Statistics and Econometrics Preliminary Examination, July 2002

There are **SIX** questions on three pages. Please answer **ALL** questions.

Question One:

Suppose that 30% of all automobile accidents are partly caused by weather conditions, and that 20% of all automobile accidents involve bodily injury. Further, of those accidents that involve bodily injury, 40% are partly caused by weather conditions.

a) What is the probability that a randomly chosen accident is partly caused by weather conditions and involves bodily injury?

b) If a randomly chosen accident was partly caused by weather conditions, what is the probability that it involved bodily injury?

c) Are the events "partly caused by weather conditions" and "involved bodily injury independent? Why or why not?

d) Suppose that you have data on 1000 randomly sampled auto accidents, but do not have the percentage information that was given at the start of this question. How would you test the hypothesis that the events "partly caused by weather conditions" and "involved bodily injury" are independent? Be specific.

Question Two:

For the model $BY + \Gamma X = U$, formally derive the rank and order conditions for identification by exclusion restrictions. Briefly explain what they mean. Verbally, if an equation is not identified by these conditions, what does it mean?

Question Three:

a) Verbally, what are the differences in design of Wald, LM and likelihood ratio tests?

b) Why are LM tests particularly useful for heteroskedasticity?

c) Why do LM tests for most models of heteroskedasticity not involve the regression coefficients? For what famous model is this not true?

Question Four:

You have been asked to estimate a single equation model

(1) $y = X\beta + u$

where X is a matrix of regressors including a constant. Assume that

(2) $u_t = \rho u_{t-1} + \varepsilon_t$,

where all of the classical assumptions hold for ε_t. Assume that ρ is nonzero and known. There are (at least) two methods (OLS and GLS) which could be used to estimate this model.

a) What are the consequences in terms of point estimates and significance tests of using OLS to estimate the model if (2) holds?

b) Formally derive the covariance matrix for the OLS estimates of β in model (1) above assuming (2) holds.

c) Which method (GLS or OLS) has the smaller variance? Why?

d) Is the conventionally calculated OLS variance smaller than the true OLS variance? Based on Davidson and MacKinnon's findings for heteroscedasticity are you likely to find much difference between conventional OLS variance and correct OLS variance? Between conventional OLS and GLS?

Question Five:

a) Find the OLS estimator for β for the model below which is given in scalar notation (i.e. there is no intercept):

$Y_t = \beta X_t + u_t$

Make assumptions about $E(u_t)$ and X_t and the model. Can you test the $E(u_t) = 0$ assumption?

b) Find $E(\hat{\beta})$ and $var(\hat{\beta})$ for the OLS estimator for β in (a).

c) Compare this model and its results with the one with an intercept which was estimated by OLS. Use standard OLS assumptions. Assume the model in (a) is true model.

d) Consider the estimator for β given by

$\hat{\beta}_1 = \bar{Y} / \bar{X}$

Find $E(\hat{\beta}_1)$ and $var(\hat{\beta}_1)$.

e) Compare the two estimators in parts (a) and (d). Which one do you like best? Why? Which estimator would you use? Explain.

Question Six:

Suppose you are interested in estimating β in the model

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

where $\varepsilon \sim N(0, \sigma^2 I)$. Note: The probability density function for a random variable x with mean μ and variance σ^2 is

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

a) Write down the likelihood function for this model.

b) Show that the maximum likelihood estimator is consistent for β. State any assumptions that you need.

c) Explain in words what it means to say that an estimator is consistent.

d) Is the maximum likelihood estimator in (b) unbiased? Justify your answer.

8/20/02

Statistics and Econometrics Preliminary Examination, Sept 2002

There are **SIX** questions on four pages. Please answer **ALL** questions.

Question One:

Suppose you are interested in testing the hypothesis that a farm would be more profitable if its manager were given a higher ownership stake in the farm. You have data from a number of farms on profits, equity structure, and background variables such as farm size and type of crop grown. Call these background variables X. Your model is

$$y = O\beta + X\gamma + \varepsilon$$

where y signifies net profit and O is the proportion of the farm that is owned by the manager. Assume that X includes a constant. Assume also that profitability (y) and ownership (O) are hereditary so that family history affects both variables. There is no measure of family history in X.

a) Suppose that ε_i is independent and identically distributed. Explain why OLS will provide an inconsistent estimate of β in this model?

b) Suppose that ε_i is independent and identically distributed. Describe how you could use instrumental variables to obtain a consistent estimate of β. What properties would your instruments need to have?

c) Suppose your sample includes 100 farms and you have five annual observations from each farm for a total of 500 observations. For almost all of the farms, the variable O does not change over the five years. Given this, your friend suggests that you need only use the most recent observation from each farm. Is your friend right? Why or why not?

d) Given the data structure in (c), the errors ε_i may not be independently distributed. Explain why this is so and describe the properties of the instrumental variables estimator in this case.

Question Two:

Set up (do not derive) the instrumental variables estimation (IVE) form of two stage least squares (2SLS) and formally show that it is consistent. What are the conditions required of a suitable instrument?

b) Verbally and informally, why is 2SLS asymptotically efficient in the class of single equation estimators?

Question Three:

Consider the linear dynamic m-equation simultaneous equation model

$$B_0 y_t + B_1 y_{t-1} + \Gamma x_t = U_t, \qquad U_t \sim N(0, \Sigma)$$

a) If the $m \times m$ matrix B_0 is diagonal, how would you estimate the parameters of the model and why? (Answer verbally with supporting logic).

b) Does your answer to question (a) above change depending on whether the $m \times m$ matrix B_1 is diagonal? Why or why not? (Answer very briefly).

c) Write the reduced form of the model algebraically for the general case of no diagonal matrices.

d) Algebraically, how would you determine whether or not the system is stable?

Question Four:

You are given two estimators, E_1 and E_2. E_1 is biased and E_2 is unbiased. We also know that $\mathrm{var}(E_1) < \mathrm{var}(E_2)$.

a) How would you choose between the two estimators?

b) Consider an estimator for the variance of a sample,

$$s^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

where $E(s^2) = \sigma^2$ and $\mathrm{var}(s^2) = 2\sigma^4/(n-1)$. Consider another estimator,

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = (n-1)s^2/n.$$

Derive the bias in $\hat{\sigma}^2$.

c) By the criterion that you gave in part (a), demonstrate which estimator is preferred, s^2 or $\hat{\sigma}^2$.

Question Five:

Below are the OLS estimates of a supply and demand model for pork based on aggregate annual time series data. The numbers below the coefficients in parentheses are the t-ratios. All regressions were run with T=30. No observations were lost.

Demand:

$$\underset{}{PPK_t} = \underset{(1.1)}{20.0} - \underset{(2.5)}{2.1}\, QPK_t + u_t$$

$R^2 = .52$ DW = 2.7 T = 30

Supply:

$$QPK_t = \underset{(2.61)}{4.0} + \underset{(2.33)}{5.2}\, PPK_t + \underset{(3.98)}{3.8}\, PC_t + \underset{(4.32)}{0.56}\, QPK_{t-1} + v_t$$

$R^2 = .68$ DW = 2.2 T = 30

The variables are: PPK-Price of Pork, QPK-Quantity of Pork, and PC-Price of corn.

DW Test critical values @ 5%:

T(or n)	k′	d_l	d_u
30	1	1.352	1.489
30	2	1.284	1.567
30	3	1.214	1.650
29	3	1.341	1.483

k′ = number of independent variables excluding the constant term.

a) Does the demand relation for pork have an autocorrelation problem? Justify your answer.

b) Assume that there is autocorrelation in the demand equation and that it is due to an omitted variable. Explain the consequences in terms of parameter estimates and their variances.

c) The supply relation has a lagged dependent variable. This implies two possible models - Partial Adjustment and Adaptive Expectations. Discuss both and indicate any statistical or specification problems you may have with either. How would you recover the true parameters? Consider the model as a single equation with no simultaneity.

d) Are there any other problems with this model? Discuss.

Question Six:

Suppose you are interested in estimating β in the model

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

where $E(\varepsilon_i \mid X) = 0$, $E(\varepsilon_i^2 \mid X) = \sigma_i^2$, and $E(\varepsilon_i \varepsilon_j \mid X) = 0$ for all $i \neq j$.

a) Show that the OLS estimator is unbiased.

b) Describe in words the properties of the OLS estimator for β and the conventionally calculated estimator for the standard error of this estimator.

c) How you would you compute an asymptotically correct t-statistic from the OLS estimate of β? Be specific.

d) Suppose you know that σ_i^2 is related to the variables in X but you do not know the correct model for σ_i^2. Describe how you would go about specifying a heteroskedasticity model and how you would construct a GLS estimator for β. Compare the properties of your GLS estimator to OLS.

University of California, Davis
Department of Agricultural and Resource Economics

July 8, 2003 4-hour exam

SPRING 2003 ECONOMETRICS PRELIMINARY EXAMINATION

1. The most efficient estimator of β, when $y_{1t} = \beta + \varepsilon_{1t}$; $y_{2t} = \varepsilon_{2t}$, and $Cov(\varepsilon_{1t}, \varepsilon_{2t}) > 0$, differs from the most efficient estimator of β when the model is the same, except that $y_{2t} = \gamma + \varepsilon_{2t}$. (Note that γ, like β, is unknown.)

   Why is this the case?

2. A woman claims that she is able to tell whether a cup of tea has been made with tea bags or with loose tea leaves. In order to test her claim, she is asked to taste eight pairs of cups of tea. One cup of each pair is made with a tea bag and the other with loose tea leaves of the same type of tea. Suppose she is able to identify correctly in six cases.

   a. State appropriate null and alternative hypotheses.

   b. What is the probability that she would correctly identify six cases, if the null hypothesis is true?

   c. Find the size of the following choices for the critical region for your test:

      i. C1: X=3
      ii. C2: X≥7
      iii. C3: X=8

   d. Suppose that she does, in fact, have some ability to detect methods for brewing tea, and that she has p=2/3 probability of finding the cup brewed with loose tea leaves.

Find the probability of a type II error for the three critical regions in part (c).

e. Give the best critical region of size $\alpha = .0352$.

3. The covariance of two random variables X and Y is defined by

$$Cov(X,Y) = E\{[X-E(X)][Y-E(Y)]\}.$$

X and Y are jointly distributed, continuous, dependent random variables.

a. Prove that the covariance between X and Y can be found using the formula below:

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

For the rest of this question, assume that you will have available n observations on the pair (X_t,Y_t), $t=1,2,\ldots,n$.

b. Give an unbiased estimator of E(XY). (Give a proof that your choice is indeed unbiased.)

c. Prove that your estimator is consistent, stating any assumptions you require.

d. Now consider estimating the covariance of X and Y, using your estimator for E(XY), and the product of respective sample means for X and Y, to estimate E(X)E(Y).

Is the estimate of the covariance that results an unbiased estimator of Cov(X,Y)?

e. If it is not unbiased, give an estimator of Cov(X,Y) that is unbiased, and justify your answer.

f. Prove that your new estimator of Cov(X,Y) is consistent.

4. Consider the regression model without an intercept term, $Y_t = \beta X_t + u_t$ (so the true value of the intercept, β_0, is zero). The assumptions of the classical linear regression model hold.

   a. What is the appropriate measure of goodness of fit in this model? Be sure to mention why the usual R^2 measure is not appropriate in this case.

   b. Now suppose the true model is $\ln Y_t = \beta \ln X_t + \varepsilon_t$ where all the classical linear model assumptions are met. The researcher, however, estimates the functional form $Y_t = \beta X_t + u_t$.

   Show that the least squares estimator obtained from this regression is biased and inconsistent. (Hint: $E(e^u) = e^{\mu + (\sigma^2/2)}$ where μ is the mean of u and σ^2 is the variance of u).

5. Consider the model $Y_t = \alpha + \beta X_t + \varepsilon_t$ where $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$. Assume that $|\rho| < 1$ and u_t satisfies all the assumptions of the classical linear regression model.

   a. Derive the conditional* maximum likelihood estimators of α and β. Assume that ρ is known. (*The conditional likelihood function is over observations from 2 to T, conditional on the first observation.)

   b. How do the conditional maximum likelihood estimators differ from the BLUE estimators of α and β?

   c. Now suppose that ρ is unknown. Derive the conditional maximum likelihood estimators of the coefficients, α and β.

6. We have used Ordinary Least Squares (OLS) to estimate the parameters of a model

$$y^{(n \times 1)} = X^{(n \times K)} \beta^{(K \times 1)} + \varepsilon^{(n \times 1)}$$

   that meets the classical assumptions except that each element of the error has a different variance, i.e., $\varepsilon_i \sim N(0, \sigma_i^2)$.

a. Formally obtain the mean of $\hat{\beta}$, the OLS estimate of β. Is $\hat{\beta}$ biased or unbiased?

b. Formally obtain the covariance matrix of the OLS estimate of β.

c. Suppose we ignore the different error variances and use the conventionally calculated covariance matrix for $\hat{\beta}$, that is, $s^2(X'X)^{-1}$; what effect will this have on hypothesis tests based on $\hat{\beta}$ and $s^2(X'X)^{-1}$?

d. Hal White became interested in the matrix $X'\Sigma X$ where $\Sigma = diagonal(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$ and proposed substituting squared residuals, $\hat{\varepsilon}_i^2$ for each of the diagonal elements σ_i^2 in Σ. For what purpose did Professor White suggest we use this matrix?

e. Since Σ grows with the sample size, we cannot obtain consistent estimates of its elements. Verbally only, how does Professor White handle this problem?

f. Suppose we know $\sigma_i^2 = \sigma^2 Z_i^2$ with Z_i known, and we estimate using this information. Verbally only, what properties can we claim for our estimates?

6. For the linear simultaneous equation model

$$By_t + \Gamma X_t = u_t \quad t = 1,2,\ldots,T$$

a. Formally derive the order condition for identification by exclusion restrictions using the Cowles Commission notation used in class.

b. Formally derive the associated rank condition for identifiability, explaining it briefly in terms of solution of the linear equations $Ax = b$.

c. Suppose we know a relation between elements of B and Γ in a given equation, say $B_{12} + 3\Gamma_{11} = 10$. How does that affect the identification in parts a and b above?

Statistics and Econometrics Preliminary Examination, Sept 2003

There are **FOUR** questions on three pages. Please answer **ALL** questions.

Question One

Let X be the length of life of a lightbulb manufactured by a certain company. Suppose X has density function corresponding to the *Exponential* distribution with parameter β:

$$f(x) = (1/\beta) \exp(-x/\beta) \quad x > 0 \text{ and } 0 \text{ otherwise.}$$

a) Find an expression for the cumulative distribution function (*cdf*) $F(a)$ for any a.

b) Find the critical region that corresponds to the most powerful test of the null hypothesis that $\beta=1$ against the alternative hypothesis that $\beta=2$. Base your test on a sample size of $n=1$.

c) Find an expression for the power of your test.

d) Describe how your answers would change if your alternative hypothesis was $\beta=4$?

e) Describe how your answer would change if (instead of changing your alternative hypothesis) you based your test on a sample of size $n=10$?

f) Explain in words how your findings, with the critical region you have proposed above, compare to corresponding results for any other critical region with the same *size*.

Question Two:

Consider the regression model $y = X\beta + \varepsilon$, where X is nonstochastic. Suppose the error term meets the classical assumptions except for an AR(1) error

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

The serially correlated error ε is assumed to be stationary.

a) Show the covariance matrix of the serially correlated error, with elements written in terms of the model parameters.

b) Write the serially correlated error in terms of the lag operator L and formally use the lag operator to solve the difference equation (invert the lag) writing ε in terms of past values of u. Do the algebra necessary to generate the first three weights in the sequence. How far back does ε depend on u?

c) Suppose β is estimated by Ordinary Least Squares despite the serially correlated error. Verbally, what are the implications for parameter estimation and hypothesis testing using the conventionally calculated OLS statistics in this case?

d) Formally derive the covariance matrix of the OLS estimator when the error follows an AR(1) process.

Question Three:

Suppose you want to find out whether electricity wholesalers in California manipulated the market between May and December 2001. A symptom of market manipulation is an excessive number of unscheduled outages. You have hourly data on unscheduled outages for the period January 1998 to December 2001 and you also have data on two explanatory variables, scheduled outages and forecast load.

Let y signify the dependent variable (unscheduled outages) and let the matrix X contain the explanatory variables and a constant. Your goal is to test whether the relationship between y and X changed after May 2001.

(Aside: The Federal Energy Regulatory Commission just found that the citizens of California were overcharged $1 million by manipulation of the market for electricity. A California commission charged with estimating the damages arrived at a number slightly in excess of $8 billion.)

a) Suppose the error term in your regression model meets the classical assumptions. Write down the regression model that you would estimate, the null hypothesis that you would test, the test statistic you would use, and the distribution of this test statistic.

b) Suppose the error term in your regression model meets the classical assumptions except for an AR(1) error

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

(The serially correlated error ε is assumed to be stationary.) Write down the conditional log likelihood function for your model. Be explicit regarding which error, ε_t or u_t, underlies the criterion.

c) Would the Maximum Likelihood Estimates be the same as the Generalized Least Squares estimates for the model with AR(1) errors? Show why or why not.

d) Given that your model has AR(1) errors, how would you test the hypothesis of interest? Briefly motivate your test choice. Write the null and alternate hypotheses, the test statistic, and its distribution. Be specific about how you would treat the autocorrelated error term and exactly which error underlies the test (ε_t or u_t).

e) Suppose that, in addition to the autocorrelated error, we wish to test the hypothesis that the error is heteroskedastic, with one constant variance before May 2001, and another, larger, constant variance for errors occurring May 2001 or later. How would you test this using a Goldfeld-Quandt test? As usual, give the null and alternate hypotheses, the test statistic, and its distribution. Be specific about how you would treat the autocorrelated error, and exactly which error you are testing (ε_t or u_t).

Question Four:

Consider the regression model $y = X\beta + \varepsilon$, where X is nonstochastic and $\varepsilon \sim N(0,\sigma^2 I)$. Suppose β is estimated using the formula $\tilde{\beta} = (Z'X)^{-1}Z'y$, where Z is nonstochastic.

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

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

c) Write down the distribution of the estimator $\tilde{\beta}$.

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

Statistics and Econometrics Preliminary Examination, July 2004

There are **FOUR** questions on three pages. Please answer **ALL** questions.

Question One:

Suppose that the following table describes the joint probability distribution of the wealth and education status of California individuals.

	Wealthy	Not Wealthy
Well Educated	0.10	0.15
Not Well Educated	0.25	0.50

a) What is the probability that a randomly chosen California individual is wealthy?

b) Is the wealth status of California individuals independent of their education status? Justify your answer.

c) Suppose you are told that the mean wealth of well-educated people is equal to the mean wealth of people who are not well educated. Is this assertion consistent with the probability distribution in the above table? Explain.

d) Suppose that you collect data from a representative sample of 1000 California individuals. These data include measurements of total wealth and of education status. Describe how you would test the assertion in part (b). Be specific and state any assumptions that you make. (Note: You do not know the exact distribution of wealth or education).

e) Using the data set in (d), describe how you would test the hypothesis that the variance of the wealth of well-educated people is equal to the variance of the wealth of people who are not well educated. Be specific and state any assumptions that you make. (Note: You do not know the exact distribution of wealth or education).

Question Two:

A candidate for a chair in agricultural economics is attempting to show great ability in applied production economics by reporting the results of a model relating annual observations on a particular California farm's pistachio yields (y) to irrigation (w), fertilization (f), and degree days (d):

$$y_t = \beta_0 + \beta_1 w_t + \beta_2 f_t + \beta_3 d_t + \varepsilon_t ,$$

assuming that the error term ε_t meets the classical assumptions. The candidate has found the rather controversial result that degree days are not important (or significant) in pistachio yield and has recommended a massive government program establishing pistachio orchards in previously agriculturally insignificant portions of Alaska.

a) The farmer supplying the data had control of irrigation and fertilization and coordinated them with each other and with the condition of the plants in the field. Is there a simultaneous equations problem here? Why or why not? Please be explicit.

b) How might your answer in (a) differ if data were available at the daily frequency rather than the annual frequency?

c) If there *is not* a simultaneous equation problem, how would you recommend estimation of the equation above? Why?

d) If there *is* a simultaneous equation problem, how would you recommend estimation of the equation above, and why?

e) Suppose that using the same data, a graduate student ran a regression relating yield to a constant and degree days and found degree days to be both significant and important. He is attempting to halt the Alaskan pistachio operation based on this information. The candidate has criticized the student for omitting obviously important variables (w and f) and reaching patently erroneous conclusions. Who is right, and why?

<u>Question Three:</u>

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

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

Question Four:

Suppose you are modeling the demand for three goods: pork (P), beef (B), and chicken (C). Consider the demand model:

$$w_{Pi} = \alpha_P + \gamma_{PP} \ln p_{Pi} + \gamma_{PB} \ln p_{Bi} + \gamma_{PC} \ln p_{Ci} + \beta_P \ln(y_i / P_i) + \varepsilon_{Pi}$$
$$w_{Bi} = \alpha_B + \gamma_{BP} \ln p_{Pi} + \gamma_{BB} \ln p_{Bi} + \gamma_{BC} \ln p_{Bi} + \beta_B \ln(y_i / P_i) + \varepsilon_{Bi}$$
$$w_{Ci} = \alpha_C + \gamma_{CP} \ln p_{Pi} + \gamma_{CB} \ln p_{Bi} + \gamma_{CC} \ln p_{Bi} + \beta_C \ln(y_i / P_i) + \varepsilon_{Ci} ,$$

where w_{ki}, is the share of individual i's meat budget devoted to good k, p_{ki} is the price of good k faced by individual i, y_i is total expenditure on the three goods by individual i, and $\ln(P_i)$ is a translog price index for individual i.

a) How does the seemingly unrelated regression estimator (SUR) compare to the ordinary least squares estimator (OLS) for the above demand system? (For the moment, you may ignore the adding-up restriction $w_{Pi} + w_{Bi} + w_{Ci} = 1$ and any theoretical restrictions on the parameters.)

b) Prove that the adding-up condition, $w_{Pi} + w_{Bi} + w_{Ci} = 1$, implies that the variance-covariance matrix of ε_{ik} is singular. Given this information, what estimation procedure would you employ? (For the moment, you may ignore any theoretical restrictions on the parameters.)

c) Suppose you wish to test whether the homogeneity condition holds, that is, test whether $\gamma_{kP} + \gamma_{kB} + \gamma_{kC} = 1$ for all k. State the null and alternative hypotheses, the test statistic and its null distribution, and the criterion for rejection of the hypothesis.

d) Suppose you wish to test whether the homogeneity condition holds, that is, test whether symmetry conditions hold, that is, test whether $\gamma_{jk} = \gamma_{kj}$ for all $k \neq j$. State the null and alternative hypotheses, the test statistic and its null distribution, and the criterion for rejection of the hypothesis.

e) What estimation procedure would you use when the homogeneity and symmetry restrictions in parts (c) and (d) are imposed? Be specific.

Statistics and Econometrics Preliminary Examination, July 2004

There are **FOUR** questions on three pages. Please answer **ALL** questions.

Question One:

Suppose that the following table describes the joint probability distribution of the wealth and education status of California individuals.

	Wealthy	Not Wealthy
Well Educated	0.10	0.15
Not Well Educated	0.25	0.50

a) What is the probability that a randomly chosen California individual is wealthy?

b) Is the wealth status of California individuals independent of their education status? Justify your answer.

c) Suppose you are told that the mean wealth of well-educated people is equal to the mean wealth of people who are not well educated. Is this assertion consistent with the probability distribution in the above table? Explain.

d) Suppose that you collect data from a representative sample of 1000 California individuals. These data include total wealth (measured in dollars) and of education (measured in years). Describe how you would test the assertion in part (c). Be specific and state any assumptions that you make. (Note: You do not know the exact distribution of wealth or education).

e) Using the data set in (d), describe how you would test the hypothesis that the variance of the wealth of well-educated people is equal to the variance of the wealth of people who are not well educated. Be specific and state any assumptions that you make. (Note: You do not know the exact distribution of wealth or education).

Question Two:

A candidate for a chair in agricultural economics is attempting to show great ability in applied production economics by reporting the results of a model relating annual observations on a particular California farm's pistachio yields (y) to irrigation (w), fertilization (f), and degree days (d):

$$y_t = \beta_0 + \beta_1 w_t + \beta_2 f_t + \beta_3 d_t + \varepsilon_t ,$$

assuming that the error term ε_t meets the classical assumptions. The candidate has found the rather controversial result that degree days are not important (or significant) in pistachio yield and has recommended a massive government program establishing pistachio orchards in previously agriculturally insignificant portions of Alaska.

a) The farmer supplying the data had control of irrigation and fertilization and coordinated them with each other and with the condition of the plants in the field. Is there a simultaneous equations problem here? Why or why not? Please be explicit.

b) How might your answer in (a) differ if data were available at the daily frequency rather than the annual frequency?

c) If there *is not* a simultaneous equation problem, how would you recommend estimation of the equation above? Why?

d) If there *is* a simultaneous equation problem, how would you recommend estimation of the equation above, and why?

e) Suppose that using the same data, a graduate student ran a regression relating yield to a constant and degree days and found degree days to be both significant and important. He is attempting to halt the Alaskan pistachio operation based on this information. The candidate has criticized the student for omitting obviously important variables (w and f) and reaching patently erroneous conclusions. Who is right, and why?

Question Three:

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

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

Question Four:

Suppose you are modeling the demand for three goods: pork (P), beef (B), and chicken (C). Consider the demand model:

$$w_{Pi} = \alpha_P + \gamma_{PP} \ln p_{Pi} + \gamma_{PB} \ln p_{Bi} + \gamma_{PC} \ln p_{Ci} + \beta_P \ln(y_i) + \varepsilon_{Pi}$$
$$w_{Bi} = \alpha_B + \gamma_{BP} \ln p_{Pi} + \gamma_{BB} \ln p_{Bi} + \gamma_{BC} \ln p_{Bi} + \beta_B \ln(y_i) + \varepsilon_{Bi}$$
$$w_{Ci} = \alpha_C + \gamma_{CP} \ln p_{Pi} + \gamma_{CB} \ln p_{Bi} + \gamma_{CC} \ln p_{Bi} + \beta_C \ln(y_i) + \varepsilon_{Ci},$$

where w_{ki}, is the share of individual i's meat budget devoted to good k, p_{ki} is the price of good k faced by individual i, and y_i is total real expenditure on the three goods by individual i.

a) How does the seemingly unrelated regression estimator (SUR) compare to the ordinary least squares estimator (OLS) for the above demand system? (For the moment, you may ignore the adding-up restriction $w_{Pi} + w_{Bi} + w_{Ci} = 1$ and any theoretical restrictions on the parameters.)

b) Prove that the adding-up condition, $w_{Pi} + w_{Bi} + w_{Ci} = 1$, implies that the variance-covariance matrix of ε_{ik} is singular. Given this information, what estimation procedure would you employ? (For the moment, you may ignore any theoretical restrictions on the parameters.)

c) Suppose you wish to test whether the homogeneity condition holds, that is, test whether $\gamma_{kP} + \gamma_{kB} + \gamma_{kC} = 1$ for all k. State the null and alternative hypotheses, the test statistic and its null distribution, and the criterion for rejection of the hypothesis.

d) Suppose you wish to test whether the symmetry conditions hold, that is, test whether $\gamma_{jk} = \gamma_{kj}$ for all $k \neq j$. State the null and alternative hypotheses, the test statistic and its null distribution, and the criterion for rejection of the hypothesis.

e) What estimation procedure would you use when the homogeneity and symmetry restrictions in parts (c) and (d) are imposed? Be specific.

2005 Ph.D. Econometrics Prelim Examination

1. Two players, A and B, alternately and independently flip a coin and the first player to obtain a head wins. Assume player A flips first.

   a. If the coin if fair, what is the probability that A wins?

   b. Suppose that $P(head) = p$, not necessarily $1/2$. What is the probability that A wins?

   c. Show, by any means necessary, that for all $p,\ 0 < p < 1,\ P(A\,wins) > 1/2.$

   d. Find the maximum likelihood estimator of p and its probability distribution.

2. a. According to the Law of Iterated Expectations,

$$E_X\left(E_Y\left(Y \mid X\right)\right) = E(Y)$$

Prove that this is correct.

b. (True or False) If the conditional mean of Y given X does not depend on X, then Y and X are uncorrelated.

c. Show that the converse is not true, i.e., Y and X may be uncorrelated, but E(Y|X) still depends on X. You may find it useful to consider the case where X and Z are independent N(0,1) random variables and Y=X^2+Z.

3. Consider the following classical linear regression model

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t \tag{1}$$

Suppose X is not observable, but we observe a closely related variable X^*, which is determined as follows:

$$X_t^* = \beta_3 + \beta_4 X_t + u_t, \tag{2}$$

where u_t is a normally and independently distributed variable with mean zero and variance σ_u^2. Furthermore, u_t and ε_t are independent for all $i, j = 1, 2, ..., n.$

By solving for X_t in terms of X_t^* and substituting this expression into the original model, we obtain

$$Y_t = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t^* \qquad (3)$$

where $\beta^* s$ are functions of the original parameters.

a. Show that the least squares estimator of β^* is <u>not</u> a consistent estimator of β_2.

b. What assumptions would you need in order to get a consistent estimator of β_2?

c. X^* is a proxy for X. What is the difference between a proxy and an instrumental variable? Be specific and state all the assumptions in both cases.

4. You are interested in modeling the process by which y is generated. The true data-generating process is

$$y = X\beta + Z\gamma + \varepsilon$$

but you estimated

$$y = X\beta + \varepsilon$$

a. Demonstrate the conditions under which OLS leads to an unbiased estimator for beta, β.

b. Suppose that those conditions are not met. Give a formula for the bias in the OLS estimator.

c. Is the OLS estimator asymptotically unbiased, for the same circumstance as in (b)?

d. Is the OLS estimator consistent, for the same circumstance as in (b)?

e. Explain why instrumental variables may be used to solve the problem in (d). State any assumptions you require for this result.

Question Five:

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 $\tilde{\beta} = \sum_{i=1}^{n} i^{-1/2} y_i \big/ \sum_{i=1}^{n} i^{-1/2} x_i$.

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

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

c) Write down the distribution of the estimator $\tilde{\beta}$.

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

e) Suppose that the model exhibits heteroskedasticity such that $\text{var}(\varepsilon_i) \sim i\sigma^2$. Now would you recommend using the estimator $\tilde{\beta}$ instead of OLS? Why or why not? (Some mathematics may be useful, but a logical answer in words is sufficient.)

Question Six:

Consider the regression model $y = X\beta + \varepsilon$, where X nonstochastic. Suppose the error term meets the classical assumptions except for an AR(1) error

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

where $0 < \rho < 1$.

a) Write down the covariance matrix of the serially correlated error term ε.

b) Verbally, what are the properties of the ordinary least squares estimate of β?

c) Derive the covariance matrix of the OLS estimate of β.

Econometrics Preliminary Examination

Department of Agricultural and Resource Economics
University of California, Davis
July 10, 2006

Answer each of the four questions.

The four questions will be given equal weight, as will each part within each question.

1. You are interested in studying the pattern of Heads and Tails when a coin is flipped. You do not assume that the coin is "fair", i.e., it need not be the case that $P(\quad eads) = P(Tai\quad)$.

   (a) Give an appropriate probability model for the outcome from a single coin flip.

   (b) Letting a 1 denote Heads and 0 denote Tails, find the moment-generating function, expected value, and variance of this random variable.

   (c) Use the moment-generating function of the random variable in (b) to find the moment-generating function, expected value, and variance of a new random variable, the number of times Heads shows when the coin in flipped in n independent trials.

   (d) Assuming that you will have n independent observations recording the outcome from flipping the coin, find the maximum likelihood estimator of the probability that the coin shows Heads.

   (e) Is your estimator unbiased? (Give a complete justification for your answer.)

   (f) Is your estimator consistent? (Give a complete justification for your answer.)

2. You are interested in a linear regression model of the form

$$y = X \qquad e$$

   where all OLS assumptions but one are satisfied.

   The one that is not satisfied concerns the X matrix; you believe that some of the variables in X are correlated with the error term.

   Hence, you consider an instrumental variables approach to estimating β.

   (a) Assuming that you make use of a matrix of intrumental variables Z, of the same dimension as X, give the instrumental estimator you will use.

(b) Prove that your estimator is consistent, stating any assumptions you require.

(c) To test a hypothesis about the vector, you require an estimated asymptotic variance-covariance matrix for your estimator.

Show how to obtain such a matrix, beginning with the asymptotic distribution of

$$\sqrt{n} \quad \hat{} \; - \quad .$$

State any assumptions you require.

(d) Suppose you wanted to test set of nonlinear hypotheses, of the form

$$h \quad =$$

Using your result from (c), write down an appropriate test statistic and state its distribution under the null hypothesis. State any assumptions you require and be specific about how you would find each component of your test statistic.

3. (a) Explain the Law of Large Numbers.

(b) What is the difference between *convergence in mean square* (also termed *convergence in quadratic mean*) and convergence in probability?

(c) What is the Central Limit Theorem, and what is its role in mathematical statistics?

(d) How might you use the Central Limit Theorem to solve the following problem?

A casino employee is asked to deliver a box containing 900 new dice to the floor of the casino. Unfortunately, the employee trips while leaving the elevator and all of the dice are tossed high into the air, landing in various places on the casino floor.

A guest at the casino happens to know some probability theory, and, with the help of others, collects all 900 dice and records whether or not they show a 6. Let X denote the number of dice showing 6. Assuming that the dice are fair, use the Central Limit Theorem to approximate

$$Pr\ 1 \quad < X \qquad 5\ .$$

(e) Returning to your answers for problem 2, explain how each assumption you made in parts (b) or (c) relates to either the Law of Large Numbers or the Central Limit Theorem.

4. You have a data set containing data on salaries and various characteristics for individuals in the labor force. The salary of individual t is Y_t and there are two potential explanatory variables, X_{1t} and X_{2t}.

In addition, the age, sex, and marital status of individual observations are identified according to the variables $AG\quad_t$, F_t (which equals 1 for women and 0 for men), and M_t (which equals 1 for married individuals).

(a) You estimated the model

$$Y_t = X \qquad e$$

using OLS. X contains three columns: a column of ones, $X\ _t$, and $X\ _t$.

It turns out, however, that

$$V \quad e_t \quad = \sigma \ AG \ _t^{\gamma}$$

The parameter γ is unknown to you.

i. State the properties of .

ii. Are the usual standard errors from your OLS printout valid? Give a justification for your answer.

(b) How would you estimate ? (Give a justification for your answer.)

(c) How would you obtain standard errors for your ?

(d) It turns out that the correct specification for the variance is now known to be

$$V \quad e_t) = \sigma_F \cdot F_t \quad \sigma_M \cdot \qquad F_t$$

i.e., there is a different variance for male and female observations. Neither variance is known.

How would you estimate σ_F and σ_M?

(e) For σ_F only, is your estimator unbiased? Give a justification for your answer.

(f) For σ_F only, is your estimator consistent? Give a justification for your answer.

(g) Suppose you want to allow for the possibility that the coefficients for men and women are different.

Explain how you would estimate to account for this possibility. Would you estimate one equation or two?

(h) How does your approach in (g) compare to the Seemingly Unrelated Regressions estimator?

(i) It turns out that half of the observations in your sample are men, half are women, and the observations come in pairs—the individuals are twins, in fact. The first 50 observations are men, and observation 51 provides data on the twin sister of the man in observation 1, observation 52 corresponds to the twin sister of the man in observation 2, and so on.

Again, how would you estimate your model, and how would it compare to seemingly unrelated regressions?

Econometrics Preliminary Examination

Department of Agricultural and Resource Economics
Department of Economics
University of California, Davis

August 24, 2006

Answer each question completely. Each question will be given equal weight, as will each part within a question.

1. X and Y are continuous random variables, with joint probability density function

$$= c \qquad \leq \ x \qquad \leq \ 3$$

and $f \qquad = \quad$ otherwise.

   (a) Find c.
   (b) Find marginal density functions for X and Y.
   (c) Find the expected value and variance of X.
   (d) Find $Pr[X > 1|Y = 1]$.
   (e) Find $Pr[X > 1|Y \geq 1]$.
   (f) Are X and Y independent random variables?

2. For a random sample from a population of normal random variables with mean μ and variance σ^2, a researcher makes use of the sample mean and sample variance.

   (a) Prove that these two statistics are
      i. unbiased
      ii. consistent
      iii. independent

(b) Prove that the sample mean is efficient.

(c) The usual t-statistic is calculated as

$$t = \frac{\sqrt{n}\,\bar{\ }}{s}$$

What hypothesis is tested by this statistic? Explain specifically how it is decided whether to reject the hypothesis.

(d) Prove that t converges in distribution to a $N \quad 1$ random variable.

3. You are interested in a particular Canadian labor market, and have gathered data on the following variables:

Y = monthly salary;

$E = 1$ if the individual speaks English (0 otherwise);

$F = 1$ if the individual speaks French (0 otherwise);

X = years of experience in the labor market;

S = years of schooling;

$W = 1$ if the individual is a woman (0 otherwise).

Given n observations on a set of Canadians, some of whom speak English only, some of whom speak French only, and some of whom speak both languages, describe in detail the models you would estimate below, and exactly how you would test particular hypotheses.

Note that there is no one in the sample that does not speak at least one of these two languages.

(a) How would you estimate the effects of each characteristic on salary and then test the hypothesis that the language variables have no effect on salary?

(b) Does it matter whether you test the hypothesis that the coefficients on E and F are equal to each other, or the hypothesis that they are both equal to zero? Explain in detail.

(c) How would you test for discrimination against women in the Canadian labor force, interpreted as a lower monthly salary, holding characteristics constant?

(d) How would you test for differences between men and women in the returns to schooling?

(e) How would you test the hypothesis that a man who speaks only French earns the same amount, for given years of schooling and experience, as a woman who speaks both languages?

(f) Suppose that some individuals in the sample are more intelligent than others. More intelligent individuals tend to have high levels of schooling and to earn high salaries. How does this information change your answer to part (d)?

4. Each part of this question includes a description of the process by which data are generated. For each part, comment on the properties of OLS in the model

$${}_t = \beta_1 \quad \beta_2 \; {}_t \quad e_t$$

(a) The true data-generating process was

$${}^*_t = \beta_1 \quad \beta_2 \; {}_t$$

and

$${}^*_t = \; {}_1 \quad \gamma_2 z_t$$

but you observe only ${}_t$ and ${}_t$ where

$${}_t = \; {}^*_t \quad e_t$$

and

$${}_t = \; {}^*_t \qquad .$$

(b) The true data-generating process was

$${}^*_t = \beta_1 \quad \beta_2 \; {}^*_t$$

and

$$= \quad {}_2 z$$

but you observe only and where

$$=$$

and

$$=$$

(c) The true data-generating process is

$$= \quad {}_2$$

and

$$\sim \text{Uniform} \quad 1$$

but you observe only $=$ and the actual , and you observe only the combinations for which > 5.

(d) The true data-generating process is

$$= \quad \_$$

and

$$= \quad \_ \quad v$$

5. You are interested in estimating the parameter that relates a dependent variable to a *single* explanatory variable in the simple linear model

$$=$$

with 's known to be a set of independent and identically distributed $N \quad \sigma^2$ random variables. The variable is non-stochastic. σ^2 is unknown.

(a) Find the least-squares estimator $\hat{}$, and show that $\hat{}$ is unbiased and that $V \hat{} = \sigma^2 / \sum {}^2$.

(b) Prove that $\hat{}$ is a consistent estimator of . Give an example of an estimator of that is unbiased but not consistent. Can an estimator be consistent without being efficient? Why or why not?

(c) Do the residuals $\hat{} = - \hat{}$ sum to zero? Explain why or why not.

(d) What happens to the properties of your $\hat{}$ if the true model has a *non-zero value* for the intercept? Does your formula for $\hat{}$ ever yield an unbiased estimator for the parameter in

$$= \alpha \qquad ?$$

Suppose you decide to use the estimator with the smallest *mean-square error*—could you ever prefer the $\hat{}$ from (a) to the usual OLS estimator for in (d), given that $\alpha \neq$?

(e) Returning to the case where $\alpha =$, and assuming that σ^2 is unknown, how would you test the hypothesis that $= 1$? (Give a complete derivation of an *exact* test; do not use an approximation, such as by taking twice the difference in log-likelihoods.)

(f) Suppose you *did not* believe that the error term in the regression model in (a) was normally distributed. Which of your properties in (a) and (b) no longer hold?

Suppose you also relax the assumption that the 's are homoscedastic. Which of your properties in (a) and (b) no longer hold?