



APPLIED RESEARCH: ISSUES, DATA, METHODS
AND PERSISTENT ESTIMATION PROBLEMS

Germano Mwabu

January, 27-31, 2008

Kampala, Uganda



1. Introduction

- Why undertake *applied* rather than *basic* research?
 - Both applied and basic research contribute to knowledge; both are valuable.
 - However, findings from applied research (based on data or theory) help address short-term *policy* concerns, whereas findings from basic research (data or theory based) may not be immediately applicable.
 - Desire to address current policy problems underlies much of applied research.



Introduction ... continued

- Obviously, we live in the midst of problems (extreme income poverty, civil conflicts, poor school performances, unemployment, HIV/AIDS, low life expectancy...)
- Research (a problem-solving activity) can generate the information needed to address these problems.



2. A quantitative framework for generating policy-relevant information

- Consider the following relationship

$$Y = f(X) \quad (1)$$

where,

Y = potential outcome of interest to policy makers, e.g., school *enrollment* rate.

X = a variable (or a set of variables) that influences Y , e.g., school fees, and household income.

- This simple equation can be used to frame and investigate numerous policy questions.



Framework ...continued

- The investigation can be conducted quantitatively *with* or *without* data.
- If, Y is viewed as *demand* for schooling, a researcher can build a mathematical model of demand for schooling (without using *data*), linking specific assumptions on *preferences* to particular demand patterns, and further exploring policy implications of the patterns.



Framework ... continued

- If, for example, parents' preferences on schooling and other goods are assumed to be *Cobb-Douglas* (a common assumption), demand for schooling, Y , has the unattractive feature that the *proportions* of income parents spend on Y and other goods are the same, irrespective of their income levels.
- A researcher may wish to build a theoretical demand model for schooling that is more policy relevant.



Framework ... continued

- This presentation focuses on applied research in which data are used to investigate policy questions, as may be framed using equation (1).
- Obviously, data alone (without theory) cannot be used to answer many policy questions.
- The applied researcher is expected to keep abreast of a fair amount of theories, in addition to *techniques of data processing and analysis*, which must be mastered.



3. Describing the policy environment

- Applied research can help policy-makers understand better the existing policy environment before implementing measures to change it.
- Descriptive statistics can be used for this purpose.
- If in equation (1), data are available on Y and X , researchers can provide information about national enrollment rates by region, gender, education and income of parents, and on school fees paid and distances covered to get to school.



Describing the policy environment...

- Nationally representative surveys on households, schools, etc, are ideal for constructing summary statistics for describing the policy environment.
- Findings from such statistics can be used to target education and anti-poverty programs to disadvantaged communities.
- However, the targeting also creates problems in evaluating impacts of the programs.



4. Measuring impact of a policy intervention

- Policy makers and donors may want to know whether an on-going intervention is having the desired effect so that it can be continued or scaled-up.
- To illustrate the key measurement issues and the general principles involved, rewrite equation (1) as:
- $Y = \mathbf{a} + \mathbf{b}X + e$ (2)

Where, in a regression context, \mathbf{a} and \mathbf{b} are parameters, and \mathbf{e} is an error term.



Measuring policy impact... continued

- Measuring impact is a special task of the more general, and demanding task of identifying *causal* effects.

In the context of equation (2),

$$Y = \mathbf{a} + \mathbf{b}X + e :$$

b = "causal effect" or "treatment effect"

X = "causal factor" or "treatment"



Measuring causal effects... continued

- (i). If in eq. (2), the $\text{cov}(X, e) = 0$,
 - \mathbf{b}_{OLS} is the causal effect of X on Y (the treatment effect (i.e., \mathbf{b}))
- (ii). If in eq. (2), the $\text{cov}(X, e) \neq 0$,
 - \mathbf{b}_{IV} identifies the causal effect of X .
- (iii). If in eq. (2), the $\text{cov}(X, e) \neq 0$,
 - \mathbf{b}_{OLS} is a biased estimator of \mathbf{b}



4. The Endogeneity problem in the estimation of treatment (causal) effects

- Consider the general model:

$$Y = \mathbf{a} + \mathbf{b}X + e \quad (2)$$

In eq (2), X is said to be endogenous to Y if $\text{cov}(X, e) \neq 0$.

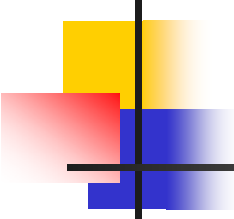
There are three sources of endogeneity (Rosenzweig and Wolpin, 2000):

- Omitted variables.
- Errors in measuring X .
- Feedback effects from Y to X .



Endogeneity... continued

- Neglected omitted variables, measurement errors and feedback effects (simultaneity bias) all appear in in the error term in eq (2), contaminating the effects of X on Y .
- Instrumental variables can be used to deal with the endogeneity problem.



5. Use of Instrumental Variables and experiments to deal with the endogeneity problem

- In eq (2), instrumental variables make X vary exogenously. A variable that makes X vary without directly affecting Y is an instrument for X .
- In the context of the school enrollment model ($Y = a + bX + e$), where X is *school fees* (which are endogenous because parents choose how much fees to pay), a centrally administered bursary scheme could be used as an instrument.



5a. Instrumental variables...

- Introducing an instrumental variable in the school enrollment model.

-- The structural model:

$$Y = a + bX + c \bullet S + e \quad (3)$$

--Reduced form model:

$$X = \alpha + \beta \bullet Bursary + \gamma \bullet S + \varepsilon \quad (4)$$

- where, X = school fees paid at a local schools; $Bursary$ = Bursary grants to local schools; S = Social economic variables (controls).



Instrumental variables...

- Eq.(4) is called the first-stage regression.
 - It is estimated first to predict a value for the endogenous variable, X .
- Eq. (3) is the second-stage regression.
 - Y is regressed against the *predicted* X (predicted bursary), controlling for observable and exogenous characteristics of students (S).
 - The IV variable (*bursary*) must be excluded from the outcome equation (i.e. eq.3).



Instrumental variables: comments

- Although the Stata command IVREG and IVREG2 (and similar commands in other softwares) facilitate estimation of eqs (3) and (4) at once, it should be noted that 2SLS (two-stage least squares) is used for estimation, with the two stages being accomplished in one command.
- Bound et al (1995) discuss problems of IV methods.
- Panel data can help some endogeneity due to fixed effects.



5b. Experimental approaches to estimation of causal (treatment) effects

- Valid IVs are difficult to find, especially when endogeneity is due to unobservables in the error term
- In using IVs researchers have incentives to choose the results they like even if the results are not true.
- For example, they can include in eqs (3 & 4) only control variables (S) and instrumental variables that produce statistically significant treatment effects or effects of certain sizes.



Experimental approaches...

- Consider eq (3) again:

$$Y = \mathbf{a} + \mathbf{b}X + \mathbf{c}S + e \quad (3a)$$

- X is “treatment” such as a bursary, income subsidy, or food supplements.
- The aim is to see how these affects school enrollment, Y .
- As in IV methods, there are two stages to estimation of treatment effects using experimental data.



Experiments...

- First stage: assign *treatment* between the treatment group and the control group randomly.
 - That is, make receipt of treatment exogenous, unrelated to the outcome variable.
- Second stage: Compare the difference between values of the outcome variable, Y for treatment and control groups.
- The difference is the treatment effect.



Experiments...

- This is equivalent to estimating eq (1):

- $Y = \mathbf{a} + \mathbf{b}X + e$

Where, X is randomly assigned so that $cov(X, e) = 0$.

- The controls, S in eq (3a) are not needed if the treatment is randomly assigned because then subjects in control sample, and in the treatment sample are the same.

--That is, *selection bias* is absent.



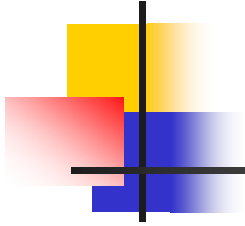
Estimation of treatment effects when Treatment Assignment Mechanism is not Perfect

- The Difference-in-difference methods (DiD).
- Consider again eq (3a):
- $Y = \mathbf{a} + \mathbf{b}X + \mathbf{c}S + e$ (3a)
- DiD corrects for bias in \mathbf{b} due to differences in subjects in control and comparison groups (because the subjects differ in S (the observables) or treatment (program is targeted)).
- Control and comparison groups not always easy to form.



Conclusion

- In practice IV and experimental methods are used together
- The challenge is to understand why an instrumental variable should work, and how to design a good experiment.
- Care is needed in collecting and processing the required data.



-
- Thank you.