Problem Set 3

Ec270c, Development Economics, Spring 2008

by

Professor Bryan Graham

Due Date: May 8th

Problem set due by 4PM in Professor Graham's mailbox on the due date. Problem sets are graded on a 1 to 5 scale with *one point off per day late.* You are free to work in groups, however each student must turn in an individual write-up and any accompanying (individually typed up, commented and executed) electronic files (if appropriate). Please e-mail me with any typos or mistakes.

1 Part A: The empirics of policy analysis

Let X_{ibc} equal the years of completed schooling of individual $i = 1, \ldots M_{bc}$, born in region $b = 1, \ldots, N$, in year $c = 1, \ldots C$. Let W_{bc} be an measure of primary school availability in region b during the period when cohort c would have been of primary school age. Consider the following statistical model relating years of completed schooling to school availability:

$$X_{ibc} = W_{bc}'\pi + \lambda_{b0} + \mu_c^* + V_{ibc}, \qquad \mathbb{E}\left[V_{ibc}\middle| W_{b1}, \ldots, W_{bC}\right] = 0$$

Define the notation

$$\mathbb{E}^*\left[\lambda_{b0}\middle| W_{b1}, \ldots, W_{bC}\right] = \eta_0 + \sum\nolimits_{c=1}^{C} W_{bc}'\eta_c,$$

where $\mathbb{E}^*\left[Y\middle| X\right]$ denotes the best (i.e., mean squared error minimizing) linear predictor of Y given X.

[1] Let $\overline{X}_{bc} = \sum_{i=1}^{M_{bc}} X_{ibc}/M_{bc}$ and show that

$$\mathbb{E}^*\left[\overline{X}_{bc}\middle| W_{b1}, \ldots, W_{bC}\right] = W_{bc}'\pi + \sum\nolimits_{c=1}^{C} W_{bc}'\eta_c + D_c'\mu,$$

where D_c is a $C \times 1$ vector of year-of-birth dummy variables and $\mu = (\mu_1, \ldots, \mu_C)'$ with $\mu_c = \mu_c^* + \eta_0$. Interpret η_c, what sign should it take if school placement is compensatory? What sign should it take if it is regressive? Relate your discussion to the Rosenzweig and Wolpin (1986, AER) paper on the syllabus.

[2] Assume your data are a random sample taken from Nicaragua in 1998. Assume that cohorts born in the years $c = 1, \ldots, T_1$ completed their primary schooling schooling prior to the end of the Somoza period. Assume that cohorts $T_2, \ldots, C$ completed their schooling during the Sandinista period, with cohorts $T_1+1, \ldots T_2-1$ attending school during portions of each regime's tenure. Given your prior knowledge of the education policies pursued by each regime, what pattern would you expect the η_c coefficients to take? Explain your answer.

[3] Show that

$$\mathbb{E}^* \left[\overline{X}_{bc} \middle| W_{b1}, \ldots, W_{bC} \right] - \mathbb{E}^* \left[\overline{X}_{bc-1} \middle| W_{b1}, \ldots, W_{bC} \right] = \Delta W_{bc}' \pi + \left(\mu_c - \mu_{c-1} \right).$$

Interpret your result and suggest is implications for consistent estimation of π.

[4] Consider the alternative model of schooling

$$X_{ibc} = W_{bc}' \pi + \lambda_{0b} + \lambda_{1b} c + \mu_c + V_{ibc}, \qquad \mathbb{E} \left[V_{ibc} \middle| W_{b1}, \ldots, W_{bC} \right] = 0.$$

Interpret λ_{1b}? Define the notation

$$\mathbb{E}^* \left[\lambda_{b1} \middle| W_{b1}, \ldots, W_{bC} \right] = \phi_0 + \sum_{c=1}^{C} W_{bc}' \phi_c.$$

Show that

$$\mathbb{E}^* \left[\overline{X}_{bc} \middle| W_{b1}, \ldots, W_{bC} \right] - \mathbb{E}^* \left[\overline{X}_{bc-1} \middle| W_{b1}, \ldots, W_{bC} \right] = \Delta W_{bc}' \pi + \phi_0 + \sum_{c=1}^{C} W_{bc}' \phi_c + \left(\mu_c - \mu_{c-1} \right).$$

Discuss the implications of your result for consistent estimation of π. Interpret the ϕ_c. What do they reveal about government education policy?

[5] Assume that $c = 1, 2$ and that $b = 1, 2$. Further assume that $W_{22} = 1$, with $W_{11} = W_{12} = W_{21} = 0$. You may think of $b = 1, 2$ as denoting urban and rural areas of Nicaragua and $c = 1, 2$ as denoting two both cohorts of individuals: the last cohort to finish primary school under Somoza and the first cohort to finish primary school un-

der the Sandinista regime. Show, using the augmented model from question 4, that the difference-in-differences

$$\pi_{\text{DID}} = \mathbb{E}\left[\left.\overline{X}_{bc}\right| b=2, c=2\right] - \mathbb{E}\left[\left.\overline{X}_{bc}\right| b=2, c=1\right] \\ - \left\{\mathbb{E}\left[\left.\overline{X}_{bc}\right| b=1, c=2\right] - \mathbb{E}\left[\left.\overline{X}_{bc}\right| b=1, c=1\right]\right\},$$

equals $\pi + \phi_2$. What bias would you expect π_{DID} to have, upward or downward? If you have more than two cohorts, how might you 'test' for whether $\pi_{\text{DID}} = \pi$? How does this related to the so-called 'common trend' assumption?

2 Part B: Nonparametric regression

Download the dataset `calories.out` from the course website. The variable Y in the dataset gives the logarithm of total daily calories available for a sample of 1,358 families from Nicaragua in 2001. The variable X gives the logarithmn of 'real' total expenditure (annualized).

[1] Compute the 5 and 95 percentiles of the empirical distribution of X. For a grid of 100 equally spaced values which span these two percentiles. For each point $m = 1, \ldots, 100$ estimate the local quadratic regression

$$\left(\widehat{\alpha}(x_m), \widehat{\beta}(x_m), \widehat{\gamma}(x_m)\right) \\ = \underset{a,b,c}{\arg\min} \sum_{i=1}^{1,358} \mathcal{K}\left(\frac{X_i - x_m}{h}\right)\left(Y_i - a - b(X_i - x_m) - c(X_i - x_m)^2\right),$$

where

$$\mathcal{K}(v) = \frac{3}{4}\left(1 - v^2\right)\mathbf{1}(|v| < 1).$$

Plot $\{x_m, \widehat{\alpha}(x_m)\}_{m=1}^{100}$ and $\left\{x_m, \widehat{\beta}(x_m)\right\}_{m=1}^{100}$ on two separate graphs. Let $h = \sigma_X \cdot 1358^{-1/5}$, one half this value and twice this value. Plot the three estimates of each object on the same graph. Which bandwidth seems best for $\alpha(x_m)$? For $\beta(x_m)$?

[2] These data were obtained via stratified random sampling. The variable `village` denotes the first strata. There are 42 such villages in the dataset. Perform the following block bootstrap procedure. Sample with replacement 42 villages from the data. For each sampled village include all sampled households. For each such sample estimate

$\left\{\widehat{\alpha}\left(x_m\right), \widehat{\beta}\left(x_m\right)\right\}_{m=1}^{100}$ using the local quadratic regression procedure and your preferred bandwidth from question 1. Repeat this a large number of times (at least 100). Form pointwise 90% confidence intervals by taking the 5 and 95 percentiles of the bootstrap distributions of $\widehat{\alpha}\left(x_m\right)$ and $\widehat{\beta}\left(x_m\right)$. Plot $\left\{x_m, \widehat{\alpha}\left(x_m\right)\right\}_{m=1}^{100}$ and $\left\{x_m, \widehat{\beta}\left(x_m\right)\right\}_{m=1}^{100}$ on two separate graphs with the corresponding confidence intervals. Deaton (1997) describes a related procedure.

[3] Discuss your results and situate them within the wider literature of the elasticity of calorie demand with respect to total household outlay.