Econometrics I - Instrumental Variables

Ryan T. Godwin

University of Manitoba

We have been assuming either that the columns of X are non-random; or that the sequence $\{\boldsymbol{x}_i, \epsilon_i\}$ is independent. Often, neither of these assumptions are tenable. This implies that $\text{plim}\left(\frac{1}{n}X'\epsilon\right) \neq \mathbf{0}$, and then the LS estimator is inconsistent.

Prove that the LS estimator is inconsistent when $\text{plim}\left(\frac{1}{n}X'\epsilon\right) \neq \mathbf{0}$.

Inconsistency of the LS estimator is a serious issue. Inconsistency means that the estimation results can be wrong, and that no matter how large n is, the problem does not go away. If the LS estimator is inconsistent, it should not be used. In this chapter, we motivate the situation through a *missing variable* that is *correlated* with a variable(s) in the X matrix. Then, we discuss instrumental variable (IV) estimation as a potential solution to the problem.

Correlation between the error term and regressors

Several ways in which to motivate the situation where X and $\boldsymbol{\epsilon}$ are correlated.[1] The very problem which motivated IV estimation is the simultaneity of price and quantity through demand and supply equations in a competitive market. For example, a linear model of demand and supply is:

$$q_i = \gamma_d p_i + \boldsymbol{x}_i^d \boldsymbol{\beta}_d + \epsilon_i^d \tag{1}$$
$$q_i = \gamma_s p_i + \boldsymbol{x}_i^s \boldsymbol{\beta}_s + \epsilon_i^s \tag{2}$$

where (1) is demand and (2) is supply, the x variables are exogenous, and the γ are the slopes of the demand or supply curves. There are two equations in two unknowns (p_i and q_i), and it is easy to solve for the equilibrium values by, for example, solving for p_i in equation 2 and substituting into equation 1. In doing so, we see that the solutions for p_i and q_i depends not only on both x_i^d and x_i^s, but also on both ϵ_i^d and ϵ_i^s. In any linear simultaneous equations model, the endogenous variables are necessarily correlated with the error terms. This is a violation of A.5, leading to inconsistency of the LS estimator.

[1] Davidson and MacKinnon (2004) discuss “Errors in Variables” (pg. 312), and “Simultaneous Equations” (pg. 314).

Another way to motivate a situation where X and $\boldsymbol{\epsilon}$ are dependent, is to consider omitted, unobservable, or *missing* variables, M. Consider also that these missing variables are correlated with the regressors X and the dependent variable $\boldsymbol{y}$. The true population model is:

$$\boldsymbol{y} = X\boldsymbol{\beta} + M\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

Since M is unobservable, the observable model that we can estimate is:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3}$$

Notice that in 3, $\boldsymbol{\epsilon}$ contains $M\boldsymbol{\gamma}$, so that X and $\boldsymbol{\epsilon}$ are not independent (X is endogenous), since X and M are correlated.

Prove that OLS is biased and inconsistent under this data generating process.

In such cases of endogeneity, we want a safe (consistent) way of estimating $\boldsymbol{\beta}$. One general family of such estimators is the family of Instrumental Variables (IV) estimators.

Instrumental variable

A variable(s), Z, qualifies as an instrument if it satisfies two conditions. An instrumental variable, Z, must be:

1. Correlated with the endogenous variables X.
   - This is sometimes called the “relevance” of an IV.
   - This condition can be tested.
2. Uncorrelated with the error term, or equivalently, uncorrelated with the dependent variable other than through its correlation with X.
   - This is sometimes called the “exclusion” restriction.
   - This restriction cannot be easily tested.

The “exclusion” restriction implies k moment conditions, which allows us to derive the IV estimator:

$$E(Z'\boldsymbol{\epsilon}) = \mathbf{0}$$

or

$$E(Z'\boldsymbol{\epsilon}) = E(Z'\boldsymbol{y} - Z'X\boldsymbol{\beta}) = E(Z'\boldsymbol{y}) - E(Z'X)\boldsymbol{\beta} = 0$$

Solving the k moment conditions for $\boldsymbol{\beta}$, replacing the expected values with sample averages, yields the IV estimator:

$$\hat{\boldsymbol{b}}_{IV} = \left(n^{-1}\sum_{i=1}^{n}\mathbf{z}_i'\mathbf{x}_i\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\mathbf{z}_i'y_i\right)$$

or, in matrix notation:

$$b_{IV} = (Z'X)^{-1}Z'\boldsymbol{y}$$

In general, this estimator is biased. We can show it is consistent, however:

$$\begin{array}{l}
\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\text{plim}\left(\frac{1}{n}X'X\right) = Q \quad ; \quad \text{p.d. and finite} \\
\text{plim}\left(\frac{1}{n}X'\boldsymbol{\epsilon}\right) = \gamma \neq 0
\end{array}$$

Full rank of the instrument matrix, the *relevancy* of the I.V., and the *exclusion restriction* imply respectively that :

$$\begin{array}{l}
\text{plim}\left(\frac{1}{n}Z'Z\right) = Q_{ZZ} \quad ; \quad \text{p.d. and finite} \\
\text{plim}\left(\frac{1}{n}Z'X\right) = Q_{ZX} \quad ; \quad \text{p.d. and finite} \\
\text{plim}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right) = \mathbf{0}
\end{array}$$

Then, the IV estimator is *consistent*:

$$\begin{aligned}
\boldsymbol{b}_{IV} &= \left(Z'X\right)^{-1} Z'\boldsymbol{y} = \left(Z'X\right)^{-1} Z'(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \left(Z'X\right)^{-1} Z'X\boldsymbol{\beta} + \left(Z'X\right)^{-1} Z'\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + \left(Z'X\right)^{-1} Z'\boldsymbol{\epsilon} \\
&= \boldsymbol{\beta} + \left(\frac{1}{n}Z'X\right)^{-1} \left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right)
\end{aligned}$$

and so:

$$\begin{aligned}
\operatorname{plim}\left(\boldsymbol{b}_{IV}\right) &= \boldsymbol{\beta} + \left[\operatorname{plim}\left(\frac{1}{n}Z'X\right)\right]^{-1} \operatorname{plim}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right) \\
&= \boldsymbol{\beta} + Q_{ZX}^{-1}\mathbf{0} = \boldsymbol{\beta}
\end{aligned}$$

Note that choosing different Z matrices generates different members of the IV family.

Although we won't derive the full asymptotic distribution of the IV estimator, note that it can be expressed as:

$$\sqrt{n}\,(\boldsymbol{b}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} N\left[\mathbf{0}, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1}\right]$$

where $Q_{XZ} = Q'_{ZX}$.

Questions: How would you estimate this asymptotic covariance matrix? How would you estimate the covariance matrix for $\boldsymbol{b}_{IV}$?

Interpreting IV as two-stage least squares (2SLS)

IV estimation is also called two-stage least squares. Before modern computers, IV estimates were calculated by two (or more) least squares estimations. Two stage least squares offers an intuitive interpretation of the IV estimator. The idea behind IV estimation is that the instrument Z may be used to “extract” the “clean” variation from the endogenous variables X. When X is endogenous, a change in X is associated with a change in $\boldsymbol{\epsilon}$, so it is impossible to *identify* how much of the observed change in X led to the observed change in $\boldsymbol{y}$. However, if we could *extract* the variation in X that is uncorrelated with variation in $\boldsymbol{\epsilon}$, then we could use this *clean* variation to estimate $\boldsymbol{\beta}$ consistently. If we find an instrument Z that is correlated with X but uncorrelated with $\boldsymbol{\epsilon}$, then we can use changes in X *due to changes in Z only*, to extract this clean variation.

1^{st} stage of 2SLS

In the first stage, we regress each column of X on Z using LS, and get $\hat{X}$. That is, we get $\hat{X} = P_Z X$.

- $\hat{X} = P_Z X$ contains the variation in X due to Z *only*.
- $P_Z X$ is not correlated with $\boldsymbol{\epsilon}$.
- Recall that $\hat{X} = P_Z X = Z\left(Z'Z\right)^{-1} Z'X$.

Question: Why is $P_Z X$ uncorrelated with $\boldsymbol{\epsilon}$?

2^{nd} stage of 2SLS

In the second stage, we estimate the model: $y = \hat{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = P_Z X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, using LS. Applying the LS formula to this model, we get:

$$\boldsymbol{b}_{IV} = \left[X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1}X'Z\left(Z'Z\right)^{-1}Z'\boldsymbol{y}$$

or just:

$$\boldsymbol{b}_{IV} = \left[X'P_Z X\right]^{-1}X'P_Z\boldsymbol{y}$$

This is the *Generalized* I.V. estimator of $\boldsymbol{\beta}$. We can actually use more instruments than regressors (the “Over-Identified” case). Why might we want to do this? (Efficiency). Note that if X and Z have the same dimensions, then the generalized estimator collapses to the simple one.

Now, let's check the consistency of this Generalized IV estimator.

$$\begin{aligned}\boldsymbol{b_{IV}} &= \left[X'P_ZX\right]^{-1}X'P_Z\boldsymbol{y} = \left[X'P_ZX\right]^{-1}X'P_Z(X\boldsymbol{\beta}+\boldsymbol{\epsilon})\\ &= \left[X'P_ZX\right]^{-1}X'P_ZX\boldsymbol{\beta} + \left[X'P_ZX\right]^{-1}X'P_Z\boldsymbol{\epsilon}\\ &= \boldsymbol{\beta} + \left[X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1}X'Z\left(Z'Z\right)^{-1}Z'\boldsymbol{\epsilon}\end{aligned}$$

So,

$$\boldsymbol{b_{IV}} = \boldsymbol{\beta} + \left[\left(\frac{1}{n}X'Z\right)\left(\frac{1}{n}Z'Z\right)^{-1}\left(\frac{1}{n}Z'X\right)\right]^{-1}\left(\frac{1}{n}X'Z\right)\left(\frac{1}{n}Z'Z\right)^{-1}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right.$$

Modify our assumptions. We have a (random) $(n \times L)$ matrix, Z, such that:

1. $\text{plim}\left(\frac{1}{n}Z'Z\right) = Q_{ZZ}$; $(L \times L)$, p.s.d. and finite
2. $\text{plim}\left(\frac{1}{n}Z'X\right) = Q_{ZX}$; $(L \times k)$, rank $= k$, and finite
3. $\text{plim}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right) = \mathbf{0}$; $(L \times 1)$

So,

$$\operatorname{plim}(\boldsymbol{b_{IV}}) = \boldsymbol{\beta} + \left[Q_{XZ}Q_{ZZ}^{-1}Q_{ZX}\right]^{-1} Q_{XZ}Q_{ZZ}^{-1}\mathbf{0} = \boldsymbol{\beta} \quad ; \quad \text{consistent}$$

Similar to before, a *consistent estimator* of σ^2 is

$$s_{IV}^2 = (\boldsymbol{y} - X\boldsymbol{b_{IV}})'(\boldsymbol{y} - X\boldsymbol{b_{IV}})/n$$

- Recall that each choice of Z leads to a *different* IV estimator.
- Z must be chosen in way that ensures consistency of the IV estimator.
- How might we choose a suitable set of instruments, in practice?
- If we have several “valid” sets of instruments, how might we choose between them?

For the “simple” IV regression model, recall that:

$$\sqrt{n}\left(\boldsymbol{b_{IV}}-\boldsymbol{\beta}\right)\xrightarrow{d}N\left[\mathbf{0},\sigma^2 Q_{ZX}^{-1}Q_{ZZ}Q_{XZ}^{-1}\right]$$

so if $k=1$,

$$Q_{ZZ}=\text{plim}\left(n^{-1}\sum_{i=1}^{n}z_i^2\right)$$

and

$$Q_{ZX}=\text{plim}\left(n^{-1}\sum_{i=1}^{n}z_i x_i\right)=Q_{XZ}$$

The asymptotic efficiency of $\boldsymbol{b}_{IV}$ will be higher, the more highly correlated are Z and X, asymptotically. We need to find instruments that are uncorrelated with the errors, but highly correlated with the regressors (asymptotically). This is not easy to do! A good instrument comes from an intimate understanding of the economics driving the regressor of interest. A good survey of some classic IV papers may be found here[2].

[2] Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4), 69-85.

Testing if IV estimation is needed

This is a good situation to be in: it means you have found a potentially valid instrument! Now, should you use it? Recall that LS is BLUE, so we should use it where possible. Consider the following.

- Why does LS fail, and when do we need IV?
- If $\text{plim}\left(\frac{1}{n}X'\boldsymbol{\epsilon}\right) \neq \mathbf{0}$.
- We can *test* to see if this is a problem, and decide if we should use LS or IV.

The Hausman test

We want to test:

$$H_0 : \text{plim}\left(\frac{1}{n}X'\epsilon\right) = \mathbf{0} \quad \text{vs.} \quad H_A : \text{plim}\left(\frac{1}{n}X'\epsilon\right) \neq \mathbf{0}$$

- If we reject H_0, we will use IV estimation.
- If we cannot reject H_0, we'll use LS estimation.
- The Hausman test is a general "testing strategy" that can be applied in many situations, not just for this particular situation!
- Basic idea: construct 2 estimators of $\boldsymbol{\beta}$:
  1. $\boldsymbol{b}$ (LS estimator): the estimator is both *consistent* and *asymptotically efficient* if H_0 is true.
  2. $\boldsymbol{b}_{IV}$: the estimator is at least *consistent*, even if H_0 is false.
- If H_0 is true, we'd expect $(\boldsymbol{b}_{IV} - \boldsymbol{b})$ to be "small", at least for large n, as both estimators are consistent in that case.

- Hausman shows that $\hat{V}\left(\boldsymbol{b}_{IV}-\boldsymbol{b}\right)=\hat{V}\left(\boldsymbol{b}_{IV}\right)-\hat{V}\left(\boldsymbol{b}\right)$, if H_0 is true.
- So, the test statistic is,
  $$H=\left(\boldsymbol{b}_{IV}-\boldsymbol{b}\right)'\left[\hat{V}\left(\boldsymbol{b}_{IV}\right)-\hat{V}\left(\boldsymbol{b}\right)\right]^{-1}\left(\boldsymbol{b}_{IV}-\boldsymbol{b}\right).$$
- $H\xrightarrow{d}\chi^2_I$, if H_0 is true.
- Here, J is the number of columns in X which may be correlated with the errors, and for which we need instruments.

Note that there are other asymptotically equivalent tests for the same null and alternative hypothesis; for example, the Durbin-Wu test.

Testing the exogeneity of instruments

The key assumption that ensures the consistency of IV estimators is that

$$\text{plim}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right) = \mathbf{0}.$$

This condition involves the *unobservable* $\boldsymbol{\epsilon}$. It is difficult to test. If there are more instruments than regressors (the over-identified case) than the Sargan-Hansen or J test may be used. In an applied economics paper, establishing the exogeneity of the instruments is more likely a matter of arguing for this key assumption through an understanding of the variables, rather than relying on a statistical test.

Weak instruments

Problems arise if the instruments are *not* well correlated with the regressors (not relevant).

- These problems go beyond loss of asymptotic efficiency.
- Small-sample bias of IV estimator can be greater than that of LS!
- Sampling distribution of IV estimator can be bi-modal!
- Fortunately, we can again test to see if we have these problems.

Tests aimed at detecting weak instruments revolve around detecting correlation between the instruments and endogenous regressors. Although it doesn't quite work, we could draw an analogy to the R^2, or the significance, of the 1$^{\text{st}}$ stage in 2SLS.

Empirical example

Let's look at data from Card (1993).[3]

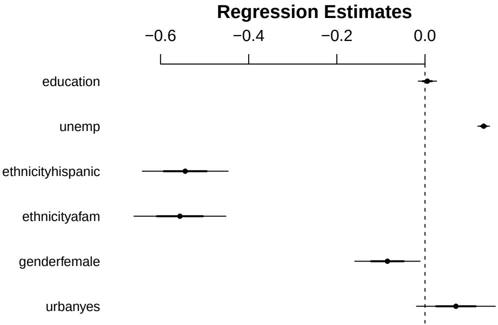
- Data contains *wage*, *years of education*, and demographic variables.
- Goal: estimate the returns to education in terms of *wage*.
- Problem: ability (intelligence) may be correlated with (cause) both wage and education.
- Since ability is unobservable, it is contained in the error term.
- The education variable is then correlated with the error term (endogenous).
- LS estimation of the returns to education may be inconsistent.

[3] Card, D. (1993). *Using geographic variation in college proximity to estimate the return to schooling* (No. w4483). National Bureau of Economic Research.

First, let's try LS (see the estimation results visualized in figure 1).

```r
library(AER)
library(arm)
data("CollegeDistance")
ls <- lm(wage ~ urban + gender + ethnicity + unemp +
    education, data = CollegeDistance)
coefplot(ls)
```

Figure: Results of LS regression using Card (1993) data. Dependent variable is *wage*. Notice that *education* is statistically insignificant.



Now let's try using *distance from college* (while attending high school) as an instrument for education. The argument for the validity of this instrument is that *distance from college* is correlated with *education*, since the closer a student is, the cheaper it is to get an education. For the instrument to be valid, we require that *distance* and *education* be correlated:

```r
summary(lm(education ~ distance, data = CollegeDistance))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.93861    0.03290 423.683  < 2e-16 ***
distance    -0.07258    0.01127  -6.441  1.3e-10 ***
---
```

While distance appears to be statistically significant, this isn't quite enough to test for validity (a testing problem we won't address here).

Now, let's estimate the model using IV.

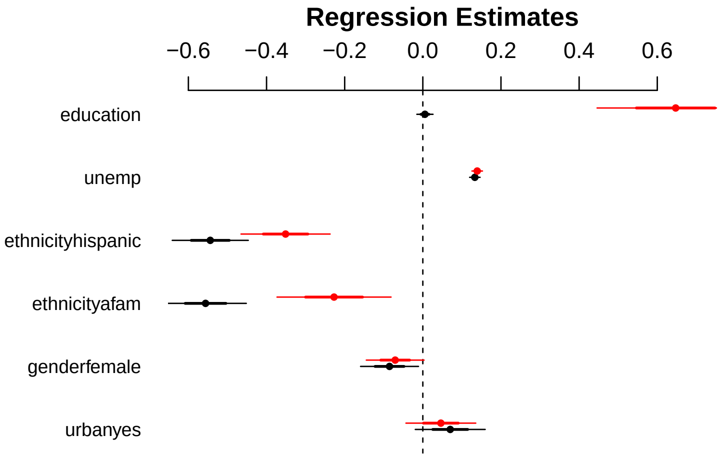
```r
library(ivreg)
iv <- ivreg(wage ~ urban + gender + ethnicity + unemp +
     education | urban + gender + ethnicity + unemp +
     distance, data = CollegeDistance)
summary(iv, diagnostics = TRUE)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -0.35903    1.90830  -0.188   0.8508    
urbanyes           0.04614    0.06039   0.764   0.4449    
genderfemale      -0.07075    0.04997  -1.416   0.1569    
ethnicityafam     -0.22724    0.09863  -2.304   0.0213 *  
ethnicityhispanic -0.35129    0.07706  -4.559 5.28e-06 ***
unemp              0.13916    0.00912  15.259  < 2e-16 ***
education          0.64710    0.13594   4.760 1.99e-06 ***

Diagnostic tests:
                  df1  df2 statistic  p-value    
Weak instruments    1 4732     50.31 1.51e-12 ***
Wu-Hausman          1 4731     41.12 1.57e-10 ***
Sargan              0   NA        NA       NA    
```

In the `ivreg` function, the population model precedes the vertical line —, and the instruments follow the — (each exogenous regressor is an instrument for itself). Notice the “Weak instruments” and “Wu-Hausman” tests. What are these *p*-values telling you? The “Sargan” test is only applicable when we have an “over-identified” IV estimator. See figure 2 for a visual comparison with LS.

Figure: Results of LS and IV (in red) regression using Card (1993) data. Dependent variable is *wage*; *distance from college* is an instrument for *education*. Notice that the returns to education are now significant!


Regression Estimates
education
unemp
ethnicityhispanic
ethnicityafam
genderfemale
urbanyes