Economics 361
Problem Set #8

Jun Ishii *
Department of Economics
Amherst College

Fall 2022

Question 1: Early Settler Mortality as an Instrument

The problems in this Question are based on "The Colonial Origins of Comparative Development: An Empirical Investigation," by D. Acemoglu, S. Johnson, & J. Robinson (*American Economic Review*, December 2001). Read the article before tackling the problems below.

The authors' key econometric model consists of equations (1)-(4) (see Section IV, p.1383).

$$\begin{aligned} logY_i &= \mu + \alpha R_i + X_i'\gamma + \epsilon_i && (1)\\ R_i &= \lambda_R + \beta_R C_i + X_i'\gamma_R + \nu_{Ri} && (2)\\ C_i &= \lambda_c + \beta_C S_i + X_i'\gamma_C + \nu_{Ci} && (3)\\ S_i &= \lambda_s + \beta_S logM_i + X_i'\gamma_S + \nu_{Si} && (4) \end{aligned}$$

The parameter of interest is α, indicating the extent to which the institution "protection against expropriation risk" contributed toward economic growth. The main econometric problem is that

$$E[logY_i \mid R, X] \quad \neq \quad \mu + \alpha R_i + X_i'\gamma$$

The above problem can be traced back to three classes of issues surrounding R_i

- "Reverse Causality" : wealthier nations can afford/prefer better institutions
- "Omitted Variables" : institutions are correlated with other **omitted** factors contributing to growth
- "Measurement Error" : institutions are measured with error

The authors argue that $logM_i$ (early settler mortality rates) can serve as an instrument for R_i. Essentially, they are arguing that

$$E[logY_i \mid logM, X] \;=\; \mu + \alpha\; E[R_i \mid logM, X] + X_i'\gamma \quad \text{or, equivalently,} \quad E[\epsilon_i \mid logM, X] \;=\; 0$$

(not exactly, but close enough for the scope of this course)

*Office: Converse Hall 315 Phone: (413) 542-2901 E-mail: jishii@amherst.edu

(a) On pp.1379-80, the authors argue that the "Omitted Variable" issue would introduce a **positive** bias into the naïve OLS estimate of α and that the "Measurement Error" issue would introduce a **negative** bias. Briefly explain the authors' assertions. Also, based on the estimates obtained by the authors, which of those two issues seem more important and why?

(b) The main 2SLS estimates of α are reported in Table 4 (p.1386). What caught your professor's eye are the R^2 for the first stage regression and the number of observations. The authors argue that their 2SLS estimates of α are *highly significant* – i.e. $H_o : \alpha = 0$ can be easily rejected. Briefly explain the authors' assertion. Also, briefly explain why your professor, respectfully, disagrees.

(c) Consider the estimates of the coefficient before $logM_i$ in the first stage regression (Table 4, p.1386, Panel B) – e.g. -0.61. Suppose the authors' econometric model is correct. In terms of $\{\mu, \alpha, \gamma, \lambda_R, \beta_R, \gamma_R, \lambda_C, \beta_C, \gamma_C, \lambda_S, \beta_S, \gamma_S\}$, what does -0.61 estimate?

(d) In order for an instrument to be considered valid and good, it must satisfy two conditions. The authors use the results in Table 3 and 4 (pp.1385-6) to argue that the Log of European Settler Mortality satisfies at least one of the conditions necessary for it to be a valid and good instrument for current institutions (R_i: protection against expropriation measure). Explain.

(e) Not everyone is convinced that $logM_i$ is a valid instrument for R_i. One of the main criticism is that $logM_i$ may be correlated with the current disease environment (e.g. prevalence of malaria) and that the current disease environment has a direct impact on $logY_i$. Briefly explain how this criticism invalidates $logM_i$ as an appropriate instrument for R_i. Briefly explain how the authors attempt to deflect this criticism.

Question 2: Linear Probability Model

Sometimes, the dependent variable of interest will be a *discrete* random variable. The standard example is of a dependent variable that is binary, realizing either the value 0 or 1. This type of 0-1 binary variable is used to indicate whether an event occurred. For example

$$Y_i = \begin{cases} 1 & \text{if } i \text{ attends college} \\ 0 & \text{otherwise} \end{cases}$$

The researcher may be interested in understanding the value of Y_i given some other data on observation i. These other variables are the explanatory variables X_i. More specifically, the researcher may be interested in estimating $P(Y_i = 1 \mid X_i)$: the *conditional* probability that some event occurs.

For example, X_i may indicate the socio-economic status of person i. Estimating $P(Y_i = 1 \mid X_i)$ would enable the researcher to study how socio-economic status affects college participation.

The **preferred** method for estimating these conditional probabilities is maximum likelihood, especially discrete choice models (e.g. probit and logit). But it is possible to estimate these conditional probabilities using regression techniques *if* $E[Y|X] = X\beta$. We explore this option – the Linear Probability Model – in this question. Note: Here, X_i refers to the i^{th} row of X.

(a) Show that $E[Y_i|X_i] = P(Y_i = 1 \mid X_i)$.
HINT: $E[Y_i|X_i]$ is a weighted sum, with weights determined by ...

Therefore, if $E[Y|X] = X\beta$ then each $P(Y_i = 1 \mid X_i)$ is a linear function of X_i. Specifically, $P(Y_i = 1 \mid X_i) = X_i'\beta$. For the rest of the problems in this question, assume $E[Y|X] = X\beta$.
Also, assume that each observation i is independent of each other and that X is full rank.

(b) Show that $\text{Var}[Y_i|X_i] = X_i'\beta\,(1 - X_i'\beta)$

(c) Suppose you regressed Y on X. Does this provide you with an unbiased estimate of β? Briefly explain.

(d) Suppose you regressed Y on X. Does this provide you with the "best" (minimum variance) unbiased linear estimator of β? Briefly explain.

(e) Suppose you wanted to estimate β using Feasible GLS (FGLS). First, explain how you would estimate $\Sigma \equiv \text{Var}(Y|X)$. Second, explain how you would obtain b^{FGLS} using that estimate of Σ.

(f) The main criticism of this Linear Probability Model, even when $E[Y|X] = X\beta$, is that the estimated $P(Y_i = 1 \mid X_i)$ may be unreasonable. Explain.
HINT: Think Kolmogorov's Axioms of Probability

(g) Read "Results on the Bias and Inconsistency of Ordinary Least Squares for the Linear Probability Model," by W. Horace and R. Oaxaca, *Economics Letters*, 2006, vol.90, pp.321-327 (`ho06.pdf`). No need to write anything. But the material discussed in the article are fair game for quiz and final exam. You may ignore the complication discussed in this article when solving **(a) - (f)**

Past Final Exam Questions

Problems in Questions 3 and 4 are based on the assigned reading "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," by Joshua Angrist, *American Economic Review*, 80(3), June 1990, pp.313-336. Question 3 problems are more conceptual and Question 4 problems more "mathematical." Question 4 is not for grade – provided as practice. (Solutions will not be provided for Q4)

For all of the problems associated with the assigned reading, **do not just quote the reading**. Explain in terms of Econ 361 lessons and discussions.

Question 3: Drafting a "Natural" Experiment – Part I

In the author's own words, the purpose of the research may be stated as

> "The goal of this paper is to measure the long-term labor market consequences of military service during the Vietnam era. Previous research comparing civilian earnings by veteran status may be biased by the fact that certain types of men are more likely to serve in the armed forces than others. For example, men with relatively few civilian opportunities are probably more likely to enlist. Estimation strategies that do not control for differences in civilian earnings potential will incorrectly attribute lower civilian earnings of veterans to military service. The research reported here overcomes such statistical problems by using the Vietnam era draft lotteries to set up a natural experiment that randomly influenced who served in the military." (p.313-4)

(3a) Using arguments explicitly discussed in Econ 361, explain why the difference in average lifetime earnings between veterans and non-veterans from the same age and socio-demographic group may provide a misleading estimate of the impact of military service on average lifetime earnings for people in that group? Would such an estimate likely over or under-estimate the impact of military service?

The author argues that whether an observed worker was draft eligible (d_i) may serve as an appropriate instrument for veteran status (s_i) in a regression seeking to measure the impact of veteran status on lifetime earnings (y_{cti}), such as the regression embodied by equation (1) in the article

$$y_{cti} = \beta_c + \delta_t + s_i\alpha + u_{it} \quad \text{where} \quad d_i = \begin{cases} 1 & \text{if } i \text{ is draft eligible} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad s_i = \begin{cases} 1 & \text{if } i \text{ is a veteran} \\ 0 & \text{otherwise} \end{cases}$$

Note that $\beta_c = \sum_{j=1}^{C} \beta_j \gamma_{ji}$ where $\gamma_{ji} = 1$ if individual i is in cohort j and 0 otherwise. Similarly for δ_t. For an observation indexed (cti), the observation concerns individual i in cohort c and period t.

(3b) Explain why the author (i) feels the need to instrument for s_i despite the inclusion of cohort and time fixed effects (β_c, δ_t) and (ii) believes d_i would serve as a good instrument for s_i

(3c) Explain why the author does not simply replace s_i with d_i and use the estimated coefficient before d_i as the estimate for α?

In section III.B, the author proposes a different set of instruments for s_i. Instead of using the Random Sequence Numbers (RSNs) to develop one dummy variable, d_i, indicating whether the worker was draft-eligible, the author proposes using the RSNs to develop 73 dummy variables, each indicating whether worker i is in one of 73 groups of consecutive lottery numbers, starting with 1-5 and ending with 360-365.

(3d) Explain why 2SLS with these 73 instruments may yield a "better" (efficient) estimator of α than 2SLS with just d_i as an instrument for s_i. **HINT:** more instruments do not necessarily mean "better" estimator

(3e) Suppose that workers varied not only in draft eligibility but also whether draft eligibility affected their military participation. For simplicity, let there be some workers who would have joined the military regardless of the draft and all others only if they were drafted. Suppose further that the impact of veteran status on lifetime earning, value of α, differed between these two groups. The author's 2SLS estimator provides, at best, a consistent estimate for only one of these two α values. Which one and why?

Question 4: Drafting a "Natural" Experiment – Part II

This question will not be graded – practice problem

The author asserts on pp.319-20 that if attention were restricted to a single cohort (only observations from a single c subindex) the instrumental variable (IV) and two-stage least squares (2SLS) estimate of α would simply be $\hat{\alpha} = \frac{\bar{y}^e - \bar{y}^n}{\hat{p}^e - \hat{p}^n}$. This is also true if we were to restrict attention to a single cohort and year, therefore data only from observations with the same (c,t) subindex. Restricting attention to a single cohort and year simplifies the main regression equation to: $y_i = \lambda + s_i\alpha + u_i$

Let N^e and N^n represent the number of observations associated with draft-eligible and draft-ineligible workers in the sample, respectively. $N = N^e + N^n$ where N is the total sample size. One can show that

$$\bar{d} \equiv \frac{1}{N}\sum_{i=1}^{N} d_i = \frac{N^e}{N} \qquad \bar{s} \equiv \frac{1}{N}\sum_{i=1}^{N} s_i = \frac{N^e\,\hat{p}^e + (N - N^e)\,\hat{p}^n}{N} = \bar{d}\,\hat{p}^e + (1-\bar{d})\,\hat{p}^n$$

(4a) Consider the first stage of the author's 2SLS estimator of α, restricting attention to a single cohort and year. Explicitly show that the first stage regression yields the following predicted value for s_i

$$\hat{s}_i = \begin{cases} \hat{p}^e & \text{if } d_i = 1 \\ \hat{p}^n & \text{if } d_i = 0 \end{cases}$$

(4b) Recall from our class discussion that 2SLS is motivated by the expectation of the dependent variable conditioned on the instruments, i.e. $E[y_i \mid d]$. Use $E[y_i \mid d]$ to explain why the result shown in **(4a)** is intuitive.

(4c) Show that $\bar{\hat{s}} = \frac{1}{N}\sum_{i=1}^{N}\hat{s}_i = \frac{1}{N}\sum_{i=1}^{N} s_i = \bar{s}$. Briefly explain how this result may be thought as a sample analog to an application of the Law of Iterated Expectations.

(4d) Now explicitly derive the author's assertion that, when restricting attention to a single cohort and year, the author's 2SLS estimator of α is $\hat{\alpha} = \frac{\bar{y}^e - \bar{y}^n}{\hat{p}^e - \hat{p}^n}$.

Hint: The math may be easier if you use the sample analogs to the *proper* definitions of variance/covariance ...