Instrumental variables in linear regression

PLSCS/NTRES 6200

D G Rossiter

NY State College of Agriculture & Life Sciences

28-March-2017

Outline

1. Determining causality
2. The problem: endogenicity
3. Instrumental variables
4. Computations
5. References

1. Determining causality
2. The problem: endogenicity
3. Instrumental variables
4. Computations
5. References

Determining causality

- Correlation $\neq$ causation
- Regression equations show *mathematical* dependency $y = f(x)$ but not causality
- Why determine causality?
  - to understand a system
  - to plan interventions, e.g., effective policy instruments

How to determine causality?

1. Controlled experiments
2. Time of intervention vs. time of observed outcomes
3. Quasi-experiments: observed outcomes ‘with/without intervention’, ‘before/after intervention’ but otherwise uncontrolled

- These are not always feasible
- Controlled experiments may too far from actual pratice, so policy instruments based on these may fail

Example: natural resources management (NRM)

- Aim: determine the effectiveness of various NRM interventions
- Example: effect of interventions on crop yield
  - Irrigation methods, weeding frequency, fertilization types and amounts ...
- These can all be determined by controlled experiments
- But we can make many more observations in farmer's fields, without experiment
  - observe agricultural practices and crop yields

Example: criminology

- Aim: determine how policing levels affect crime incidence
- No way to do a controlled experiment, but there is plenty of data for an observational study

Example: education policy

- Aim: determine how tutoring programmes affect graduation rates
- Can compare matched groups with/without program
  - ethical problem: denying the programme to some students
- Can offer the programme and observe outcomes
  - intensity of tutoring, number of sessions . . . vs. outcome
  - Problem: self-selection, not a cross-section of abilities
  - Problem: “inherent ability” as such can not be measured directly

1. Determining causality
2. The problem: endogenicity
3. Instrumental variables
4. Computations
5. References

The problem: endogenicity

- The chosen predictor variables may not be (fully) **exogenous** to (outside of) the system
- The intervention may not be independent of (1) unobserved context or (2) the result
- The predictor may not be directly **observable** but influences the regression.
- This results in **correlation** with the error term of the regression; this is termed **endogenicity** (inside the system)
- Therefore the regression coefficents do *not* represent the true effect of the intervention

NRM example

- use of fertilizer increases crop yield – by how much? We want the dose–response curve, however:
  - increased fertilizer use may also result in the farm taking more care with weeding (unobserved)
  - increased fertilizer use may be the result of favourable early-season weather or favourable soil conditions
  - farmers differ in their skill of fertilizer application (timing, placement . . .)
  - these would in any case increase yields, even if fertilizer had not been increased
- If all possible production factors were observed we could build a multivariate regression
  - problems with colinearity, over-determination (n observations vs. p predictors), interaction . . .
  - how do we know we've included "all" factors?

Criminology example

Problem: **interdependence** of result (crime) and predictor (police effort)

- “Economic theory suggests police and crime are negatively correlated. However, it is surprisingly difficult to demonstrate this relation empirically, as areas with greater numbers of crimes tend to hire more police.”[1]
- So, the OLS estimate would be biased upward (less negative): more police → less crime, but a smaller effect than is actually the case
- A naïve analysis might argue that police do not deter crime

[1] Lin (2009)

Education example

- Problem: “ability” can not be measured
- Problem: students with different abilities may preferentially use (or not) the service
- Example: only weak students use the service: inflated results because students make big advances from a low base (?) or poor results because students can’t benefit (?)

Assumptions of linear regression

- model: $y = BX + u$; univariate case $y = \beta_0 + \beta_1 x + u$
  - y response (dependent, target) variable
  - X **design matrix** of predictor (independent) variables
  - B vector of regression coefficients to be estimated
  - u residuals: i.i.d., uncorrelated errors (noise)
- Assumption: **regressors** X **are uncorrelated with errors** u – i.e., X is **exogenous** to the response
  - "ex"ogenous = "outside" influence
  - "external" effect of X on y

$$\begin{tikzcd}
x \arrow[r] & y \\
u \arrow[ru] &
\end{tikzcd}$$

- This is guaranteed in **controlled** experiments, with all non-experimental factors held constant

Endogenicity

- If a regressor x is corrrelated with the error term u it is called **endogenous**.
  - "endo" genous = "inside" influence

$$\begin{tikzcd}
x \arrow[r] & y \\
u \arrow[u] \arrow[ur] &
\end{tikzcd}$$

- The error ("noise") affects the predictor as well as the predictand

How can endogenicity occur?

1. **Reverse causality**: the “dependent” variable also has an effect on the “independent” one.
2. **Omitted variable** that helps explain y and is also correlated with x, i.e., it makes x endogeneous
   - Some omitted variables are **unobservable**, e.g., “aptitude”, “skill”, “motivation”, “degree of optimism”
   - Can we argue that there are no omitted **observable** variables that affect both predictors and target?
3. **Autocorreled errors** (spatial, temporal, repeated measurements)
4. **Measurement errors** in the covariates

Effects of endogenicity

- Actual model is $y = BX + (\gamma X + \varepsilon)$
  - i.e., the error term u is a linear function of the predictor
- This **biases** the estimates of B
  - $\frac{dy}{dx} = \beta + \frac{du}{dx}$
  - if no relation between u and x, then $\frac{du}{dx} = 0$, so fitted $\widehat{\beta} = \beta$
  - if a relation, fitted $\widehat{\beta}$ estimates $\beta + \frac{du}{dx}$
  - this can be greater or less than the true β
- The regression *does* give a correct estimate of the **endogenous** effect of x on y and can be used for **prediction**
- It *does not* give a correct estimate of the **exogenous** effect of x on y – this is what is needed for **policy decisions**

Solutions

1. Conduct a **controlled experiment** where x and its effect on y are directly observed, controlling for all other possible predictors
   - rarely feasible and can be unethical
   - does not solve the problem of unobservable covariables
2. Build a **multivariate regression** with “all” possible predictors
   - Not possible with **unobservable** variables such as “farmer skill”
   - Many practical problems with large numbers of predcitors
3. Use an **instrumental variable**

1 Determining causality

2 The problem: endogenicity

3 Instrumental variables

4 Computations

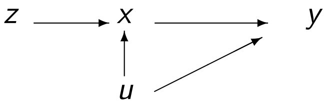
5 References

Instrumental variables – Objective

- We want to estimate the true **causal** effect, in isolation, of some intervention: $y = f(x)$
- This will allow us to decide about policy which could modify the intervention, i.e., level of x
- We set up a regression equation $y = \beta x$ and estimate β_{OLS} from the data.
- If there is **endogenicity**, the regression coefficient β_{OLS} from linear regression does *not* express causality
- Instrumental variables is a technique to solve this problem; it will produce β_{IV} which will be the **exogeneous** effect of x on y.

Instrumental variables

- z is an exogenous **instrument** that substitutes for the endogeneous predictor x
- z has a strong linear relation with the endogeneous predictor x
- z has *no* correlation with the error term u, i.e., is itself **exogenous**



- i.e., z such that Δz results in Δx but not directly Δy

Is it a valid instrument?

- Must *not* be correlated with the error term u: $\mathrm{Cov}(z, u) = 0$, i.e., must be exogenous
  - This is called the **exclusion restriction**
  - This *can not* be tested, it must be argued from knowledge of the system
- Must be correlated with the endogenous explanatory variable x: $\mathrm{Cov}(z, x) \neq 0$
  - This can be tested; the results of the first stage of two-stage regression give the strength of this relation: $x = \gamma_0 + \gamma_1 z + \varepsilon$ solved by OLS
  - Prefer a **strong first stage**, otherwise the second stage is quite inefficient in estimating β_{IV}

Criminology example

Instrumental variable for policing intensity: **variations in state or local tax rates**: changes in the 1-year lagged sales tax rate

- considered as an exogeneous “shock” to the system
- correlated with police presence (more tax increase → more budget for police)
- **un**correlated with the error term: affect crime only through police presence
    - Question: how could changes in tax rate be otherwise correlated with crime?

1. Determining causality

2. The problem: endogenicity

3. Instrumental variables

4. Computations

5. References

R packages and functions

AER	“Applied Econometrics with R”	`ivreg`
ivpack	“Instrumental Variable Analyses”	`ivpack`
sem	“Structural Equation Models”	

Computations: Two-stage regression

This is the easiest to understand, although it gives slightly incorrect values for the standard error.

First look at the univariate case:

1. **Model** the predictor as a linear function of the instrumental variable: $x_i = \gamma_0 + \gamma_1 z_i + \varepsilon_i$ and fit the $\widehat{\gamma}$
2. **Substitute** the fitted values from this equation for the original values of the predictor: $x_i^* = \widehat{\gamma_0} + \widehat{\gamma_1} z_i$
3. **Model** the predictand as a linear function of the of the substituted predictor: $y_i = \beta_0 + \beta_1 x_i^* + u$

These β_{IV} estimate the exogenous effect of x on y.

Two-stage regression: general case

1. Model the **endogenous** variable by all the **exogenous** variables and the **instrumental** variables:

$$x_k = \gamma_1 + \gamma_2 x_2 + \ldots + \gamma_{k-1} x_{k-1} + \theta_1 z_1 + \ldots + \theta_L z_L$$

2. Compute the fitted variables for the (now exogenous) variable:

$$\widehat{x}_k = \widehat{\gamma}_1 + \widehat{\gamma}_2 x_2 + \ldots + \widehat{\gamma}_{k-1} x_{k-1} + \widehat{\theta}_1 z_1 + \ldots + \widehat{\theta}_L z_L$$

3. Model the predictand usin these fitted values:

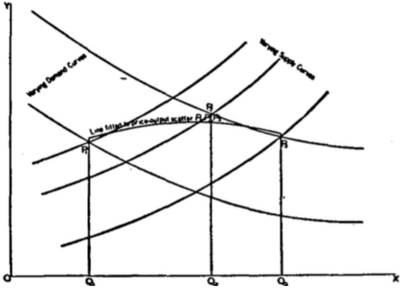
$$y = \beta_1 + \beta_2 x_2 + \ldots + \beta_K \widehat{x}_k$$

1. Determining causality
2. The problem: endogenicity
3. Instrumental variables
4. Computations
5. References

Historical note

- Wright, Phillip G. (1928). "The tariff on animal and vegetable oils". New York: The Macmillan company; Appendix B
- "[E]lasticity supposes different experiments with prices in the same market at a single instant of time. . . . Actual observations must be made at different times and during the period between observations both of supply and demand may change."

FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.



References I

- CRAN Task View: Econometrics: `https://cran.r-project.org/web/views/Econometrics.html`
- Baiocchi, M., Cheng, J., & Small, D. S. (2014). "Instrumental variable methods for causal inference". **Statistics in Medicine**, 33(13), 2297–2340. `https://doi.org/10.1002/sim.6128`

  > *"The IV method seeks to find a randomized experiment embedded in an observational study and use this embedded randomized experiment to estimate the treatment effect."*

- Kleiber, C., & Zeileis, A. (2008). **Applied econometrics with R**. UserR! Series. New York, Springer.

References II

- Worrall, J. L., & Kovandzic, T. V. (2010). "Police levels and crime rates: An instrumental variables approach." **Social Science Research**, 39(3), 506–516. https://doi.org/10.1016/j.ssresearch.2010.02.001
- Lin, M.-J. (2009). "More police, less crime: Evidence from US state data". **International Review of Law and Economics**, 29(2), 73–80. https://doi.org/10.1016/j.irle.2008.12.003