Comprehensive Exam in Econometrics
Suffolk University
September 2006

Answer all questions. You have four hours to take this test. Time yourself wisely.

A1. Consider $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{X}$ represents a non-stochastic T x k data matrix, including a constant.

a) Let $\mathbf{X}\beta = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$ where $\mathbf{X}_1$ and $\mathbf{X}_2$ are non-stochastic T x k_1 and T x k_2 matrices respectively. Discuss the consequences of erroneously (i) excluding X_2 when it should have been included, (ii) including X_2 when it should have been left out.

b) In a) above, how do you test if β_2 is significantly different from zero. If it is found insignificant, should you drop X_2 and re-estimate the model. In general, explain the selection criteria that are used to identify the most appropriate set of explanatory variables.

c) Although economic theory provides some information on the choice of **y** and **X**, it is generally silent on its functional form. Briefly compare a linear model with an analogous log-linear model. How do you choose between the two?

A2. Consider the classical linear regression model, $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{X}$ is non-stochastic. Let $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon')=\sigma^2\Omega$ where Ω is symmetric positive definite matrix.

a) Consider the following model for real estate values:

$$\text{Price}_i = \beta_0 + \beta_1\text{SQFT}_i + \beta_2\text{Yard}_i + \beta_3\text{Pool}_i + \varepsilon_i \quad \text{where } i = 1,2,\ldots,N \text{ and:}$$

Price = Price in thousands of dollars
SQFT = Living area in square feet
Yard = Size of yard in square feet
Pool = Dummy variable taking value 1 if the house has a pool.

You suspect heteroscedasticity in this cross sectional study. Discuss the structure of the Ω matrix under heteroscedasticity and the implications of ignoring it. Describe the test for heteroscedasticity if you feel that it may be caused by the variations in SQFT. If detected, describe the GLS (WLS) procedure to take care of the problem.

b) Consider $y_t = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \gamma y_{t-1} + \varepsilon_t$ where $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$. Why is the Durbin-Watson test not valid anymore? Describe the Durbin's h and the LM tests for testing for serial correlation in this dynamic model. Describe the method for estimating the parameters efficiently if the 1st order serial correlation is detected?

(B1) Consider the following model for a simple economy:

$$C_t = \alpha_0 + \alpha_1 Y_t + \alpha_2 D_t + u_{t1}$$
$$I_t = \beta_0 + \beta_1 Y_t + \beta_2 G_t + \beta_3 X_t + u_{t2}$$
$$Y_t = C_t + I_t + G_t$$
$$X_t = (D_t)^2$$

where C is consumption spending, Y is total income, D is a dummy variable that takes the value 1 during years of drought and 0 in other years, I is investment spending, G is government spending, and u_1 and u_2 are serially uncorrelated error terms with zero means and constant variances.

(a) Which variables in the model are endogenous? predetermined? Also determine the identification status of the consumption and investment equations according to rank and order conditions.

(b) Specify the *unrestricted* reduced form equation for I_t in general (you do not need to solve for the parameters explicitly). Describe how this equation would be estimated.

(c) Specify which parameters in the structural model above can be consistently estimated and how this would be done.

(B2) Consider the following linear regression model

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \quad i = 1, \ldots, 100 \tag{1}$$

where Y is the hourly wage of individual i, D is a dummy variable that takes the value 1 if the individual is female and 0 if male, and $u \sim \mathrm{N}(0, 1)$. The data is so that the first 25 observations in the sample are all females and the remaining observations are all males. You are also given the following information: $\sum_{i=1}^{100} Y_i = 1010.4134$, $\sum_{i=1}^{25} Y_i = 267.6273$, $Var(Y) = (100-1)^{-1} \sum_{i=1}^{100} (Y_i - \bar{Y})^2 = 1.1623$, and based on OLS estimation of (1) $\hat{u}^\top \hat{u} = 103.0316$, $\hat{\beta}_0 + \hat{\beta}_1 = 10.7051$.

(a) What is the value of $\hat{\beta}_1$?

(b) Test the null hypothesis that there is no systematic wage difference between males and females at 5% level of significance using the likelihood ratio test.

(B3) Consider the following wage equation for working women

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + u$$

where *wage* is the estimated wage from earnings per hour, *exper* is actual labor market experience, and *educ* is years of schooling. The error term u is thought to be correlated with *educ* because of omitted ability, as well as other factors, such as quality of education and family background. We have data on mother's years of schooling (*motheduc*), father's years of schooling (*fatheduc*), and husband's years of schooling (*huseduc*).

(a) Under what circumstances do *motheduc*, *fatheduc*, and *huseduc* make good instrumental variables?

(b) Explain in detail how you would estimate the wage equation by 2SLS and how you would obtain the correct standard errors?

(c) Briefly explain in words the idea behind the Durbin-Wu-Hausman test. Also, explain in detail how you would test for endogeneity of education in the wage equation by the Durbin-Wu-Hausman test.