ECN 726: Econometrics II

Nick Kuminoff Arizona State University: Fall 2013

EXAM

Due by 8am on Sunday, November 10th

All answers must be prepared on your own. You may consult books, notes, journal articles, and online material in developing your answers. You may not discuss any part of the exam with anyone. Any evidence of cooperation will be treated as a violation of ASU's *Student Code of Conduct* and handled according to university policy (https://students.asu.edu/srr/code).

Please submit your work in an email to Nastya containing: (i) a pdf file with your solutions to all problems, and (ii) Matlab m-files containing code that will allow her to reproduce your work. To obtain full credit, all derivations must be shown and all work must be legible. Exams submitted after the 8am deadline will be charged a tax of 1 point per minute late.

Good luck!

PART I: PROPERTIES OF ESTIMATORS

A) Suppose you have n observations on the following model: $y = \beta_1 x_1 + \beta_2 x_2 + \mu$, where β_1 and β_2 are scalars, and x_1 and x_2 are $nx1$ vectors. Both explanatory variables are correlated with μ. Your goal is to recover β_1. You have a valid instrument, z, for x_1. Unfortunately, you do not have an instrument for x_2. Suppose you run a two-stage least squares model that uses z as an instrument for x_1 but ignores the endogeneity of x_2.

Derive an equation expressing your second-stage estimator, $\hat{\beta}_1^{2SLS}$, as a function of the data on y, x_1, x_2, and z.[1] Use your expression to define additional restrictions on the data under which $\hat{\beta}_1^{2SLS}$ is a consistent estimator for β_1.

(*15 points*)

B) Consider a latent variable model where the observed outcome, y_i, depends on the latent variable c_i such that:

$$y_i = 1 \text{ if } c_i > 0, \qquad \text{and} \qquad y_i = 0 \text{ if } c_i \le 0,$$

$$\text{where} \quad c_i = X_i\beta + u_i, \quad \text{with} \quad (u_i \mid X_i) \sim N(0, \kappa^2) \quad \text{and} \quad \kappa^2 \neq 1.$$

The vector X_i contains variables that are observed by the econometrician. The scalar u_i represents the composite effect of variables that affect behavior but are not observed by the econometrician.

(*15 points*: *10/5*)

i) Demonstrate that probit estimation of this model identifies β/κ rather than β.

ii) "Unobserved variables are a serious problem for the probit model because even if u_i is uncorrelated with *X*, the probit model does not identify β." Do you agree or disagree with this statement? Please explain your reasoning.

[1] You may find it helpful to invoke results from the Frisch-Waugh-Lovell Theorem.

C) Consider the following wage equation: $w_{it} = \beta w_{it-1} + \gamma d_{it} + \delta x_{it} + \alpha_i + \varepsilon_{it}$. In words, worker *i*'s wage in year *t* depends on her wage last year (w_{it-1}), an indicator for whether she participates in a job training program (d_{it}), and a vector of demographic attributes such as her education and experience (x_{it}). Wages also depend on latent human capital, which can be divided into a time-constant component (α_i) and a serially uncorrelated time-varying component (ε_{it}). Furthermore, assume that $E[\alpha_i | d_{it}] \neq 0$ and $E[\varepsilon_{it} | w_{it-1}, d_{it}, x_{it}, \alpha_i] = 0$.

Suppose you have the ability to collect data on w_{it}, d_{it} and x_{it} for several years. How many years of data would you require, at a minimum, to develop a consistent estimator for the model parameters (β, γ, δ)? Write down the model you would estimate and explain why it is consistent.

(*10 points*)

PART II: ECONOMETRIC RESEARCH DESIGN

Some jobs are more dangerous than others. Table 1 summarizes the 5-year average fatal risk rates for 19 industries in the United States in 2010. In the mining industry, for example, an average of 3.5 workers per year (out of 10,000) died due to work-related accidents between 2006 and 2010. In comparison, there were no deaths in the management industry. Workers must be paid more to accept dangerous jobs. This is well known.[2]

Table 1: 5-Year Average Fatal Risk Rates by Industry, 2010

Industry Code	Industry Group	% of workers	Fatal risk rate (out of10,000)
2100	mining	0.4	3.49
2200	utility	1.2	0.49
2300	construction	7.1	1.09
3123	manufacturing	21.0	0.26
4200	wholesale trade	6.3	0.45
4445	retail trade	9.3	0.23
4800	transportation and warehousing	3.0	1.73
4923	currier and messengers, warehouse and storage	0.9	0.36
5100	informaiton	2.4	0.24
5200	finance and insurance	6.4	0.05
5300	real estate and rental and leasing	1.6	0.27
5400	professional, scientific, and technical services	5.2	0.15
5500	management of companies and enterprises	0.1	0
5600	administrative and support and waste management remediation services	2.3	0.56
6100	educational services	11.8	0.06
6200	health care and social assistance	13.7	0.06
7100	arts, entertainment, and recreation	1.3	0.32
7200	accomodation and food services	3.1	0.17
8100	other services	3.2	0.35

Imagine you are hired by the Department of Labor (DOL) to predict how a marginal decrease in the fatal risk rate would affect wages. DOL gives you annual data on a panel of 10,000 workers. For each worker, you have the following information for each year from 2000 through 2010:

- Wage, 5-year average fatal risk rate, education, years of experience, industry, occupation.

In general, of these variables may change over time. Workers gain experience and may also increase their education by attending school part time. However, virtually all of the variation in risk is cross-sectional. That is, virtually all of the variation occurs *between* industry-occupation pairs. This can be seen from tables 1 and 2. Table 2 provides an example of the cross-section variation across occupations within a single industry—mining. Not surprisingly, the fatal risk rate is high for people who go down into mines (code 45: construction and extraction) and low for managers and secretaries of mining companies (codes 11 and 41). In contrast, the time-series variation in fatal risk within each industry-occupation pair is only slightly above zero. For example, the fatal risk rate for construction and extraction occupations in the mining industry ranges from 12.69 in 2000 to 12.84 in 2010. This lack of time-series variation is due, in part, to the fact that the Census Bureau reports fatal risk rates as 5-year averages. Finally, note that most workers remain in the same industry

[2] The literature begins with T.C. Shelling's 1968 paper "The Life You Save May be Your Own" in Problems in Public Expenditure Analysis by Samuel B. Chase, Jr. and a 1976 book chapter by Richard Thaler and Sherwin Rosen entitled "The Value of Life Saving" in Household Production and Consumption by Nestor E. Terleckyj. For reviews of the literature see Kip Viscusi's 1993 JEL article "The Value of Risks to Life and Health" and the 2002 JPAM article by Janusz Mrozek and Laura Taylor titled "What Determines the Value of Life? A Meta Analysis".

and occupation throughout the study period. Only 1% of workers change industries and/or occupations each year.

Table 2: 5-year Average Fatal Risk Rates by Occupation within the Mining Industry, 2010

Occupation Code	Occupation within the Mining Industry	Fatal risk rate (out of10,000)
11	management	0.00
13	business and financial operations	0.00
15	computer and mathematical	0.00
17	architecture and engineering	0.05
23	legal	0.00
25	arts, design, entertainment, sports, and media	0.00
27	healthcare practitioners and technical	0.79
29	healthcare support	0.51
31	protective service	0.80
33	food preparation and service	0.06
35	building and grounds cleaning and maintenance	0.04
39	sales and related	0.08
41	office and adminstrative support	0.01
45	construction and extraction	12.84
47	installation, maintenance, and repair	1.51
49	production	0.96
51	transportation and material moving	3.71

A standard concern with estimating wage compensation for job risk is that unobserved job skill will bias the estimator. Specifically, the concern is that workers with higher unobserved job skill will choose to work in jobs that have a lower fatality rate.[3]

Propose a research design to estimate the wage effect of a marginal change in fatal risk that mitigates the concern about omitted variable bias as much as possible, using only the data described above. Write a brief memo explaining your proposal. For full credit, your memo must include: (i) the equation you plan to estimate, with a clear description of all notation; (ii) an explanation of the sources of variation in the data that allow you to identify the effect of fatality risk on wages; and (iii) a discussion of any important caveats to using your results to predict the wage effect of a marginal reduction in risk.

(20 points)

[3] This was formalized in a 1992 JPE paper by Hae-shin Hwang, W. Robert Reed, and Carlton Hubbard titled "Compensating Wage Differentials and Unobserved Productivity."

PART III: MULTI-SAMPLE GMM[4]

It is not uncommon to find there is no single data set that will allow us to identify the parameters we seek to estimate. Yet, it may be possible to identify subsets of the parameters from different data sets describing the same underlying population. This idea motivates the multi-sample approach to GMM estimation. For example, Arellano and Meghir (1992) combine two different micro data sets drawn from the same population to identify $\theta = [\theta_1, \theta_2]$. Moment conditions based on the first sample identify θ_1, moment conditions based on the second sample identify θ_2, and combining both sets of moment conditions allows them to identify θ.

Even if θ is identified by a single data set, adding more moment conditions based on complementary data sets may improve the efficiency of the estimation. Imbens and Lancaster (1994) demonstrate this for a situation where θ is identified by micro data. They show that adding additional moment conditions based on macro data for the same population can improve the efficiency of their estimator.

This problem asks you to derive a two-sample GMM estimator (2SGMM) and implement it. Consider a situation where it is possible to derive non-redundant moment conditions from two different samples drawn from the same underlying population. Arellano & Meghir (1992) derive consistency and asymptotic normality results for the estimator in (1)-(2):

$$(1) \qquad \hat{\theta}_{2SGMM} = \arg\min_{\theta \in \Theta} Q(\theta) = \left(\frac{N_1}{N_1 + N_2}\right) Q_1(\theta) + \left(\frac{N_2}{N_1 + N_2}\right) Q_2(\theta)$$

$$(2) \qquad \hat{V}\left(\hat{\theta}_{2SGMM}\right) = \left[N_1 A_1\left(\hat{\theta}\right) + N_2 A_2\left(\hat{\theta}\right)\right]^{-1}\left[N_1 B_1\left(\hat{\theta}\right) + N_2 B_2\left(\hat{\theta}\right)\right]\left[N_1 A_1\left(\hat{\theta}\right) + N_2 A_2\left(\hat{\theta}\right)\right]^{-1},$$

$$\text{where } A_j(\theta) = \frac{\partial^2 Q_j(\theta)}{\partial\theta\partial\theta'}, \text{ and } B_j = N_j\left(\frac{\partial Q_j(\theta)}{\partial\theta}\frac{\partial Q_j(\theta)}{\partial\theta'}\right) \text{ for } j = 1,2.$$

Q_1 and Q_2 denote objective functions based on the moment conditions associated with each data set, and N_1 and N_2 denote the sample sizes associated with Q_1 and Q_2.

[4] Arellano, Manuel and Costas Meghir. 1992. "Female Labor Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets." *Review of Economic Studies.* Vol. 59. 537-559.

Imbens, Guido W. and Tony Lancaster. 1994. "Combining Micro and Macro Data in Microeconometric Models." *Review of Economic Studies.* Vol. 61. 655-680.

(A) Arellano & Meghir describe Q as a “criterion function” that could be the objective function for an m-estimator or for GMM. Demonstrate that when Q_1 and Q_2 are GMM objective functions associated with two sets of independent moments, $h_1(v_i,\theta)$ and $h_2(v_i,\theta)$, and the optimal weights matrix is used, the Arellano-Meghir estimator can be expressed as (3)-(4):

$$(3)\qquad \hat{\theta}_{2SGMM} = \arg\min_{\theta\in\Theta} Q(\theta) = \left(\frac{N_1}{N_1+N_2}\right) g_1' W_1 g_1 + \left(\frac{N_2}{N_1+N_2}\right) g_2' W_2 g_2$$

$$(4)\qquad \hat{V}\left(\hat{\theta}_{2SGMM}\right) = \left[N_1 G_1' W_1 G_1 + N_2 G_2' W_2 G_2\right]^{-1},$$

where $g_j = N_j^{-1}\sum_i h_j(v_i,\theta)$ and $G_j = N_j^{-1}\sum_i \frac{\partial h_j(v_i,\theta)}{\partial \theta'}$, for $j = 1,2$.

(*10 points*)

(B) Consider the model in equations (5)-(6). Equation 5 expresses the demand for a good as a function of its price, quality, and consumer income. Equation 6 gives a consumer’s Marshallian surplus for a quality change in the good. Both equations are consistent with the same underlying structure for consumer preferences and, therefore, share some of the same parameters, $\theta = [\alpha,\beta,\delta,\gamma]$.

$$(5)\qquad Q = \alpha + \beta \text{ price} + \gamma \text{ quality} + \delta \text{ income} + u\,.$$

$$(6)\qquad MCS = -\frac{1}{2\beta}\left[a_1^2 - a_0^2\right] - \gamma \text{ price}\cdot\left(\text{quality}_1 - \text{quality}_0\right) + \varepsilon\,,$$

where $a_i = \alpha + \delta \text{ income} + \gamma \text{ quality}_i$ for $i = 0,1$.

The Matlab data file *data* contains two independent data sets on the same population:

$$\underset{200x4}{v1} = [Q, quality, price, income] \text{ and } \underset{50x5}{v2} = [MCS, price, income, quality_0, quality_1].$$

Use the model in (5)-(6) to develop a 2SGMM estimator for $\theta = [\alpha,\beta,\delta,\gamma]$ based on the objective function in (3). Use your results to fill in the following table:

Parameters	(i) 1st Step GMM Estimates	(ii) 2nd Step GMM Estimates	(iii) Iterated GMM Estimates
α	____	____	____
β	____	____	____
δ	____	____	____
γ	____	____	____

(i) Calculate 1st step GMM estimates from setting the weights matrices equal to the identity matrix. Report your results in column (i).

(*7 points*)

(ii) Calculate 2nd step GMM estimates based on the optimal weights matrix. Report your results in column (ii).

(*7 points*)

(iii) Calculate iterated GMM estimates and report them in column (iii).

(*7 points*)

(vi) Discussion question. What can explain the differences between your results in columns (i) through (iii)?

(*9 points*)