Exam Econometrics II

August 24, 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)

The Kernel density estimator for the density at a particular point (x_0) is given by

$$\hat{f}_K(x_0) = \frac{1}{Nh} \sum_i K\left(\frac{x_i - x_0}{h}\right)$$

where N denotes sample size, h bandwidth, and $K(\cdot)$ the Kernel. The bias and variance of this estimator are given by

$$BIAS = \frac{1}{2}h^2 f''(x_0) \int z^2 K(z) dz$$

$$VAR = \frac{1}{Nh} f(x_0) \int [K(z)]^2 \, dz$$

where $z = (x_i - x_0)/h$

a) What are the conditions for point-wise consistency?

b) Define a criterion for choosing the bandwidth optimally. Use this criterion to show how the optimal bandwidth relates to sample size. (The full derivation is not required. But you should use a couple of key equations to show the precise relationship between the optimal bandwith and sample size)

c) Does the optimal bandwidth satisfy the conditions for point-wise consistency? Explain.

Question 2 (10p)

Consider the following model

$$Y_i = \beta_0 + \beta_1 x^*_{1i} + u_i$$

x^*_{1i} is measured with error. Rather we observe

$$x_{1i} = x^*_{1i} + v_i$$

where the measurement error is purely random (such that $COV(x^*_{1i}, v_i) = COV(u_i, v_i) = 0$). We thus estimate the model

$$Y_i = b_0 + b_1 x_{1i} + e_i$$

a) Derive an expression that relates the estimable regression coefficient (b_1) to the true regression coefficient (β_1)

b) Suppose you add another regressor (x_2) to the model. Does adding another regressor increase or reduce measurement error bias? Explain. (assume that x_2 is not measured with error and that $COV(x_{2i}, v_i) = 0$)

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)

Suppose you want to estimate the average effect of treatment on the treated (ATT)

$$ATT = E(Y_i(1) - Y_i(0)|\, D_i = 1)$$

You are prepared to assume that you can observe and control for all factors (X) that confound the causal relationship between Y_i and D_i. A common empirical approach in this scenario is matching.

a) State formally the minimal assumption that allows you to estimate ATT in this setting

b) Compare matching to traditional regression. To what extent are these two approaches different? What are the advantages and disadvantages of matching relative to regression?

Question 5 (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 6: Evaluation of empirical stragies I (20p)

You are interested in estimating the (causal) return to college education. Thus, you want to estimate the relationship:

$$\ln Y_i = \alpha + \beta College_i + X'_i \gamma + \epsilon_i$$

where i indexes individuals, $\ln Y$ denotes (the log of) annual earnings, *College* years of college education, and X a set of control variables.

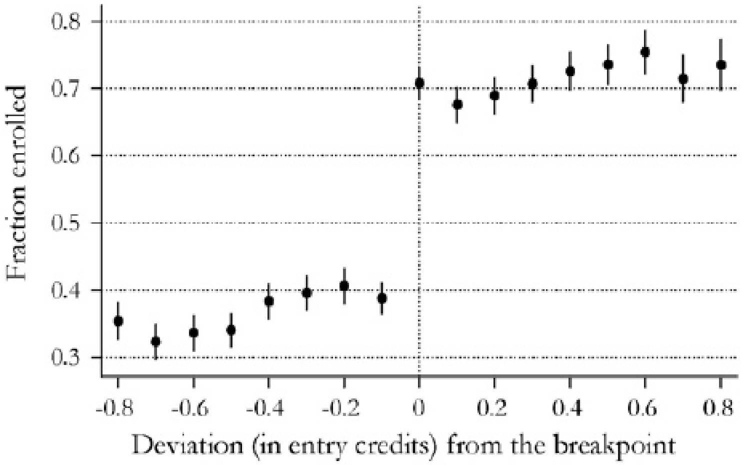
a) Consider estimating the above equation by OLS. Why is the OLS-estimate of β likely biased? What is the likely sign of the bias?

b) An alternative approach is to use an admission rule. Admission to a given program is typically based on entry credits (for instance the grade point average from upper secondary school). Those who are above the entry credit breakpoint are admitted to the program; those who are below the breakpoint are not admitted to the program (but may be admitted to another program). Figure 1 comes from a recent paper using admissions data from 1982. Explain how the admission rule can potentially help you in solving the identification problem you identified under a). Explicitly specify any regressions that you would estimate.

c) The remaining figures (Figures 2-4) come from the same paper. Explain (briefly) why each of the three figures is important and what you would conclude from each of the three figures. (Please make sure to specify which figure you are talking about. For instance: "Figure X shows... It is important because... From Figure X I conclude...").

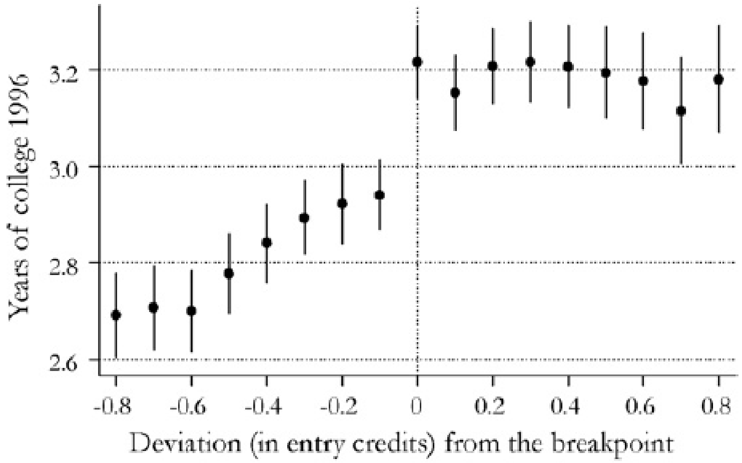
d) Do you think that this empirical strategy can be used to estimate the causal return to college education? Why or why not?

Figure 1: College enrollment in 1982



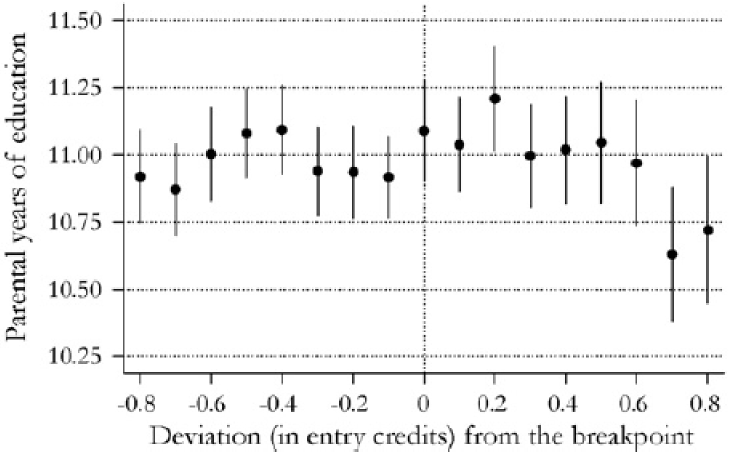
Note: Circles are local averages. Solid lines around each circle show the 95% confidence interval.

Figure 2: Years of College 1996



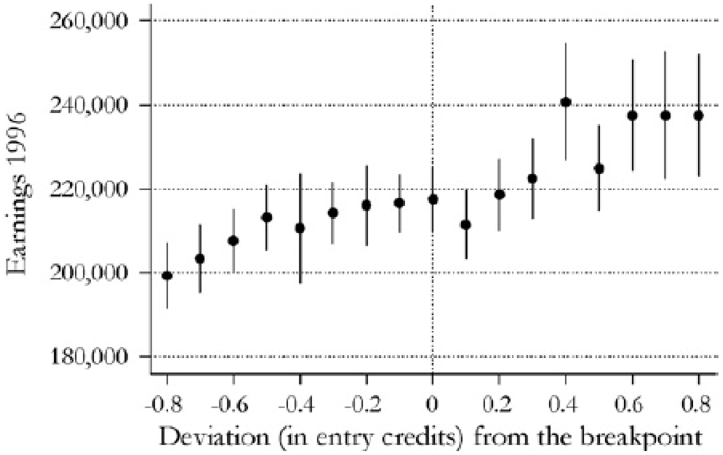
Note: Circles are local averages. Solid lines around each circle show the 95% confidence interval.

Figure 3: Parental years of education



Note: Circles are local averages. Solid lines around each circle show the 95% confidence interval.

Figure 4: Earnings 1996



Note: Circles are local averages. Solid lines around each circle show the 95% confidence interval.

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

A recent paper examines whether reading performance has a causal impact on math performance. To examine this question the authors use data from (different waves of) PISA and focus attention on first-generation immigrants.

The structural equation of interest is the following:

$$MATH_{iodt} = \beta READ_{iodt} + \gamma AA_i + \delta LD_{od} + X'_{iodt}\phi + \alpha_o + \alpha_d + \alpha_t + \epsilon_{iodt} \quad (3)$$

$MATH$ denotes math performance for student i, from country o, in destination country d and PISA wave t; $READ$ is a measure of reading performance which is defined in analogy with $MATH$. AA denotes the age of arrival of the immigrant student, LD the linguistic distance between origin and destination countries, and X a vector of control variables. α_o, α_d, and α_t are fixed effects for origin country, destination country, and PISA wave. The coefficient of interest is β.

The measure of linguistic distance comes from an algorithm comparing the pronounciation of common words. The greater the distance between the PISA test language (in the destination country) and the majority language of the student's country of birth, the higher is LD.

The authors worry that $READ$ is endogenous to math performance. As an instrument for $READ$ they use the interaction between linguistic distance and age at arrival. The first-stage equation in their IV-strategy thus is

$$READ_{iodt} = b(AA_i \times LD_{od}) + cAA_i + dLD_{od} + X'_{iodt}f + a_o + a_d + a_t + e_{iodt}$$

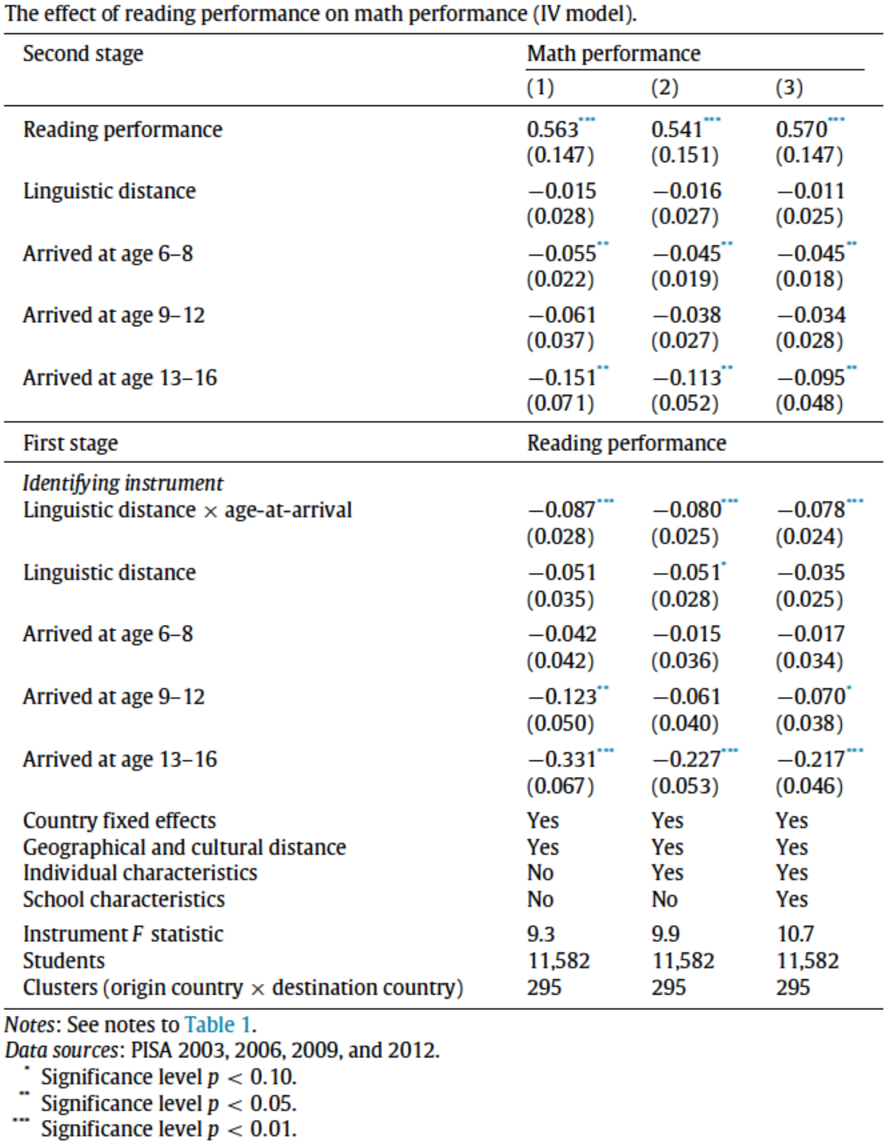
Table 1 reports the main results. Across columns you see estimates with differenct sets of control variables (they are defined in the table note).

a) Directly estimating equation (3) by OLS yields an estimate of 0.771 (standard error: 0.010) in a specification that corresponds to column (1) of Table 1. Does the difference between the OLS and IV estimates comply with your priors regarding the bias of OLS? Motivate your answer.

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) Any differences particular to a pair of origin and destination countries are accounted for by controlling linearly for linguistic, cultural, and geographic distance. Suggest a more flexible way of controlling for such differences that would still allow you to identify the coefficient of interest.

Table 1: Main results



The effect of reading performance on math performance (IV model).

Second stage	Math performance		
	(1)	(2)	(3)
Reading performance	0.563*** (0.147)	0.541*** (0.151)	0.570*** (0.147)
Linguistic distance	−0.015 (0.028)	−0.016 (0.027)	−0.011 (0.025)
Arrived at age 6–8	−0.055** (0.022)	−0.045** (0.019)	−0.045** (0.018)
Arrived at age 9–12	−0.061 (0.037)	−0.038 (0.027)	−0.034 (0.028)
Arrived at age 13–16	−0.151** (0.071)	−0.113** (0.052)	−0.095** (0.048)
First stage	Reading performance		
Identifying instrument			
Linguistic distance × age-at-arrival	−0.087*** (0.028)	−0.080*** (0.025)	−0.078*** (0.024)
Linguistic distance	−0.051 (0.035)	−0.051* (0.028)	−0.035 (0.025)
Arrived at age 6–8	−0.042 (0.042)	−0.015 (0.036)	−0.017 (0.034)
Arrived at age 9–12	−0.123** (0.050)	−0.061 (0.040)	−0.070* (0.038)
Arrived at age 13–16	−0.331*** (0.067)	−0.227*** (0.053)	−0.217*** (0.046)
Country fixed effects	Yes	Yes	Yes
Geographical and cultural distance	Yes	Yes	Yes
Individual characteristics	No	Yes	Yes
School characteristics	No	No	Yes
Instrument *F* statistic	9.3	9.9	10.7
Students	11,582	11,582	11,582
Clusters (origin country × destination country)	295	295	295

Notes: See notes to Table 1.
Data sources: PISA 2003, 2006, 2009, and 2012.
* Significance level $p < 0.10$.
** Significance level $p < 0.05$.
*** Significance level $p < 0.01$.

Notes: Country fixed effects include origin country fixed effects and destination country fixed effects. Geographical distance is the distance of the capitals between origin and destination country. Cultural distance is based on genetic proximity measure between populations (see Spolaore and Wacziarg, 2009). Individual characteristics include student age and gender as well as highest education of parents, highest occupational status of parents, and number of books at home. School characteristics include school location, indicator for private school, weekly math instruction time, enrollment, shortage of math and language teachers, and three autonomy measures (content, personnel, and budget).