DRE 7006
Panel data / Microeconometrics, Lecture 1

Christian Brinch

BI, January 2023

Lecture outline

- Course presentation and plan:
  - Modern applied micro and the "credibility revolution".
  - Potential outcomes, heterogeneous effects and the interpretation of estimates.
  - Large-N panel data and similarity to cross-sections.
  - Course outline and STATA.
- Ordinary least squares and statistical inference
  - Derivation of and consistency of OLS.
  - Asymptotic inference for OLS.
  - Heteroskedasticity-robust inference.
  - Testing and confidence intervals.
  - Bootstrap and bootstrap standard errors in regression.

"Let's take the con out of econometrics"

- You know a great deal about how to do statistical inference.
- The trouble is that we traditionally have observational rather than experimental data: statistical inference is inherently suited for experimental data.
- In practice, we need to assume more than we want to get to any conclusions (whimsical assumptions?)
- How does the credibility (fragility, robustness) of conclusions depend on the credibility of assumptions?
- Leamer (1983, AER) suggested: "Let's take the con out of econometrics":
  - "Hardly anyone takes data analyses seriously."
- But a lot has happened since 1983...

Analysis of experimental versus observational data

- Lalonde (1986) compares an early RCT in economics (the effect of a labor market program) to other econometric evaluation methods.
- Regression methods, methods based on selection on observables.
- Models involving selection models, but selection models without credible exclusion restrictions.
- In brief, Lalonde (1986) found that we were not able to reconstruct the results from the RCT using the best practice econometric techniques of the day.

The credibility revolution in empirical economics

- The credibility revolution is described by Angrist and Pischke (2010, JEP) - origins are much older.
- Emphasis is on experiments - either actual experiments, thought-experiments or natural experiments.
- Design-based approach - either by explicit design or because "nature" or policies unintendedly came up with a good design.
  - by design or by chance - we want to know why we should think we have exogenous variation.
  - turned the burden of proof - it is not sufficient that I am not sure why my variation in "x" is exogenous.
- Quasi-experimental and experimental approaches - but many of the same methods as "before".
- Extreme focus on robustness to model specification - reliance on whimsical assumptions "forbidden".

Randomized controlled trials in economics

- The history of RCTs in economics goes back a long way:
  - RAND health insurance experiment.
- RCTs obviously has a role to play in the evaluation of programs.
- Social programs: Progresa.
- Labor market programs.
- An explosion in the use of RCTs in development economics. With the 2019 Nobel laureates.
- So we have experimental data. Question is more: Is there more to economics than "program evaluation"?

Why are RCTs the "Gold Standard"?

- The potential outcomes framework (more or less the same way as we think about causality in economic theory).
  - Y_1 Observed outcome if individual is treated ($D = 1$).
  - Y_0 Observed outcome if individual is not treated ($D = 0$).
  - Think about wages with and without labor market program.
- A comparison of means with and without labor market programs estimates:

$$E(Y|D=1) - E(Y|D=0) = E(Y_1|D=1) - E(Y_0|D=0)$$

$$= \left[E(Y_1|D=1) - E(Y_0|D=1)\right] + \left[E(Y_0|D=1) - E(Y_0|D=0)\right]$$

$$= \text{causal effect} + \text{selection effect}$$

- Randomization ensures that selection effect disappears:

$$E(Y_1|D) = E(Y_1)$$

Panel data? Well, large-N panel data

- The techniques taught in this course apply to cross-sectional or panel data.
- Panel data is combination of cross-section and time-series, right.
- Very little time series in this course. The reason is that we can often get what we want from cross-sectional variation.
- Textbooks: Angrist and Pischke: "Mostly harmless econometrics", Wooldridge: "Econometric analysis of cross-sectional and panel data"

How can you do panel data without doing time series modelling?

- You have time series data $(Y_{1,1}, X_{1,1}) \ldots, (Y_{T,1}, X_{T,1})$
- You are interested in the relationship (regression) between Y and X.
- OLS is consistent under a wide variety of autocorrelation structures.
- Ordinary OLS standard errors are not valid - you somehow need to model the autocorrelation structure to get valid standard errors.
- With panel data: If you also have$(Y_{1,2}, X_{1,2}) \ldots, (Y_{T,2}, X_{T,2}), \ldots, (Y_{1,N}, X_{1,N}) \ldots, (Y_{T,N}, X_{T,N})$ then you can conceptually do the OLS N times and assess the uncertainty based on the cross-sectional variation in regression coefficients.
- Operationally, we do this using "clustered standard errors".

Estimating models and interpreting the results more broadly - an example

- This will be a recurring topic.
- As alluded earlier, we can estimate treatment effects by comparing population means in treated and untreated samples.
- We can also estimate treatment effects in a linear regression model and get the same result, having treatment D as the right hand side variable.
  - A regression with just a dummy variable is an example of a "fully saturated" regression model, estimated parameters simply reflect the conditional averages.
- But what if treatment effects differ, such that $Y_1 - Y_0$ varies between individuals?
  - A violation of the literal regression model. $Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$.
  - Still $\hat{\beta}$ will estimate an average treatment effect because: $E(Y_1 - Y_0) = E(Y_1) - E(Y_0) = E(Y|D = 1) - E(Y|D = 0)$ under randomization.

Wrapping up this intro

- We need to learn the techniques and craft involved in the "new" applied micro economics/econometrics.
- A bit of this is the same as the old techniques. A bit is craftmanship - what do we expect to see when someone presents a diff-in-diff analysis?
- We will spend quite a bit of time on technical stuff. And then a bit of time on topics that are important to applied micro. (The furious five?)
- In this course we will be interested in causal effects (as they are defined in the potential outcomes framework).
- Although the models usually suggest that we are interested in what happens to y conditional on x, both variables are typically treated as random.

Linear regression model

- Model of the distribution of the random variable y, conditional on x (index on unit number is suppressed, this is generic)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u,$$

or

$$y = x\beta + u,$$

where $x = [\ 1 \quad x_1 \quad x_2 \quad \cdots \quad x_K\ ]$ and $\beta = [\ \beta_0 \quad \beta_1 \quad \cdots \quad \beta_K\ ]'$

- β are the parameters of the model
- u is a random variable
- We intend to have N draws from this distribution. When N is fixed, it is sometimes useful to stack variables in matrix form.

$$Y = X\beta + U$$

Derivation of OLS

- Assumptions

$$\mathrm{E}(x'u) = 0$$

$$rank\mathrm{E}(x'x) = K$$

1. Premultiply $y = x\beta + u$ with x' : $x'y = x'x\beta + x'u$
2. Take expectations $E(x'y) = E(x'x\beta) + E(x'u)$ which gives $E(x'y) = E(x'x)\beta$
3. Solve for β: $\beta = [E(x'x)]^{-1}E(x'y)$

- Identified (not definition, but sufficient condition): can be written in terms of population moments of observable variables.
- Estimation by analogy principle, method of moments: Estimate β by replacing population moments with sample moments:

$$\hat{\beta} = \left(N^{-1} \sum_{i=1}^{N} x_i' x_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} x_i' y_i \right) = (X'X)^{-1}(X'y)$$

Convergence in probability

- A sequence of random variables $\{x_N : N = 1, 2, \ldots\}$ converges in probability to the constant a if for all $\varepsilon > 0$, $P(|x_N - a| > \varepsilon) \to 0$ and $N \to \infty$. We write $\text{plim} x_n = a$ and $x_N \xrightarrow{p} a$.
- If $a = 0$ we say that $\{x_N\}$ is $o_p(1)$.
- A sequence of random variables $\{x_N : N = 1, 2, \ldots\}$ is bounded in probability if and only if for every $\varepsilon > 0$, there exists a constant b_ε and an integer N_ε such that

$$P(|x_N| \geq b_\varepsilon) < \varepsilon$$

  for all $N \geq N_\varepsilon$. We write $x_N = O_p(1)$.
- If $x_N \xrightarrow{p} a$ then x_N is $O_p(1)$

Some rules (simplified)

- $o_p(1) + o_p(1) = o_p(1)$
- $O_p(1) + O_p(1) = O_p(1)$
- $O_p(1)O_p(1) = O_p(1)$
- $o_p(1)O_p(1) = o_p(1)$
- Slutsky's theorem: $\text{plim} g(x_N) = g(\text{plim} x_N)$ if g is continuous at $\text{plim} x_N$.
  - Includes $\text{plim} \frac{1}{x_N} = \frac{1}{\text{plim} x_N}$ when $P(x_N = 0) \to 0$.

Consistency of OLS

- An estimator $\hat{\theta}$ for the parameter θ is consistent if $\hat{\theta} \xrightarrow{p} \theta$.
- Under the assumption that $y = x\beta + u$,

$$\begin{aligned}
\hat{\beta} &= \left(N^{-1}\sum_{i=1}^{N} x_i' x_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} x_i' x_i \beta\right) + \left(N^{-1}\sum_{i=1}^{N} x_i' x_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} x_i' u_i\right) \\
&= \beta + \left(N^{-1}\sum_{i=1}^{N} x_i' x_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} x_i' u_i\right)
\end{aligned}$$

- $\left(N^{-1}\sum_{i=1}^{N} x_i' x_i\right)^{-1}$ approaches $E(x'x)^{-1}$ in probability (by Slutsky's theorem).
- $\left(N^{-1}\sum_{i=1}^{N} x_i' u\right)$ is $o_p(1)$.
- OLS is consistent: $\hat{\beta} = \beta + o_p(1)$

Convergence in distribution

- A sequence of random variables $\{x_N : N = 1, 2, \ldots\}$ with distribution functions F_N converges in distribution to the continuous random variable x with distribution function F if and only if

$$F_N(\xi) \to F(\xi)$$

  as $N \to \infty$ for all $\xi \in \mathbb{R}$.
- Continuous mapping theorem: $x_N \xrightarrow{d} x$ and g continuous implies: $g(x_N) \xrightarrow{d} g(x)$.
- $z_N \xrightarrow{d} \text{Normal}(0, V) \implies A' z_N \xrightarrow{d} \text{Normal}(0, A' V A)$
- $z_N \xrightarrow{d} \text{Normal}(0, V) \implies z_N V^{-1} z_N \xrightarrow{d} \chi^2_K$
- Convergence in distribution $\implies O_p(1)$

Central limit theorem

- Family of theorems: This one is useful, but the conditions are not necessary.
- Let $\{w_i : i = 1, 2, \ldots\}$ be a sequence of independent, identically distributed $Gx1$ random vectors such that $E(w_{ig}^2) < \infty$, $g = 1, \ldots, G$ and $E(w_i) = 0$. Then

$$N^{-1/2} \sum_{i=1}^{N} w_i \xrightarrow{d} Normal(0, B)$$

- B is the variance (matrix) of w_i, that is $E(w_i' w_i)$

Distribution of $\sqrt{N}(\hat{\beta}-\beta)$

- $$\begin{aligned}\sqrt{N}(\hat{\beta}-\beta) &= \left(N^{-1}\sum_{i=1}^{N}x_i'x_i\right)^{-1}\left(N^{-1/2}\sum_{i=1}^{N}x_i'u_i\right)\\ &= \left(A^{-1}+o_p(1)\right)\left(N^{-1/2}\sum_{i=1}^{N}x_i'u_i\right)\\ &= A^{-1}\left(N^{-1/2}\sum_{i=1}^{N}x_i'u_i\right)+o_p(1)\end{aligned}$$
- Applying the central limit theorem
$$\left(N^{-1/2}\sum_{i=1}^{N}x_i'u_i\right)\xrightarrow{d} Normal(0,B),$$
where B is $E(u_i^2x_i'x_i)$.
- Thus
$$\sqrt{N}(\hat{\beta}-\beta)\xrightarrow{d} Normal(0,A^{-1}BA^{-1})$$
- Where $A=E(x'x)$.

Asymptotic variance of $\hat{\beta}$ with homoskedasticity

- When $\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} Normal(0, A^{-1}BA^{-1})$ we say that $\hat{\beta}$ has asymptotic variance $\frac{1}{N}A^{-1}BA^{-1}$.
- Assumption: Homoskedasticity, - the variance of u is independent of x: $B = E(u^2x'x) = \sigma^2 E(x'x)$
- Then $B = \sigma^2 A$ and the variance matrix $A^{-1}BA^{-1} = \sigma^2 A^{-1}$ (Obviously: A cancels out).
- We estimate σ^2 by the sample variance of the residuals and A by the corresponding sample moment and arrive at the estimated asymptotic variance

$$s^2(X'X)^{-1}$$

- The standard errors are the square roots of the diagonals of this matrix.

Heteroskedasticity-robust variance estimator:

- We estimate B by the sample moment

$$\frac{1}{N}\sum \hat{u}_i^2 x_i' x_i$$

- And the asymptotic variance matrix of $\sqrt{N}(\hat{\beta}-\beta)$ becomes (when we also replace the A with the sample moments)

$$(\frac{1}{N}X'X)^{-1}\left(\frac{1}{N}\sum \hat{u}_i^2 x_i' x_i\right)(\frac{1}{N}X'X)^{-1} = N(X'X)^{-1}\left(\sum \hat{u}_i^2 x_i' x_i\right)(X'X)^{-1}$$

  and the asymptotic variance of $\hat{\beta}$is estimated as

$$(X'X)^{-1}\left(\sum \hat{u}_i^2 x_i' x_i\right)(X'X)^{-1}$$

- It seems dubious to apply the homoskedasticity assumption. It buys us little.
- Note that we only have an estimate of the variance matrix and standard errors. We do not know the true variance matrix.

Instructive comparison of robust and nonrobust formulas:

- These are versions of the formulas for the simple regression model:
- Non-robust:

$$n^{-1}\frac{\sum_i \hat{u}_i^2}{\sum_i (x_i-\bar{x})^2}$$

- Robust:

$$n^{-1}\frac{\sum_i \hat{u}_i^2 \frac{(x_i-\bar{x})^2}{\sum_j (x_j-\bar{x})^2}}{\sum_i (x_i-\bar{x})^2}$$

Inference: Tests

- There are three very useful ways of testing a restriction, going beyond OLS estimation, this is really ML stuff:
  - Estimating the unrestricted model and performing a Wald test.
  - Estimating only the restricted model and analysing the relationship between residuals from the estimated model and additional variables.
  - Estimating both restricted and unrestricted models and comparing the goodness-of-fit: R-squared in this context. (Hard to do with robust-style standard errors.)

Wald test

- We want to test the null hypothesis: $\beta = \beta_0$
- Under the null hypothesis:

$$\chi = \left(\hat{\beta} - \beta_0\right)' V^{-1} \left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \chi^2_K,$$

  where V is the estimated asymptotic variance of $\hat{\beta}$
- So a natural test is to reject the null hypothesis if we get something in the tail area of the Chi-square - distribution. This is the Wald test.
  - More precisely: Reject H_0 at significance level α if the test statistic $\chi > F_\chi^{-1}(1-\alpha)$ or the p-value of the test is given by $p = 1 - F_\chi(\chi)$where F is the CDF of the Chi-square distribution (with the appropriate degrees of freedom).
- The test can directly be applied to a subset of the parameters in β - this is just a question of picking out the corresponding terms of the variance matrix.
- The standard t-test is the special case where we test one parameter. (We here use the squared t-stat.)

Testing linear restrictions with Wald test

- A set of linear restrictions can be specified as $C\beta = a$. C is here a matrix and a i a vector. The number of rows in C (and a) is the number of restrictions you want to test. This should obviously not be larger than the dimension of β - and the restrictions would more generally have to be consistent (and not linear dependent).
- Still: If $C\beta = a$, then $C\hat{\beta} - a$ is $o_p(1)$, $\sqrt{N}\left(C\hat{\beta} - a\right)$ is asymptotically normal, and $C\hat{\beta} - a$ has asymptotic variance equal to $CAvar(\hat{\beta})C'$ and we can just apply the Wald test to this expression.

Wald style confidence interval/region

- The following is a pivot (some object that has a known distribution that does not depend on the unknown parameter β):

$$\chi = \left(\hat{\beta} - \beta\right)' V^{-1} \left(\hat{\beta} - \beta\right) \xrightarrow{d} \chi^2_K.$$

- A $1-\alpha$ percent confidence interval can then be constructed by picking all values of β such that $\chi < F_\chi^{-1}(1-\alpha)$.
- This reduces to the usual $\hat{\beta} \pm 1.96 s.e.(\hat{\beta})$ confidence interval in the univariate case with $\alpha = 0.05$.

Parametric bootstrap

- Assume that we know the model. For our purposes we know β and we know the joint distribution of u and x.
- We can then simulate data from the model, and we can compute $\hat{\beta}$. If we do this repeatedly, we can assess the sampling variance of $\hat{\beta}$.
  - The standard deviations in parameters from such a procedure correspond to the standard errors that we produced above using asymptotic theory.
- A parametric bootstrap is the corresponding exercise where we use the estimated $\hat{\beta}$ from the data as a starting point.
- Similarly, if we were interested in assessing the sampling variance of the sample mean or the sample median in a univariate analysis, and we knew the distribution function, we could simply simulate from the distribution function and assess the variability.
- This is not extremely useful in our context. We rarely know the model. It is useful as a thought experiment.

The bootstrap. Nonparametric.

- Stay univariate for the moment. Suppose we do not know the true CDF. But we do know the empirical distribution function. With a large sample, the empirical distribution function will closely estimate the true CDF.
- So what if we try to sample from the empirical distribution function in place of the true distribution function? With large data sets we get approximately the same results!
- What does it mean to sample from the empirical distribution function? Sampling with replacement!
  - Assume that you have 6 observations: [2,5,1,10,3,9]. You want to assess the variability of the sample mean:
  - You roll a die 6 times and get (1,4,1,6,5,1) - so you pick the data set [2,10,2,9,3,2] - and compute the mean 4,67.
  - You roll a die 6 times again and get (2,2,5,4,3,2) - so you pick the data set [4,4,5,6,1,2] - and compute the mean 3,67.
  - And then you do this many more times and compute the standard deviation in your generated data of sample means.

Nonparametric bootstrap in regression analysis.

- We draw observations based on sampling with replacement, just as for univariate data. One observation is one "line" in the data set used for our regression analysis.
- We repeatedly draw new datasets of the same size as our original data set and run the regression analysis - keeping our result $\hat{\beta}$. Finally, we compute the variance matrix of these results.
- The variance matrix from this exercise estimates the same object as the variance matrix from asymptotic theory.
- It is interesting to cross-validate the standard errors.
- Note that there is also another way to bootstrap regressions - based on the empirical distribution of residuals.