PhD/MA Econometrics Examination

January 2013

MA: Complete Parts A and B Only

PhD: Complete all 3 Parts – A, B, and C

PART A

ANSWER ANY <u>TWO</u> QUESTIONS IN THIS SECTION

---

1. Suppose x and y follow the joint probability density function:

   $f(x,y) = 7.5\, x\,(2 - x - y) \qquad 0 \le x \le 1,\ 0 \le y \le 1,$ and 0 *Otherwise*

   a. Find the marginal pdf $f(x)$.

   b. Find the conditional pdf $f(x|y)$.

   c. Are x and y independent? Show proof.

   d. Calculate the conditional expectation $E(x|y)$.

2. Consider the following linear model in two forms:

   Matrix: $Y = X\beta + U \quad (1)$

   A scalar slope only model: $y_t = \alpha_1 x_t + u_t \quad (2)$

   a. State the OLS classical linear model (CLM) assumptions.

   b. Derive the OLS estimator for model 1 or 2.

   c. Derive the MLE estimator for either 1 or 2.

   d. Derive the variance of the slope coefficient either for 1 or 2.

   e. Prove the Gauss Markov Theorem for either of the two models.

   f. While applying OLS, we assume $E(U) = 0$. What could be the economics meaning of such an assumption? Can you give one practical example under which such an assumption could be violated?

PART A, CONTINUED

3. **For this question,** use the table below, and note the following:
"HWSEI is a constructed variable that assigns a Hauser and Warren Socioeconomic Index (SEI) score to each occupation using the modified version of the 1990 occupational classification scheme available in the OCC1990 variable. The HWSEI variable is a measure of occupational status based upon the earnings and educational attainment associated with each category in the 1990 occupational scheme."

Source	SS	Df	MS			
				Number of obs	=	12104
				F(13, 12090)	=	498.53
Model	1662649.59	13	127896.1	Prob > F	=	0
Residual	3101628.43	12090	256.5449	R-squared	=	???
				Adj R-squared	=	0.3483
Total	4764278.02	12103	393.6444	Root MSE	=	16.017

HWSEI	Coef.	Std.	t	P>t	[95% Conf.	Interval]
Age	1.4378	???	20.55	???	???	???
Age^2	-0.0174	0.0009	-20.44	0.0000	-0.0190	-0.0157
Female	-2.2691	---	---	---	-2.8422	-1.6959
Nursery school to grade 4	-1.4976	2.4374	-0.61	0.5390	-6.2754	3.2801
Grade 5, 6, 7, or 8	-2.3946	1.7404	-1.38	0.1690	-5.8061	1.0169
Grade 9	-1.9035	1.7752	-1.07	0.2840	-5.3831	1.5761
Grade 10	-1.0185	1.7148	-0.59	0.5530	-4.3798	2.3429
Grade 11	0.8919	1.6985	0.53	0.6000	-2.4373	4.2211
Grade 12	8.0669	1.5618	5.17	0.0000	5.0056	11.1283
1 year of college	12.1170	1.5784	7.68	0.0000	9.0230	15.2109
2 years of college	17.0264	1.6285	10.46	0.0000	13.8342	20.2186
4 years of college	26.0401	1.5930	16.35	0.0000	22.9176	29.1626
5+ years of college	35.9438	1.6119	22.3	0.0000	32.7843	39.1033
Constant	-8.8311	2.0393	-4.33	0.0000	-12.8284	-4.8337

Omitted groups: "male" and "no school or N/A"

a. Interpret the coefficient (and other relevant information) on "Grade 11." Grade 11 is a dummy variable for people who have only completed through 11th grade, i.e., one year short of finishing high school. Now, looking at the coefficient on the variable Grade 12, what can we learn?

b. Calculate R^2. Explain your answer. What does the number mean?

c. What is the F value that is reported above? What does it tell you? Can you calculate it from the information given above? Use it to test the joint significance of the slope coefficients.

d. Now, suppose you want to test the joint significance of just the schooling variables. How would you do it? Show all the steps.

e. Calculate the missing values in the row labeled "Age."

f. Looking at the row labeled "Female," what can you say about the significance of the coefficient?

g. What is the "Dummy Variable Trap?"

PART B

ANSWER ANY TWO QUESTIONS IN THIS SECTION

4. Suppose you have a linear model, given in matrix notation as:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{u}$$

Assume the first column of the data matrix **X** is entirely 1's. For parts (a)-(c) only, further assume that the data are from a representative, simple random cross-section sample.

a. What is heteroskedasticity? Define the term, and explain what problem it causes for OLS estimation. Then, provide at least two different practical, real-world examples where heteroskedasticity is likely to be a concern.

b. Explain in words and show mathematically how the GLS estimator for **b** differs from the OLS estimator. Present the variance-covariance matrix for **u** (the GLS $\boldsymbol{\Omega}$ matrix), assuming only simple heteroskedasticity (no autocorrelation, clustering, etc.).

c. White (1980) introduced a testing procedure for heteroskedasticity. Wooldridge provided an alternative but arguably equivalent testing procedure, that addresses a practical problem with implementation of White's test. Choose **one** of these two tests (**OR,** if you prefer, BP or GQ), identify it by name, then describe the test algorithm step-by-step. Specifically: What are the steps? What information is required, and how can you obtain it in practice? What is the test statistic used, and to what distribution should you compare it? **Bonus:** explain the motivation for Wooldridge's alternative to White's test.

For parts (a)-(c) above, you assumed cross-sectional data. Now, for parts (d) and (e), you should instead assume a panel dataset. Suppose N=1000 and T=5. Let the subscript *i* index individuals, and let *t* index time. For illustration, suppose the outcome variable, y_{it}, represents individual *i*'s income in year *t*. Finally, assume that, after controlling for the explanatory variables **X**, income is still correlated within individuals over time; however, individuals are still drawn randomly.

d. Three common modeling strategies one might consider using for panel data include:

(i) random effects (ii) fixed effects (iii) first-differencing

For each of these three modeling strategies, write the corresponding linear regression equation using the subscript notation given above. For example, using subscripts, the original model shown above would be: $y_i = \mathbf{x}_i\mathbf{b} + u_i$

Your new models should account both for time-invariant individual-specific heterogeneity, and for time shocks that affect all individuals. Please be sure to define any new variables or notation you introduce. For each modeling strategy, note at least one condition, situation, and/or assumption that would make that strategy preferable to the other two.

e. **True or False, and Explain**: If we include individual fixed effects in our model, then the usual OLS standard errors for **b** will be correct, and we don't need to use GLS.

PART B, CONTINUED

5. Background: Gruber (*JPE*, 2000) investigates the effects of changes in Canadian disability insurance benefits on the labor supply. Gruber writes: "The difficulty of appropriately identifying disability and the generous levels of benefits available have led many observers to claim that disability insurance is largely distorting work decisions and in essence subsidizing the early retirement of the older workers for whom appropriately defining a career-ending disability is most difficult." (p. 1163)

He estimates:

$$NE_i = \alpha + \beta_1 CPP + \beta_2 AFTER + \beta_3 (CPP \times AFTER) + \gamma \mathbf{x}_i + \varepsilon_i$$

where NE_i is a dummy variable equal 1 if individual i is unemployed, *CPP* is a dummy variable equal 1 for individuals living in Canadian provinces where there was a policy change, and 0 for Quebecois (for whom there was no policy change), *AFTER* is a dummy variable equal 1 for years after the policy change, and 0 before, and $\mathbf{x}_i$ is a vector of individual demographics such as marital status, number of dependents, and educational attainment. Note, the data are repeated cross-sections, not panel data. In the paper, Gruber uses logit estimation. However, for simplicity, for this question assume a linear probability model (LPM).

a. Explain how each of the β coefficients above – that is: β_1, β_2, and β_3 – should be interpreted. What variation does this identification strategy exploit (be specific)? More generally, what is the fundamental identifying assumption for difference-in-difference models?

b. A reviewer argued the policy change is endogenous, i.e., the provinces that decided to change their disability policies are systematically different from Quebec (the control). Gruber responds by constructing a synthetic instrumental variable (IV).

   i. In general, what are the two requirements (conditions) for a valid IV? Can these conditions be evaluated using statistical tests? If so, how?

   ii. In this specific context, what would Gruber need to demonstrate to persuade a reviewer that this IV is valid?

   iii. Provide the algorithm (step-by-step procedure) for the Hausman test, a statistical test to show that IV estimation is necessary. What key assumption does the Hausman test make?

c. In Gruber's model, the level of observation is the individual, but the level of variation—Canadian province—is much more aggregated. As a result, as Bertrand et al. (2004) discuss, repeated cross-sections are likely to have "grouped" errors, resulting in over-rejection of the null hypothesis as T grows large. Three possible solutions are suggested: parametric correction, for example presuming AR(1) errors; empirical variance-covariance correction robust to arbitrary correlation within group (commonly called clustered standard errors); and block bootstrap.

   i. Identify one advantage and one disadvantage associated with each of these suggested solutions, as compared with the others.

   ii. Choose one of the corrections listed above – AR(1) parametric, **OR** cluster-robust, **OR** block bootstrap – and explain how it is implemented in practice.

PART B, CONTINUED

6. **LPMs and Asymptotic Normality**

   a. In general, what concerns do you have about using a LPM? For example, one reason for concern might be violating the CLM assumption of normally-distributed errors. Provide two additional (different) reasons why an LPM estimated via OLS is generally considered problematic.

   b. Discuss the weighted least squares (WLS) estimation technique in this context. In particular:

      i. How does WLS differ from OLS and from GLS?

      ii. From a practical perspective, how would you implement WLS for a LPM? Provide the procedure step-by-step.

   c. LPMs' violation of the normality assumption above can be ignored, provided some other assumptions hold.

      i. Define **consistency** (as it applies to estimators). How does the notion of consistency differ from **unbiasedness**?

      ii. If $\hat{\beta}_{OLS}$ is consistent, what will happen to its distribution as N grows large?

      iii. How does the asymptotic normality assumption differ from the standard CLM normality assumption? In particular, <u>what</u> is assumed to be normally distributed in each case?

      iv. Finally, what does the normality assumption, whether in its CLM finite-sample or asymptotic form, contribute? For example, do we need it to prove unbiasedness or consistency of the estimator $\hat{\beta}_{OLS}$?

PART C

ANSWER ANY TWO QUESTIONS IN THIS SECTION

7. Consider the following structural model:

$$BabyBMI_t = \alpha_0 + \alpha_1 MotherAntenatalVisit_t + \alpha_2 WeeklyCalorie_t + \alpha_3 Female_t + u_t$$
$$MotherAntenatalVisit_t = \beta_0 + \beta_1 TravelTimetoClinic_t + \beta_2 Education_t + \beta_3 Work_t + v_t \quad (1)$$

a. Identify the endogenous and predetermined variables.

b. Are these equations identified?

c. What sign do you expect for these slope coefficients?

d. Write the model in a grand matrix notation:

$$Y = ZB + U \quad (2)$$

Present the elements/variables in Y, Z, B, and U

e. Derive the variance covariance matrix, using notation: $E[\mathbf{uu'}] = \Omega \otimes \mathbf{I}_T$.

f. Pulling the first equation from the grand set-up, let us write it as follows,

$$y_1 = Y_1\gamma + X_1\Delta + u_1 \quad (3)$$

Where Y_1 and X_1 are endogenous and predetermined variables, respectively.

g. Now, taking the equation in 1.f in its generic form, use one of the four options presented below to estimate this equation. Be sure to show all the steps in detail.

i) 2SLS estimation of equation 1 as presented in equation (3) **OR**

ii) Instrumental Variable estimation of equation 1 as presented in equation (3) **OR**

iii) Full Information Maximum Likelihood (FIML) estimation of the model as presented in (1) or in grand notation as in equation (2). Use the joint normality and the Jacobian term. **OR**

iv) 3SLS –GLS—estimates of the 2-equation system as presented in model (1)
(You can rewrite and/or reformulate the system of equations to showcase your understanding the estimation method.)

h. Present the variance-covariance matrix for the estimator of your choice in 7.g.

i. Compare and contrast the Limited Information Method versus the Full Information Method.

PART C, CONTINUED

8. Background: In Kathmandu Valley, one of my students conducted a stated preference survey, and asked the 1200 residents of the valley about their willingness to pay (one-time payment into a trust fund) to clean up the Bagmati River. It was an open ended question. The response ranged from 0 to Rs. 100,000. 25% of the resident gave a value of 0 perhaps reflecting their unwillingness to support such a contribution believing that the money will be misused by the government. So, there may have been al lot of "negative preference." As such, the student decided to estimate this model as a Tobit model with the assumption of a censoring from below 0. We also noticed that there were many people (11%) who gave Rs. 100,000 (the maximum contribution), demonstrating perhaps there is a right-hand censoring too. Also, make note of the fact that the River has a very strong religious significance for the Hindus.

The Model:

$$WTP_t^* = \alpha_0 + \alpha_1 Income_t + \alpha_2 HouseDistance\_KM\_FromRiver_t + \alpha_3 Female_t + \alpha_3 Hindu_t + u_t$$

Where $WTP_t^* = WTP_t$ if $WTP_t^* > 0$, else $WTP_t = 0$.

a. What sign would you expect on these coefficients?

b. Set up a left censored maximum likelihood model likelihood function assuming the normal distribution function.

c. Discuss ways to estimate the mean willingness to pay.

d. Now, suppose we create a three-category ranking of their preferences: WTP values of those who said they would pay 0. 0, 1-100,000 and greater than 100,000. Now, you have two methods to estimate this model – i) Multinomial Logit or ii) Ordered Logit Model. Pick one of the two and derive the likelihood function.

e. Discuss the estimation algorithm to optimize a typical likelihood function of your choice, b or d.

f. How would you obtain the standard errors of these coefficients?

g. **Bonus:** Assuming the following model:
$WTP_t^* = \alpha_2 HouseDistance\_KM\_FromRiver_t + u_t$

Demonstrate that an OLS estimator of the slope coefficient would be biased.

PART C, CONTINUED

9. Assume the amount of waste produced (in tonnage per year) and city size have a nonlinear relationship:

$$Waste_t = \emptyset_0 CitySize_t^{\emptyset_1}$$

You have a choice to introduce the error term as per your estimation choice below, either additively

$$Waste_t = \emptyset_0 CitySize_t^{\emptyset_1} + u_t$$

or multiplicatively

$$Waste_t = \emptyset_0 CitySize_t^{\emptyset_1} e^{u_t}$$

a. How would you estimate such a model? Present the method – OLS, **OR** MLE, **OR** non-linear least squares.

b. Now, assume that the error term is heteroskedastic and is driven by the number of business establishments. Assuming a heteroskedastic form of your choice, set up the model and present the estimation method: iterative GLS **OR** multiplicative heteroskedastic maximum likelihood.

c. How would you test for the heteroskedasticity problem in this model?

d. What kind of magnitude and the sign would you expect for this particular slope? Explain. Does it have to do anything with the notion of the power law?

e. If the heteroskedasticity problem does not go away, it is recommended that we use White's (aka Huber-Eicker-White, HEW, "sandwich") error correction, often known as the robust standard error. Explain the theory behind this correction.