MIT [14.661] **Philip Oreopoulos**
Fall 2004-2005

Problem Set 2
Labor Economics

You may work in groups, but you must do the coding and write-up on your own. Include your code with your writeup. Do not submit STATA output in raw form.

This problem set provides an introduction and summary to issues related to intergenerational mobility. The related papers you may find helpful consulting are:

Haider, Steven, and Gary Solon, "Life-Cycle Variation in the Association between Current and Lifetime Earnings," mimeo, March 2004-10-09

McCrary, Justin, and Healther Royer, 'Does Maternal Education Affect Infant Health? A Regression Discontinuity Approach Based on School Entry Age Laws," mimeo

Oreopoulos, Philip, Marianne Page, and Ann Stevens, "Does Human Capital Transfer from Parent to Child? The Intergenerational Effects of Compulsory Schooling", NBER WP#10164

Part I

The model of interest is:

$y_{1i} = \beta y_{0i} + e_i$,

where y_{1i} is average (age detrended) lifetime log earnings for sons, $y_{1i} = \bar{y}_{1it}$. and y_{0i} is average (age detrended) lifetime log earnings for fathers, $y_{0i} = \bar{y}_{0it}$. β is the intergenerational mobility coefficient that provides an omnibus measure of how father's labor market outcomes relate to son's eventual labor market outcomes.

1) Suppose that log earnings for fathers at age s is $y_{0is} = y_{0i}$, and log earnings for sons at age t can be decomposed into three components: $y_{1it} = y_{1i} + w_{1it} + u_{1it}$, where w_{1it} reflects transitory deviations from trend, and u_{1it} is measurement error. Both the transitory and measurement error component have mean zero, and $\text{cov}(w_{1it}, y_{1i}) = \text{cov}(u_{1it}, y_{1i}) = \text{cov}(w_{1it}, u_{1it}) = 0$.

You only have sons and father's earnings data for one year. Show that a linear regression of annual sons earnings at age t on annual fathers earnings at age s generates a consistent estimate for β.

2) Now define the transitory and measurement error components of father's log earnings at s as: $y_{0is} = y_{0i} + w_{0is} + u_{0is}$, and assume $\text{cov}(w_{0is}, y_{0i}) = \text{cov}(u_{0is}, y_{0i}) = \text{cov}(w_{0is}, u_{0is}) = 0$. Calculate the attenuation bias for your estimate of β

3) Assume no serial correlation in log earnings across periods. What is the attenuation bias using an average of father's log earnings over T years? Assuming the share of the variance in annual earnings is accounted by permanent factors by .5, transitory factors by .3, and measurement error by .2, how much is the bias reduced averaging over 5 years compared to just using one year? What about over 10 years?

Bonus: Assume serial correlation:

$w_{0is} = \rho w_{0is-1} + \xi_{is}$.

What is the attenuation bias, and how much is the attenuation bias reduced by averaging over 5 years and 10 years if $\rho = .8$ (you need to recall rules of summation to do this). Ignore the fact that persistence in annual shocks affects the average lifetime earnings if the shocks do not die off over the course of an individual's working life.

4) The model above assumes log earnings differs by level rather than growth. Everyone has the same age profile, and so we can detrend by age without any concern. For any given age, the variance in log earnings is the same. Now consider the implications if sons differ in earnings growth. This is a natural extension if we believe wage growth differs by initial skill or education level. To contrast with above, suppose there are no transitory shocks and no measurement error. Father's earnings are $y_{0is} = y_{0i}$, and that initial son's log earnings are the same and increase linearly:.

$y_{1it} = x_{10} + \gamma_i t$

Then $y_{1it} = \lambda_t y_{1i}$, where $\lambda_t = 1$ only when $t = T/2$. $\lambda_t < 1$ and $y_{1it} < y_{1i}$ for younger ages and vice versa for older ages. At $t = 0$, earnings are the same for everyone, and the intergenerational mobility coefficient is zero. We can rewrite $y_{1it} = y_{1i} + v_{1it}$, where $v_{1it} = y_{1i}(\lambda_t - 1)$. Notice we have non-classical measurement error.

What is the bias from regressing y_{1it} on y_{0is}? Compare this with question 1.

5) Finally, assume non-classical measurement error for fathers too: $y_{0is} = \lambda_s y_{0i}$. What is the bias from regressing y_{1it} on y_{0is}?

Part II

Part II provides a brief example of a regression discontinuity design.

The dataset bw_educ_data.dta contains cell mean data for all births in Texas and California between 1989 and 2001 among mothers aged 23 years or less (see McCrary and Royer for details). Data are aggregated by birth day of mother (365 cells).

Variable definitions

Rel_bdate: number of days born before or after the closest school entry day cut-off (December 1st in California, September 1st in Texas).
Educ_cal: mean years of schooling for mothers from the California sample
Educ_texas: mean years of schooling for mothers from the texas sample
Bw_cal: fraction of children born low birth weight from the California sample
Bw_texas: fraction of children born low birth weight from the texas sample

Children age 6 a day before the school entry law are required to enter Grade 1, while children age 6 a day after the entry law do not enter Grade 1 until the following school year. These latter children typically have one year less education before having the legal option to drop out of high school.

We are interested in estimating the following model: $y_i = \beta S_i + e_i$, where y_i is birthweight of child from mother i, and S_i is years of schooling for mother i. Suppose relative birth date, x_i, is small enough an interval that we can treat it as continuous. A discontinuity exists at: x^*. Children with $x_i \leq x^*$ enter school one year earlier than children with $x_i > x^*$. A child is legally allowed to leave school when they turn 16. Thus, assuming the existence of left and right limits at x^*

$$S^+ \equiv \lim_{x \to x_+^*} E(S_i \mid x_i) > S^- \equiv \lim_{x \to x_-^*} E(S_i \mid x_i)$$

Around birth dates arbitrarily close to x^*,

$$E(y_i \mid x^* + \varepsilon) - E(y_i \mid x^* - \varepsilon) = E(\beta S_i \mid x^* + \varepsilon) - E(\beta S_i \mid x^* - \varepsilon)$$

The limit of this difference when $\varepsilon \to 0$ is:

$$y^+ - y^- \equiv \lim_{\varepsilon \to 0} E(y_i \mid x^* + \varepsilon) - \lim_{\varepsilon \to 0} E(y_i \mid x^* - \varepsilon) = \beta(S^+ - S^-)$$

So the parameter of interest is identified by:

$$\beta = \frac{y^+ - y^-}{S^+ - S^-}$$

1) Estimate the 'first stage', $S^+ - S^-$, by assuming schooling is a polynomial function of date of birth and an indicator for $x_i > x^*$. Assume the polynomial takes on the same functional form to the left and right of the discontinuity (you may also consider other functional forms or kernel estimators if you choose). Graph

the fitted first stage on scatter graph, with 95 percent confidence regions around the predicted fit, and a line at the discontinuity.

2) Do the same for the reduced form.

3) What is your IV-regression discontinuity estimate for the effect of schooling on child's birth weight? Interpret this as a LATE.

4) Why is it useful that we have both Texas and California data?

Part III

The dataset repeat_educm.dta contains observations from the 1% samples of the 1960,70, and 80 U.S. Censuses. The sample includes all children ages 7 to 15 born in the United States matched to mothers also born in the United States.

Variable definitions

State: current state of household residence
Bpl2mom: birth state of mom
Year14m: year mom was age 14
Dropagem: minimum school leaving age at state mom was born in, when mom was age 14
Age: age of child
Censyear: census year
Clmom: predicted number of years mom required to stay in school based on state of birth, and year mom was 14
Agemom: age of mom
Iwagemom: mom's wage and salary income
Higr2mom: mom's highest grade completed
Ftotinc: mom's total family income
Famsize: number of own family members in household
Blac: black indicator for child
Female: female indicator for child
Repeat: indicator whether child behind at least once grade (based on school entry age at current state of residence and age of child)
Higrres50: indicator for whether child's grade below median for age/state group
Regionm: region of mom's state of birth (defined on IPUMS web site)

We're interested in examining the intergenerational effect of mother's compulsory schooling on the probability that a child repeats a grade.

1) Estimate the first stage, both with the individual sample and with aggregated cell means (if STATA runs out of RAM, find another computer to work on):

$$MothEd_{ylm} = \eta_0 + \eta_1 CL7_{l,m+14} + \eta_2 CL8_{l,m+14} + \eta_3 CL9_{l,m+14} + \eta_4 X_{ylm} + v_y + v_l + v_m + v_{ylm}$$

where $MothEd_{ylm}$ represents mother's education level for the group of youths observed in census year *y,* with mothers from state l born in year m, and X is a vector of variables that capture the child group's average race, gender and age. CL7, CL8, and CL9 which are dummy variables that denote required years

of schooling prior to obtaining a work permit of 7, 8, or 9 or more years. You may use an alternative compulsory schooling measure using the dropagem variable or assuming a linear single variable for CLmom or dropagem.

How are the concerns about the validity of the instrument similar to concerns that arise from a difference-in-differences analysis? What are the advantages with working with aggregated cell means?

2) Produce a table that shows the first stage and reduced form for the full sample and the sample of mothers with more than 12 years of school. Why does the second sample provide a sensitivity check to the analysis?

3) Produce a table that shows the OLS and IV results for the dependent variables repeat, log mother's income, and log family income, with and without regional trends. (If you're working with cell means, you will have to aggregate up to include mother variation by higr2mom in order to run the OLS).

4) We can examine whether a discontinuous break in average schooling and the probability of repeating a grade occurs the year after a compulsory school law restriction. The dataset leadlag.dta contains indicator variables for whether an increase in the CL variable occurred or is about to occur. The variable m10, m9, m8, …, are indicator variables that a the CL variable will increase 10 years from the year a mother was 14, 9 years, 8 years, etc… The variables p1, p2, p3, … are indicator variables that the CL variable increased the year a mother was age 14, 1 year earlier, 2 years earlier, etc… The data include years and states only for 20 year ranges where one change occurred. Thus, if there was no law change in a 20 year period, the lead lag indicators were dropped.

Merge the leadlag data to your collapsed dataset. Plot or make a table of the lead lags using mother's schooling and the repeat variables as the dependent variables, including, at least, fixed effects for birth cohort and mother's state of birth. Try this just for the sample of mothers with at least 12 years of school. Why is this a good check on the validity of the instrument? Do you find these results convincing?

5) summarize your findings in Part I and Part II in one paragraph.