

Topic 5 Simulation: Proving causality using IV or Panel Data

We often look at relationships between two variables and we are interested in correlation, a descriptive link between one factor and another. However, it is often more important for us to understand a causal relationship. Not just is there a relationship between A and B but does A cause B?

There are often issues in identifying a causal relationship, as rather than A causing B we might have B causing A, or even an outside factor driving both A and B, with no true link between A and B. This is a well known concept in philosophy and makes up one of the famous laws of logic “post hoc ergo propter hoc”.

What examples do we have to illustrate things such as this? A simple example might be something such as every time I go to sleep, the sun goes down, therefore, my going to sleep causes the sun to set. This is obviously not the direction of causality here. Likewise on certain days I may see more people with umbrellas and also see more people taking taxi rides and think there is some relation between taxis and umbrellas, when really the fact that it is raining is causing both!

Some of these issues can easily be solved by thinking about a problem in detail and trying to outline a logical mechanism through which we believe A impacts on B. To test these things however we do need to be quite careful. There are two good tools we should have to establish causality. In a cross sectional environment we can use IV regressions to establish if A changes B or vice versa. In a panel setting we can use alternative tools such as the Granger Causality test. We will try to discuss both in these notes.

Cross Sectional Causality

Eli Berman has an interesting paper¹ on how development spending has an impact on violence in a given region during the Iraq war. It is basically investigated whether you can buy hearts and minds during a conflict and Iraq was one of the first major conflicts where there was a large amount of data collection so that questions such as this could be answered. This he was trying to test for a relationship between development spending and violence. We could write this in a simplified form below:

$$Violence_i = \beta_0 + \beta_1 Development\ Spending_i + \varepsilon_i$$

If we actually tried to estimate this however rather than finding that development spending had a negative impact on violence we would likely find the reverse, a positive relationship. This is not because development spending is causing the higher levels of violence, but rather that regions that have higher levels of violence are likely to have more development spending on them. Thus we have causality acting in the wrong direction. We should be able to show this issue in a relatively simple manner using simulation.

We generate a dataset as usual, however, this time we will create x to be dependent on y, rather than y being dependent on x. The causality runs the other way. We can then run our standard

¹ Eli Berman, Jacob N. Shapiro, and Joseph H. Felter, "Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq," *Journal of Political Economy* 119, no. 4 (August 2011): 766-819. <https://doi.org/10.1086/661983>

regression and we may be led to believe that the x is impacting on y because of the correlation between them.

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

Figure 1: Regression Output for x=5y

```
. reg y x
```

Source	SS	df	MS	Number of obs	=	1,000
Model	768.869351	1	768.869351	F(1, 998)	=	4538.00
Residual	169.090176	998	.169429034	Prob > F	=	0.0000
				R-squared	=	0.8197
				Adj R-squared	=	0.8195
Total	937.959527	999	.938898425	Root MSE	=	.41162

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	.1677989	.0024909	67.36	0.000	.1629109	.1726869
_cons	.0068563	.0130188	0.53	0.599	-.018691	.0324035

We know that this is not the correct form of causality here, we created x to depend on y, not y to depend on x. If we introduce an instrument, something that impacts on x but does not impact on y we can start to find out the direction of causality. X still depends on y here, but we now have z, our instrument, which we can use to determine if changes in x impact upon y.

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

Here we can straight away see that there is no causal relationship whereby changes in x cause changes in y.

Figure 2: Regression Ouput for IV, $x=5y$

```
. ivreg y (x=z)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs	=	1,000
Model	-265.535168	1	-265.535168	F(1, 998)	=	1.85
Residual	1203.49469	998	1.20590651	Prob > F	=	0.1742
				R-squared	=	.
				Adj R-squared	=	.
Total	937.959527	999	.938898425	Root MSE	=	1.0981

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	-.0268303	.0197288	-1.36	0.174	-.0655451	.0118844
_cons	.0258126	.0347793	0.74	0.458	-.0424363	.0940616

Instrumented: x

Instruments: z

What happens if we run IV and the direction of causality is correct? Below is the same idea but with y actually being caused by x, and not vice versa. We still have an instrument z which causes changes in x and we can thus still run an IV regression here.

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

Below we can see that there is a positive and significant causal relationship between x and y, where a one unit increase in x increases y by 5 units.

It should be noted that in the example below IV is not actually needed to find an unbiased estimate. There is no bias and thus OLS would have found the correct estimate of β_1 , IV was only useful here to establish the causality of the relationship. IV therefore has two key uses: 1) Remove bias when there may be unobserved variation that impacts both upon x and y, 2) to establish causality in a cross sectional data setting.

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

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs	=	1,000
Model	120651.889	1	120651.889	F(1, 998)	=	99739.93
Residual	950.592638	998	.952497633	Prob > F	=	0.0000
				R-squared	=	0.9922
				Adj R-squared	=	0.9922
Total	121602.482	999	121.724206	Root MSE	=	.97596

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	4.98722	.0157915	315.82	0.000	4.956232	5.018209
_cons	-.0076114	.0308891	-0.25	0.805	-.0682265	.0530036

Instrumented: x

Instruments: z

Panel Data/Time Series Causality

When we have panel data or time series data then we have a few more useful tools to deal with the issue of causality. The Granger causality test is built upon the concept that the future cannot determine the past, and thus a lagged variable cannot be caused by a variable measure in the current time period. Continuing with the previous example we can imagine rather than observing violence and development spending in one time period for each region (i) we can observe this numerous times and thus we can add the subscript t.

$$Violence_{it} = \beta_0 + \beta_1 Development\ Spending_{it} + \varepsilon_{it}$$

The Granger causality test relies upon the idea that past values cannot influence current values. Thus if we find a relationship between lagged development spending and violence then it must be that development spending is causing violence and not vice versa. This can be seen below:

$$Violence_{it} = \beta_0 + \beta_1 Development\ Spending_{i,t-1} + \varepsilon_{it}$$

Thus a significant β_1 would suggest that the link runs from development spending to violence.... Or would it?!

The issue that we have here is that violence in time period t is likely to be related to the level of violence in period t-1. Thus we should really see the model as more like:

$$Violence_{it-1} + Change = \beta_0 + \beta_1 Development\ Spending_{i,t-1} + \varepsilon_{it}$$

Thus a significant coefficient on development spending may just show that development spending and violence both in period t-1 are related, and we lose the ability to comment on causality. How do we actually deal with this issue then? We control for the past value of violence on the right hand side of the equation.

$$Violence_{it} = \beta_0 + \beta_1 Development\ Spending_{i,t-1} + \beta_2 Violence_{i,t-1} + \varepsilon_{it}$$

Now if we find a positive and significant coefficient on development spending we should feel relatively confident that the causal relationship is between development spending to violence. A more formal definition of Granger Causality in the case of two time-series variables, X and Y is thus: "X is said to Granger-cause Y if Y can be better predicted using the histories of both X and Y than it can by using the history of Y alone." The formal test therefore is whether any of the coefficients on the X terms are statistically different to zero.

Sometimes we do have feedback loops in economics, where A impacts B and B impacts upon A. There is no real quick fix to this issue and thus I don't really want to discuss it in any detail. All I will say on the matter is that sometimes it is worth running the opposite version of the above equation to ensure that violence is not also driving development spending in a feedback loop. If causality is only found in one direction then we should be relatively happy with the causal nature of the relationship.

$$Development\ Spending_{it} = \beta_0 + \beta_1 Development\ Spending_{i,t-1} + \beta_2 Violence_{i,t-1} + \varepsilon_{it}$$

Can we simulate a situation where we can prove causality through a Granger set up? I think we can! I have generated a panel with 50 regions and 5 years of data below. Here y is dependent upon x, and x is highly correlated to its past value. This is just so that when we create y to be dependent on x, y will now also be highly dependent upon its past value.

```
clear
set obs 50
set seed 101
gen x= rnormal()
label var x "development spending"
gen id = _n
label var id "region"
expand 5
bysort id: gen year=_n
label var year "Year of observation"
tab year
gen u = rnormal()
replace x = (0.5*x[_n-1]) + rnormal() if _n>1
gen y = 2 + 5*x + 2*u
reg y x
reg x y
```

Looking at the regression with y as the dependent and x as the independent variable, we are estimating the correct coefficient on x and finding it to be statistically significant. We may however be unsure of whether it is a causal relationship. The next figure highlights this with x as the dependent variable and y as the independent, we still find a significant correlation between the two.

Figure 4: Regression Output for $y=5x$, Y as the dependent variable

Source	SS	df	MS	Number of obs	=	250
Model	11540.35	1	11540.35	F(1, 248)	=	3275.48
Residual	873.767068	248	3.52325431	Prob > F	=	0.0000
				R-squared	=	0.9296
				Adj R-squared	=	0.9293
Total	12414.1171	249	49.8558919	Root MSE	=	1.877

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
y						
x	5.040488	.0880714	57.23	0.000	4.867024	5.213951
_cons	2.170965	.1187888	18.28	0.000	1.937002	2.404929

Figure 5: Regression Output for $y=5x$, X as the dependent variable

Source	SS	df	MS	Number of obs	=	250
Model	422.257165	1	422.257165	F(1, 248)	=	3275.48
Residual	31.9708159	248	.12891458	Prob > F	=	0.0000
				R-squared	=	0.9296
				Adj R-squared	=	0.9293
Total	454.227981	249	1.82420876	Root MSE	=	.35905

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x						
y	.1844296	.0032225	57.23	0.000	.1780826	.1907765
_cons	-.4037589	.0235442	-17.15	0.000	-.450131	-.3573869

Let's try and test for a Granger causal link here then. First I have tsset the data so that STATA knows we are dealing with time series data and numerous individuals. I have run a correlation of y on the lag of y (l.y) to check for a correlation and the two do seem to be strongly correlated. Due to this we do need to control for the lag of y in our regression as well as the lag of x. Thus we run "reg y l.x l.y" to see if the lagged value of x contributes towards our prediction of y, even when we control for the lag of y. I have then run the opposite test to check if there is a causal relationship in the reverse manner, with y impacting future values of x. The results can be seen in Figures 6 and 7 below.

```
tsset id year
corr y l.y
reg y l.x l.y
reg x l.y l.x
```

Looking at the results in Figure 6 we see that the lagged value of x is highly significant in determining y, even when controlling for the lagged value of y. We would thus say that x has a Granger causal relationship with y. The same cannot be said of the reverse in Figure 7, lagged y values do not influence future values of x. The causality runs one way from x to y, just how we created it!

Figure 6:

```
. reg y l.x l.y
```

Source	SS	df	MS	Number of obs	=	200
Model	3722.18511	2	1861.09255	F(2, 197)	=	57.11
Residual	6420.11701	197	32.5894264	Prob > F	=	0.0000
				R-squared	=	0.3670
				Adj R-squared	=	0.3606
Total	10142.3021	199	50.9663423	Root MSE	=	5.7087

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x						
L1.	3.735998	1.134498	3.29	0.001	1.498678	5.973317
y						
L1.	-.1018753	.2174818	-0.47	0.640	-.5307665	.3270159
_cons	2.254006	.631182	3.57	0.000	1.009265	3.498747

Figure 7:

```
. reg x l.y l.x
```

Source	SS	df	MS	Number of obs	=	200
Model	140.381424	2	70.1907119	F(2, 197)	=	57.97
Residual	238.526772	197	1.2107958	Prob > F	=	0.0000
				R-squared	=	0.3705
				Adj R-squared	=	0.3641
Total	378.908196	199	1.90406129	Root MSE	=	1.1004

x	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
y						
L1.	-.0156534	.0419199	-0.37	0.709	-.0983227	.0670159
x						
L1.	.7049943	.2186759	3.22	0.001	.2737482	1.13624
_cons	.0205566	.1216611	0.17	0.866	-.2193687	.2604819

While we are happy now that in a panel setting we can determine causality is our estimate of the impact of x on y accurate? No. We know the coefficient should be 5. Why is this? Well we have built endogeneity into our regression due to the relationship between the lagged value of y and the dependent variable y. If you are not sure about this look back on your notes on the IV week and see

if you can understand why. The fix to this issue results in one of the most popular forms of estimation in empirical econometrics, the Arellano-Bond estimator which is too complex to discuss in this course but I am happy to discuss informally with anyone who is interested.

When else should we be careful in using Granger causality tests? When we are looking at non-stationary variables. Non-stationarity is the subject that we are covering in Topic 7 but it is worth bearing in mind that the material from week 6 is designed for stationary data, and once you have seen a bit more about non-stationary it is worth reviewing these materials again. If one or more series is non-linear then a method such as that proposed by Toda & Yamamoto (1995) may be the best bet². There is a detailed walk through of this process on one of my favourite econometrics blogs that you can find here: <http://davegiles.blogspot.co.uk/2011/04/testing-for-granger-causality.html>. This is of course much more advanced than the material we have considered in the course and is just for your own knowledge and if you have an interest in this area!

I hope this was interesting....

David Morris

² Full reference should be available at: <http://www.sciencedirect.com/science/article/pii/S0304407694016168>