am and takes between 40 and 50 minutes to get there. Let the random variable *X* denote his/her time of departure and the random variable *Y* the travel time. Assume that these variables are independent and uniformly distributed.

a. Plot the joint probability density function (pdf).
b. Find the probability that the person arrives at work before 9 am.

QUESTION TWO

Consider the probability density function (pdf) defined by

$$f(x,y)=\begin{cases} A(x+2y) & \text{if } 0<y<1 \text{ and } 0<x<2 \\ 0 & \text{otherwise} \end{cases}$$

a. Find the value of *A*
b. Find the marginal distribution of *X*
c. Find the joint cumulative distribution function of *X* and *Y*
d. Find the pdf of the random variable *Z*=3*X*+1.

QUESTION THREE

Consider the model

$$S_i = \beta_1 + \beta_2 Y_i + \beta_3 A_i + \varepsilon_i$$

where *S* is the sales of a firm in the i^{th} state, *Y* is total income in the state, *A* is the amount of money spent by the company advertising in that state, and ε is a random error (*i*=1,2,…,50). You are deep in the backwoods of the world and only have a spreadsheet program like Excel with Ordinary Least Squares (OLS) available to use in the computations below.

a. You suspect that the random error ε_i is heteroskedastic with a standard deviation that depends on the size of the state population P_i. Describe step by step how you will go about testing for this heteroskedasticity. Explicitly detail the mechanics of any calculations you will make and the OLS regressions you will run. Be sure to state the null and alternate hypotheses to be considered, the test statistic you will use and its distribution, and the general form of the criterion for acceptance or rejection (e.g. smaller/larger than some identified critical value).

b. Suppose you find that there is heteroskedasticity but ignore it and use OLS to estimate the parameters of the model. Without formal proof, what can you claim for your coefficient estimates?

c. Suppose you find that there is heteroskedasticity but ignore it and use OLS to estimate the parameters of the model. You are interested in the statistical significance of the advertising variable. The Student's t statistic reported by the OLS program is 2.00. What, if anything, can you conclude regarding significance from this result?

d. Assuming that $\sigma_i = \sigma^2 P_i$ where P_i is the population of the i^{th} state, describe step by step how you would estimate the coefficients of the model. What advantages, if any, could you claim for the coefficient estimates and t and F statistics based on this estimation?

e. You have predictions of population and income and advertising expenditures for Washington D.C. (not in the original sample of 50 states). How would you forecast sales in Washington D.C.? Which coefficients would you use, the OLS coefficients from part b above, or the alternative coefficients from part d above (why, very briefly?). Would you use the original data on S and A, or some transformation of it in making your forecasts?

QUESTION FOUR

Consider a demand curve $q_t = \alpha_0 + \alpha_1 p_t + \varepsilon_t$, where q_t denotes the log of quantity and p_t denotes the log of price. You want to estimate the elasticity of demand (α_1). Suppose you have an additional variable x_t, which you believe is not an argument in the demand function. Your research assistant presents you with the following Stata output.

```
. regress q p

      Source |       SS       df       MS              Number of obs =     207
-------------+------------------------------           F(  1,   205) =    2.54
       Model |  4.63304973     1  4.63304973           Prob > F      =  0.1124
    Residual |  373.661173   205  1.82273743           R-squared     =  0.0122
-------------+------------------------------           Adj R-squared =  0.0074
       Total |  378.294223   206  1.83637972           Root MSE      =  1.3501

------------------------------------------------------------------------------
           q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           p |   -.075116   .0471152    -1.59   0.112    -.1680085    .0177765
       _cons |   .1336847   .0938377     1.42   0.156     -.051326    .3186954
------------------------------------------------------------------------------
```

```
. ivregress 2sls q (p=x)

Instrumental variables (2SLS) regression              Number of obs =      207
                                                      Wald chi2(1)  =    10.01
                                                      Prob > chi2   =   0.0016
                                                      R-squared     =        .
                                                      Root MSE      =   3.8612

------------------------------------------------------------------------------
           q |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           p |  -1.892644   .5982455    -3.16   0.002    -3.065184   -.7201047
       _cons |   .1383316    .268376     0.52   0.606    -.3876758    .6643389
------------------------------------------------------------------------------
Instrumented:  p
Instruments:   x
```

```
. regress q p x

      Source |       SS       df       MS              Number of obs =     207
-------------+------------------------------           F(  2,   204) =   66.72
       Model |  149.596451     2  74.7982256           Prob > F      =  0.0000
    Residual |  228.697772   204  1.12106751           R-squared     =  0.3955
-------------+------------------------------           Adj R-squared =  0.3895
       Total |  378.294223   206  1.83637972           Root MSE      =  1.0588

------------------------------------------------------------------------------
           q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           p |   .0220188   .0379245     0.58   0.562    -.0527555    .0967932
           x |  -.8147542   .0716495   -11.37   0.000    -.9560227   -.6734856
       _cons |   .0829824    .073727     1.13   0.262    -.0623823    .2283471
------------------------------------------------------------------------------
```

```
. regress p x

      Source |       SS       df       MS              Number of obs =     207
-------------+------------------------------           F(  1,   205) =   10.96
       Model |   41.656688     1   41.656688           Prob > F      =  0.0011
    Residual |  779.454744   205  3.80221826           R-squared     =  0.0507
-------------+------------------------------           Adj R-squared =  0.0461
       Total |  821.111432   206  3.98597782           Root MSE      =  1.9499

------------------------------------------------------------------------------
           p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .4255339   .1285613     3.31   0.001      .172062    .6790059
       _cons |    .028908    .135763     0.21   0.832    -.2387627    .2965788
------------------------------------------------------------------------------
```

a. Interpret the estimated coefficient on *p* in each of the first three regressions. Which of these estimates do you prefer? Explain.

b. Based on your chosen estimate in part (a), write down a 95% confidence interval for α_1.

c. Comment on the bias of the estimate you chose in part (a). Be specific.

d. For this part, assume the data are cross sectional. Suppose you are concerned about bias in the estimated standard error of your chosen estimate (a). Describe the possible sources of such bias and the steps you would take to generate an unbiased standard error estimate.

e. For this part, assume the data constitute a time series. Suppose you are concerned about bias in the estimated standard error of your chosen estimate (a). Describe the possible sources of such bias and the steps you would take to generate an unbiased standard error estimate.

f. Suppose you have another variable *z* that is correlated with price and may or may not be an argument in the demand function. Describe in detail how you would incorporate this variable into your analysis. Include in your answer any tests you would run to ascertain the appropriate use of *z*.

QUESTION FIVE

Suppose you have a sample of size n on the random variables y_i and x_i, and you model the distribution of $y_i \mid x_i$ as

$$y_i \mid x_i \sim N\left(x_i'\beta_0, (x_i'\beta_0)^2\right).$$

Note: the probability density function for a random variable $z \sim N(\mu, \sigma^2)$ is $f(z) = (2\pi\sigma^2)^{-1/2}\exp(-0.5\sigma^{-2}(z-\mu)^2)$.

a. Write down the log likelihood function for the unknown parameters β_0. State any assumptions that you make.

b. Suppose your likelihood model is correctly specified, but you estimate β_0 by ordinary least squares. Show that your estimate is consistent for β_0. State any assumptions that you make.

c. Suppose your likelihood model is correctly specified. Show that the maximum likelihood estimate $\hat{\beta}$ is consistent for β_0. State any assumptions that you make.

d. Suppose the true distribution of $y_i|x_i$ is nonnormal. How would you interpret the parameter estimate obtained by maximizing the likelihood in (1)?

e. Write down the information equality and explain the implications for standard error estimation if it doesn't hold.

Econometrics Preliminary Examination

August 16, 2010

Answer each part of each question. All questions are weighted equally. Within each question, each part receives equal weight. Some distribution tables are included on the last two pages of the exam.

QUESTION ONE

Let Y_1, Y_2, Y_3, Y_4 be independent, identically distributed random variables from a population with mean μ and variance σ^2. Let $\overline{Y} = (Y_1 + Y_2 + Y_3 + Y_4)/4$ denote the average of these four random variables.

a. Find the expected value and variance of $\overline{Y}$ in terms of μ and σ^2.

b. Now, consider a different estimator of μ:

$$\overline{W} = Y_1/8 + Y_2/8 + Y_3/4 + Y_4/2 .$$

Show that $\overline{W}$ is an unbiased estimator of μ.

c. Show that $\overline{W}$ is a less efficient estimator for μ than $\overline{Y}$.

d. Show that $\overline{Y}$ is the best linear unbiased estimator for μ.

QUESTION TWO

A sample of 100 men and a sample of 64 women are selected at random from the employees at a large firm. The sample draws only from the set employees with a particular job description, and the firm employs a large number of people with this job description. The following table shows summary statistics for monthly salaries of the sampled workers.

	Average Salary ($\overline{Y}$)	Standard Dev. (S_Y)	n
Men	\$3,100	\$200	100
Women	\$2,900	\$320	64

a. What do these data suggest about wage differences in the firm? Do they represent statistically significant evidence that wages of men and women are different at the 5% significance level? To answer this question, first state the null and alternative hypothesis and second, compute the relevant t-statistic. Assume that, conditional on gender, salaries are independent across individuals.

b. Construct a 95% confidence interval for the difference in average salaries between men and women at this firm.

c. Provide a precise interpretation of your 95% confidence interval in (b).

d. Can you conclude that gender discrimination exists in the firm based on your results in parts (a) and (b)? Explain.

QUESTION THREE

Consider the following cross-section regression model based on 43 geographic markets in August 2006, in which Y_i is the monthly price in dollars of basic cable television service, X_i is the number of channels provided (a quality measure), and D_i is a dummy variable that is equal to 1 in the competitive markets (those with two or more cable providers) and is 0 otherwise:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i \quad u_i \sim iid\ N(0, \sigma^2),\ i = 1, 2, \dots, n.$$

Suppose the classical assumptions hold.

a. What are $E(u_i)$, $E(\beta_3)$, and $Var(u_i)$?

b. Algebraically, what is $E(Y_i \mid X_i, D_i)$ for the competitive markets? For the non-competitive markets? What is the difference between the two expected prices?

c. What do the classical assumptions imply about $Var(Y_i \mid X_i, D_i)$ for the competitive and non-competitive markets? Formally obtain the variance in both cases.

d. How would you impose the restriction that the competitive and monopolistic quality-adjusted expected prices are equal? (Just explain.)

Ordinary Least Squares estimation of the model produced the following estimates:

$$Y_i = \underset{4.32}{12.17} + \underset{2.84}{0.54}X_i \underset{-3.21}{-14.38}D_i \qquad \hat{\sigma}^2 = 3.17,\ F_c = 26.51,\ R^2 = 0.57$$

(Student's *t* statistics are below the coefficient estimates; F_c is the calculated regression F statistic.)

For parts e, f, g, and h below, assume that the classical assumptions hold.

e. At the 1% significance level, formally test the hypothesis that competition *reduces* prices (note the italics). Be sure to state the null and alternate hypotheses, the test statistic and its distribution, and draw a plain-language formal conclusion suitable for a formal report, *i.e.* as strong as possible while not overstepping what you actually have determined about the effect of competition on cable TV prices.

f. At the 5% significance level, formally test the hypothesis associated with Overall Goodness of Fit, also known as the Test for the Existence of a Relationship. State the null and alternate hypotheses, the test statistic and its distribution, and draw a plain-language formal conclusion.

g. You are advising a potential entrant into a new cable TV market. The company plans to offer 70 channels. What monthly price could they expect to charge each subscriber if they are the only supplier in the market? What price could they expect if another firm also enters simultaneously?

h. Suppose now that we are not sure that all the classical assumptions hold in this case. In particular, we want to know whether monopolists might have larger error variances than competitors. Accordingly, we separate the 43 data points into 26 characterized by monopoly and 17 characterized by competition and obtain the two regressions

$$Y_i = \underset{2.23}{20.75} + \underset{3.49}{0.65}X_i \qquad \hat{\sigma}^2 = 4.47,\ F_c = 22.15,\ R^2 = 0.48$$

$$Y_i = \underset{3.92}{-2.33} + \underset{3.29}{0.41}X_i \qquad \hat{\sigma}^2 = 2.47,\ F_c = 27.86,\ R^2 = 0.65$$

Formally state the hypothesis in question, test it, and draw the conclusion.

i. Suppose that, regardless of how the test in part (i) above turns out, we want to assume that all competitors have one error variance σ_1^2 and all monopolists have another σ_2^2 in the model

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i \quad u_i \sim iid\ N(0,\sigma^2),\ i = 1,2,\ldots,N.$$

Describe the steps you would take to estimate the model in this case. Be explicit about any constructions you employ in estimation.

QUESTION FOUR

Spending on social services varies widely across states in the United States, as does the unemployment rate. Suppose you are interested in estimating the causal effect of social services spending on unemployment. You are testing a theory that the availability of social services reduces labor supply.

You begin by specifying the model:

$$U_i = \beta_0 + \beta_1 S_i + \varepsilon_i$$

where U_i denotes the unemployment rate in state *i* and S_i denotes spending per resident on social services in state *i*. (Note that S_i is measured in dollars per resident of the state, not dollars per recipient of social services.)

a. Suppose some states in your sample have experienced recent economic shocks that have raised unemployment rates, thereby increasing the demand for (and therefore spending on) social services. Given this information, what will be the properties of the OLS estimate of β_1 in the above regression model?

b. Continuing from (a), suppose you observe for each state a variable X_i that measures the size of recent state-level economic shocks. How would you incorporate X_i into your regression model? Justify your answer. Would the resulting estimate of β1 capture the causal effect of social services spending on unemployment?

c. Continuing from (a), suppose you observe an instrumental variable Z_i. What properties must the instrumental variable possess for it to be useful in producing a consistent estimate of β_1 with a small finite-sample bias?

d. Suppose you learn that your data do not come from the United States as we know it. Instead, they come from an imaginary country that assigns the social services budget based on the letter that starts the state name (i.e., states that start with A-M are assigned high budgets and the rest are assigned low spending). These budgets are assigned as dollars per resident in the state. Given this information, how would you estimate β_1? Justify your answer.

e. Continuing from (d), suppose that in spite of the government assignment of spending levels, many states spend a different amount on social services than they were assigned. Explain how you would test whether these deviations cause bias in your estimate of β_1. Write down the test statistic you would use and state its asymptotic null distribution. You may assume that you observe the variables Z_i and X_i from parts (b) and (c).

QUESTION FIVE

Suppose you have a sample of size *n* on the random variables y_i and x_i, and you model the distribution of $y_i \mid x_i$ as

$$y_i \mid x_i \sim N\left(x_i'\beta_0, \exp(x_i'\beta_0)\right).$$

Note: the probability density function for a random variable $z \sim N(\mu, \sigma^2)$ is $f(z) = (2\pi\sigma^2)^{-1/2}\exp(-0.5\sigma^{-2}(z-\mu)^2)$.

a. Write down the log likelihood function for the unknown parameters β_0. State any assumptions that you make.

b. Suppose the true distribution of $y_i|x_i$ is nonnormal. How would you interpret the parameter estimate obtained by maximizing the likelihood in (a)?

c. Would OLS produce a consistent estimate of β_0? Explain. A well-reasoned answer in words is sufficient.

d. Explain why maximum likelihood provides an asymptotically efficient estimate of β_0 relative to OLS.

F test of restrictions
(where the number of restrictions is $q = k\text{-}m$):

$$F_c = \frac{(ESS_R - ESS_U) \div q}{ESS_U \div (n - k)}$$

Goldfeld-Quandt F test:

$$F_c = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{ESS_2 \div (n_2 - k)}{ESS_1 \div (n_1 - k)}$$

Overall goodness of fit:

$$F_c = \frac{(ESS_R - ESS_U) \div (k-1)}{ESS_U \div (n - k)}$$

$$= \frac{RSS_U \div (k-1)}{ESS_U \div (n - k)}$$

$$= \frac{R^2 \div (k-1)}{(1 - R^2) \div (n - k)}$$

TABLE A.4b **Upper 5% Points of the *F*-Distribution**

n \ *m*	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22

Note: m = degrees of freedom for the numerator
n = degrees of freedom for the denominator

Source: Handbook of Tables for Mathematics, edited by Robert C. West and Samuel M. Selby, 1970. Reprinted with the permission of the CRC Press

Percentage Points of the *t*-Distribution

d.f.	1T = 0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
	2T = 0.8	0.5	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Note: 1T = area under one tail; 2T = area under both tails.
For 25 degrees of freedom (d.f.), $P(t > 2.060) = 0.025$ and $P(t < -2.060 \text{ or } t > 2.060) = 0.05$.

Source: Biometrika Tables for Statisticians, Vol. I. Edited by E. S. Pearson and H. O. Hartley, 3rd edition, 1966. Reprinted with the permission of the Biometrika Trustees.