Sample Problems for the Final Examination

Economics 321

May 10, 2005

Please answer the following questions.

1. Prove the fundamental result known as "iterated expectations," which is as follows. Let $f(y,x)$ be the joint probability density function (pdf) of a pair of random variables Y and X, and let $h(x)$ be the marginal pdf of X. The expected value of Y conditional on $X = x$ is

$$\mathrm{E}(Y|X=x) = \int_Y y \frac{f(y,x)}{h(x)} dy,$$

and in general this quantity depends on x. You are to prove that

$$\mathrm{E}Y = \int_X \mathrm{E}(Y|X=x)\, h(x) dx,$$

or, when expressed in an equivalent way, that

$$\mathrm{E}Y = \mathrm{E}_X \left[\mathrm{E}(Y|X=x)\right].$$

2. Consider the simple linear model

$$Y_i = X_i\beta + \varepsilon_i,$$

in which X_i is a single explanatory variable and there is no constant term. Multiply this equation through by X_i to obtain

$$X_iY_i = X_i^2\beta + X_i\varepsilon_i,$$

and then add and take averages across observations,

$$\frac{1}{n}\sum_{i=1}^{n} X_iY_i = \frac{1}{n}\sum_{i=1}^{n} X_i^2\beta + \frac{1}{n}\sum_{i=1}^{n} X_i\varepsilon_i.$$

From this point, please explain how to obtain the ordinary least squares estimator $\hat{\beta}$. Be sure to state the assumptions that are needed to obtain this estimator, discuss where in your argument each such assumption is needed, and say why it is needed.

3. Consider the linear model with two explanatory variables (but no constant term)

$$Y_i = X_{i,1}\beta_1 + X_{i,2}\beta_2 + \varepsilon_i.$$

To analyze this model, multiply the equation through by $X_{i,1}$ to obtain

$$X_{i,1}Y_i = X_{i,1}^2\beta_1 + X_{i,1}X_{i,2}\beta_2 + X_{i,1}\varepsilon_i,$$

and then add and take averages across observations to create a new equation expressed in terms of averages. Then create a second new equation by multiplying through the model by $X_{i,2}$ and again adding and averaging.

With these two new equations in hand, please explain how to obtain the ordinary least squares estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$. Be sure to state the assumptions that are needed to obtain this estimator, discuss where in your argument each such assumption is needed, and say why it is needed.

4. Consider the formula for the ordinary least squares estimator

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

for the case of the simple linear model $Y_i = X_i\beta + \varepsilon_i$ with X_i being a single explanatory variable (there is no constant term), $\mathbf{X}$ an $n \times 1$ column vector with all such X_i values, and $\mathbf{Y}$ an $n \times 1$ column vector with all the Y_i's. Show that this formula can be derived by finding the value $\hat{\beta}$ that minimizes the sum of squares

$$S = \sum_{i=1}^{n} (Y_i - X_i\beta)^2.$$

5. With reference to the simple linear model $Y_i = X_i\beta + \varepsilon_i$, in which X_i represents a single explanatory variable and the model contains no constant term, consider the estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Find the expected value of $\hat{\beta}$ and its variance. Be sure to state carefully any assumptions you need to make, and say why you need to make them. Be sure to use the method of iterated expectations in your answer.

6. Consider the simple linear model $Y_i = X_i\beta + \varepsilon_i$, in which X_i is a single explanatory variable (there is no constant term). Using the estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ of its β parameter, define the i-th residual as $e_i = Y_i - X_i\hat{\beta}$ and let $\mathbf{e}$ be the $n \times 1$ vector of such residuals. Prove that $\mathbf{X}'\mathbf{e} = 0$.

   Note that this is an *algebraic result* and the proof does not require you to find expected values.

7. With reference to the simple linear model $Y_i = X_i\beta + \varepsilon_i$ (in which X_i is a single explanatory variable, and there is no constant term) consider the estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. The residual for observation i is $e_i = Y_i - X_i\hat{\beta}$, the $n \times 1$ vector of such residuals is denoted by $\mathbf{e}$, and s^2, the estimator of σ^2, is

$$s^2 = \frac{1}{n-1}\mathbf{e}'\mathbf{e}.$$

   Prove that $\mathrm{E}s^2 = \sigma^2$. Be sure to state carefully any assumptions you need to make, and say why you need to make them. Be sure to use the method of iterated expectations in your answer, and note that you will also need to make use of the trace of a matrix.

8. Consider the multivariate linear model $Y_i = \mathbf{X}_i'\beta + \varepsilon_i$ in which $\mathbf{X}_i$ is a $k \times 1$ vector of explanatory variables for observation i and β has k elements. The ordinary least squares estimator takes the form $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, in which $\mathbf{X}$ is an $n \times k$ matrix with $\mathbf{X}_i'$ in the i-th row.

   Consider the i-th residual $e_i = Y_i - X_i'\hat{\beta}$, the $n \times 1$ vector of such residuals $\mathbf{e}$, and s^2, the estimator of σ^2,

$$s^2 = \frac{1}{n-k}\mathbf{e}'\mathbf{e}.$$

   Prove that $\mathrm{E}s^2 = \sigma^2$. Be sure to state carefully any assumptions you need to make, and say why you need to make them. Be sure to use the method of iterated expectations in your answer, and you will also need to make use of the trace of a matrix.

9. For the simple linear model $Y_i = X_i\beta + \varepsilon_i$, which contains one explanatory variable and no constant term, make the assumption that ε_i is *normally distributed* with mean zero and variance σ^2. Using this assumption, explain how to test the null hypothesis $H_0 : \beta = \beta_0$ against the alternative hypothesis $H_A : \beta \neq \beta_0$. Derive the test statistic, say how it is distributed if the null

hypothesis is true, and explain how to calculate the rejection region for the test.

10. Consider the multivariate linear model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, in which $\mathbf{X}$ is of dimension $n \times k$, and the ordinary least squares estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Prove that the sum of squares of the dependent variable, $\sum_{i=1}^{n} Y_i^2$, can be partitioned into one part that is "explained" by $\mathbf{X}\hat{\beta}$ and a remaining part that is the sum of squares of the residuals. In other words, prove that

$$\mathbf{Y}'\mathbf{Y} = \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} + \mathbf{e}'\mathbf{e}.$$

Note that this is an *algebraic result* and the proof does not require you to find expected values.

You will recall that this formula provides the basis for one of the popular definitions of R^2,

$$R^2 = \frac{\hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}}{\mathbf{Y}'\mathbf{Y}},$$

which is referred to as the "uncentered" version of R^2.

11. Suppose that you specify your regression model as $Y_i = X_i\beta + \varepsilon_i$, including only one explanatory variable X_i and no constant term, and then estimate β using ordinary least squares, yielding $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Unfortunately, the true data-generating process is not what you have specified, but rather $Y_i = X_i\beta + Z_i\gamma + u_i$, with Z_i being a single additional explanatory variable that you have mistakenly omitted from your specification. Assume that $\mathrm{E}(\mathbf{u}|\mathbf{X},\mathbf{Z}) = \mathbf{0}_{n\times 1}$. Using this assumption, find the expected value of the $\hat{\beta}$ estimator that comes from a regression of $\mathbf{Y}$ on $\mathbf{X}$ alone.

12. Consider a multivariate linear model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ in which $\mathbf{X}$ is of dimension $n \times 4$ (it includes a column of ones associated with the regression constant term β_1 and three explanatory variables). You have estimated β by ordinary least squares, and have the estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and its estimated variance matrix $s^2(\mathbf{X}'\mathbf{X})^{-1}$.

Explain in detail how to test the null hypothesis $H_0 : \beta_1 = 5, \beta_3 + \beta_4 = 1$ using a Wald test. Be sure to show exactly how to calculate the test statistic, and describe the distribution of the test statistic when the null hypothesis is true. Also discuss how to calculate the critical region for the test using a significance level of 0.01.

13. Consider two regression models in which wage rates are the dependent variables, one model being estimated using a sample of data for women, and the other estimated with a sample of data for men. Assume that the two samples are independent.

    The regression models for women and men can be written as

$$\mathbf{Y}_W = \mathbf{X}_W \beta_W + \varepsilon_W$$
$$\mathbf{Y}_M = \mathbf{X}_M \beta_M + \varepsilon_M$$

    and we will assume that for women, $\mathrm{E}(\varepsilon_W | \mathbf{X}_W) = \mathbf{0}$, $\mathrm{E}(\varepsilon_W \varepsilon_W' | \mathbf{X}_W) = \sigma_W^2 \mathbf{I}$, and likewise for men, $\mathrm{E}(\varepsilon_M | \mathbf{X}_M) = \mathbf{0}$, $\mathrm{E}(\varepsilon_M \varepsilon_M' | \mathbf{X}_M) = \sigma_M^2 \mathbf{I}$. Note that we are allowing σ_W^2 and σ_M^2 to be different.

    Explain in detail how to test the null hypothesis $H_0 : \beta_W = \beta_M$ using a Wald test. Be sure to show exactly how to calculate the test statistic, and describe the distribution of the test statistic when the null hypothesis is true. Also discuss how to calculate the critical region for the test using a significance level of 0.01.

14. You have total cost and output data for a sample of firms, and specify the total cost function as

$$C_i = \beta_1 + \beta_2 Q_i + \beta_3 Q_i^2 + \beta_4 Q_i^3 + \varepsilon_i$$

    with C_i being the total costs for the i-th firm and Q_i being the level of output for that firm. Assume that $\mathrm{E}(\varepsilon | \mathbf{X}) = \mathbf{0}$ and $\mathrm{E}(\varepsilon \varepsilon' | \mathbf{X}) = \sigma^2 \mathbf{I}$, where $\mathbf{X}$ is composed of a column vector of ones, associated with the constant term, and three additional columns containing levels of output Q, the square of output, and the cube of output.

    You estimate β by ordinary least squares, obtaining $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$. Using these estimates, you would like to test whether marginal cost equals average cost at an output level of 100. (Recall that marginal cost is the derivative of the total cost function with respect to Q_i, and average cost is C_i/Q_i. In calculating average cost, you should ignore the term ε_i/Q_i.)

    Devise a test statistic for the null hypothesis that at output level $Q = 100$, marginal cost equals average cost. Explain how the test statistic is distributed if the null hypothesis is true, and discuss how to calculate the critical region for the test using a significance level of 0.05.

15. Consider a simple regression model $Y_i = \beta X_i + \varepsilon_i$ in which X_i is a single explanatory variable. Note that the model contains no constant term. We assume $\mathrm{E}(\varepsilon_i|\mathbf{X}) = 0$ as usual, but we allow ε_i to be heteroskedastic. That is, we allow $\mathrm{E}(\varepsilon_i^2|\mathbf{X}) = \sigma_{ii}$, so that the variance of the disturbance term can differ from one observation to the next. However, we do not know anything about the nature of this heteroskedasticity.

    Answer the following: (a) What is the variance of the ordinary least squares estimator $\hat{\beta}$ when ε_i is heteroskedastic? (b) How can you estimate this variance? (c) Explain how (with heteroskedasticity present) you could use the ordinary least squares estimator $\hat{\beta}$ to test the null hypothesis $H_0 : \beta = b$, where b is a constant, against the alternative hypothesis $H_A : \beta \neq b$. Present the test statistic, discuss how it is distributed when the null hypothesis is true, and describe how you would determine the rejection region for the test.

16. Imagine that you are given a set of data about U.S. zip code areas, in which each zip code data point represents an average taken over the adults who are residents of that zip code area. Zip code areas differ in population, and your data set contains information on the number of adults in each zip code. That is, zip code area 1 has n_1 adults over whom the averages have been taken; zip code area 2 has n_2 adults; and, in general, zip code area i has n_i adults. All these n_i values are given in your aggregated zipcode dataset.

    For the i-th zip code, Y_i represents the average income of the adult residents of that zip code, and X_i is the average education level of these adults. To be precise, if we let Y_{ia} denote the income of adult a in zip code i, then $Y_i = (1/n_i)\sum_{a=1}^{n_i} Y_{ia}$, and similarly, $X_i = (1/n_i)\sum_{a=1}^{n_i} X_{ia}$, the average of adult educational levels.

    Now consider the regression model $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$, where, as above, the i subscript represents the i-th zip code area, and Y_i, X_i, and ε_i all represent averages taken over the n_i adult residents of the zip code area. In this set-up, the zip code disturbance term ε_i is (like Y_i and X_i) an average of n_i disturbance terms for the n_i adults who live in that zip code. The disturbance term specific to each such adult has mean zero and variance σ^2.

    Explain why the zip code disturbance term ε_i is heteroskedastic. Then explain how to estimate β_1 and β_2 by the method of feasible generalized least squares (FGLS).

17. You have a time-series of macroeconomic data on investment Y_t and interest rates X_t, and have specified your regression model as $Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$. You suspect, however, that ε_t is autocorrelated, with $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$.

    Assume that $-1 < \rho < 1$, that the $\{\varepsilon_t\}$ series is stationary, and that $\mathrm{E}\, u_t \varepsilon_{t-j} = 0$ for all $j \geq 1$. (By "stationary," we mean that the variance of ε_t is the same in every period t and that $\mathrm{E}\,\varepsilon_t\varepsilon_{t-k}$ may depend on k but does not depend on t.) Also u_t has variance σ_u^2 and u_t is not autocorrelated.

    Using these assumptions, derive the variance of ε_t (it will be a function of σ_u^2 and ρ) and find $\mathrm{E}\,\varepsilon_t\varepsilon_{t-2}$.

18. Consider a simple regression model $Y_i = X_i\beta + \varepsilon_i$, in which X_i is a single covariate and the model contains no constant term. Unfortunately, $\mathrm{E}[\varepsilon_i|X_i] \neq 0$, that is, X_i and the disturbance term ε_i are correlated. Luckily, though, you have an instrumental variable Z_i available in your data, which is correlated with X_i but uncorrelated with ε_i, in that $\mathrm{E}[\varepsilon_i|Z_i] = 0$.

    Multiply the regression equation through by Z_i to obtain

    $$Z_i Y_i = Z_i X_i \beta + Z_i \varepsilon_i,$$

    and then add and take averages across observations,

    $$\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i = \frac{1}{n}\sum_{i=1}^{n} Z_i X_i \beta + \frac{1}{n}\sum_{i=1}^{n} Z_i \varepsilon_i.$$

    From this point, please explain how to obtain the instrumental variables estimator $\hat{\beta}_{IV}$. Be sure to state the assumptions that are needed to obtain this estimator, discuss where in your argument each such assumption is needed, and say why it is needed.

19. You have a time-series of macroeconomic data on investment Y_t and interest rates X_t, and have specified your regression model as

    $$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + \varepsilon_t.$$

    Note that Y_{t-1}, the lagged value of investment, appears on the right-hand side as an explanatory variable.

    You strongly suspect that ε_t is autocorrelated, with $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$. You feel safe in assuming, however, that interest rates X_t and the disturbance ε_s are uncorrelated for all time periods t and s that are covered in your data.

Answer the following: (a) Explain why this model suffers from "statistical endogeneity," that is, why a right-hand side explanatory variable is correlated with the disturbance term. (b) What instrumental variables are available to estimate the model? Why are they valid instrumental variable? (c) Suppose that $\rho = 1$. How could you estimate β_2 and β_3 without recourse to instrumental variables?