ECON 7130 - Microeconomics III

Spring 2016

Notes for Lecture #4

Today:

- Overview of Identification
- Structural vs. Atheoretic Approaches to Econometrics
- Overview of some common public data sources
- Introduction to Panel Estimators: Individual Fixed Effects

Indentification:

- Goal of econometric work is to identify *causal* effects
  - Need to be careful here - claims of causation imply an economic model (hence our focus on modeling)
  - Lots of assumptions of often implicit in identification strategy that help to produce causal inference
    - Statistical assumptions about conditional independence (exogeneity)
    - Functional form assumptions
    - But don't be afraid of assumptions - all scientists need to make them - key is to know what you are doing so you can justify
- Technical:
  - The econometric model satisfies the technical conditions insuring a unique global maximum for the statistical objective function
    - e.g., Rank and order conditions satisfied
    - Order condition: the number of excluded exogenous variables not less than the number of endogenous variables minus 1 (necessary condition)
    - Rank condition: at least many unique equations as endogenous variables (necessary and sufficient)
- Intuitive, How are parameters identified:
  - What are the key features of the data, or the key sources of (assumed) exogenous variation in the data, or the key a priori theoretical or statistical assumptions imposed in the estimation, that drive the quantitative values of the parameter estimates, and strongly influence the substantive conclusions drawn from the estimation exercise?
- Ways to get exogenous variation needed to identify causal effects:
  - Randomized, controlled experiment (the gold standard?)
    - Randomize treatment and control groups
    - Control all aspects of treatment
  - "Natural", natural experiments
    - Randomized outcomes as a result of nature, e.g., weather events, twin births, human cloning, date of birth, realizations of child gender
    - Rosenzweig and Wolpin (2000) identify 20 papers that use these 5 instruments to estimate "effects" of education and experience on earnings, "effects" of children on female labor supply, and elasticities of consumption with respect to permanent and transitory income shocks

    - Surely other, more recent examples with these five instruments
    - Randomize who gets treatment, but can't control all aspects of treatment
    - As random as you can get, outside a lab
  - Quasi-natural experiments
    - Using variation in policies to uncover treatment effects
    - Keane: "[quasi-experiment] has come to mean any policy intervention that affects two groups differently, regardless of whether those two groups look similar initially"
    - Keane gives examples of:
      1. Small firms used as a control group for large firms, in estimating the effect of changes in credit laws that only affect large firms' output
      2. Men used as a control group for women, in estimating the effect of changes in welfare rules (affecting only women) on interstate migration
      3. Upland counties in the deep South used as a control group for coastal counties, to estimate the impact of treatment for helminth infection, given only to residents in the latter areas, on child outcomes
      4. People under 65 used as a control group for people over 65 to estimate the effect of Medicare on mortality rates, etc.
    - More recent work has tried to emphasize the choosing of similar treatment and control groups through, e.g. regression discontinuity design, panel estimators

- While randomization seems ideal for identification, it's not a panacea
  - e.g., Angrist (*AER*, 1990), who used Vietnam era draft lottery numbers - which were randomly assigned but influenced the probability of "treatment" (i.e., military service) - as an instrument to estimate the effect of military service on subsequent earnings (draft number below some ceiling means potentially drafted - but have to go through physical etc.)
  - The estimates imply that military service reduced annual earnings for whites by about \$1500 to \$3000 in 1978\$ (with no effect for blacks), about a 15% decrease
  - Since random treatment, is causal effect of service identified correctly?
  - Heckman (*JHR*, 1997): when effects of service are heterogeneous in the population, the lottery number may not be a valid instrument, despite the fact that it is randomly assigned (this is known as *heterogeneous treatment effects*)
    - Note that people with high lottery numbers (less likely to be drafted) may still choose to join because they expect a positive return
    - In draft, also get those with negative returns
    - So two groups - one whose earnings benefit from service, another who don't
    - This means that the estimator used by Angrist will overstate the negative effect on the randomly chosen person (or population as a whole)
    - e.g., Suppose there are two types of people, both of whom would have subsequent earnings of \$100 if they do not serve in the military. Type 1's will have a 20% gain if they serve, and Type 2's will have a 20% loss.
    - Say Type 1's are 20% of the population, and Type 2's 80%.
    - So the average earnings loss for those who are drafted into service is -12% = (0.8*-20%)+(0.2*20%).
    - Now, let's say that 20% of the draft eligible group is actually drafted (while the Type 1's volunteer regardless).
    - Then, the Wald estimator gives: $\hat{\beta} = \frac{(\bar{y}^E - \bar{y}^N)}{(P^E - P^N)} = \frac{(100*(1+((0.8*0.2*-0.2)+(0.2*1*0.2))))-(100*(1+(0.8*0*-0.2)+(0.2*1*0.2)))}{((0.8*0.2)+(0.2*1))-((0.8*0)+(0.2*1))} = \frac{100.8-104}{.36-.2} = -20\%$
    - Effect is just that on the Type 2's who only enter if drafted - not the average effect for those in service

        - If volunteering not possible would get, $\hat{\beta} = \frac{(\bar{y}^E - \bar{y}^N)}{(P^E - P^N)} = \frac{(100*(1+((0.8*0.2*-0.2)+(0.2*0.2*0.2)))) - (100*(1+(0.8*0*-0.2)+(0.2*0*0.2)))}{((0.8*0.2)+(0.2*0.2)) - ((0.8*0)+(0.2*0))} = \frac{97.6-100}{.20-0} = -12\%$, the correct average effect of those who serve
    - Even if identified this effect for this treatment group, what does it mean?
    - Quantitative magnitude of the estimate cannot be interpreted without further structure - why the negative effect? PTSD? Missing education? Missing experience in private sector?

- Structuralist approach to identification a little different
    - Model of behavior implies certain economic outcomes
    - Changes in model parameters change these economic outcomes
    - Identify parameters by matching model outcomes to data
    - Necessary condition is that different model parameters have differential effects on outcomes

Structural vs. Atheoretic Approaches to Econometrics:

- I've linked to a number of papers that are critical of the "experimentalist" approach to econometrics.
- While we'll spend most of our time on cutting edge reduced form tools of this school (and I have made *Mostly Harmless Econometrics* a recommended text), I want you to be aware of some of the perhaps exaggerated claims of this school of thought
- We've talked about structural modeling before - where you explicitly model the optimization problem of agents
- Structural models rely on economic theory to build the models
- "Atheoretic" models claim to not rely so heavily on assumptions about economic theory
    - But theory implicit in choice of controls, functional forms
        - e.g. Donohue and Wolfers (*Stanford Law Review*, 2005) - 1965, death penalty abolished in Canada. Homicide rates in US for 1965-1971 increase from 5 to 8.5 per 100,000. In Canada, rates increase from 1.3 to 2.2 per 100,000. i.e., rates increase 70% in both countries over the time period. Diff in Diff implies no effect of death penalty. But if use levels rather than logs, get that abolishing the death penalty lowers murder rates by 2.6 per 100,000 ((8.5-3)-(2.2-1.3)=2.6)
        - Leamer: "Data alone cannot reveal the relationship between crop yield and fertilizer... we must resort to prior info."
        - This is not unique to economics (e.g. Wangensteen surgery experiments)
    - Further, to the extent you draw policy relevant conclusions, you need to be using theory
        - *Salesman*: Ma'am, this vacuum cleaner will cut your work in half. *Customer*: Terrific! Give me two!
- Moreover, both approaches rely heavily on statistical assumptions
    - e.g., distributions of errors, conditional independence, etc.
- Pischke says structural has lead to lack of consensus in IO and macro
    - Keane says that not true
    - Labor has lack of consensus - see how elasticities all over board - can't even agree on basic things like direction of effect of min wage or tax increase

  - Point to marketing where people largely agree - mostly due to structural approach and knowing mechanisms behind results
  - Says IO just suffers because don't use dynamics as much as should and therefore getting some weird results

- Beware of the Lucas Critique when dealing with reduced form analyses
  - If want to predict outcomes of policies, need policy invariant parameters
  - Lucas talking specifically about using macro aggregates, but this mostly generalizes
  - Suggests forecasting accuracy may be completely separate than ability to predict outcomes of policy
  - E.g., Observing that Fort Knox has never been robbed, is it safe to assume that if take guards away, no one will rob? But likely to correctly predict that won't be robbed if no policy change.

- What I think
  - Both approaches valuable
  - Reduced form better to uncovering relationships know nothing about
  - Structural more useful - policy experiments, welfare analyses
  - But structural approach generally takes a lot more effort:
    - Specify model
    - Solve model
    - Estimate model
  - Structural also may face publishing problems - mostly used in IO and macro - not as much in public/labor
  - Career concerns have had me doing more reduced form work, but hope to get back to structural

Data Sources:

- Health
  - Medical Expenditure Panel Survey (MEPS)
  - Survey of Income and Program Participation (SIPP)
  - Health and Retirement Study (HRS)

- Socio-Economic
  - Current Population Survey (CPS)
  - Panel Study of Income Dynamics (PSID)
  - Consumer Expenditure Survey (CEX)
  - American Community Survey (ACS)
  - General Social Survey (GSS)
  - Intergrated Public Use Microdata Series (IPUMS)
  - National Longitudinal Survey of Youth (NLSY)
  - American Time Use Survey (ATUS)
  - Survey of Consumer Finances (SCF)

- Financial/Firm

  - Computstat
  - Center for Research in Security Prices (CRSP)
  - Job and Labor Turnover Survey (JOLTS)
  - Census Survey of Small Business Owners (SBO)
  - Longitudinal Business Database (LBD)

- Crime
  - FBI Uniform Crime Reports

- Political
  - Federal Election Committee Campaign Finance Data
  - Congressional Biographical Database
  - Federal Assistance Awards Data System (FAADS)

- Administrative Data
  - Social Security Administration (SSA)
    - Work with SSA employee
    - Some without SSA employee, but limited access data (e.g. SSA supplement to the HRS)
  - Internal Revenue Service (IRS)
    - Work with IRS or Treasury employee
    - Become a contractor with IRS (IRS Research has call for proposals every 1-2 years, usually in October) - very difficult, must work in secure datacenter
    - Use public use files (cost $4,000 per year for recent data, Univ. of Michigan has 1979-1990 public use panel)
  - Census (including BEA)
    - Work with Census employee
    - Become a contractor with Census (easier than with IRS, but still not easy, must work in secure data center)

- Webpages as data sources:
  - Lots of information out there: prices of goods/services, financial data, sports, news/trends
  - Can "scrape" webpages to acquire data using tools in languages such as R and Python.

- Census data: http://www.census.gov/ces/dataproducts/economicdata.html

- NBER data: http://www.nber.org/data/

- *AER* (and some other journals) require release of data with publication (see webpage for previous issues - links to download data)

- Also, let me mention that it's worth spending time to identify the most appropriate data source. I feel like every paper I saw in Fall of 2012 used the NLSY - even if not suitable, didn't utilize panel, etc.

Individual Fixed Effects:

- Panel Data: NLSY, PSID, SIPP, Compustat, Administrative Data (IRS, SSA, BEA), LBD,...

- Main idea of using individual fixed effects models:

  - Control for time-invariant, unobserved heterogeneity
  - Identify causal effects from changes in treatment *within* an individual/group

- Example (from Angrist and Pischke): Estimating earnings as a function of union status and other observables

- Want to estimate: $Y_{it} = \alpha + \lambda_t + X_{it}^{'}\beta + \rho D_{it} + \varepsilon_{it}$
  - Problem: union status may be biased because it's correlated with an omitted variable that is also correlated with earnings
  - e.g., If those with higher ability join unions at higher rates but are also more productive (conditional on everything else)

- Observe either union or non-union earnings for each individual in each year: Y_{1it} or Y_{0it}

- Let A_i be unobserved ability

- Assume that $E[Y_{0it}|A_i, X_{it}, t, D_{it}] = E[Y_{0it}|A_i, X_{it}, t]$
  - That is, conditional on ability and other covariates, union status is as good as randomly assigned

- We're also going to want to make two further assumptions:
  1. A_i appears without a time subscript in the linear model: $E[Y_{0it}|A_i, X_{it}, t] = \alpha + \lambda_t + A_i\gamma + X_{it}^{'}\beta$ ** This is key**
  2. That the causal effect of union membership is additive and constant across individuals and time: $E[Y_{1it}|A_i, X_{it}, t] = E[Y_{0it}|A_i, X_{it}, t] + \rho$

- This implies that $E[Y_{it}|A_i, X_{it}, t, D_{it}] = \alpha + \lambda_t + A_i\gamma + X_{it}^{'}\beta + \rho D_{it}$

- Which implies the fixed effects regression equation: $Y_{it} = \alpha_i + \lambda_t + X_{it}^{'}\beta + \rho D_{it} + \varepsilon_{it}$
  - Where, $\varepsilon_{it} \equiv Y_{0it} - E[Y_{0it}|A_i, X_{it}, t]$
  - and, $\alpha_i \equiv \alpha + A_i\gamma$
  - Note that α_i and λ_t (the individual fixed effects and the time effect) are parameters to be estimated - dummies for individuals or year

- Now can get unbiased, causal effect of union status on earnings by estimating FE model

- Estimated FE model is equivalent to estimating model of deviations from mean
  - To show:
  - Calculate individual averages: $\bar{Y}_i = \alpha_i + \bar{\lambda} + \rho\bar{D}_i + \bar{X}_i^{'}\beta + \bar{\varepsilon}_i$
  - Then, if subtract these means from FE regression above, get: $Y_{it} - \bar{Y}_i = \lambda_t - \bar{\lambda} + \rho(D_{it} - \bar{D}_i) + (X_{it} - \bar{X}_i)^{'}\beta + (\varepsilon_{it} - \bar{\varepsilon}_i)$
  - Note how the individual fixed effect is differenced out since it picks up the mean earnings for that person, conditional on all covariates

- Relation between FEs and Differencing data:
  - Differencing estimator: $\Delta Y_{it} = \Delta\lambda_t + \rho\Delta D_{it} + \Delta X_{it}^{'}\beta + \Delta\varepsilon_{it}$
  - Note that if just have two periods, differencing is exactly the same as using deviations from means
  - For more than two periods, it is not, but it is a way to get rid of the unobserved heterogeneity
  - Potential problem with more than two periods is that error terms become serially correlated, so need to adjust for (e.g., use robust std errors)
  - Don't have to use first differences - can use longer lags.

        - ∗ Some show that longer lags have less problems with measurement error.

- Potential problems with FE:
  - Attenuation bias
    - ∗ FE estimates very susceptible to noise from measurement error
    - ∗ e.g., in union example, union status relatively stable, so if use FE estimator, where rely on changes in union status for each individual, misreported union status can create a lot of noise
    - ∗ Will bias estimates towards zero
    - ∗ Cure: Instrumental Variables or external validation to reduce measurement error
  - Remove "too much" variation
    - ∗ It's not just the variation from the omitted variable that fixed effects remove
    - ∗ You now lose all cross-sectional variation in control variables
    - ∗ Result:
      - · More difficult to identify parameters precisely/find exogenous variation
      - · Potential for larger bias from further omitted variables (e.g. time varying ability)
    - ∗ Cure: Instrumental Variables
- Stata:
  - Use `xtreg` (Stata calculates the estimates using deviations from means) OR
  - Use `reg` with dummy variables for individual fixed effects

Creating Pseudo Panels from Pooled Cross-Sections:

- Seminal paper: Deaton (*Journal of Econometrics*, 1985)
- Identifying causal effects often relies on controlling for unobserved heterogeneity (e.g., fixed effects), many economic questions have a dynamic component
- Problem: Data researchers want is not always available as a panel (e.g., because panels are costly to collect)
- There are some very good repeated cross-sections: Current Population Survey in the U.S.A., and the Family Expenditure Survey in the United Kingdom
- Idea: Create a "synthetic panel" or "psuedo-panel" out of the series of cross sections
- Basic idea: group individuals in each cross section together based on like characteristics, then estimate the model treating each of these groups ("cohorts") like an individual observed over time
- Does this work?
  - Think of characteristics that define cohorts as instruments:
    - ∗ Need these characteristics to be correlated with control variables
    - ∗ Need these characteristics to be uncorrelated with unobservables in the model (e.g., ability)
  - Need to observe these characteristics for everyone (don't want selection problems)
  - These characteristics must be fixed over time
  - Requires the assumption that unobservable individual effects are drawn from the same population distribution across periods before and after the treatment. Otherwise there is the possibility of compositional bias.
    - ∗ Essentially, assume the true causal relationship is stable over time.

- This assumption is not required for panel analysis because you do observe the same individuals over time

- How it works in practice:
  - Variables to define cohorts: birth year, birth region, gender, parental characteristics
  - If fixed effect model - use cohort fixed effect (ends up being like the mean fixed effect for those in cohort)
  - If dynamic model - can use differencing, etc. like normal panel

- Example: Bank, Blundell, and Tanner (*AER*, 1998)
  - Question: What happens to the standard of living of the elderly? Or Is consumption behavior consistent with the Permanent Income/Lifetime Consumption Hypothesis?
  - There has been observed a large decline in consumption upon retirement. Termed the Retirement Consumption Puzzle, because inconsistent with PIH
  - So need to look at changes in a person's consumption over time
  - Problem: There do not exist panel data on a comprehensive measure of consumption.
  - Solution: Use Family Expenditure Survey (UK), which is a time series of cross-sections, and create a synthetic panel
  - Model:
    - Lifetime utility $= U = \sum_{t=1}^{T}\left[\frac{1}{1+(\delta_0+\delta' z_{1it})}\right] u(c_{it})$
    - Per period utility $= u(c_{it}) = \frac{exp(\theta_0+\theta' z_{2it})}{1-(\rho_0+\rho' z_{3it})} c_{it}^{(1-(\rho_0+\rho' z_{3it}))}$
      - Where,
      - $\delta_0 + \delta' z_{1it}$ = subjective rate of time preference
      - $\rho_0 + \rho' z_{3it}$ = coeff of relative risk aversion
      - $exp(\theta_0 + \theta' z_{2it})$ = heterogeneity of within period utility function
    - Euler condition from optimization of HH's consumption/savings problem ⇒

$$\Delta ln(c_{it}) = \alpha_1' \Delta(z_{3it} * ln(c_{it})) + \alpha_2' \Delta z_{2it} + \alpha_3' z_{1it} + \alpha_4 r_t + \alpha_5 + \varepsilon_t$$

      - Where,
      - $\alpha_1 = \frac{\rho}{\rho_0}$
      - $\alpha_2 = \frac{-\theta}{\rho_0}$
      - $\alpha_3 = \frac{-\delta}{\rho_0}$
      - $\alpha_4 = \frac{-1}{\rho_0}$ = - the coeff or relative risk aversion for the baseline
      - α_5 = adjustment for log-linear approximation error and δ_0
    - z's are demographic controls
    - Obviously, need data with dynamics, hence constructing cohorts for synthetic panel
    - Create synthetic panel with cohorts based on 4-year date of birth bands
    - Instrument for endogeneity of multiple adults and real interest rate variables
      - Use lags of key control variables
      - The instrument set contains age of head of household, per capita real GDP at age 20, a lag of the inflation rate, two-period lags of the interest rate, consumption growth, income growth, proportion of households with children, and change in mortality rate, and two- and three-period lags of proportion of households with multiple adults, proportion of heads and proportion of second adults unemployed, and proportion of heads and proportion of second adults retired
    - Findings: Much of drop has to do with complementarities between consumption and labor supply, but still model can't generate full drop in consumption. Likely candidate is that people overestimate pension/retirement savings and only realize when retire