Introduction to Econometrics: Recitation 9

Seung-hun Lee

1 Instrumental Variable Regression

1.1 Motivation

Recall that when there is a possibility of three of the internal validity threat factors - namely omitted variable bias, measurement error, and simultaneity bias - then we run into the case where the OLS estimation gives us biased estimates. Specifically, these cases bias our result because the error term u is no longer expected to be 0 conditional on X ($E[u|x] \neq 0$).

This is where **instrumental variables regression** comes in handy. Instrumental variables allow us to eliminate bias in the following sense: When you have an independent variable X, there are parts of this variable that are correlated with u and the other parts that are independent of u. If we are able to find Z that is correlated with X but not with u, using this Z variable would allow you to sort variable X into what is correlate with u and what is not. Then, using the part that is not correlated with u, variable Z allows us to get unbiased estimates.

This approach can be very useful in so many settings. Most of our independent variables are **endogenous** in the sense that X is not a given, but determined as a result of *choice* of individuals, along with dependent variable Y. For instance, If Y variable is test score and X is time spent studying, it is possible that the estimates of the effect of time spent studying can be biased. Think about the case where you see your friend (or rival, whichever you prefer) get a really high score. After you know that this person spends a lot of time studying, you decide to study more. Effectively, you see your friend's Y and that affects your selection of X - simultaneity bias. Alternatively, there might be factors uncontrolled by this regression affecting your choice of study hours - maybe hours spent commuting to schools. Lastly, the measure of X might be inaccurate - for instance, there might be a case

where you think you are studying but in reality, it is not. These cases can be addressed by finding some variables that are related with time spent studying but is not correlated with the error term.

1.2 IV conditions

Unfortunately, not every IV is found easily (in fact, you might be able to rebut my examples above). Finding variables that qualify as IV is challenging. The conditions they should satisfy are

Relevance : Variable Z satisfies relevancy condition if $cov(X, Z) \neq 0$

Exogeneity : Variable Z satisfies exogeneity condition if $cov(Z, u) = 0$

In words, variable Z should be somewhat correlated with the original independent variable X. Additionally, variable Z should not be correlated with u. The last part seems a bit challenging to conceptualize. So I will introduce another way of thinking about this condition.

Exclusion : Variable Z satisfies exclusion condition if it affects Y only through X. In other words, when X is controlled for, Z alone should not affect Y.

So how does this condition come in? Suppose that your model is (I skip subscript i for convenience) $Y = \beta_0 + \beta_1 X + u$. Then $cov(Z, u) = 0$ can be re-written as

$$\begin{aligned} cov(Z, u) &= cov(Z, Y - \beta_0 - \beta_1 X) \\ &= cov(Z, Y) - cov(Z, \beta_0) - cov(Z, \beta_1 X) \end{aligned}$$

By imposing $cov(Z, u) = 0$ and the fact that the covariance of a random variable with a fixed constant is 0, I can get that

$$cov(Z, Y) = cov(Z, \beta_1 X)$$

This condition means that Z is correlated with Y *only through* X. If there is something else other than $cov(Z, \beta_1 X)$ in the right hand side, then this is a violation of the exclusion condition because there is a covariance between Z and Y that is not explained by X. If this happens, the IV estimation fails to be accurate.

1.3 IV estmiation

There are many ways to estimate IV. I will be introducing **Two stage least squares**, **covariance method**, and **derivations from the reduced form**.

1.3.1 Two stage least squares

From the disussion in the beginning, IV estimation allows us to separate X into the part that is correlated with u and uncorrelated with u. This methods explicitly shows how that is done. The steps for carrying out this approach is as follows:

1. **Regress with X as dependent, Z as independent variable.** You will get an equation that looks like
$$X = \delta_0 + \delta_1 Z + v$$
From this regression, obtain the predicted values of X, denoted as $\hat{X} = \hat{\delta}_0 + \hat{\delta}_1 Z$. This $\hat{X}$ is the part that is related with Z but is uncorrelated with X.

2. **Regress with Y as dependent, $\hat{X}$ as independent variable.** Regressing with this $\hat{X}$ will satisfy the $E[u|\hat{X}] = 0$ condition, as the $\hat{X}$ is uncorrelated with u, as obtained in the previous step. Thus, your regression equation looks like this:
$$Y = \beta_0 + \beta_1 \hat{X} + u$$
Then you run a OLS regression on the above equation and get the two stage least squares estimator $\hat{\beta}_{\text{TSLS}}$.

There is one caveat to this method. In particular, you should be careful about what standard errors you are using. If you take the second stage standard error as is, then you end up with wrong standard errors. This is because the second stage standard error does not take into account that you obtained $\hat{X}$ using a separate (first stage) regression. The following example should make things clear. Going back to the regression on cigarette demand in 1995, we are interested in how demand for cigarettes is affected by income and price of cigarettes. In other words,

$$\ln Q_i^d = \beta_0 + \beta_1 \ln P_i^d + u_i$$

Because quantity demanded can affect price of cigarettes, there is a simultaneity bias. To get around this, we use general sales tax per pack. The following figures demonstrate two different standard errors.

```
. regress lpackpc xhat if year==1995, vce(robust)

Linear regression                               Number of obs     =         48
                                                F(1, 46)          =      10.54
                                                Prob > F          =     0.0022
                                                R-squared         =     0.1525
                                                Root MSE          =     .22645

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        xhat |  -1.083586   .3336948    -3.25   0.002    -1.755279   -.4118936
       _cons |   9.298537   1.597118     5.82   0.000     6.083705    12.51337
------------------------------------------------------------------------------

. ** ivregress
. ivregress 2sls lpackpc (lravgprs=rstax) if year == 1995, vce(robust)

Instrumental variables (2SLS) regression          Number of obs   =         48
                                                  Wald chi2(1)    =      12.05
                                                  Prob > chi2     =     0.0005
                                                  R-squared       =     0.4011
                                                  Root MSE        =     .18635

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.083587   .3122037    -3.47   0.001    -1.695495    -.471679
       _cons |    9.29854   1.496144     6.22   0.000     6.366152    12.23093
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   rstax
```

Notice that the standard errors are different. Therefore, if you want to work with correct standard errors, you must use `ivregress 2sls y x1 (x2 = z)` command.

1.3.2 Covariance method

Go back to the part where we discussed the exclusion condition. From there, get

$$cov(Z, Y) = cov(Z, \beta_1 X) \implies cov(Z, Y) = \beta_1 cov(Z, X)$$

From this, we can get

$$\hat{\beta}_1 = \frac{cov(Z, Y)}{cov(Z, X)}$$

where division is possible because we require a relevancy condition ($cov(Z, X) \neq 0$)

1.3.3 Derivations from the reduced form

This requires some amount of algebra. Denote the two equations as

$$X = \pi_0 + \pi_1 Z + v \text{ (where } cov(Z, v) = 0)$$
$$Y = \gamma_0 + \gamma_1 Z + w \text{ (where } cov(Z, w) = 0)$$

When you rewrite the first equation in terms of Z, you get

$$Z = \frac{X}{\pi_1} - \frac{\pi_0}{\pi_1} - \frac{v}{\pi_1}$$

Then plug this into the second equation. Reorganizing this equation, you should get

$$Y = \left(\gamma_0 - \frac{\pi_0 \gamma_1}{\pi_1}\right) + \left(\frac{\gamma_1}{\pi_1}\right) X + \left(w - \frac{\gamma_1}{\pi_1} u\right)$$

What this gets you is that β_1 from the equation where we had X as independent variable is $\beta_1 = \frac{\gamma_1}{\pi_1}$. This also implies that when Z rises by 1 unit, X exogenously changes by π_1 units and Y changes by γ_1 as a result. In other words, 1 unit of exogenous change in X translates to $\beta_1 = \frac{\gamma_1}{\pi_1}$ units of change in Y.

1.4 Properties of IV estimators

Note that the IV estimator can be written in terms of the ratios of the two covariances. In sample analogs, it is written as

$$\hat{\beta}_{1,\text{TSLS}} = \frac{s_{zy}}{s_{zx}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})}$$

Note that $\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})$ is not an unbiased estimator for the covariance of Z, Y (and similar for $\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})$) If you have large samples, however,

then $\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})$ is a **consistent** for covariance of Z, Y[1]. In mathematical terms, $s_{YZ} \xrightarrow{p} cov(Y, Z), s_{XZ} \xrightarrow{p} cov(X, Z)$. With these properties, we can say that the IV estimator of β_1, which is $\hat{\beta}_{1,\text{TSLS}}$ is a consistent estimator for the true β_1

The notion of IV estimation does not change much even in multiple regression settings. Suppose that we have

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \delta_1 W_{1i} + ... + \delta_l W_{li} + u_i$$

where X variables are endogenous and W variables are exogenous. Assume that we have found a total of m (not necessarily, equal. We'll save the discussion to overidentification test) variables that could qualify as IVs. Additionally, IV regression requires these assumptions

IV1 $E[u_i|W_{1i}, ..., W_{li}] = 0$ (At least for exogenous variables, this is satisfied)

IV2 $(Y_i, X_{1i}, .., X_{ki}, W_{1i}, .., W_{li}, Z_{1i}, ..., Z_{mi})$ are IID

IV3 The Y, X, W, Z variables all have nonzero finite 4th moments

IV4 The instruments are valid. That is $cov(Z_{ji}, u_i) = 0$ for all $j = 1, ..., m$ and relevancy conditions are satisfied for all Z's.

1.5 Identification issues

In a general sense, a parameter is **identified** if different values of the parameter produce different distributions of the data. In other words, there is a one-to-one matching of the parameters and the distributions. For instance, A normal distribution with mean μ and variance σ^2 is identified because different values of μ, σ^2 produce different distributions. If it is the case that the same distribution can be obtained from different parameter values, we say that the parameters are not identified.

In the context of IV's, the parameters we are interested are coefficients for the independent variables. This depends on how much instrument variables (m in the previous section) we have relative to the endogenous variables (k). There are three possible cases

[1]However, $\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})$ is unbiased. For more information refer to https://math.stackexchange.com/questions/2019122/unbiased-estimate-of-the-covariance

Just-identified : When $m = k$. There are just enough instruments to identify k endogenous variables

Overidentified When $m > k$. There are more than enough instruments. We later test whether the instruments are valid using **overidentification test**.

Underidentified When $m < k$. There are not enough instruments. The coefficients for X's will not be identified

So the key takeaway here so far is that when you are using IV, you need at least as much instrumental variables as the number of endogenous regressors you have. Now the issue is having too much IV bad? That depends on the answer we get from the overidentification test.

Overidentification test applies to the case where $m > k$ and aims to test whether the instrument variables we found are valid. To nail this idea more clearly, suppose that we only have one independent variable that is also endogenous -X_i. Also suppose that we have two instrumental variables - Z_{1i}, Z_{2i}. Now, we run two two-stage least squares regression with each one of the instrumental variables. The estimates for the coefficient for X are

$$\hat{\beta}_{Z_1} = \frac{cov(Z_1, Y)}{cov(Z_1, X)}, \ \hat{\beta}_{Z_2} = \frac{cov(Z_2, Y)}{cov(Z_2, X)}$$

The numbers look different, although both $\hat{\beta}$'s are suppose to be the same coefficient estimating impact of X_1 on Y. The overidentification test checks whether the differences between $\hat{\beta}_{Z_1}, \hat{\beta}_{Z_2}$ are large. If the two estimates are similar in the sense that the sample analogs of both $\frac{cov(Z_1,Y)}{cov(Z_1,X)}$ and $\frac{cov(Z_2,Y)}{cov(Z_2,X)}$ converge to the same thing in probability, we don't really have to worry. If they converge to something different, then we may have problem with either one or potentially both instrumental variables. Therefore, the takeaway from overidentification test is that it checks whether there exists a faulty IV variable in the case where we found too many candidates for instrumental variables.