Regression, Causality and Identification Issues

Kamiljon T. Akramov, Ph.D.
IFPRI, Washington, DC, USA

Training Course on Applied Econometric Analysis

June 4, 2015, WIUT, Tashkent, Uzbekistan

Standard OLS Model: Summary

- Consider a simple regression model

  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

- It is assumed to satisfy the following four conditions
  - The model is linear in the parameters β
  - n observations are drawn randomly from populations of interest
  - $E(u|X_1, X_2)=0$, i.e., no endogeniety in the true model
  - No perfect collinearity
- These four assumptions guarantee that the OLS estimates of parameters will be unbiased

Standard OLS model: Summary (cont.)

- Standard OLS model provides an estimate of the effect on *Y* of arbitrary changes in independent variables (ΔX)
- It resolves the problem of omitted variable bias, if an omitted variable can be measured and included
- It can handle certain nonlinear relations (effects that vary with the *X*'s)
- Still, standard OLS might yield a *biased* estimator of the true *causal* effect

Motivation

- The most challenging empirical questions in economics involve causal-effect relationships:
  - Will mandatory health insurance really make people healthier?
  - How does an additional year of education change earnings?
  - What is the effect of farm size on agricultural productivity?
  - How does agricultural diversity impact nutritional outcomes?

Does Chocolate Consumption enhance cognitive function?

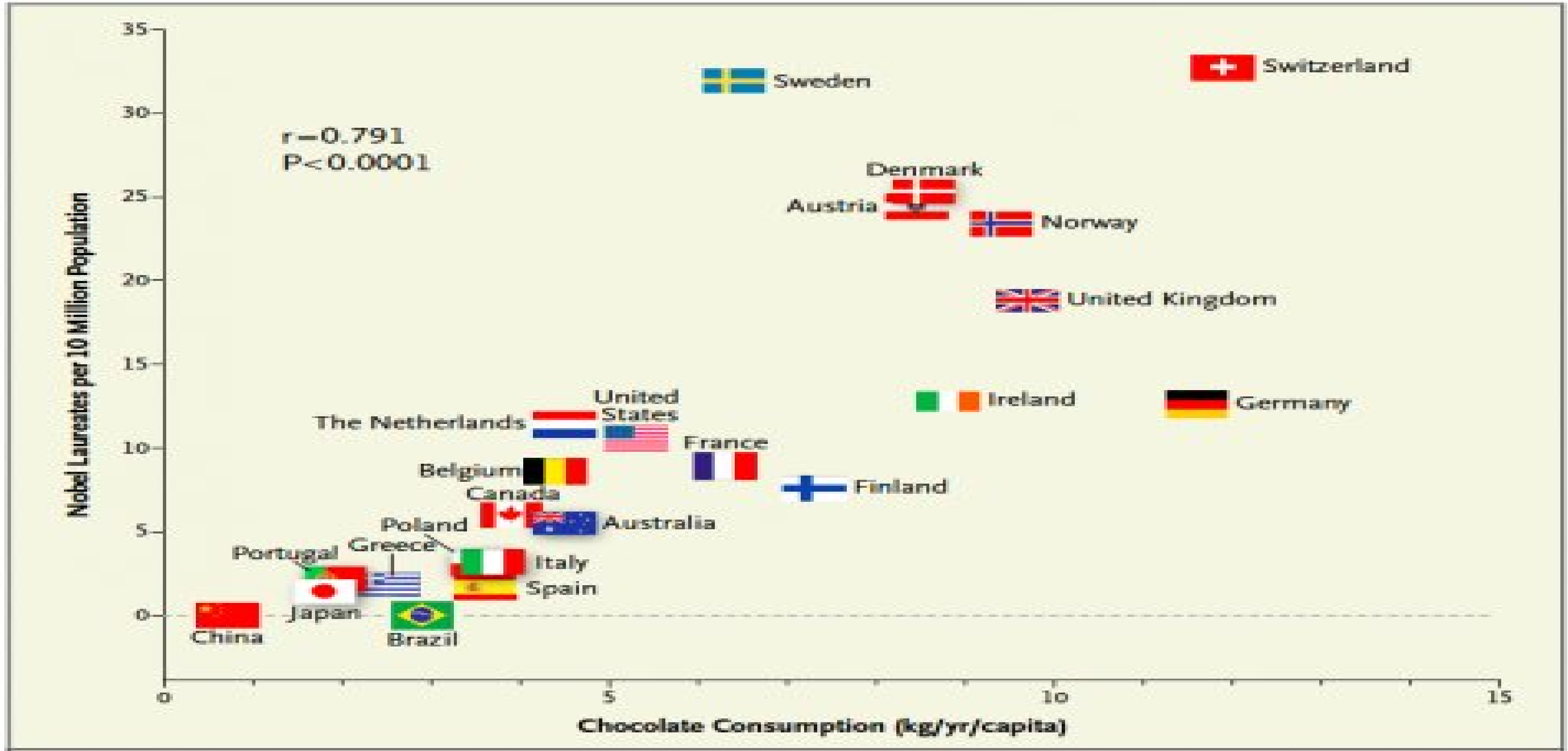

Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Source: Messerli (2012)

Regression analysis

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

  George E. P. Box (1987)
- All regression (statistical) 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 1: impact of hormone replacement therapy on the risk of coronary heart diseases (Hully et al. 1998; Velickovic 2015)
  - Example 2: Relationship between levels of government debt and rates of economic growth (Reinhart & Rogoff controversy)

Regression and causality

- The aim of standard regression analysis is to infer parameters of a distribution from samples drawn of that distribution
- With the help of such parameters, one can:
  - Infer association among variables
  - Estimate the likelihood of past and future events
  - As well as update the likelihood of events in light of new evidence or new measurement
- Causal analysis goes one step further:
  - Its aim is to infer aspects of the data generation process
  - With the help of such aspects, one can deduce not only the likelihood of events under static conditions, but also the dynamics of events under *changing conditions*

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 $Y_{1i} and\ Y_{2i}$, 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, i.e.,

Observed difference=Average treatment effect on treated + Selection bias

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

- 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 calculated, they can be estimated.
- Most 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 biase 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

Internal Validity

- Recall our definition of internal validity:
  - An econometric analysis is internally valid if the statistical inferences about causal effects are valid for the population being studied
- We know that internal validity hinges on two things:
  - The estimator of the causal effect should be consistent (unbiased would be nice too, but it’s not always feasible)
  - Hypothesis tests should have the desired significance level (i.e. you should be using the correct standard errors)

Threats to Internal Validity

- Omitted variable bias
- 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 and observed, include it as a regressor in multiple regression
    - But be careful about bad control variables
- Possibly, use panel data in which each entity (individual) is observed more than once
    - Fixed effects to control for unobserved heterogeneity
- 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. (4) 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)
  - 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?
- 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