Ec331: Research in Applied Economics
Spring term, 2022

Wk.2 class: The practise of Econometric analysis

Term Overview

Wk.2 General Econometric Concerns

Wk.3 Discrete Choice, Ordered Choice (optional)

Work in Progress - Presentations (10%, 8 slides, 15 mins max)

Wk.4 Ethan, Patrick (plus if demand, discussion on Panel data issues)

Wk.5 Andy, Freddie, Isaac

Wk.6 Milana, Anna, Alex

Wk.7 Vasily, James

Wks. 8-10 Analysis development

Resources

Stata search function and pdf documentation

Ec331 Moodle page, Helpdesk

Textbooks (examples):

Cameron and Trivedi (2009) "Microeconometrics using Stata"

Baum (2006) “An Introduction to Modern Econometrics”

Mitchell (2012) “Interpreting and Visualising Regression Models Using Stata”

Mitchell (2012) “A Visual Guide to Stata Graphics”

Long and Freese (2005) “Regression Models for Categorical Dependent Variables”

Greene and Hensher (2010) "Modeling Ordered Choices"

Wider frames

In Economics:

Caplin & Schotter (2008) "The Foundations of Positive and Normative Economics"

Gintis (2009) "The Bounds of Reason: Game Theory and the Unification of the Behavioural Sciences"

The historical:

Diamond and Robinson (2010) "Natural Experiments of History"

The personal:

Kahneman (2011) "Thinking, fast and slow"

The normative:

Sen (2009) "The Idea of Justice"

The Practise of Econometric Analysis

Today:

(1) Introduction

(2) The experimental ideal and the selection problem

(3) Regression fundamentals

Appendix: Cameron and Trivedi (2009) - overview

For detail on (1) - (3), see: Angrist and Pischke (2009)

"Mostly Harmless Econometrics: an empiricist's companion"

(1) Introduction

"The modern menu of econometric methods can seem confusing, even to an experienced number cruncher. Luckily, not everything on the menu is equally valuable or important. Some of the more exotic items are needlessly complex and may even be harmful. On the plus side, the core methods of applied econometrics remain largely unchanged, while the interpretation of basic tools has become more nuanced and sophisticated."

Angrist and Pischke (2009)

Introduction

Underlying beliefs:

(1) That research is "most valuable when it uses data to answer specific causal questions, *as if* in a randomized clinical trial... In the absence of a real experiment, we look for well-controlled comparisons and/or natural quasi-experiments."

(2) The principle that the estimators "in common use almost always have a simple interpretation that is not heavily model dependent", implying a "conceptual robustness of basic econometric tools."

> e.g. linear regression providing useful information about the conditional mean function, regardless of shape

Angrist and Pischke: 4 FAQs

(1) What is the causal relationship of interest?

- the purely descriptive has a role; but causality facilitates prediction

(2) What experiment might ideally capture the causal effect of interest?

- helps formulate causal question precisely, highlighting concerns for manipulation and those hopefully held constant

(3) What is your identification strategy?

- the means by which observational data is used to approx. an experiment

(4) What is your mode of statistical inference?

- describing the population, the sample, and the assumptions in constructing standard errors

(2) The Experimental Ideal: The Selection Problem

Treatment: $D_i = \{0,1\}$

Observed outcome: $Y_i = \begin{cases} Y_{1i} \text{ if } D_i = 1 \\ Y_{0i} \text{ if } D_i = 0 \end{cases}$

$$= Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$$

$$= E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1]$$

$$+ E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]$$

Random Assignment Solves the Selection Problem

$$
\begin{aligned}
&E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \\
&\quad = E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1] \\
&\qquad + E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0] \\
&\quad = E[Y_{1i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 1] \\
&\quad = E[Y_{1i} - Y_{0i} \mid D_i = 1] \\
&\quad = E[Y_{1i} - Y_{0i}]
\end{aligned}
$$

"Benchmark", but problem: logistical and ethical issues in implementation of randomized trials for policy evaluation

- need more readily available sources of variation, if possible natural or quasi-experiments that mimic a randomized trial

Regression Analysis of Experiments

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

$$*ass.\ Y_{1i} - Y_{0i} = \rho\ *$$

$$= \alpha + \rho D_i + \varepsilon_i$$

so

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$$

$$= \rho + E[\varepsilon_i \mid D_i = 1] - E[\varepsilon_i \mid D_i = 0]$$

Since the selection bias term disappears under random assignment, a regression of observed outcomes on a treatment dummy estimates causal effect. The further inclusion of controls that are uncorrelated with the treatment reduces the residual variance, in turn lowering the standard error of the regression estimates.

Natural Experiments in History

Some inspiration:

Diamond and Robinson (2010) "Natural Experiments in History"

Ch1 - P.V. Kirch, "Controlled Comparison and Polynesian Cultural Evolution"

Ch3 - S. Haber, "Politics, Banking and Economic Development: Evidence from New World Economies"

Ch6 - A. Banerjee and L. Iyer, "Colonial Land Tenure, Electoral Competition and Public Goods in India"

Ch7 - A. Acemoglu, D. Cantoni, S Johnson and J.A. Robinson, "From Ancien Regime to Capitalism: the Spread of the French Revolution as a Natural Experiment"

(3) Regression Fundamentals

Even before considering causality, a (treatment) variable may predict the dependent variable in a narrow statistical sense, summarised by the conditional expectation function.

CEF for discrete variables:

$$E[Y_i \mid X_i = x] = \sum_t tP(Y_i = t \mid X_i = x)$$

CEF for continuous variables:

$$E[Y_i \mid \mathrm{X}_i = x] = \int_{-\infty}^{\infty} y \mathrm{f}_{Y|X}(\mathrm{y} \mid \mathrm{x})\,\mathrm{dy}$$

Regression Anatomy

Population regression coefficient vector defined as soln. to least squares problem:

$$\beta = \arg\min E[(Y_i - X_i'b)^2]$$

Slope coefficient for *k*th regressor:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all other covariates (i.e. ***beta k*** is the the bivariate slope coefficient after partialling out)

Linear regression and the CEF

Regression estimates provide a valuable baseline for most empirical research because regression is tightly linked to the CEF, a natural summary of empirical relationships.

Three reasons why vector of population regression coefficients is of interest:

(1) The Linear CEF Theorem: If the CEF is linear, the population regression function is it (limited empirical relevance, special cases only)

(2) The Best Linear Predictor Theorem: The function $X_i'\beta$ is the best linear predictor of Y given X, in a MMSE sense.

(3) The Regression CEF Theorem: The function $X_i'\beta$ provides the MMSE linear approximation to $E[Y_i \mid X_i]$

Note: (3) is Angrist and Pischke's favourite way to motivate regression, to extent distribution of Y focus of interest (rather than individual prediction).

Further concerns

Asymptotic OLS inference

- including use of heteroskedasticity-consistent standard errors
- valid inference under weak assumptions (relying on large-sample theory)
- 'traditional' inference requires stronger assumptions (but applicable to a sample of any size)

Regression and Causality - "A regression is causal when the CEF it approximates is causal... The CEF is causal when it describes differences in average potential outcomes for a fixed reference population."

Causal interpretation is justified under the conditional independence assumption (CIA), which asserts that conditional on observed characteristics, X, selection bias disappears.

$$Y_{si} \perp s_i \mid X_i, \text{ for all } s$$

Omitted Variables Bias

Consequence of omitting A variables from regression:

$$if \quad Y_i = \alpha + \rho X_i + A_i'\gamma + e_i \quad but\ omit\ A_i,$$

$$\frac{Cov(Y_i, X_i)}{V(X_i)} = \rho + \gamma\delta_{AX}$$

where δ_{AX} is a vector of coefficients from a regression of A on X, and γ is the coefficient of A in the "long" regression model for Y.

If claiming an absence of omitted variables bias, suggesting that regression is one you want i.e. you're prepared to rely on the CIA for a causal interpretation.

Issue: when is the CIA a plausible basis for empirical work? Best case scenario, random assignment of treatment, conditional on X. Otherwise, need to consider process that determines the treatment; is selection only based on measurables?

Controls and Pragmatic Concerns

Controls can increase the likelihood that regression estimates have causal interpretation. Good controls are those that are fixed when the treatment is determined. Bad controls are themselves outcome variables, leading to a version of selection bias.

Some final Pragmatic Concerns:

(i) Be careful, systematic, and document progress.

(ii) Graphical visualisation always v. useful: histograms, scatter-plots, line graphs over time (e.g. for different groups), graphs of marginal effects etc.

(iii) Key output is tables of coefficients, with different specifications side by side for ease of comparability. Print these off as soon as possible, and discuss widely!

(iv) Keep an eye on missing observations / sample size, and sensitivity to outliers.

(v) Analysis <u>not</u> just about getting a specific result, but understanding its context: sensitivity to changes in specification and careful consideration of assumptions underlying results, and associated qualifications to inference (esp. re: possible bias).

Appendix: Overview of Cameron and Trivedi (2009)

Cameron and Trivedi (2009) "Microeconometrics Using Stata"

1) Stata basics
2) Data management and graphics
3) Linear regression basics
    inc. specification analysis, prediction, sampling weights
4) Simulation
5) GLS regression
6) Linear IV regression
7) Quantile regression
8) Linear panel-data models
9) Linear panel-data models: extensions
10) Non-linear regression methods

Appendix: Overview of Cameron and Trivedi (2009)

11) Nonlinear optimisation methods
12) Testing methods
13) Bootstrap methods
14) Binary outcome models
15) Multinomial models
16) Tobit and selection models
17) Count-data models
18) Nonlinear panel models