Instrumental Variables Regression

NICHOLAS R. JENKINS

Overview

1. Why do we need IV regression?
2. IV Model
3. Estimation Techniques
4. Instrument Validity
5. Applications

Why do we need IV regression?

Case Studies:

1. Does Putting Criminals in Jail Reduce Crime?
2. Does Aggressive Treatment of Heart Attacks Prolong Lives?

Does putting criminals in jail reduce crime?

- This was a question that was investigated by the famous economist, Steven Levitt
- To answer this question, we could regress crime rates on incarceration rates and control variables such as economic conditions and demographics using OLS

$$Crime\ Rate = \beta_0 + \beta_1 Incarceration\ Rate_i + \beta_2 X_i + \cdots + \beta_k X_i$$

- Will this give us an accurate answer to the question? No
- There is a major threat of simultaneous causality bias
  - An increased incarceration rate reduces the crime rate
  - However, an increase in the crime rate increases incarceration rates
- Leaves us unable to parse out the true effect of incarceration on crime
  - Moreover, this cannot be solved by adding more control variables
- Levitt eliminated the simultaneous causality bias by finding an instrumental variable and estimating an IV regression using Two Stage Least Squares (TSLS)

Does aggressive treatment of heart attacks prolong lives?

- Researchers compared patients who received treatment to patients that did not
  - Specifically, the regressed the length of survival of the patient on a binary variable for the treatment plus control variables
- By simply estimating a regression with OLS we are at risk of biased estimators (due to a possibly unobservable omitted variable)
  - This is the case because patients don't randomly receive treatment for heart attacks
  - They receive treatment because the doctor recommended it to an at-risk patient
  - If they opt for the treatment based on unobserved factors relevant to health outcomes not in the data set, then the treatment decision will be correlated with the error term
  - If the healthiest patients are the ones that receive the treatment, then the OLS estimator will be biased and the treatment will appear more effective than it really is
- Using valid instruments in an IV regression they avoided any risk of bias

So, why do we need IV regression?

- IV regression allows us to solve problems where the regressor (independent variable) is correlated with the error term and helps establish internal validity
  - Omitted variables
  - Errors-in-variables (measurement errors in regressors)
  - Simultaneous causality

The IV Regression Model

The general IV model has 3 variables in addition to the dependent variable:

1. Endogenous Variables: *X*
   - Problematic variables (independent variables that are correlated with the error term)
2. Exogenous Variables: *W*
   - Variables that are uncorrelated with the error term
   - Can be control variables
3. Instrumental Variables: *Z*
   - Relevant: correlated with the endogenous (problematic) variable
   - Exogenous: not correlated with the error term

There must be at least as many Instruments (*Z*) as there are endogenous regressors (*X*)

- When the number of instruments equals the number of endogenous regressors, the coefficients are **exactly identified**
- When the number of instruments exceeds the number of endogenous regressors, the coefficients are **overidentified**

The IV Regression Assumptions

Only slight modifications to the multiple regression assumptions

- Modification #1: The conditional mean assumption applies only to exogenous variables

$$E(u_i|W_{1i}, \dots, W_{ri}) = 0$$

- Modification #2: The two conditions for valid instruments hold
  - Relevance
  - Exogeneity

Estimation Techniques

1. Two Stage Least Squares (TSLS)
   - Carries out the regression in two stages
   - This is what allow us to correct the problem of correlation with the error term
2. Limited Information Maximum Likelihood (LIML)
   - This should be used when the problem of weak instruments cannot be avoided (more on this later)
3. Generalized Method of Moments (GMM)
   - This technique should be used when the model is nonlinear in the parameters

Two Stage Least Squares (TSLS)

The idea is to separate the endogenous (problematic) variable into two parts: the problematic component that is correlated with the error term and an unproblematic component that is not correlated with the error term.

In the first stage, the endogenous variable *X* is regressed on the instrument *Z* to obtain a predicted value of *X*.

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

- This regression separates the endogenous variable *X* into the part that can be predicted by *Z* ($\pi_0 + \pi_1 Z_i$) and the problematic part of *X* (v_i) which is correlated with the error term

In the second stage, the dependent variable Y is regressed on the predicted value of X from the first stage.

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

- This means that our second stage regression only contains the unproblematic component of *X*.

TSLS in the General IV Model

With a single endogenous regressor, our equation of interest is:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i$$

The first stage regression regresses the endogenous variable *X* with the instruments *Z* and the exogenous variables *W*.

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i$$

With multiple endogenous variables, each one requires its own first stage regression.

STATA does all this for you!

TSLS Standard Errors

1. The standard errors reported in the second-stage regression will be incorrect because it is the second stage of a two stage process
   - STATA and other software programs will account for this and correct it if you use a TSLS specific command
2. Just as always, heteroskedasticity is a threat to our estimators thus it is important to use heteroskedasticity-robust calculations of the standard errors
   - Just a reminder, because homoskedasticity is a special case of heteroskedasticity, the heteroskedasticity-robust standard errors produce valid estimators whether the errors are heteroskedastic of homoscedastic

$$\text{Heteroskedasticity: } \sigma^2_{\hat{\beta}_1} = \frac{\text{var}\left[\left(X_i - \mu_X\right)u_i\right]}{n\left[\text{var}\left(X_i\right)\right]^2}$$

$$\text{Homoskedasticity: } \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{n\sigma^2_X}$$

Valid Instruments

There are 2 conditions for valid instruments:

1. Instrument relevance: corr(Z, X) ≠ 0
   - The instrument must correlated with the endogenous variable
2. Instrument exogeneity: corr(Z, u) = 0
   - The instrument must not be correlated with the error term (not correlated with any omitted variables)
   - We are trying to identify the variation in *X* that is uncorrelated with the error term!

Checking Instrument Relevance

Checking instrument relevance

- Instruments that do not explain very much of the variation in *X* are called **weak instruments**
  - Weak instruments lead to biased estimators, even in large samples
  - Ultimately, it means that you're not solving the problem that you set out to

The First-stage F-statistic (Specific to TSLS Estimation)

- When there is a single endogenous variable, we can compute the first-stage F-statistic
- The first-stage F-statistic tests the null hypothesis that coefficients on all of the instruments are equal to zero in the first stage
- The larger the F-statistic the better
- At least how large should the first-stage F-statistic be? 10.
- Anything lower indicates a weak instrument

Checking Instrument Relevance

Why does the first-stage F-statistic need to be greater than 10?

If $\beta_1^{OLS} - \beta_1$ is the bias of the OLS estimator, then we can show that the bias of the TSLS estimator is:

$$\frac{(E(\hat{\beta}_1^{TSLS}) - \beta_1) \approx (\beta_1^{OLS} - \beta_1)}{[E(F) - 1]}$$

Where the expected value of F is the first-stage F-statistic.

If $E(F) = 10$ then the bias of the TSLS estimator relative to the bias of the OLS estimator is $\frac{1}{9}$ which is small enough for reasonable conclusions.

Checking Instrument Exogeneity

If the instruments are not exogenous, then the estimators will not be consistent

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{cov(Z,u)}{cov(Z,X)} \neq \beta_1$$

- The estimate will converge in probability to something other than the population value
- Simply put, recall that the idea of IV regression is to capture the variation in *X* that is unrelated to the error term

Can we statistically test for exogeneity?

- **Only if** the coefficients are overidentified
- This is called the **overidentifying restrictions test** (The J-statistic)
- Chi-squared distribution with $m - k$ degrees of freedom (m is the number of instruments; k is the number of endogenous variables)
  - This is why we cannot preform the test if the coefficients are exactly identified

If the model is exactly identified then we must rely on our “expert knowledge”

The Overidentifying Restrictions Test

$$J - statistic = mF$$

- If you had two instruments, then you could run two different regressions using different instruments in each regression
- If the instruments are exogenous, then the two estimates would be very similar
- If the estimates are not similar, then there would be a problem
- The J-statistic makes this comparison for you
  - Under the null hypothesis all the instruments are **exogenous**
  - Rejecting the null of the overidentifying restrictions test means that we have invalid instruments

Does putting criminals in jail reduce crime?

$$Crime\ Rate = \beta_0 + \beta_1 Incarceration\ Rate_i + \beta_2 X_i + \cdots + \beta_k X_i$$

- There is a major threat of simultaneous causality bias

He used an instrument to fix the issue.

- What is something that affects incarceration but does not effect crime?
- His answer? The capacity of existing prisons. Why?
  - It takes time to build prisons and short-term capacity limitations force states to release prisoners prematurely (reduces incarceration rates)
  - He used lawsuits targeted at reducing prison overcrowding
  - He did not report the first-stage F-statistic but it did show an effect on incarcerations in his data
  - He used multiple instruments for different types of incarceration which allowed him to use the overidentifying restrictions test, which he failed to reject suggesting that his instruments are exogenous

Using IV Regression and TSLS, his estimated effect was three times larger than the effect estimated using OLS

Does aggressive treatment of heart attacks prolong lives?

Aggressively treating heart attack victims could save lives (McClellan, McNeil, Newhouse 1994)

- Specifically, the regressed the length of survival of the patient on a binary variable for the treatment plus control variables that affect mortality
- By simply estimating a regression with OLS we are at risk of biased estimators (due to a possibly unobservable omitted variable)
  - If they opt for the treatment based on unobserved factors relevant to health outcomes not in the data set, then the treatment decision will be correlated with the error term

We can eliminate any threat of bias by using an instrument that affects treatment but not the health outcome

- The researchers suggested using the difference between the distance from the heart attack patient's house to the treatment hospital and the distance to the nearest hospital of any kind
- If this relative distance affects the probability of receiving treatment, then is it relevant
- If it is randomly distributed across heart attack victims, then it is exogenous

Their TSLS estimates suggested that the treatment has a small, and possibly zero, effect on health outcomes where OLS suggested a large positive effect

Does aggressive treatment of heart attacks prolong lives?

- Their OLS regression used actual treatment as an independent variable, however they argued that actual treatment is itself the outcome of a decision by a patient and doctor, thus it is correlated with the error term
- TSLS, on the other hand, used **predicted treatment** as an independent variable, where the variation in predicted treatment is a function of the variation in the instrument; patients closer to a hospital offering the treatment are more likely to receive the treatment

This is interesting to note because…

- The IV regression estimates the effect of the treatment on a patient for whom distance is an important factor in the treatment decision
- This effect is likely to be different from that of a "typical" randomly selected patient
  - This could explain the greater estimated effectiveness in the OLS regressions
- It also suggests that we should look for instruments that affect the probability of treatment, but for reasons that are uncorrelated with the health outcomes

Thank you!

Source Information:

Stock, James H., and Mark W. Watson. *Introduction to Econometrics*. 3rd ed. Boston: Addison-Wesley, 2011.