Mgmt 469

Causality and Identification

As you have learned by now, a key issue in empirical research is identifying the direction of causality in the relationship between two variables. This problem often arises when it is unclear in a regression whether X causes Y or Y causes X. An example from my own research experience helps to clarify the problem and its solution.

The Problem of Supplier Induced Demand

Historically, MDs have controlled the medical process. Analysts have long noted that medical prices are higher in markets that have more MDs per capita. Some have hypothesized that when MDs who face competition "induce demand" so as to maintain their incomes. The theory of demand inducement is almost dogma among health policy analysts, especially in Canada and Europe. The supplier induced demand (SID) hypothesis may be stated as follows:

> *The SID hypothesis:* When the supply of physicians increases, the demand for their services increases. That is, increased supply causes increased demand.

Early proponents of SID proposed testing it with the following model:

$$Q^d = g(X^d, P, Q^s) + \varepsilon, \quad \text{where}$$

Q^d = the quantity of physician services provided (e.g., number of surgeries),
X^d = one or more demand shifters,
P = price
Q^s = the number of physicians per capita, and
ε = the error term.

One might want to test this model by gathering the required data and running OLS. A positive, statistically significant coefficient on Q^s might be considered evidence of SID.

It turns out that this would be a very bad approach. The reason: doubts about causality

Consider the following, somewhat related example:

There is a positive correlation between construction workers and the number of office buildings under construction.

· What is the assumed direction of causality?

The fundamental causality question in supplier induced demand is: Do more MDs cause high demand (X causes Y), or does high demand cause more MDs (Y causes X)?

· Unless this causality is disentangled, you cannot interpret the coefficient on Q^s.

· X is an endogenous regressor; i.e., the value of X is correlated with the underlying error in the model.

More on endogeneity

· The concern is this problem is that Y might cause X

· Suppose you have an observation for which underlying error is positive. This will give a larger value for Y

· Y causes X and if Y is larger, then X will be larger

· This implies that X and the error are positively correlated; thus, there is endogeneity bias

Identification in a model with endogeneity bias

The trick to resolving causality in a model with endogeneity bias is to find one or more variables that have two features:

· They must be correlated with X

· They must be uncorrelated with the underlying error

- This implies that they are not caused by Y.
- Nor can they be variables that you might have included as control variables on the RHS of your original regression

· These variables are called *instruments*. The technique for using them in regression is known as instrumental variables (IV) regression or two-stage least squares (TSLS) regression.

The following thought experiment may clarify the intuition behind this approach

· Suppose we could randomly allocate MDs to different places, making sure to allocated more MDs to some places than others.

· We are now sure that Q^s is not related to any factor that might affect Y (and therefore affect X). Thus, Q^s is truly exogenous

· If we run our regression and find that per capita utilization is higher in the places that have more MDs, we have confirmed SID. Causality is not in doubt.

IV and TSLS seek the real world equivalent of such randomization.

· In other words, we need some instruments that cause real world MDs to locate in certain places for reasons that have nothing to do with demand.

· If demand is high in places that score high on these instruments, then we conclude that MDs do cause demand (instruments→MD supply→demand)

· The hard part is finding these instruments

- Perhaps we could examine the number and quality of golf courses. It seems doubtful that golf courses are directly related to demand.

- If we saw that areas with high golf courses have high demand, we would blame inducement.

- Our logic would then be: (golf courses→MD supply→demand).

If we want to use IV/TSLS estimate $Q_d = g(\underline{X}_d, Q_s, P)$ and resolve causality, we need to proceed as follows:

Stage one: regress $Q_s = f(\underline{X}_s, \underline{X}_d, P)$

· We include the identifiers $\underline{X}_s$. These are variables that plausibly shift Q_s but are otherwise unrelated to Q_d. To maximize efficiency, we also include any other variables on the RHS of the original model, in this case X_d and P.

· Recover the *predicted values* of Q_s. Call the predicted values $\mathbf{Q_s}$.

Stage two: regress $Q_d = g(\underline{X}_d, \mathbf{Q_s}, P)$

· The coefficient on $\mathbf{Q_s}$ indicates the extent to which Q_s affects Q_d. It is purged of any other relationship between Q_d and Q_s thanks to using instruments in stage one.

· Note: the stage 2 regression must be adjusted to account for the fact that $\mathbf{Q_s}$ is a noisy estimate of Q_s. A good software package like Stata will do this for us.

In Stata, we type

regress Q_d $\underline{X}_d$ P ($\underline{X}_s$ $\underline{X}_d$ P)

To summarize:

· TSLS/IV is necessary if causality is in doubt

· TSLS/IV requires identifiers for the RHS variable whose causative impact is in question.

· Good identifiers have the following characteristics

- They are correlated with the problematic RHS variable
· They have no causative relationship with the LHS variable - that is, they should not belong in $\underline{X}_d$ or be caused by Y.

· Second stage replaces the actual value of the problematic variable with its predicted value from the first stage

Victor Fuchs used TSLS in his study of inducement.

· His first stage estimates predicted physician supply by location.

· He uses predicted supply in the second stage estimate of demand, and obtains an inducement elasticity of .28. (That is, for every 10 percent increase in the supply of surgeons, the number of surgeries per capita would increase by 2.8 percent.)

· Later studies using more control variables find elasticities of about .10. These findings have led to calls for controls on MD supply and specialist training

The SID literature has subsequently reexamined these studies, assessing whether the instruments are valid

· The key identifiers are $ hotel/capital (Fuchs) and weather conditions (later authors)

· Are these identifiers?

- Most of the time, we rely on theory to be our guide. Do these cause MD locations, but are otherwise unrelated to demand. It is difficult to see why these are any more related to MD supply than they are to demand.

- In cases such as these, we can employ econometric techniques to test our instruments. Econometrician Jerry Hausman has developed several tests for identifiers. I describe Hausman tests below.

· I published a paper that showed flaws in the Fuchs' identifiers.

- I tested SID in a market where we would be surprised to see it – the market for childbirths

- I used the same identifiers as Fuchs and others and found an elasticity of about .10. More OB/GYN MDs appeared to lead to more childbirths! The method appears to find inducement even when there should not be any.

- This is not a problem with TSLS *per se*, but does show the need for good identifiers.

Implementation Issues (optional)

· Minimum requirements for TSLS

- General rule of thumb is that R^2 from the identifiers should be about .20 or higher.
- Should have at least 50 observations

· Testing the identifiers

- To test for endogeneity, you can perform a *Hausman test*
- Hausman tests generally examine whether certain excluded variables are correlated with dependent variables.

1) Regress $Y = a + bX + cZ + dX^*$ (where X are the original values of the predictors whose causality is in doubt, and X^* are the predicted values)

2) Test significance of *d* using t or F statistic. If *d* is significant, then you should reject hypothesis that X variables are exogenous, and seek out better identifiers.

· Here is another Hausman test that tells you whether you have chosen good identifiers

1) Regress Y on the endogenous RHS variable (e.g., #MDs), X_d and those variables in X_s that are not identifiers.

2) Take the residuals and correlate them with the identifiers.

- If the correlation is significant, then the identifiers are predictors of Y, and are not really identifiers - they belong in X_d

- I did this with weather and hotels and found that they belonged in X_d

TSLS and Omitted Variable Bias (optional)

Because simultaneity bias and omitted variable bias are different flavors of the same problem (endogeneity bias), they share the same solution.

· Suppose that you are regressing Y on X but are concerned that X is correlated with some omitted factor Z. Suppose further that you cannot collect data on Z. Your coefficient on X may suffer from omitted variable bias.

· If you can find one or more instruments for X - variables that cause X but have no causative relationship with Z - then you can eliminate the bias.

· You would use the methods described above, using the predicted value of X instead of the actual value.

· The strength of TSLS and IV regression is that they avoid bias. The weakness is that you need good instruments.

Never trust OLS results when IV/TSLS is required. But do not always expect to be able to perform IV/TSLS