Assessing Studies Based on Multiple Regression Analysis

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

Regional Training Course on Applied Econometric Analysis

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

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. There is random sampling of observations
3. No exact linear relationship between two explanatory variables and number of observations greater than number of explanatory variables
4. Explanatory variables must be exogenous (zero conditional mean), i.e., $E(\varepsilon | X_1, X_2, \ldots, X_n)=0$
5. 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
6. 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 5 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

Introduction

- Multiple regression has some key virtues:
  - It provides an estimate of the effect on dependent variable (*Y)* of arbitrary changes in independent variable (ΔX*)*
  - It resolves the problem of omitted variable bias, if an omitted variable can be measured and included
  - It can handle nonlinear relations (effects that vary with the *X*'s)

- However, violations of one or more classical assumptions may produce biased and/or inconsistent parameter estimates

Framework for assessing empirical research

- Construct validity: to what extent are the constructs of interest successfully operationalized in the research?
- Internal validity
  - The statistical inferences about causal effects are valid for the population being studied
- External validity
  - The statistical inferences can be generalized from the population and setting studied to other populations and settings, where the “setting” refers to the legal, policy, and physical environment and related salient features

Threats to External Validity

- How far can we generalize class size results from California school districts?
- Differences in populations
  - California in 2005?
  - Massachusetts in 2005?
  - Mexico in 2005?
- Differences in settings
  - Different legal requirements concerning special education
  - Different treatment of bilingual education
- Differences in teacher characteristics

Threats to Internal Validity

- Five threats to the internal validity of regression studies:
  - Omitted variable bias
  - Wrong functional form
  - Errors-in-variables bias
  - Sample selection bias
  - Simultaneous causality bias
- All of these imply that $E(u_i|X_{1i},\ldots,X_{ki}) \neq 0$

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

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

where α, ρ, and γ are population regression coefficients and ξ_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?

Wrong functional form (cont.)

- If the functional form is incorrect, then inferences on causal effects will be biased

- Potential solutions to functional form misspecification
  - Continuous dependent variable: use the "appropriate" nonlinear specifications in *X* (logarithms, interactions, etc.)
  - Discrete (*example*: binary) dependent variable: need an extension of multiple regression methods ("probit" or "logit" analysis for binary dependent variables).

Measurement Error or Errors-in-Variables Bias

- We assume that *X* is measured without 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?
- Intentionally false response problems with surveys
  - What is the current value of your financial assets?
  - How often do you drink and drive?

Measurement Error Bias: Illustration

- Suppose

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

  is "correct" model in the sense that the standard OLS assumptions, including $E(\varepsilon_i|X_i)$=0, hold

- Let

  X_i = unmeasured true value of X

  X_i' =imprecisely measured version of X

Measurement Error Bias: Illustration (cont.)

- Then

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$= \beta_0 + \beta_1 X_i' + [\beta_1 (X_i - X_i') + \varepsilon_i]$$
$$= \beta_0 + \beta_1 X_i' + \varepsilon_i'$$

where

$$\varepsilon_i' = \beta_1 (X_i - X_i') + \varepsilon_i$$

If X_i' is correlated with ε_i' then $\hat{\beta}_1$ will be biased:

$$\begin{aligned} cov(X_i', \varepsilon_i') &= \text{cov}(X_i', \beta_1(X_i - X_i') + \varepsilon_i) \\ &= \beta_1 \text{cov}(X_i', (X_i - X_i') + cov(X_i', \varepsilon_i) \\ &= \beta_1[\text{cov}(X_i', X_i) - var(X_i')] + 0 \neq 0 \end{aligned}$$

because

$$\text{cov}(X_i', X_i) \neq var(X_i')$$

Potential solutions to measurement error bias

- Obtain better data
- Develop a specific model of the measurement error process.
  - This is only possible if we know the nature of the measurement error – for example a subsample of the data are cross-checked using administrative records and the discrepancies are analyzed and modeled
- Instrumental variables regression

Sample selection bias

- We have assumed simple random sampling of the population. 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

Example 1: Returns to education

- What is the return to an additional year 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?

Example 2: Institutional quality and economic growth

- 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

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

Simultaneous causality bias in equations

- Causal effect of X on Y

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Causal effect of Y on X

$$X_i = \gamma_0 + \gamma_1 Y_i + \xi_i$$

- Large ε_i means large Y_i, which implies large X_i (if γ_1>0), i.e., corr(X_i, ε_i) ≠ 0
- Thus, estimated β_1 is biased and inconsistent
- Example: A district with particularly bad test scores given the STR (negative ε_i) receives extra resources, thereby lowering its STR; so X_i and ε_i are correlated

Potential Solutions to Simultaneous Causality Bias

- Randomized controlled experiment
  - Because X_i is chosen at random by the experimenter, there is no feedback from the outcome variable to Y_i (assuming perfect compliance)
- Develop and estimate a complete model of both directions of causality: Large macro models (e.g. Federal Reserve Bank-US)
- Use IV regression to estimate causal effect of interest

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 week we will focus on modern tools of applied econometrics that help to detect causal relationships