Exam Econometrics II

June 1, 2016

Instructions: Write your identification number on each paper and on the unnumbered cover. Answer each question on separate sheets of paper. If you think that a question is vaguely formulated, specify the conditions used for answering it.

Good luck! Peter Fredriksson

Question 1 (10p)

Consider the logit model

$$\Lambda(X_i'\beta) = \frac{\exp(X_i'\beta)}{1+\exp(X_i'\beta)}$$

where β is estimated using Maximum Likelihood.

a) What is the log-likelihood function corresponding to this problem?

b) Derive the first-order condition for maximization of the likelihood.

c) Derive the asymptotic variance of $\hat{\beta}$. (Hint. Probably easiest to use $\hat{\Omega} = \left[-\sum_i E\left[H_i(\hat{\beta})\middle| X_i\right]\right]^{-1}$)

Question 2 (10p)

You are interested in estimating the effect of a job-training program on employment and hourly wages. To estimate these effects you have access to experimental variation where treatment and control status have been randomized.

a) Suppose there is perfect compliance with randomization. Under what condition(s) can you interpret the wage effect as a causal effect?

b) Suppose now that there is imperfect compliance, and that the compliance problem only applies to the treatment group. How would you estimate the causal effect of the job-training program on employment? What “treatment effect” is estimated?

Question 3 (10p)

Consider the following model:

$$y_i = \beta x_i + u_i \tag{1}$$

$$x_i = Z_i'\gamma + v_i \tag{2}$$

All variables are deviated from their means. The (population) regression errors are potentially correlated: $COV(u_i, v_i) \neq 0$. Equation (1) corresponds to the structural equation and equation (2) to the first-stage regression. In this setting, the small sample bias of 2SLS approximately equals

$$E(\hat{\beta}_{2SLS} - \beta) \simeq \frac{COV(u_i, v_i)}{VAR(v_i)} \left[\frac{1}{F+1}\right]$$

where F is the population F-statistic in the first stage.

a) Derive an expression for the bias of OLS.

b) Discuss the bias of 2SLS when the instruments are weak: When are instruments weak? What happens in the extreme case when the instruments have no predictive value? What are possible solutions to a weak-instruments problem?

Question 4 (15p)

Consider the following setting. You are interested in the effect of a binary "treatment" (D_i) on an outcome (Y_i). The treatment is potentially endogenous, however. To estimate the effect of the treatment on the outcome, you have access to a binary instrument (Z_i). All causal responses are allowed to be heterogeneous in the population.

a) What assumptions do you have to impose in order to estimate a meaningful treatment effect in this setting?

b) Use these assuptions to derive an expression that illustrates what instrumental variables estimate in this setting.

Question 5 (15p)

The Regression Discontinuity (RD) and the Regression Kink (RK) designs are two approaches to estimating causal effects. Describe and compare these two approaches. When are these two designs applicable? What are the fundamental identifying assumptions? How do you validate these identifying assumption? How would you specify the regression equations in each of the two approaches? Which of these two approaches do you think are more credible?

Focus on the sharp versions of the two designs and parametric approaches to estimating the causal effects.

Question 6: Evaluation of empirical stragies I (20p)

A recent paper examines whether the incidence of debt is affected by the income of your neighbors. To answer this question, the authors run regressions of the following kind

$$Debt_{iz} = \alpha + \gamma y_i + \beta y_z + X_i'\phi + \epsilon_{iz}$$

Debt is a binary indicator for having a loan, y_i denotes individual monthly income, y_z the average monthly income in the zip-code (postal code) where the individual resides, and X_i a vector of control variables. The coefficient of interest is β.

Table 1 below reports a sub-set of the results. It shows marginal effects along with standard errors that are clustered on zip-code (the underlying specification is a Logit). For both income variables, the marginal effect is evaluated at an increase Euro 500 (average income equals Euro 1383).

Table 1: The relationship between debt and income of neighbors

	Dependent variable:	
	Has a colleteralized house loan	Use of overdraft facility
	(1)	(2)
individual income	0.001	0.006
	(0.0001)	(0.0002)
zip-code income	0.011	0.020
	(0.0021)	(0.0014)
Control variables		
Individual wealth	yes	yes
Use of internet banking	yes	yes
Gender	yes	yes
Age fixed effects (FE)	yes	yes
Marital status FE	yes	yes
Occupation FE	yes	yes
Nationality FE	yes	yes
pseudo-R^2	0.134	0.182
# observations	446,765	446,765

The dependent variable in column (1) equals 1 if the individual has a colleteralized house loan (mean of dependent variable is 0.040); in column (2) it equals 1 if the individual uses an overdraft facility on the account (mean of the dependent variable is 0.083). To estimate the relationship, the authors use data from a large bank that has a substantial share (40%) of the market.

On the basis of the results, the authors conclude that the income of neighbors has a positive effect on individual demand for credit.

Your job is scrutinize the empirical strategy. What underlying assumption(s) is (are) the strategy built on? Do you think that this (these) assumption(s) is (are) credible? Can you think of ways of validating the empirical strategy?

Question 7: Evaluation of empirical strategies II (20p)

A recent paper examines whether giving birth at the hospital or at home affects the health of newborns. In the Netherlands, parents can choose (ex ante) whether to give birth at a hospital or at home, provided that the birth is projected to be low-risk.

Low-risk births are always supervised by a mid-wife (no matter if the birth takes place in the home or in the hospital). If complications arise during delivery, if the delivery takes too long, or there is need for pain medication, the mid-wife refers the woman to an obstetrician (i.e. a doctor specializing in pregnancy problems); any of these problems would imply a transfer to a hospital in case the delivery starts at home.

High-risk births, on the other hand, are always supervised by an obstetrician and always take place at a hospital.

The empirical analysis is based on some 686,000 first births, of which some 356,000 are low-risk pregnancies and some 330,000 are high-risk pregnancies. The main analysis focuses on low-risk pregnancies.

The structural equation of interest is the following:

$$Y_{izt} = \alpha + \beta Hospital_{izt} + X'_{izt}\phi + \epsilon_{izt}$$

Y denotes an outcome for infant i who is born in year t to a mother residing in zip-code z. *Hospital* is a binary variable indicating that the birth took place at a hospital, and X is a vector of control variables. The coefficient of interest is β.

The authors worry that *Hospital* is endogenous to the health outcomes for the infant. As an instrument for *Hospital* they use the distance between the home and the nearest obstetric ward.

Table 2 reports the main results for low-risk pregnancies. Across columns you see different indicators of child health (the Apgar score is a summary indicator of child health, with higher scores indicating better health). The distance variables are binary variables for the indicated distances.

a) Compare the magnitudes of the OLS (Panel A) and IV-estimates (Panel D). What kind of selection process would give rise to the differences between the OLS and the IV?

b) Evaluate the IV-strategy from an a priori point of view. Do you think that the conditions for doing IV are fulfilled? Why or why not?

c) The authors also report the results of regressions that are equivalent to panel C for high-risk pregnancies. A test of the joint significance of the distance dummies has p-values of 0.55, 0.52, and 0.77. Why is this important for their empirical strategy?

d) In supplementary tables, the authors report that 74% of mothers are of Dutch decent when Distance: < 1 km, 90% when Distance: 4-7 km, and 93% for Distance $>$ 11 km. Do such differences represent a soucre of concern? Why or why not?

Table 2: Main results

	7-day mortality (1)	28-day mortality (2)	Apgar score (3)
Panel A. OLS (dependent variable: newborn health)			
Hospital	−0.001	−0.072	−0.061***
	(0.155)	(0.163)	(0.004)
Panel B. First stage (dependent variable: hospital birth)			
Distance: < 1 km	0.075***	0.075***	0.074***
	(0.009)	(0.009)	(0.009)
Distance: 1–2 km	0.073***	0.073***	0.073***
	(0.008)	(0.008)	(0.008)
Distance: 2–4 km	0.060***	0.060***	0.060***
	(0.007)	(0.007)	(0.007)
Distance: 4–7 km	0.037***	0.037***	0.036***
	(0.007)	(0.007)	(0.007)
Distance: 7–11 km	0.030***	0.030***	0.030***
	(0.008)	(0.008)	(0.008)
F-statistic	27.979	27.979	28.031
Panel C. Reduced form (dependent variable: newborn health)			
Distance: < 1 km	−0.701**	−0.853**	0.020**
	(0.324)	(0.341)	(0.009)
Distance: 1–2 km	−0.702**	−0.770***	−0.003
	(0.282)	(0.299)	(0.008)
Distance: 2–4 km	−0.554**	−0.718**	0.006
	(0.276)	(0.293)	(0.007)
Distance: 4–7 km	−0.330	−0.500*	0.004
	(0.286)	(0.301)	(0.008)
Distance: 7–11 km	−0.548*	−0.629**	0.016**
	(0.294)	(0.309)	(0.008)
Panel D. IV (dependent variable: newborn health)			
Hospital	−8.287***	−9.219***	−0.018
	(3.157)	(3.353)	(0.088)
Observations	356,412	356,412	355,761
Mean fraction hospital birth	0.678	0.678	0.678
Mean health outcome	1.779	1.978	9.660

Notes: Each column in each panel lists estimates from separate regressions. All regressions control for year, month, and day-of-week of birth, maternal age, ethnicity, gestational age, a third degree polynomial in birth weight, newborn gender, multiple birth, obstetrician supervision, breech birth, and average income in the postal code of mother's residence (see Section III). The excluded distance category comprises postal codes at least 11 km away from an obstetric ward. The *F*-statistic corresponds to a test of joint significance of the distance indicators. Robust standard errors clustered at the postal code level are shown in parentheses.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.