Regression, Causality and Identification Issues

Dr. Kamiljon T. Akramov

IFPRI, Washington, DC, USA

Regional Training Course on Applied Econometric Analysis

June 15, 2016, 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.)

- The most challenging empirical questions in economics involve cause-and-effect relationships
  - Does obtaining a college degree increase an individual's labor market earnings?
  - If so, is this particular effect large relative to the earnings gains that could be achieved only through on-the-job training?
  - What is the causal effect of manpower training on earnings? (Morgan & Winship 2007)
  - What is the causal effect of institutions on economic growth (Acemoglu, Johnson, and Robinson 2001)
  - What is the causal effect of ODA on economic growth (Rajan & Subramanian 2008; Akramov 2012)

Introduction (Cont.)

- These types 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

Counterfactual model of causality

- 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
- Example: 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
  - If only two treatments are considered, they are referred to as treatment and control

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

Counterfactual model of causality (cont.)

- 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

Standard OLS Model: Summary

- Consider a simple regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

- Standard OLS model provides an estimate of the effect on *Y* of arbitrary changes in independent variables (ΔX)
- The meaning of regression coefficients is the impact of one-unit increase in a given explanatory variable (X_i) on the dependent variable Y, holding constant other explanatory variables
- It can handle certain nonlinear relations (effects that vary with the *X*'s)

Assumptions of Classical Linear Regression Models

1. The regression model is linear in parameters β, is correctly specified, and has additive error term
2. No exact linear relationship between two explanatory variables and number of observations greater than number of explanatory variables
3. Explanatory variables must be exogenous (zero conditional mean), i.e., $E(\varepsilon|X_1, X_2,\ldots, X_n)=0$
4. Independently and identically distributed (iid) error terms, i.e., $\varepsilon \sim iid(0, \sigma^2)$
   - Expected value of the error term in population is zero
   - The error term has a constant variance
   - Observations of the error term are uncorrelated with each other
5. The error term is normally distributed

Best Linear Unbiased Estimator (BLUE)

- The Gauss-Markov theorem states that OLS estimator is BLUE if the assumptions 1 through 4 listed above are fulfilled
- Unbiased means that the OLS estimates of the coefficients are centered around the true population values of the parameters estimated
- Consistent means that as the sample size approaches infinity, the estimates converge to the true population parameters
- Violations of one or more classical assumptions will produce biased and/or inconsistent parameter estimates

Causal analysis: schooling and earnings

- Causal relationship between schooling and earnings tells us what people would earn, on average, if we could either
  - Change their schooling in a perfectly controlling environment or
  - Change their schooling randomly so that those with different levels of schooling would otherwise comparable
- *Conditional independence assumption (CIA)* requires that we must hold a variety of control variables fixed for causal inferences to be valid
  - Selection on observables
  - Covariates to be fixed are assumed to be known and observed

Causal analysis: schooling and earnings

- Assume schooling is a binary decision, C_i
- Two potential earnings variables

$$Outcome = \begin{cases} Y_{1i} & \text{if } c_i = 1 \\ Y_{0i} & \text{if } c_i = 0 \end{cases}$$

- We would like to know the difference between $Y_{1i}\ and\ Y_{0i}$, which is causal effect of schooling on individual i

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

Causal analysis: schooling and earnings

- Comparison of average earning conditional on schooling status is formally linked to the average causal effect

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

- If selection bias is positive, the naïve comparison of earnings exaggerates the benefits of schooling
- CIA asserts that conditional on observed characteristics selection bias disappears and comparisons of average earnings across schooling levels have a causal interpretation

Fundamental problem of causal inference

- It is impossible to observe the value of Y_{1i} and Y_{0i} on the same individual and, therefore, it is impossible to directly observe the effect of schooling on earnings
- Another way to express this problem is to say that we cannot infer the effect of schooling because we do not have the *counterfactual* evidence, i.e., what would have happened in the absence of schooling
- Given that the causal effect for a single individual cannot be observed, we aim to identify the **average causal effect** for the entire population or for sub-populations

Fundamental problem of causal inference: solution

- The econometric solution replaces the impossible-to-observe causal effect of treatment on a specific unit with the possible-to-estimate *average causal effect* of treatment over a population of units
- Although $E(Y_{1i})$ and $E(Y_{0i})$ cannot both be observed, they can be estimated
- Econometrics methods attempt to construct from observational data consistent estimates of

$$\overline{Y}_{1i} \text{ and } \overline{Y}_{0i}$$

Causal analysis: additional issues

- In most circumstances, there is simply no information available on how those in the control group would have reacted if they had received the treatment instead
- This is the basis for an important insight into another potential bias of standard regression analysis – treatment heterogeneity
- Thus, two sources of biases need to be eliminated from estimates of causal effects from observational studies
    1. Selection Bias: Baseline difference
    2. Treatment Heterogeneity
- Most of the methods available only deal with selection bias, simply assuming that the treatment effect is constant in the population or by redefining the parameter of interest in the population

Macro example

- What explains income differences across countries?
- Hypothesis: the quality of institutions explains the variation in per capita income across countries
- How would you establish causal link between institutions and income?
- Higher levels of economic development may cause higher levels of institutional quality
- Unobserved variable may jointly determine both high levels of institutional quality and high levels of income

Threats to Classical Assumptions

- Omitted variables
- Model misspecification or wrong functional form
- Measurement error
- Selection bias
- Simultaneous causality bias
- All of these imply that $E(u_i|X_1,X_2) \neq 0$

Omitted Variable Bias

- The bias in the OLS estimator that occurs as a result of an omitted factor is called omitted variable bias
- For omitted variable bias to occur, the omitted factor “Z” must be:
  - a determinant of Y; and
  - correlated with the regressor X but unobserved, so cannot be included in the regression
- Both conditions must hold for the omission of Z to result in omitted variable bias

Omitted Variable Bias Formula

- Regression of wages on schooling

$$Y_i = \alpha + \rho S_i + \gamma A_i + e_i$$

where α, ρ, and γ are population regression coefficients and e_i is a regression residual that is uncorrelated with all regressor

- What are the consequences of leaving ability out of regression?
- OVB formula

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

Where δ_{AS} is the vector of coefficients from regressions of the elements of A_i and S_i

Potential Solutions to Omitted Variable Bias

- If the variable can be measured, include it as a regressor in multiple regression
- Possibly, use panel data in which each entity (individual) is observed more than once
- If the variable cannot be measured, use instrumental variables regression
- Run a randomized controlled experiment

OVB Example: estimates of the returns to education for men in the NLSY

Controls	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
	None	Age dummies	Col. (2) and additional control variables (mother’s and father’s years of schooling, and dummies for race and census region)	Col. (3) and AFQT score	Col. (4) and occupation dummies
	0.132 (0.007)	0.131 (0.007)	0.114 (0.007)	0.087 (0.009)	0.066 (0.010)

Table reports the coefficient on years of schooling in a regression of log wages on years of schooling and the indicated controls. Source: Angrist and Pischke (2009).

Misspecification or Wrong Functional Form

- Arises if the functional form is incorrect
    - If an interaction term is incorrectly omitted, then inferences on causal effects will be biased
    - Variable transformations (logarithms)
        - Log-level, Level-log, Log-log models
    - Discrete dependent variables
- For example, the effect of dietary diversity on nutritional outcomes may depend on children’s age
- Other examples?

Measurement Error

- In reality, economic data often have measurement error
  - Data entry errors in administrative data
  - Recollection errors in surveys
    - when did you start your current job?
  - Ambiguous questions problems
    - what was your income last year?
  - Variations in perceptions
- Intentionally false response problems with surveys
  - What is the current value of your financial assets?
  - How often do you drink and drive?

Measurement Error (cont.)

- If X_i is measured with error, it is in general correlated with the error term, so estimated parameter ($\hat{\beta}$) is biased and inconsistent
- Potential solutions
  - Obtain better data
  - Develop a specific model of the measurement error process
  - Use IV approach

Sample selection bias

- Standard OLS assumes that the data is collected through simple random sampling of the population
- However in some cases, simple random sampling is thwarted because the sample, in effect, “selects itself”
- *Sample selection bias* arises when a selection process
  - Influences the availability of data and
  - That process is related to the dependent variable
- Correlation between the independent variable and other variables that are correlated with the outcome of interest render selection into the “Treatment group” non-random
- Instead, assignment to the treatment group is a function of some other factor and, more importantly, that other factor may be correlated with an outcome

Selection Bias (example 1)

- Institutional quality and economic development
    - There are both observed and unobserved processes that lead to the adoption and perpetuation of institutions across countries
    - These factors are correlated with economic development
    - Thus they need to be neutralized to avoid inducing a biased calculation of the treatment effects of institutions on growth
    - Otherwise, they will engender a difference in the baseline measures of the outcome of interest between the control and treatment group before exposure to the treatment
    - Thus, any difference in the control and treatment groups after exposure to treatment need to be adjusted to account for the preexisting differences

Selection Bias (example 2)

- Returns to education: What is the return to an additional years of education?
- Empirical strategy:
  - Sampling scheme: simple random sampling of **workers**
  - Data: earnings and years of education
  - Estimator: regress ln(*earnings*) on *years of education*
- Ignore issues of omitted variable bias and measurement error – is there sample selection bias?

Potential Solutions to Sample Selection Bias

- Institutions and economic development
  - IV (Acemoglu and Robinson, etc.)
- Returns to education
  - Sample college graduates, not workers including unemployed
- RCTs
- Construct a model of the sample selection problem and estimate that model

Simultaneous Causality

- X causes Y, but what if Y causes X, too
- Example: Class size effect
  - Initial hypothesis: Low STR results in better test scores assuming that there is a causal relationship running from STR to Test Scores through a better learning environment
  - But what if the school board responds to low average test scores by hiring more teachers for those school districts?
  - Then the causality runs both ways. But why is this a problem?
  - It leads to correlation between STR and the error term
- Estimation of demand and supply functions

Potential Solutions to Simultaneous Causality Bias

- Randomized controlled experiment
- Develop and estimate a complete model of both directions of causality: Large macro models (e.g. Federal Reserve Bank-US)
- IV approach

Summary

- Framework for evaluating regression studies:
    - Internal validity
    - External validity
- Threats to internal validity of causal analysis:
    - Omitted variable bias
    - Misspecification or wrong functional form
    - Measurement error or errors-in-variables bias
    - Sample selection bias
    - Simultaneous causality bias
- Next few days of the course will focus on modern tools of applied econometrics that help to detect causal relationships