14.32 Econometric Data Science
Professor Anna Mikusheva
September, 2019

Lectures 19-20

Instrumental Variable Regression

Instrumental Variable (IV) regression is a powerful and effective method to address multiple issues with OLS regression, such as OVB, simultaneous causality and error-in-variable bias.

Basic idea

Imagine that our goal is to estimate a causal effect from a change in X on Y and a proposed regression is

$$Y = \alpha + \beta X + e, \tag{1}$$

but the OLS in this setting would deliver invalid results if X is endogenous, that is, if X is correlated with the error term e.

An instrumental variable is a variable that satisfies two conditions:

(1) relevance: $cov(X, Z) \neq 0$

(2) exogeneity: $cov(Z, e) = 0$

We can use an instrumental variable in the following way: let us consider $cov(Y, Z)$ and use equation (1):

$$cov(Y, Z) = \beta cov(X, Z) + cov(e, Z) = \beta cov(X, Z).$$

Thus,

$$\beta = \frac{cov(Y, Z)}{cov(X, Z)}.$$

If we have an i.i.d. sample $(X_i, Z_i, Y_i), i = 1, ..., n$, then we can use this equation for estimation:

$$\widehat{\beta}_{IV} = \frac{s_{YZ}}{s_{XZ}} = \frac{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})Y_i}{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})X_i}. \tag{2}$$

Another idea: to isolate in an endogenous regressor X, the exogenous part which is due to variation in Z, and use it as a regressor.

Step 1. Regress X on Z:

$$X = \pi_0 + \pi_1 Z + v,$$

by OLS and produce estimates $\widehat{\pi}_0$ and $\widehat{\pi}_1$. Generate the exogenous part of X:

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

Step 2. Run an OLS regression of Y on $\widehat{X}$:

$$Y = \alpha + \beta \widehat{X} + e^*.$$

Call this estimator $\widehat{\beta}_{TSLS}$ (two-stage least squares)

Cite as: Anna Mikusheva, course materials for 14.32 Econometric Data Science, Fall 2019. Massachusetts Institute of Technology.

Claim Estimators $\widehat{\beta}_{IV}$, defined in equation (2), and the one obtained by the two-step procedure are identical.

Indeed,

$$\widehat{\beta}_{TSLS} = \frac{s_{Y\widehat{X}}}{s^2_{\widehat{X}}} = \frac{\widehat{\pi}_1 s_{YZ}}{\widehat{\pi}_1^2 s_Z^2}.$$

This is due to the fact that all variation in $\widehat{X}_i = \widehat{\pi}_0 + \widehat{\pi}_1 Z_i$ is due to the variation in Z_i. Also notice that $\widehat{\pi}_1 = \frac{s_{XZ}}{s_Z^2}$. Thus

$$\widehat{\beta}_{TSLS} = \frac{s_{YZ}}{\widehat{\pi}_1 s_Z^2} = \frac{s_{YZ}}{\frac{s_{XZ}}{s_Z^2} s_Z^2} = \frac{s_{YZ}}{s_{XZ}} = \widehat{\beta}_{IV}.$$

□

Statement. Under relevance and exogeneity of the instrument Z, estimator $\widehat{\beta}_{IV}$ is consistent and asymptotically gaussian.

Indeed,

$$\widehat{\beta}_{IV} = \frac{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})Y_i}{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})X_i} = \frac{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})(\alpha + \beta X_i + e_i)}{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})X_i} = \beta + \frac{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})e_i}{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})X_i}.$$

Exogeneity of instruments and the Law of Large numbers would guarantee that $\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})e_i \to^p 0$, while the relevance and the Law of Large Numbers give $\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})X_i \to^p cov(Z, X) \neq 0$. Thus the estimator is consistent.

Now,

$$\sqrt{n}(\widehat{\beta}_{IV} - \beta) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_i - \bar{Z})e_i}{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z})X_i}.$$

We can show that due to the Central Limit Theorem, the numerator is asymptotically gaussian. □.

Attention! If you run an estimation in STATA as two separate stages, the automatically produced standard errors are incorrect, as they do not take into account the uncertainty of the first stage.

General IV Model

We will add to the previous model multiple regressors, multiple instruments and multiple controls. Assume a regression of interest is:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \beta_{k+1} W_1 + ... + \beta_{k+r} W_{k+r} + e$$

Assume that $Z_1, ..., Z_m$ are instruments, that is, they satisfy the following two conditions:

(1) exogeneity: $E[eZ_1] = 0, ..., E[eZ_m] = 0$;

(2) relevance: $E\left[(1, Z_1, ..., Z_m, W_1, ..., W_r)'(1, X_1, ..., X_k, W_1, ..., W_r)\right]$ is matrix of size $(1 + m + r) \times (1 + k + r)$ and of rank $1 + k + r$. It means that the instrument Z's can produce a full rank variation in X' even after conditioning on W.

The model is just identified if $m = k$ and over-identified if $m > k$.

Plan for two stage least squares (TSLS) :

Step 1. Run OLS regressions of each of the X's on all of the Z's and all of the W's. That is:

$$X_1 = \pi_0^1 + \pi_1^1 Z_1 + ... + \pi_m^1 Z_m + \pi_{m+1}^1 W_1 + ... + \pi_{m+r}^1 W_r + u^1$$
$$...$$
$$X_k = \pi_0^k + \pi_1^k Z_1 + ... + \pi_m^k Z_m + \pi_{m+1}^k W_1 + ... + \pi_{m+r}^k W_r + u^k.$$

Denote all estimates as $\widehat{\pi}_i^j$. Produce the predicted values:

$$\widehat{X}_j = \widehat{\pi}_0^j + \widehat{\pi}_1^j Z_1 + ... + \widehat{\pi}_m^j Z_m + \widehat{\pi}_{m+1}^j W_1 + ... + \widehat{\pi}_{m+r}^j W_r.$$

Step 2. Run an OLS regression of Y on $\widehat{X}_1, ..., \widehat{X}_k$ and on $W_1, ..., W_r$:

$$Y = \beta_0 + \beta_1 \widehat{X}_1 + ... + \beta_k \widehat{X}_k + \beta_{k+1} W_1 + ... + \beta_{k+r} W_{k+r} + e.$$

All estimates collected into $k + r + 1$-dimensional vector β are referred to as TSLS estimate $\widehat{\beta}_{TSLS}$.

Assumptions of IV model

Assumption 1 $E[e|W_1, ..., W_r] = 0$;

Assumption 2 $(Y_i, X_{1,i}, ..., X_{k,i}, W_{1,i}, ..., W_{r,i}, Z_{1,i}, ..., Z_{m,i})$ are i.i.d. draws from joint distribution;

Assumption 3 No ouliers;

Assumption 4 No perfect multicollinearity;

Assumption 5 Relevance and exogeneity of instruments hold (conditions (1) and (2) above).

Statement Under the assumptions above, $\widehat{\beta}_{TSLS}$ is consistent and asymptotically gaussian.

Relevance: weak instruments

There are two important assumptions about instruments: exogeneity and relevance. Let's discuss whether we can in some way test them and what happens if they are violated. We start with exogeneity.

In the very simplistic case with 1 endogenous regressor, 1 instrument and no controls we had

$$\widehat{\beta}_{TSLS} = \frac{s_{YZ}}{s_{XZ}}.$$

It is hugely important both for consistency and for asymptotic gaussianity that the limit of the denominator $cov(Z, X) \neq 0$, as the devision by zero invalidates all derivations. Aparently, even if $cov(X, Z)$ is not zero, but is close to zero we may experience problems with statistical inferences. In particular, if the estimation mistake in s_{XZ} is large compared to its asymptotic value $cov(X, Z)$, then random variable $\frac{1}{s_{ZX}}$ will be very volatile (as s_{XZ} will often give values close to zero). Such a case is called 'weak instruments', and often implies invalid classical inferences. In particular, if instruments are weak we may have:

- a very biased TSLS estimator;
- tests based on t-statistics may be of the incorrect size;
- standard confidence sets may have low coverage.

Similar phenomena may present themselves in a general model (with multiple instruments/controls) as well. We have a partial test for these phenomena.

Assume that the regression of interest involves a single endogenous regressor and several controls:

$$Y = \beta_0 + \beta_1 X + \beta_2 W_1 + ... + \beta_{1+r} W_r + e,$$

with instruments $Z_1, ..., Z_m$. The proposed test for weak identification is based on the first stage regression:

$$X = \pi_0 + \pi_1 Z_1 + ... + \pi_m Z_m + \pi_{m+1} W_1 + ... + \pi_{m+r} W_r + u.$$

First, run OLS for such a first-stage regression, then calculate the F-statistic for testing the null hypothesis

$$H_0 : \pi_1 = ... = \pi_m = 0.$$

However, we do not use the usual critical values for this test, but rather compare the F-statistic with 10. If it is larger than 10, your instruments are strong enough for reasonable inferences. However, if $F < 10$ you may be in danger of encountering weak instruments.

What to do if you find weak instruments? There are weak-instrument-robust procedures which produce more reliable results, but they are outside of the scope of this course.

Test for exogeneity: J-test

Can we test for exogeneity? We can test it partially. In the just identified case it is untestable, but if the model is over-identified (more instruments than endogenous variables), then we can test whether the instruments agree.

Idea. Imagine a simple model with one endogenous regressor and two instruments:

$$Y = \alpha + \beta X + e,$$

and instruments Z_1 and Z_2. One instrument would have been enough to obtain a consistent estimator for β. Let's construct two different estimators: $\widehat{\beta}_1$ is the IV estimator using only Z_1 as an instrument, and $\widehat{\beta}_2$ is the IV estimator using only Z_2 as an instrument. If both instruments are exogenous, then $\widehat{\beta}_1 \approx \widehat{\beta}_2$. We can formally test that they estimate the same value. That is what is formally (and efficiently) done in a J-test.

Plan for the J-test

1. Estimate $\widehat{\beta}$ via TSLS.
2. Calculate residuals from the IV regression:

$$\widehat{e}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1i} - ... - \widehat{\beta}_k X_{k,i} - \widehat{\beta}_{k+1} W_{1i} + ... + \widehat{\beta}_{k+r} W_{ri}$$

   (notice, it is NOT a residual of the second stage regression, as is the latter one using $\widehat{X}$'s)
3. Run OLS (with homoskedastic errors) on the following regression:

$$\widehat{e} = \gamma_0 + \gamma_1 Z_1 + ... + \gamma_m Z_m + \gamma_{m+1} W_1 + ... + \gamma_{m+r} W_r + u.$$

4. Calculate in the last regression the F-statistic for the null $H_0 : \gamma_1 = ... = \gamma_m = 0$.
5. $J = m \cdot F$. Compare it to the critical values of χ^2_{m-k}. If it exceeds the critical value, then the hypothesis of exogeneity is rejected.

Word of caution.

- If the$J$-test rejects, it does not say which instrument is endogenous. It only says that there is a disagreement between the instruments
- If the J-test does not reject, it does not mean that all instruments are exogenous. The power of the J-test is poor in some directions.
- It does not make sense at all to choose instruments by doing multiple J-testing and then selecting the not-rejected set. This is a very poor strategy, and should not be applied.
- It is common to report the results of a J-test for overidentified models as a sanity check.