

Topic 4 Simulation: Instrumental Variables

This week we will be tackling the issue of instrumental variables (IV) using our simulation tools. IV is one of the most commonly used and important tools in econometrics, with a wide range of applications and many extensions in panel data settings such as the Arellano–Bond estimator and the Blundell–Bond estimator.

Unfortunately, even though IV is such a useful and commonly used tools it is often poorly understood, and like all models, it does have it's own limitations. Today I just want to show how IV does in a simulation framework to hopefully help aid the understanding of the method, as well as talk through a few of the potential limitations of IV/the situations we should be wary of.

First of all we need to create a dataset. With IV we are typically concerned with situations where there is a relationship between an independent variable (x) and the error term (u), and thus we will build this into our model. We also need to create an instrument (z) to use in our IV regression and this instrument needs to be related to x but not to y. The following code should give us what we want:

```
clear
set obs 1000
set seed 101
gen u = rnormal()
gen z = rnormal()
gen x= rnormal()*5 + z*4 + u*2
gen y = 3*x + 5*u
reg y x
```

Our standard regression is estimating the coefficient on x to be 3.2, whereas the true value should be 3. From the 95% confidence interval though we can tell that we are 95% certain that the point estimate lies between 3.15 and 3.24, thus we can start to see that we are relatively confidently getting the wrong answer. We were expecting this as x and u are related and thus the assumptions that make OLS BLUE are violated. In situations where we believe there is this relationship between x and u we should be using IV so let's move on and do this.

Figure 1: Regression Output for OLS $y=3x$, endogenous x

Source	SS	df	MS	Number of obs	=	1,000
Model	461761.549	1	461761.549	F(1, 998)	=	21284.38
Residual	21651.4672	998	21.6948569	Prob > F	=	0.0000
				R-squared	=	0.9552
				Adj R-squared	=	0.9552
Total	483413.016	999	483.896913	Root MSE	=	4.6578

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	3.199631	.0219316	145.89	0.000	3.156594	3.242668
_cons	.1427066	.147321	0.97	0.333	-.1463877	.431801

There are different ways of running IV, either using the IV command in STATA or running the two step model ourselves. I prefer the second option for teaching purposes so we can see what

is going on behind the scenes! We will use our instrument z , that we believe has a relationship with x but no relationship with y in our IV. The first step is to regress our instrument on the independent variable that we believe is endogenous. We then predict the independent variable from this regression (predict $\hat{x}$) as this should give us the part of x that is uncorrelated with y . We use this new estimate then in our second step in replacement of the endogenous variable x . This should give us:

```
reg x z
predict xhat
reg y xhat
```

The coefficient estimates from this should be the same as using the inbuilt STATA command, `ivreg y (x=z)`, but the standard errors are likely to be slightly larger using this approach. The main thing we are interested in here is the coefficient on x , which we can now see is estimated to be 3.05, much close to the true value of 3. So our IV regression did the trick! We now have a much better estimate of x when there is a relationship between x and u .

Figure 2: Regression Output for IV $y=3x$, endogenous x

Source	SS	df	MS	Number of obs	=	1,000
Model	150861.708	1	150861.708	F(1, 998)	=	452.74
Residual	332551.308	998	333.217744	Prob > F	=	0.0000
				R-squared	=	0.3121
				Adj R-squared	=	0.3114
Total	483413.016	999	483.896913	Root MSE	=	18.254

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
xhat	3.053182	.1434919	21.28	0.000	2.771602	3.334762
_cons	.1231126	.5775693	0.21	0.831	-1.010277	1.256502

What issues are common with IV then? The first issue I want to talk about is the issue of weak instruments. Weak instruments means that the relationship between our instrument (z) and our endogenous variable (x) in this case is not very strong and thus z only explains a small part of x . In our previous example this was not really a concern, we created x to have a reasonably strong relationship with z : $\text{rnormal}() \cdot 5 + z \cdot 4 + u \cdot 2$.

Here we will show what happens if that is not the case and there is only a small link between the two. First we drop x and y and then we create the new x with a smaller link to z , and recreate y using the same line as above. Let's look to see how a standard OLS regression performs here before looking at our IV results.

```
drop x
drop y
gen x= rnormal()*5 + z + u*2
gen y = 3*x + 5*u
reg y x
ivreg y (x=z)
```

What do the results suggest? Figure 3 shows the OLS results and we are quite far away from 3 this time with an estimate of 3.3. Figure 4 where we use our instrument shows that we are getting close to true x value (i.e. our IV regression is reducing the bias), but the estimate of 3.2 is still not great. The true value of x does fall inside the confidence intervals for the IV regression however, while it is not close for the OLS regression. IV is still better than OLS when the instrument is weak.... but a stronger instrument is better than a weaker one!

Figure 3: Regression Output for OLS $y=3x$, endogenous x

Source	SS	df	MS	Number of obs	=	1,000
Model	312797.201	1	312797.201	F(1, 998)	=	15320.46
Residual	20376.1237	998	20.4169576	Prob > F	=	0.0000
				R-squared	=	0.9388
				Adj R-squared	=	0.9388
Total	333173.325	999	333.506832	Root MSE	=	4.5185

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	3.33006	.026904	123.78	0.000	3.277265	3.382854
_cons	.123922	.1428894	0.87	0.386	-.1564761	.4043201

Figure 4: Regression Output for IV $y=3x$, endogenous x and with weak instrument

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs	=	1,000
Model	312510.7	1	312510.7	F(1, 998)	=	438.57
Residual	20662.6254	998	20.7040335	Prob > F	=	0.0000
				R-squared	=	0.9380
				Adj R-squared	=	0.9379
Total	333173.325	999	333.506832	Root MSE	=	4.5502

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	3.229277	.1542009	20.94	0.000	2.926682	3.531872
_cons	.1215022	.1439366	0.84	0.399	-.1609509	.4039553

Instrumented: x
Instruments: z

So weak instruments don't help, but it does not seem to be too much of a problem above. The problem with weak instruments is that if our instrument has even a very small correlation with u, then IV becomes very bad, very quick! Why would this happen? If z has even a small link to y then it would end up having a relationship with u. Let's build this into our data and see what happens...! To do this we will introduce another explanatory variable v, which is unobserved but will partially explain the instrument z. If we make z somewhat correlated with v and the error term u also related to v this means there must be a link between z and u, but a weak one. Our x term now is related to u (this makes the variable endogenous and is why we are using IV)

and now has a relationship with z too. We have a weak instrument (z) and we have a relationship between z and u.

```
gen v = rnormal()
gen z2 = rnormal() - v
gen u2 = 10*rnormal() + v/4
gen x2 = rnormal()*10 + z2/4 + .1*u2
gen y2 = 3*x2 + 5*u2
ols y2 x2
ivreg y (x2=z2)
```

What do our results show? Figure 5 shows the OLS results with Figure 6 showing our IV estimates. This is the real issue with weak instruments, if there is any relationship between the instrument and the error term and a weak instrument our IV results are way off. There is a very large downward bias on x, our confidence intervals are so large we cannot reject the null hypothesis that the coefficient on x is equal to zero and our IV model is worse than just running a standard OLS.

Figure 5: Regression Output OLS for $y=3x$, endogenous x

Source	SS	df	MS	Number of obs	=	1,000
Model	312797.201	1	312797.201	F(1, 998)	=	15320.46
Residual	20376.1237	998	20.4169576	Prob > F	=	0.0000
				R-squared	=	0.9388
				Adj R-squared	=	0.9388
Total	333173.325	999	333.506832	Root MSE	=	4.5185

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	3.33006	.026904	123.78	0.000	3.277265	3.382854
_cons	.123922	.1428894	0.87	0.386	-.1564761	.4043201

Figure 6: Regression Output IV for $y=3x$, endogenous x with weak instrument and relationship between z and u

Source	SS	df	MS	Number of obs	=	1,000
Model	-3122.60409	1	-3122.60409	F(1, 998)	=	0.04
Residual	336295.929	998	336.969869	Prob > F	=	0.8393
				R-squared	=	.
				Adj R-squared	=	.
Total	333173.325	999	333.506832	Root MSE	=	18.357

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x2	.1902179	.937623	0.20	0.839	-1.649721	2.030157
_cons	.1124161	.6714241	0.17	0.867	-1.205149	1.429981

Instrumented:	x2
Instruments:	z2

The above problem is not really caused by the slight relationship between z and u, it is more due to the weak instrument. If we increase the explanatory power of z on x this issue disappears! I show this below by only changing the strength of the relationship between z and x and the results in Figure 7 improving dramatically and outperforming OLS again!

```
gen x3 = rnormal()*10 + 10*z2 + .1*u2
reg x3 z2
gen y3 = 3*x3 + 5*u2
ivreg y3 (x3=z2)
```

Figure 7: Regression Output IV for $y=3x$, strong instrument and relationship between z and u

Source	SS	df	MS	Number of obs	=	1,000
Model	2736591.14	1	2736591.14	F(1, 998)	=	590.83
Residual	2601137.61	998	2606.35031	Prob > F	=	0.0000
				R-squared	=	0.5127
				Adj R-squared	=	0.5122
Total	5337728.75	999	5343.07182	Root MSE	=	51.052

y3	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x3	2.80789	.1155173	24.31	0.000	2.581205	3.034575
_cons	-1.117771	1.616927	-0.69	0.490	-4.290738	2.055196

Instrumented: x3

Instruments: z2

There were a few tricky sections today to work through but hopefully the main points come through. IV is very useful but we need to be aware of it's limitations and that is what these notes tried to work through today. To recap these key takeaway points:

- 1) IV regressions are a great way of obtaining accurate estimates of coefficients when we believe an independent variable is endogenous (i.e. is related to u).
- 2) IV outperforms OLS when there is an endogenous independent variable and a strong instrument.
- 3) Weak instruments do not cause issues if there is not relationship between the instrument and the error term, IV will still outperform OLS.
- 4) If there is a weak instrument and a relationship between the instrument and the error term then IV will be a poor model to use and will perform much worse than OLS.

One final point we did not show but some more (slightly complex) simulation can show is that while point 3 holds here with quite a large sample size of 1000, as the sample size becomes small this is not the case. In fact weak instruments become much more of an issue with the sample size is small and with sample sizes around 500 or less OLS is often seen to outperform IV with weak instruments. With very small sample sizes (less than 250) IV estimates with weak instruments should not be trusted at all! That's all from me today. Remember if you are using IV, ensure you trust that the relationship between your instrument and the endogenous variable is strong.... Unless you have a large sample size and an instrument not related to the error term!

David Morris