Econometrics Preliminary Examination

July 30, 2011

Answer each part of each question. All questions are weighted equally. Within each question, each part receives equal weight. You should have 6 pages of questions.

Question One

Let $Y \sim N(\mu, \sigma^2)$. Consider the following two estimators of σ^2:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

The latter is the well-known variance estimator that is unbiased.

Hints: The variance of $\hat{\sigma}^2$ is $2\sigma^4(n-1)/n^2$. Note that $s^2 = \frac{n}{n-1}\hat{\sigma}^2$.

a) Formally compare the variance of the two variance estimators, *i.e.*, compare $V(\hat{\sigma}^2)$ to $V(s^2)$.

b) Compare the mean squared errors of the two estimators, *i.e.*, compare $MSE(\hat{\sigma}^2)$ to $MSE(s^2)$.

c) Are either or both of the estimators consistent? (Be explicit about your definition of consistency.)

d) We use the standard deviation much more often than we use the variance, *e.g.*, in *t* tests and confidence intervals, and so on. Informally, given $E(s^2) = \sigma^2$, what can we say about $E(s)$?

Question Two

A population of farmers i=1,…,N cultivates grain (Q_i) using labor (L_i), capital (K_i), and land (T_i). Suppose you specify a Cobb-Douglas production function of the following form:

$$Q_i = AL_i^{\beta_L} K_i^{\beta_K} T_i^{\beta_T} \exp(u_i)$$

which implies

$$\ln Q_i = \alpha + \beta_L \ln L_i + \beta_K \ln K_i + \beta_T \ln T_i + u_i$$

where the vector of parameters is $\begin{bmatrix} \alpha & \beta_L & \beta_K & \beta_T \end{bmatrix}$, and u_i is an error term with mean zero and constant variance.

Suppose each farmer's land is one of three types, and define t_i such that

$$t_i = \begin{cases} 1 & \text{farmer i's land is of type 1} \\ 2 & \text{farmer i's land is of type 2} \\ 3 & \text{farmer i's land is of type 3} \end{cases}$$

It is unknown which (if any) type is more productive, but it is known that β_L, β_K, and β_T do not depend on land type. You observe Q_i, K_i, L_i, and t_i.

a. Describe in detail how you would test econometrically whether land type matters in grain production. Be sure to tell us how you would formulate your test, what your null hypothesis would be, and what criterion you would use as a basis for rejecting it.

b. Suppose land type does not matter in grain production and farmers can adjust labor and capital allocation in response to weather events that affect crop yield. Would OLS applied to the regression

$$\ln Q_i = \alpha + \beta_L \ln L_i + \beta_K \ln K_i + \beta_T \ln T_i + u_i$$

produce unbiased estimates of the production function parameters? Why or why not? State any assumptions that you make, and provide a mathematical proof.

c. Suppose land type does matter in grain production and farmers cannot adjust capital and labor through the growing season. At planting time, they know the type of their land and allocate labor and capital for the entire growing season. Four months later, they observe output. Would OLS applied to the regression

$$\ln Q_i = \alpha + \beta_L \ln L_i + \beta_K \ln K_i + \beta_T \ln T_i + u_i$$

produce unbiased estimates of β_L, β_K, and β_T? Why or why not? State any assumptions that you make, and explain clearly your reasoning.

d. Suppose land type does not matter in grain production and farmers cannot adjust labor and capital allocation in response to weather events that affect crop yield. You are concerned that the error term may have greater variance on large farms. What effect would this variance pattern have on OLS estimates of β_L, β_K, and β_T?

e. Continuing from (4), describe the action you would take to see whether this is the case and explain the action you would take if you conclude that there is such variance pattern. Justify your answer.

f. Suppose land type does not matter in grain production and farmers cannot adjust capital and labor through the growing season. You wish to test the hypothesis of constant returns to scale (i.e., $\beta_L+\beta_K+\beta_T = 1$). Your research assistant will perform the test, but the only estimation method she knows how to employ is OLS. Describe in detail how you would test this hypothesis using only OLS estimation a standard *t*-test. Be sure to tell us how you would formulate your test, what your null hypothesis would be, and what criterion you would use as a basis for rejecting it.

Question Three

The food stamp program gives vouchers to poor household to use in purchasing food. Prior research has suggested that, by reducing the average price of food, this program causes an increase in obesity. You collect data on the following variables:

dW_i =the change in the weight of individual *i* over a one-year period

FS_i= the number of months during that year that individual *i* was on food stamps

x_i= an additional variable that you believe causes weight gain directly

Your research assistant presents you with the following Stata output. The option ", robust" calculates heteroskedasticity robust standard errors.

REGRESSION 1

```
. reg dW FS

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =   32.06
       Model |   107.33735     1   107.33735           Prob> F       =  0.0000
    Residual |  662.954615   198  3.34825563           R-squared     =  0.1393
-------------+------------------------------           Adj R-squared =  0.1350
       Total |  770.291965   199   3.8708139           Root MSE      =  1.8298
------------------------------------------------------------------------------
          dW |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          FS |   .5316136   .0938923     5.66   0.000     .3464563    .7167708
       _cons |   .0245024   .1293882     0.19   0.850    -.2306534    .2796582
------------------------------------------------------------------------------
```

REGRESSION 2

```
. reg dW FS, robust

Linear regression                                      Number of obs =     200
                                                       F(  1,   198) =   10.96
                                                       Prob> F       =  0.0011
                                                       R-squared     =  0.1393
                                                       Root MSE      =  1.8298
------------------------------------------------------------------------------
             |               Robust
          dW |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          FS |   .5316136   .1605573     3.31   0.001     .2149918    .8482353
       _cons |   .0245024   .1294937     0.19   0.850    -.2308614    .2798662
------------------------------------------------------------------------------
```

REGRESSION 3

```
. reg dW x, robust

Linear regression                                      Number of obs =     200
                                                       F(  1,   198) =   21.06
                                                       Prob> F       =  0.0000
                                                       R-squared     =  0.3544
                                                       Root MSE      =  1.5848
------------------------------------------------------------------------------
             |               Robust
          dW |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.257964   .2741206     4.59   0.000     .7173936    1.798535
       _cons |   .0115524   .1132124     0.10   0.919    -.2117044    .2348092
------------------------------------------------------------------------------
```

REGRESSION 4

```
. reg dW x FS, robust

Linear regression                                      Number of obs =     200
                                                       F(  2,   197) =   13.00
                                                       Prob> F       =  0.0000
                                                       R-squared     =  0.3558
                                                       Root MSE      =  1.5871
------------------------------------------------------------------------------
             |               Robust
          dW |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.331777   .2632172     5.06   0.000     .8126918    1.850862
          FS |  -.0737512   .0891095    -0.83   0.409    -.2494822    .1019798
       _cons |   .0108643   .1132372     0.10   0.924    -.2124484    .2341771
------------------------------------------------------------------------------
```

REGRESSION 5

```
. ivregress 2sls dW (FS=x), robust first

First-stage regressions
-----------------------

                                                       Number of obs   =        200
                                                       F(   1,    198) =     161.32
                                                       Prob> F         =     0.0000
                                                       R-squared       =     0.4549
                                                       Adj R-squared   =     0.4522
                                                       Root MSE        =     1.0225
------------------------------------------------------------------------------
             |               Robust
          FS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.000833    .078798    12.70   0.000     .8454417    1.156224
       _cons |  -.0093295   .0721699    -0.13   0.897    -.1516497    .1329908
------------------------------------------------------------------------------
```

```
Instrumental variables (2SLS) regression              Number of obs =     200
                                                      Wald chi2(1)  =   18.46
                                                      Prob> chi2    =  0.0000
                                                      R-squared     =       .
                                                      Root MSE      =   2.077
------------------------------------------------------------------------------
             |               Robust
          dW |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          FS |   1.256917   .2925705     4.30   0.000     .6834899    1.830345
       _cons |   .0232788   .1469567     0.16   0.874     -.264751    .3113085
------------------------------------------------------------------------------
Instrumented:  FS
Instruments:   x
```

a. Compare the first two regressions and describe what the differences between them reveal to you about the data.

b. Interpret the estimated coefficient on *FS* in each of regressions 2, 4, and 5. Which of these estimates do you prefer? Explain.

c. Based on your chosen estimate in part (b), write down a 95% confidence interval for the effect of food stamps on weight gain. Interpret this interval in words.

d. Comment on the bias of the estimate you chose in part (b). Be specific.

e. Suppose you have another variable *z* that does not affect weight gain directly and may or may not be correlated with food stamp participation. 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*.

f. Suppose your research assistant informs you that the data do not represent a cross-section of individuals. Rather, they represent annual data on a single individual over her lifetime. Now, what are your conclusions about the properties of your chosen estimate in part (b) and the effect of food stamps on weight gain?

Question Four

The random variable y_i measures the number of stock trades made by individual *i* in a particular month. Let x_i denote a vector of explanatory variables including a constant. You decide to model $y_i \mid x_i$ using a Poisson distribution with independence across *i*.

The PDF for the Poisson distribution is

$$f(y_i;\mu)=\frac{\mu^{y_i}\exp(-\mu)}{y_i!}$$

and the Poisson distribution has the property that $E(y_i)=\mu$ and $\text{var}(y_i)=\mu$. You choose the functional form $\mu_i=\exp(x_i'\beta)$.

a. Write down the log-likelihood function for the unknown parameters β.

b. Show that the information matrix equality holds for your model. State any assumptions you make.

c. Suppose that the model is misspecified causing the information matrix equality to fail to hold. Describe the implications of this failure for the properties of $\hat{\beta}$. Specifically, what are the implications for (i) consistency and asymptotic normality, (ii) efficiency, and (iii) interpretation of $\hat{\beta}$?

d. Suppose the Poisson likelihood model is correctly specified, but instead you generate the parameter estimate

$$\tilde{\beta}=\arg\min_{\beta}\sum_{i=1}^{n}\left(y_i-\exp(x_i'\beta)\right)^2.$$

Describe the properties of your estimate. Your answer should address (i) consistency and asymptotic normality, (ii) efficiency, and (iii) interpretation of the parameter estimate.

e. Suppose the Poisson likelihood model is correctly specified, but instead you run a linear least squares regression of y_i on x_i. Describe the properties of your estimate. Your answer should address (i) consistency and asymptotic normality, (ii) efficiency, and (iii) interpretation of the parameter estimate.

Econometrics Preliminary Examination

August 15, 2011

Answer each part of each question. All questions are weighted equally. Within each question, each part receives equal weight. You should have 6 pages of questions.

Question One

A Harvard study sought to decide the effectiveness of aspirin in reducing heart attacks. A large number of individuals -- 22,071 independent volunteer doctors – were randomly assigned to two groups. Group one took a placebo, a pill identical to aspirin but containing no aspirin, while group two received one aspirin a day. Over a period averaging nearly five years, the investigators recorded the responses, heart attack or no heart attack. The results (combining fatal and non-fatal heart attacks) are:

	Attack	**No Attack**	***n***	**Attack Rate**
Placebo	239	10,795	11,034	$\overline{p}_1 = \frac{239}{11,034} = .0217$
Aspirin	139	10,898	11,037	$\overline{p}_2 = \frac{139}{11,037} = .0126$

Incidentally, these are real data. The placebo and aspirin group observations are independent samples from two binomial populations.[1] For clarity, we refer to a heart attack as a "success" (!). In the placebo population, the chance of success is p_1 while in the aspirin population the chance of success is p_2. The objective is to estimate the unobserved true difference in heart attack rates, $p_1 - p_2$, and draw some statistically valid conclusions about it.

a. Notice that the attack rates are sample means, i.e., for each group they are 1/n times the sum of n Bernoulli outcomes. Write down formal expressions for the expected value of the placebo group attack rate, $E(\overline{p}_1)$, and the aspirin group attack rate, $E(\overline{p}_2)$.

---

[1] *From Wikipedia:* If X ~ B(n,p) (that is, X is a binomially distributed random variable), then the expected value of X is $E[X] = np$ and the variance is $\text{var}[X] = np(1-p)$.

This fact is easily proven as follows. Suppose first that we have a single Bernoulli trial. There are two possible outcomes: 1 and 0, the first occurring with probability p and the second having probability $1 - p$. The expected value in this trial will be equal to $\mu = 1 \cdot p + 0 \cdot (1-p) = p$. The variance in this trial is calculated similarly: $\sigma^2 = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = p(1-p)$.

The generic binomial distribution is a sum of n independent Bernoulli trials. The mean and the variance of such distributions are equal to the sums of means and variances of each individual trial:

$$\mu_n = \sum_{k=1}^{n} \mu = np, \qquad \sigma_n^2 = \sum_{k=1}^{n} \sigma^2 = np(1-p)$$

b. Again recognizing that the heart attack rates are sample means, *i.e.*, for each group they are 1/n times the sum of n Bernoulli outcomes, write down formal expressions for the true variance of the placebo group, $V(\overline{p}_1)$ and of the aspirin group, $V(\overline{p}_2)$.

c. How would you estimate the true difference in heart attack rates?

In order to do anything statistically interesting with this difference in heart attack rates we will need its variance and the form of its distribution. Consider the variance first:

d. What is the variance of the difference of the two sample success means, $V(\overline{p}_1 - \overline{p}_2)$? Be explicit about any assumptions you use. How would you estimate this variance and its associated standard error?

e. One question we would like answered is whether or not a zero difference in heart attacks between the two groups is within the 95% confidence bounds calculated from our estimates. Unfortunately, you find yourself at your plantation on Lake Kivu in Rwanda without your trusty laptop. Fortunately, you recognize that 22,071 is a large number of observations and, being a skilled statistician, you recognize that the large sample allows you to approximate the confidence interval easily. Identify the large-sample approximation, in the absence of a calculator approximate any square root you may need, and write the resulting 95% confidence bounds for the difference in heart attacks between the two groups.

f. More formally, we wish to perform a hypothesis test regarding the heart attack rates in the aspirin and placebo groups. **Formally** state the null and alternate hypotheses, the test statistic you would calculate, its distribution and its value, and draw a plain-language conclusion that embodies exactly what you can report, no more and no less, suitable for filing with the Food and Drug Administration.

Question Two

The United States currently pays around $20 billion per year to farmers in direct subsidies. These subsidies aim to achieve "farm income stabilization" and were redesigned in 1996 to be "decoupled," i.e., program benefits do not depend on farmers' current production decisions. Many researchers have attempted to test the null hypothesis that these farm subsidies do not affect crop output.

You have cross-section data from a random sample of U.S. farmers in 2005 on each of the following variables:

FS_i	the farm subsidy (in dollars) received by farmer *i*
Z_i	a variable that determines farmer *i*'s eligibility for the subsidy, such that $FS_i = \delta Z_i + \eta_i$, where η_i denotes an error term
Q_i	the value of farmer i's production of program crops
X_{ji}	a vector of inputs *j=1,…,J* used by farmer *i* to produce these crops, including land

Consider the following model to test the effect of farm subsidies on the production of program crops:

$$\text{(i)} \qquad q_i = \alpha + \sum_{j=1}^{J} \beta_j x_{ji} + \gamma FS_i + \varepsilon_i$$

where the lower case q_i and x_{ji} denote natural logarithms of Q_i and X_{ji}, respectively; α, β_j (j=1, 2, …, J), and γ are parameters, and ε_i is a stochastic error. Throughout this problem, assume that all input choices are made at the beginning of the crop season and cannot be adjusted thereafter.

a. State whether the following statement is true or false (please show why): If the input variables are left out of the model, then ordinary least squares estimates of the parameter γ will necessarily be biased.

b. State whether the following statement true or false (please show why): Ordinary least squares estimates of the parameters in (i) will be unbiased.

c. You are concerned that the error term may have greater variance on large farms. What effect would this variance pattern have on ordinary least squares estimates of the parameters in (i)?

d. Continuing from (c), describe the action you would take to see whether the error term has greater variance on large farms, and explain the action you would take if you conclude that there is such variance pattern. Justify your answer.

e. Suppose you estimate the parameters in (i) using instrumental variables with Z_i as an instrument for FS_i. State precisely, both in words and mathematically, the conditions that Z_i must satisfy for this estimator to be consistent

f. Prior to 1996, subsidies were linked to output of program crops. When you present your results at the AAEA meetings, someone in the audience points out that, in practice, there was significant continuity of subsidy payments after decoupling; farmers who got subsidies before 1996 continued to get them after 1996. Given this new information, how would you want to change your estimation strategy, and what identification problems would you encounter?

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_i = ($gcriv_i$ $gpris_i$) as a 1×2 vector and X_i = (1 $gincpc_i$ $final1_i$ $final2_i$) as a 1×4 vector, and specify the moment condition $E(y_i \mid X_i) = -X_i B \Gamma^{-1}$, 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
------------------------------------------------------------------------------


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

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. Levitt uses the variables final1 and final2 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 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.

Econometrics Preliminary Examination

July 5, 2012

Answer each part of each of the six questions. All questions are weighted equally. Within each question, each part receives equal weight.

Question One

You have been given the task of writing a computer program that can generate random draws from various probability distributions. Unfortunately, the software you have can only generate draws from the normal distribution with mean 0 and variance 1. Otherwise, your software contains standard math functions, but cannot generate random variables from other probability distributions.

Explain how you could still use this software to generate draws from the following probability distributions:

a) N(10, 400)

b) Chi-squared with 10 degrees of freedom

c) A *t* distribution with 9 degrees of freedom

d) The *F* distribution with 1 numerator and 9 denominator degrees of freedom

e) The binomial distribution with parameters n=20 and p=0.5

f) The truncated normal(10,400), where the distribution is truncated from below at 9

Question Two

a) Suppose that an independent random sample is taken from a $N(\mu,\sigma^2)$ distribution. Prove directly, without invoking any weak or strong laws of large numbers, that the sample mean $\bar{X}$ is a consistent estimator of the population mean, μ.

b) Given the classical simple linear regression model, $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$, show that the ordinary least squares estimator of β_2 is consistent. Specifically state all the assumptions needed to prove this.

c) What can we say about the small sample properties of a consistent estimator? Be explicit.

Question Three

a) In a large group of heterosexual married couples, the standard deviation of the husbands' ages is 4 years and the standard deviation of the wives' ages is 3 years. Your friend suggests that the standard deviation of the difference in ages of married couples is therefore 5 years. Instead, the standard deviation of the difference equals 2 years. Explain why your friend thinks the standard deviation should be 5 and explain what feature of the data causes the standard deviation to actually be 2.

b) You are interested in two measurements of some socioeconomic characteristics associated with the husbands in this group. You denote them by y_{1i} and y_{2i} for the $i = 1,2,\ldots, n$ husbands in your sample.

Your interest is in the expected values of these two random variables, in the population from which the sample of husbands was drawn. Suppose you think that the data follow this process:

$$y_{1i} = \beta_1 + u_{1i}$$
$$y_{2i} = u_{2i}$$

The errors are distributed as bivariate normal random variables with mean 0, constant variances (not assumed to be equal), and a positive covariance between the two errors for a given husband in the sample. The observations are independent across the individual men in the sample.

Professor Blue states that you should calculate the sample mean of the y_{1i} values to find the most efficient estimator of β_1. Is this a correct statement? Why or why not?

c) What if your model is changed to one where the second equation is

$$y_{2i} = \beta_1 + u_{2i}$$

where--that is not a typo---the parameter in the second equation is believed to be the same as the parameter in the first equation? Is Professor Blue correct about estimating β_1? Why or why not?

d) What if your model is changed to one where the second equation is

$$y_{2i} = \beta_2 + u_{2i}$$

instead? Is Professor Blue correct about estimating β_1? Why or why not?

Question Four

Your understanding of the market for fresh strawberries produced in California is that week *t*'s quantity produced (*Q*) depends on the previous week's weather (*W*) and the previous week's price:

$$Q_t = \beta_1 + \beta_2 P_{t-1} + \beta_3 W_{t-1} + \varepsilon_t \qquad (1)$$

Market prices are determined by

$$P_t = \alpha_1 + \alpha_2 Q_t + \alpha_3 A_t + \alpha_4 P_{t-1} + u_t \qquad (2)$$

where *P* denotes market price and *A* is a variable measuring retailer advertising expense. You believe that *A* measures the expense incurred in time *t*, but that the relevant decisions about advertising are made weeks in advance.

a) Indicate whether or not you would be willing to estimate equation (1) using OLS, and justify your answer.

b) Indicate whether or not you would be willing to estimate equation (2) using OLS, and justify your answer.

c) Do your answers to (a) and (b) change if it is lagged quantity, instead of lagged price, that appears in equation (1)? Explain with as much detail as is needed to justify your answers.

d) Do your answers to (a) and (b) change if it is current price, instead of lagged price, that appears in equation (1)? Explain with as much detail as is needed to justify your answers.

Now assume for the remaining parts (e)-(g) of this question that it is current price, instead of lagged price, that appears in equation (1).

e) Suppose you believe that $\alpha_4=0$. Comment on the identification status of each equation.

f) Describe the two-stage least-squares estimator for your part (e) model, and indicate how you would change that estimator if you believed that $\alpha_4 \neq 0$.

g) Is your estimator consistent? State all conditions you require, and demonstrate that your answer is correct.

Question Five

This question continues from question four. In all parts, assume as in (4e-4g) that it is current price, instead of lagged price, that appears in equation (1).

a) Consider the following alternative estimator for the second equation.

   Q_t is regressed on A_t alone, to construct predicted values denoted by $\hat{Q}_t$. Then equation (2) is estimated as is, using OLS, except that $\hat{Q}_t$ replaces Q_t. (Continue to assume α_4=0 if you would like to do so.) Does such a procedure produce consistent estimates? State any conditions required for this to be correct.

b) Repeat the previous part assuming that Q_t is regressed on W_{t-1} alone and everything else remains the same.

c) The following page contains some regression output generated by your research assistant for equation (2) (assuming α_4=0). What do you conclude about the likely bias in the two-stage least squares estimates? Justify your answer.

d) From the regression output, what do you conclude about the parameter β_2? Justify your answer.

e) Suppose you suspect that prices are much more volatile when quantity produced is low. Explain how this data feature would affect the output your research assistant generated.

f) Explain to your research assistant how she should modify the analysis to account for the feature described in part (e). Justify your recommendations.

```
. reg P Q A

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =  358.54
       Model |  713.646975     2  356.823487           Prob > F      =  0.0000
    Residual |  96.5349358    97  .995205524           R-squared     =  0.8808
-------------+------------------------------           Adj R-squared =  0.8784
       Total |  810.181911    99  8.18365566           Root MSE      =   .9976

------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Q |  -1.034809   .0399634   -25.89   0.000    -1.114126    -.955493
           A |   .5332983   .0979118     5.45   0.000     .3389705    .7276261
       _cons |  -.0503691   .0998265    -0.50   0.615    -.2484972    .1477589
------------------------------------------------------------------------------

. ivregress 2sls P (Q=lagW) A, first

First-stage regressions
-----------------------

                                                Number of obs   =        100
                                                F(   2,     97) =     240.63
                                                Prob > F        =     0.0000
                                                R-squared       =     0.8323
                                                Adj R-squared   =     0.8288
                                                Root MSE        =     1.0395

------------------------------------------------------------------------------
           Q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           A |  -.0290731    .101988    -0.29   0.776    -.2314912    .1733449
       lagW |   2.142014   .0978051    21.90   0.000     1.947898     2.33613
       _cons |   .0711839   .1042168     0.68   0.496    -.1356577    .2780255
------------------------------------------------------------------------------

Instrumental variables (2SLS) regression               Number of obs =     100
                                                       Wald chi2(2)  =  601.12
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.8806
                                                       Root MSE      =  .98362

------------------------------------------------------------------------------
           P |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Q |  -1.016176   .0432044   -23.52   0.000    -1.100855   -.9314966
           A |   .5357126   .0965672     5.55   0.000     .3464443    .7249808
       _cons |  -.0486768    .098441    -0.49   0.621    -.2416177     .144264
------------------------------------------------------------------------------
Instrumented:  Q
Instruments:   A lagW

. estat endogenous

  Tests of endogeneity
  Ho: variables are exogenous
```

```
Durbin (score) chi2(1)          =  1.10827  (p = 0.2925)
Wu-Hausman F(1,96)              =  1.07586  (p = 0.3022)
```

Question Six

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

$$y_i \mid x_i : \text{Gamma}\left(k, \exp(x_i'\beta_0)\right).$$

Suppose that k is known and you aim to estimate β_0.

Note: the probability density function for a random variable z : $\text{Gamma}(k,\theta)$ is

$$f(z;k,\theta) = \frac{1}{\Gamma(k)\theta^k} z^{k-1} \exp\left(-\frac{z}{\theta}\right).$$

where $\Gamma(k)$ denotes the gamma function. The mean is $E[z] = k\theta$ and the variance is $\text{var}[z] = k\theta^2$.

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

b) Suppose you estimate β_0 by maximum likelihood assuming k=1 when the true value of k is 2. Would your estimator be consistent for β_0? Justify your answer. How would you interpret the probability limit of your maximum likelihood estimate?

c) Would OLS regression of y_i on x_i produce a consistent estimate of β_0? Justify your answer.

d) Would the estimator $\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \left(y_i - \exp(x_i'\beta)\right)^2$ produce a consistent estimate of β_0? Justify your answer.

Econometrics Preliminary Examination

August 20, 2012

Answer each part of each of the five questions. All questions are weighted equally. Within each question, each part receives equal weight.

Question One

Given a classical linear regression model with the usual assumptions,

$$Y_i = \beta_1 + \beta_2 X_{i2} + ... + \beta_K X_{iK} + \varepsilon_i \quad i = 1,2,...,n,$$

suppose that we want to test the following hypothesis:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

a. For the above hypothesis test, define the type II error and illustrate the concept with a graph.

b. Define the power of the test. Draw a typical power function for the above test.

c. Suppose we have cross sectional data on a set of children and the dependent variable in our regression model is the BMI (body mass index). Explanatory variables consist of age, gender, etc. There may be more than one child per family. Do you believe that the error term is correlated within families? Why or why not?

d. Following from (c), what effect, if any, would correlation among the errors have on the power of a test? Explain. What would you recommend to increase the power of the test?

e. What are some other factors that affect the power of the hypothesis test? Explain, and recommend a remedy for each example.

Question Two

Suppose the true model is:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

where all the classical assumptions hold. The researcher, however, estimates the following model:

$$Y_i = \beta_1 + \beta_2 X_{i2} + u_i \ .$$

a. Derive the ordinary least squares estimator for β_2 in the researcher's model.

b. Is the researcher's OLS estimator biased? Consistent? Provide proofs in your responses.

c. Suppose that a researcher wants to estimate a demand function using scanner data from a supermarket. (Scanner data is data collected electronically at the point of purchase.) The data include prices and quantities, but not consumer income or other individual-specific variables. The researcher regresses quantities demanded on own prices, prices of substitutes, and prices of complements. What can you conclude about the properties of OLS estimators of the price coefficients? What would you recommend the researcher to do to overcome these problems?

d. Consider the assertion, "All models are misspecified because relevant variables are omitted; therefore OLS estimators are always biased and inconsistent." Comment.

Question Three

You are interested in the linear model:

$$y = X\beta + u,$$

where it is known that u is a vector of normal errors with expected value 0 and variance-covariance matrix $\sigma^2 V$. The matrix V is known, diagonal, and not equal to the identity matrix. X is an n by K matrix with columns representing variables and rows associated with the elements in y; the variables in X are assumed to be non-random.

a. You estimate the vector of K parameters in β using OLS. What is the probability distribution of your estimator?

b. Explain why, in this situation, you cannot rely on the results from a standard OLS printout. Be precise about which results are useful and which are potentially misleading.

c. You have three choices for what to do instead. One approach uses generalized least-squares and the other two continue to use your OLS parameter estimates. Explain each, citing advantages and disadvantages.

d. Explain how you could construct a reliable 95% confidence interval for an individual coefficient in β, without re-estimating your model. (Be specific about what you mean by "reliable".)

e. Suppose you did not make use of any of the alternative approaches in part (c). Does OLS produce an unbiased set of estimates of the elements in β? Are they asymptotically unbiased? Are they consistent? Are they asymptotically normal? Are they efficient? (For each property, give a justification for your answer.)

Question Four

For the same setup as in Question Three, you would like to test a set of linear hypotheses of the form H:Rβ=r. R is a J by K matrix of known constants and r is a J$\times$1 vector of known constants.

a. Demonstrate how to construct a Wald test of your hypotheses.

b. Demonstrate how to construct a likelihood-ratio test of the same set of hypotheses.

c. Explain why each test reduces to a t-test when J=1, or an F-test when J>1.

Question Five

Suppose y_i is an *iid* random variable measuring the amount of time taken for some event to occur (e.g., the length of time until an individual adopts a new technology). Let x_i denote a vector of explanatory variables. The true distribution of $y_i \mid x_i$ is exponential, i.e.,

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

where $\lambda_i = \exp(x_i'\beta)$. The probability density function for the exponential distribution is $f(y_i \mid x_i, \beta) = \lambda_i \exp(-y_i\lambda_i)$. Under the exponential distribution, the mean and variance of $y_i|x_i$ are $E[y_i \mid x_i] = \exp(-x_i'\beta_0)$ and $\text{var}[y_i \mid x_i] = \exp(-x_i'\beta_0)^2$.

a. Write down the log likelihood function for the unknown parameters β_0.

b. Suppose you fit a linear regression model, i.e., you specify the model:

$$E\left[x_i\left(y_i - x_i'\alpha_0\right)\right] = 0$$

Interpret the parameter α_0.

c. It is possible to show that $\sqrt{n}\left(\hat{\alpha} - \alpha_0\right) \xrightarrow{d} N(0, V_0)$, where n denotes the sample size, $\hat{\alpha}$ is the estimator in (b), and $V_0 = (E[x_i x_i'])^{-1} E[x_i x_i'(y_i - x_i'\alpha_0)^2](E[x_i x_i'])^{-1}$. Write down a consistent estimator for V_0.

d. Would the moment condition $E\left[x_i(y_i - \exp(-x_i'\beta_0))\right] = 0$ yield a consistent estimate of β_0? Justify your answer. A detailed proof is not necessary and an intuitive answer in words is sufficient.

e. Suppose you estimate β_0 using GMM with the moment conditions:

$$E\left[x_i\left(y_i - \exp(-x_i'\beta_0)\right)\right] = 0$$

$$E\left[\left(y_i - \exp(-x_i'\beta_0)\right)^2 - \exp(-x_i'\beta_0)^2\right] = 0$$

to estimate β_0. Would your estimator be consistent for β_0 and would it be more efficient than an estimate based on the first moment condition only? A detailed proof is not necessary and an intuitive answer in words is sufficient.