Research Methods

Carlos Noton

Term 2 - 2012

Outline

Main Assumptions

Suppose God's model is like this:

$$Y = f(X, \varepsilon)$$

Nature or God set X. A different God played with a roulette and drew values of ε. Based on these values, Y is uniquely defined through function f.
Formally, ε is the non-systematic component, but in practise ε is everything but X.
Values of X are **independent** of shocks ε (i.e. $X \perp \varepsilon$).

Linear Specification

Outcome Y is a function of the **Observable** K characteristics denoted by $X = \{X_1, X_2, .., X_K\}$; and ε.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_K X_{Ki} + \varepsilon_i$$

for $i \in \{1, .., N\}$. Shocks have mean zero, $E(\varepsilon_i) = 0$, and constant variance, $V(\varepsilon_i) = \sigma_\varepsilon^2$.

Last week we were very pleased thinking of ε as an **unobservable** random shock.

Can ε be an **unobservable** characteristic?

Linear Specification

Recall our matrix notation:

$$Y = X\beta + \varepsilon$$

where Y and ε are column vectors of dimension $N \times 1$,X is a matrix of dimension $N \times K$and β is a vector of dimension $K \times 1$.

In empirical work, we only observe Y and X, and we have to estimate vector β and the variance σ^2_ε.

Identification in this case

But, what if ε is an **unobservable** characteristic?

What happen if unobserved ε is uncorrelated with X?
If ε is uncorrelated with X: NO PROBLEM.
If $cov(X, \varepsilon) = 0$, we still have all the good properties we need!

What happen if X and ε are correlated?

What did we learn last week about systematic correlation between explanatory variables?

YES, identification problem!. Estimation of slope β is going to be poorly identified. Moreover, if $cov(X, \varepsilon) \neq 0$, we will find a systematic bias in our estimates.

Not assuming exogeneity

Think of the following Examples in Economics:

- Outcome Y is wages; Control X education, ε is Unobservable ability of the individual.
- Outcome Y is ice-cream quantity; Control X is ice-cream price, ε is weather shock observed by the producer.
- Outcome Y is technology investments; Control X is size of the firm, ε is unobservable manager's ability.
- Outcome Y is health status; Control X is drinking or smoking dummy, ε is unobservable stress level.

Critique to the Exogeneity Assumption

We think that people choose their level of education based on ability, firms choose ice-cream prices based on weather shocks, firms choose size of the firm based on manager's ability, people choose drinking/smoking based on their stress level.
Therefore, we have theoretical models that argues that the exogeneity assumption is wrong. X is chosen based on ε, thus $cov(X, \varepsilon) \neq 0$
This seems very natural when you have agents making rational decisions based on their unobserved characteristic ε or expectations about ε. We have a problem when X is altered in a systematic way due to ε, which creates correlation between X and ε.

This is the so-called Endogeneity problem. Think the link with identification issue of last week?

Formal Terms

$$\begin{aligned}
\widehat{\beta}_{OLS} &= (X'X)^{-1}(X'Y) \\
&= (X'X)^{-1}(X'[X\beta + \varepsilon]) \\
&= (X'X)^{-1}(X'X)\beta + (X'X)^{-1}(X'\varepsilon) \\
&= \beta + (X'X)^{-1}(X'\varepsilon)
\end{aligned}$$

The last term will not vanish if there is systematic correlation between X and the unobserved characteristic ε.

Testing

Last week, we plot the two characteristics to check any potential correlation.

Now, ε is unobservable. Hence, any test is very indirect and the best arguments (in my opinion) comes from Economic Theory or from a particular feature of the analysis.

For example: if the ice-cream prices are set a month in advance, then we can argue that weather shocks are not important. If manager's ability can be captured through a firm fixed effect, then the unobservable characteristic, ε will not have this component.

Notice that fixed effects are meaningful if we see changes in outcome over time (panel data). For example, if individual i always smokes. Fixed effect trying to capture his *fixed* stress level lead to an identification problem as we saw last week.

Solution I: Instrumental Variables

Remember we are trying to estimate a marginal causal effect summarized in vector β.

Suppose only one explanatory variable X_i

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

for $i \in \{1, .., N\}$. This estimation tries to fit the best linear function to two variables, but suppose we have the endogeneity problem: $cov(X_i, \varepsilon_i) \neq 0$.

Solutions: Instrumental Variables

Suppose that characteristic X_i can be explained by an exogenous variable Z_i that is not correlated with ε_i. Hence, the following linear model meets all the OLS assumptions:

$$X_i = \gamma_0 + \gamma_1 Z_i + v_i$$

for $i \in \{1, .., N\}$. Here, we assume $cov(Z_i, v_i) = 0$. Hence, standard OLS gives the estimate of $\widehat{\gamma} = (Z'Z)^{-1}(Z'X)$.

IV Solution

An important assumption is that the instruments are not correlated with the unobserved characteristic ε, i.e. $cov(Z_i, \varepsilon_i) = 0$. If that is true, then we can compute predicted values of $\widehat{X}_i$ that are not contaminated with ε.

$$\widehat{X}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 Z_i$$

IV Solution

Replace

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ &= \beta_0 + \beta_1(\gamma_0 + \gamma_1 Z_i + v_i) + \varepsilon_i \\ &= \underbrace{\beta_0 + \beta_1\gamma_0}_{\text{new intercept } \delta_0} + \underbrace{\beta_1\gamma_1}_{\text{new slope } \delta_1} Z_i + \underbrace{\beta_1 v_i + \varepsilon_i}_{\text{new random term } u_i} \\ &= \delta_0 + \delta_1 Z_i + u_i \end{aligned}$$

You can show that: $\widehat{\beta}_1 = \frac{\widehat{\delta}_1}{\widehat{\gamma}_1}$

IV Solution

This is equivalent to run the original regression with the estimated $\widehat{X}_i = \widehat{\gamma}_0 + \widehat{\gamma}_1 Z_i$. Replace

$$Y_i = \beta_0 + \beta_1 \widehat{X}_i + \varepsilon_i$$

You can show that the slope is the same β_1 but now $cov(\widehat{X}_i, \varepsilon_i) = 0$ since $cov(Z_i, \varepsilon_i) = 0$.

Intuitively, Z_i only matters through X_i, that is why it helps to identify β_1.

Linear Model with K explanatory variables

Suppose K explanatory variables $X_1, X_2, .., X_K$ in matrix notation:

$$Y = X\beta + \varepsilon$$

where Y and ε are column vectors of dimension $N \times 1$, X is a matrix of dimension $N \times K$ and β is a vector of dimension $K \times 1$.

$$X = Z\delta + v$$

Now as dependent variable you have K different components so you at least need K different instruments.

If only a subset $K_1 < K$ is suspicious of being endogenous, then only need K_1 instruments and the exogenous variables are part of Z_1.

IV Example

"Information Technology and Economic Change: The Impact of the Printing Press" by Dittmar (QJE 2010)
Did printing imply a faster growth of cities?

$$Y_{it} = X'_{it}\beta + T_i \underbrace{\left(\sum_t \alpha_t D_t \right)}_{\lambda_t} + \varepsilon_{it}$$

Y_{it} is log city growth for city i in time t, T_i is an indicator variable capturing whether city i was an early adopter of print technology, X_{it} is a vector of covariates (including time effects and city effects).

Problems if the positive association between the adoption of print technology and city growth is due to printers selecting cities that were already bound to grow quickly; i.e. $cov(T_i, \varepsilon_{it}) \neq 0$.

IV Example

We need some variable Z_i that affects T_i (i.e. $cov(Z_i, T_i) \neq 0$); and only affects Y_i through T_i such that $cov(Z_i, \varepsilon_{it}) = 0$.

He used the distance, Z_i, to Mainz in Germany, where the movable type printing press was developed by Johannes Gutenberg around 1450. In subsequent decades entrepreneurial printers spread the technology to other European cities.

$$T_i = Z_i\gamma + v_i$$

Some tests to validate the instrument: Z_i does not explain nothing else but T_i. Especially, Y_{it} before Gutenberg.

Natural Experiments or Quasi-Experimental Approach

This approach seeks for episodes where the changes in X are reasonable exogenous, then β can be identified!

For example,

- Unexpected changes in legislation
- Natural shocks like: earthquakes, plagues, new discoveries
- Wars or some unexpected change in regimes.

It is very important that agents do not expect this change to happen and do not have time to choose X again, otherwise, it does not meet the assumptions.

Example

"Quality Matters: The Expulsion of Professors and the Consequences for PhD Student Outcomes in Nazi Germany" by Fabian Waldinger (JPE 2010)
Looking for the effect of faculty quality on PhD student outcomes:

$$Outcome_{idt} = \beta_1 + \beta_2(Avg.FacultyQuality)_{dt-1}$$
$$+\beta_3(Student/FacultyRatio)_{dt-1} + \beta_4 X_{idt} + \varepsilon_{idt}$$

Do we have an endogeneity problem?

Waldinger (JPE 2010)

To address the endogeneity of faculty quality, he uses exogenous variation provided by the expulsion of mathematics professors in Nazi Germany.

$$Outcome_{idt} = \beta_1 + \beta_2(Avg.FacultyQuality)_{dt-1}$$
$$+\beta_3(Student/FacultyRatio)_{dt-1} + \beta_4 X_{idt} + \varepsilon_{idt}$$

Hence, he can identify β_2 and β_3!!

Showing these phenomena in STATA

See Program "IVExercise.do"