Regression and Causal Inference

Dr. Kamiljon T. Akramov
IFPRI, Washington, DC, USA

Regional Training Course on Applied Econometric Analysis

June 4-15, 2018, WIUT, Tashkent, Uzbekistan

Introduction

- "Essentially, all (statistical) models are wrong, but some are useful"

  George E. P. Box (1987)
- All econometric models are description of real world phenomenon using mathematical concepts, i.e., they are just simplifications of reality
- Regression analysis can be very useful if it is carefully designed
  - In accordance with current good practice guidelines, and
  - A thorough understanding of the limitations of the methods used
- If not, it can be not only inaccurate but also potentially damaging by misleading policymakers, practitioners and public
  - Example: Relationship between levels of government debt and rates of economic growth (Reinhart & Rogoff controversy)

Introduction (Cont.)

- This type of questions are simple cause-and-effect questions of the form
  - Does X cause Y?
  - If X causes Y, how large is the effect of X on Y?
  - Is the size of this effect large relative to the effects of other causes of Y?
- Simple cause-and-effect questions are the motivation for much empirical work in economics
- Definitive answers to such questions may not always be possible to formulate due to data constraints

Causal Inference

- Causal effect of program or policy interventions
- Some examples:
  - Job training programs on earnings and employment
  - Class size on test scores
  - Minimum wage on employment
  - Military service on earnings and employment
  - Tax-deferred saving programs on savings accumulation
  - Promotion of nutrition sensitive food value chains on nutritional outcomes
  - Farm size on agricultural productivity or income

Causal Inference (cont.)

- Causal effect of economic and behavioral variables
- Some examples:
  - Interest rate on credit card usage
  - Incentive scheme on employer productivity
  - Remittances on household consumption
  - International prices on domestic prices
  - Terrorist risk on economic behavior

Causes of Effects vs. Effects of Causes

- Important distinction between cause and effect:
  - Cause: an event that generates some phenomenon
  - Effect: the consequence (or one of the consequences) of the cause
- Crucial asymmetry in the difficulty of learning about the cause of an effect versus learning about the effect of a cause.
- Goal: understand why this asymmetry exists and what are its consequences for conducting research in economics as well as in other social sciences

Causal Inference Framework

- Potential Outcomes: each individual has a different outcome corresponding to each level that the treatment takes
  - Potential outcomes can be random or non-random

- Assignment Mechanism: each individual is assigned treatment based on some mechanism, and this mechanism guides how estimation and inference will be conducted
  - Assignment Mechanism will generate a random “treatment status” for identification purposes

Potential Outcomes: Causation as Manipulation

- Causal analysis: must have ability to expose or not expose each unit to action of cause
- Essential "each unit be potentially exposable to any one of the causes" (Holland, 1986)
  - If units could have been exposed to cause but they were not in practice: no problem
  - If units could not have been exposed to cause in any state of world: our cause might not really be a cause
  - Example: worker's education level versus worker's gender

Potential Outcomes: Causation as Manipulation (Cont.)

- Each unit has as many potential outcomes as different possible treatments there are
  - Called "potential" outcomes because only one of them is observed
  - Observed outcome is the one that corresponds to level of the treatment actually selected by (or assigned to) the unit.
- This introduces the idea of counterfactual: what would the outcome of this unit look like if the unit had been exposed to a different treatment?
- Key ideas:
  - (Non-manipulable) attributes and (manipulable) causes
  - Pre-exposure ("pre-treatment") and post-exposure ("post-treatment")

Potential Outcomes: Causation as Manipulation (Cont.)

- Basic Binary Treatment Setup
  - Each unit i is exposed to a binary treatment
    - $T_i = \ 1$ if unit i received treatment cause
    - $T_i = 0$ if uni$i$ received the control cause
- Each unit i is has two potential outcomes:
  - $Y_i(1)$: outcome that would be observed if unit i were exposed to treatment cause
  - $Y_i(0)$: outcome that would be observed if unit i were exposed to control cause
- Observed data: (Y_i, T_i) where

$$Y_i = T_i * Y_i(1) + (1 - T_i) * Y_i(0)$$

Stable Unit Treatment Value Assumption (SUTVA)

- Key (implicit) assumption: $Y_i(t)$ depends only on unit i's treatment status
- Implies that potential outcomes for unit i are unaffected by the treatment of unit j
- Rules out “interference” and “spillovers” across units
- Examples:
  - Effect of fertilizer or chemicals on crop yield
  - Effect of flu vaccine on hospitalization
- This assumption may be problematic:
  - Choose the units of analysis to minimize interference across units!
  - Address “interference” and “spillovers” explicitly.

Causality with Potential Outcomes

- Treatment Effect with Binary Treatments:

$$\tau_i \coloneqq Y_i(1) - Y_i(0)$$

- Effect of the treatment cause (relative to the control cause) on unit i
- τ_i depends on potential outcomes, not observed outcomes

Fundamental Problem of Causality

- For each unit i, we observe either $Y_i(1)$ or $Y_i(0)$, but never both!
- At the individual level, there is simply no way to learn about the causal effect (unless large amount of homogeneity in population)
- We can define aggregate estimands of interest about which we will be able to learn

- If we have multiple units, we can estimate average treatment effect (ATE)

$$\tau_{ATE} \coloneqq E[Y_i(1)] - E[Y_i(0)]$$

Assignment Mechanism

- Crucial ingredient in causal inference is the process by which each unit is selected or was assigned the particular treatment condition that it received
- It is a conditional probability of receiving treatment as a function of potential outcomes and covariates
- Two important cases
  - Random assignment: Known, independent of potential outcomes
  - Unconfounded assignment: Unknown, conditionally independent of potential outcomes
- Individualistic, probabilistic, and unconfounded assignment mechanisms

Key Ideas

- Assignment mechanism is the procedure that determines which units are selected for treatment intake
    - Random assignment
    - Selection on observables
    - Selection on unobservables
- Typically, treatment effects models attain identification by restricting the assignment mechanism in some way
- Causality is defined by potential outcomes, not by realized (observed) outcomes
- Observed association is neither necessary nor sufficient for causation
- Estimation of causal effects of a treatment (usually) starts with studying the assignment mechanism

Counterfactual model of causality (example)

- Causal states and relationship between potential and observed outcome variables
  - Two alternative states of a cause with a distinct set of conditions, exposure to which potentially affects an outcome of interest
- College degree and earnings
  - Outcome of interest: labor market earnings
  - Two states: whether or not an individual has obtained a college degree
  - Population of interest: adults between the ages 30 and 50
  - The causal effect of a college degree is about 40% higher wages on average (Angrist and Pischke 2009)
- Alternative causal states are referred to as alternative treatments
  - Treatment: college degree
  - Control: no college degree

Counterfactual model of causality (cont.)

- Key assumption:
  - each individual in the population of interest has a potential outcome under each treatment state, even though each individual can be observed in only one treatment state at any point in time
- Causal effect of college degree
  - Adults who have completed only high school degrees have theoretical what-if earnings under the state “have a college degree”
  - Adults who have obtained a college degree have theoretical what-if earnings under the state “have only a high school degree”
  - These what-if potential outcomes are counterfactuals

Counterfactual model of causality (cont.)

- Potential outcomes of each individual are defined as true values of outcome of interest that would result from exposure to alternative causal states
- Potential outcomes of each individual i are y_i^1 and y_i^0, where superscript 1 signifies treatment state and superscript 0 signifies control state
- In theory, an individual level causal effect can be defined as a simple difference

$$y_i^1 - y_i^0$$

- However, it is impossible to observe both y_i^1 and y_i^0 for any individual, thus, causal effect cannot be observed and directly calculated at the individual level
- Researcher must analyze observed outcome variable Y that takes on values y_i^1 and y_i^0 for those in treatment and control states
- y_i^0 is unobservable counterfactual outcome for individual i in treatment group, and y_i^1 is unobservable counterfactual outcome for individual i in control group

Some general comments

- In empirical research, we focus on estimating average causal effect for groups of individuals defined by specific characteristics
- To effectively estimate average causal effect, the process by which individuals of different types are exposed to the cause of interest have to be modelled
- Doing so requires plausible assumptions that allow for the estimation of average unobservable counterfactual values for specific groups of individuals
- If assumptions are plausible and appropriate methods of estimation and statistical inference are used, then an average difference in the values $y_i$can be given a causal interpretation

Causal analysis using experimental versus observational data

- Randomized experiments
  - Assignment mechanism is known and controlled, so estimating causal effect is straightforward in this case
  - Randomization is called the “gold standard” for causal inference because it balances observed and unobserved confounders
- Cannot always randomize so we do observational studies, where we need to adjust for the observed and unobserved covariates
  - Assignment mechanism not known, usually depends on covariates
  - Need to model for dependency and take this into account
- We have to design observational studies that approximate experiments
  - In an observational study researcher should always ask himself: How would the study be conducted if it were possible to do it by controlled experimentation (Cochran 1965)

Approximating Experiments

- It is important to distinguish between:
  - Covariates: Pre-treatment variables, potential confounders
  - Outcomes: Variables potentially affected by the treatment
- Randomized Experiment: Well-defined treatment, clear distinction between covariates and outcomes
- Better Observational Study: Well-defined treatment, clear distinction between covariates and outcomes
- Poorer Observational Study:
  - Hard to say when treatment began or what treatment really is
  - Distinction between covariates and outcomes is blurred
  - No baseline survey

Observational Studies: Key Questions

- How were treatments assigned?
  - Randomized Experiment: Random assignment.
  - Better Observational Study: Assignment is not random, but assignment mechanism is clearly described. Try to find “natural experiments”, where assignment is “as good as random”
  - Poorer Observational Study: No attention given to the assignment mechanism
- Were treated and controls comparable?
  - Randomized Experiment: Balance table for observables
  - Better Observational Study: Balance table for observables and ideally sensitivity analysis for unobservables
  - Poorer Observational Study: No direct assessment of comparability is presented

Main identification strategies for causal analysis using observational data

- Difference-in-differences: unobservables may differ, but their effect may not change much in time
- Instrumental variables: find variables that “randomize” some units into treatment
- Regression discontinuity designs: exploit (sharp or fuzzy) discontinuities in probability of treatment assignment
- Matching methods: match treatment and control groups using their observable characteristics
- Next week we will learn more about causal inference using experimental and observational data

Thank you