POLS571 - Longitudinal Data Analysis
September 25, 2001

1 Causality

You all have already discussed causality at some length in other classes, so we won't get all philosophical here. The important thing to remember is that time-series data provide both opportunities and challenges for addressing causality.

1.1 Granger Causality: The Concept

"Granger causality" is a term for a specific notion of causality in time-series analysis.[1] The idea of Granger causality is a pretty simple one:

A variable X *Granger-causes* Y if Y can be better predicted using the histories of both X and Y than it can using the history of Y alone.

Conceptually, the idea has several components:

- Temporality: Only *past* values of X can "cause" Y.
- Exogeneity: Sims (1972) points out that a necessary condition for X to be exogenous of Y is that X fails to Granger-cause Y.
- Independence: Similarly, variables X and Y are only independent if both fail to Granger-cause the other.

Granger causality is thus a pretty powerful tool, in that it allows us to test for things that we might otherwise assume away or otherwise take for granted.

[1] Clive Granger, the UCSD econometrician, gets all the credit for this, even though the notion was apparently first advanced by Weiner twenty or so years earlier.

1.2 Granger Causality Testing

Freeman (1983) discusses two sets of tests for determining Granger causality.

1.2.1 ARIMA models/Cross-Correlations

If the series in question are stationary ARMA(p,q) processes:

$$\phi_{p_Y} L^{p_Y} Y_t = \theta_{q_Y} L^{q_Y} u_{Yt} \tag{1}$$

$$\phi_{p_X} L^{p_X} X_t = \theta_{q_X} L^{q_X} u_{Xt} \tag{2}$$

then we can consider the cross-correlation functions of the two series. In particular, under the null hypothesis of independence (no Granger causality in either direction), the cross-correlations of the innovations u_{Xt} and u_{Yt} will be zero at all positive and negative lags.

To implement this approach is simple; one:

1. Estimates an appropriate ARIMA model for each series, then
2. estimates the cross-correlations of the estimated $\hat{u}$s.

In Stata, the cross-correlation command is `-xcorr-`. The approximate standard errors of the cross-correlations are just $\frac{1}{\sqrt{T}}$. Cross-correlation values larger than ± 2 standard errors from zero indicates the presence of Granger causality (i.e., a lack of *Granger*-independence).

While the ARIMA/cross-correlation approach is fine, it has a few drawbacks:

- The method is sensitive to the choice of lag length for the cross-correlations,
- The test can't tell you the directionality of causality, only the presence or absence of it;
- The statistic lacks power, as compared to the regression-based tests discussed below.

So, we generally don't use this approach a lot.

1.2.2 The "Direct Granger Method"

As the name suggests, we can also assess Granger causality in a more direct way: by regressing each variable on lagged values of itself and the other, e.g.:

$$Y_t = \beta_0 + \sum_{j=1}^{J} \beta_j Y_{t-j} + \sum_{k=1}^{K} \gamma_k X_{t-k} + u_t \tag{3}$$

We can then simply use an F-test or the like to examine the null hypothesis $= 0$ Critical is the choice of lags J and K; insufficient lags yield autocorrelated errors (and incorrect test statistics), while too many lags reduce the power of the test. This approach also allows for a determination of the causal direction of the relationships, since we can also estimate the "reverse" model:

$$X_t = \beta_0 + \sum_{j=1}^{J} \beta_j X_{t-j} + \sum_{k=1}^{K} \gamma_k Y_{t-k} + u_t \tag{4}$$

Also, it is important to remember that Granger causality testing should take place int he context of a fully-specified model. If the model isn't well-specified, "spurious" relationships may be found, despite the fact of no actual (conditional) relationship between the variables. We'll talk more about Granger causality when we discuss VAR models later in the course.

2 Time Series and Spurious Regressions

2.1 What it is

Consider the regression of two I(1) series:

$$Y_{1t} = \beta_0 + \beta_1 Y_{2t} + e_t \tag{5}$$

where:

$$\begin{aligned} Y_{1t} &= Y_{1t-1} + u_{1t} \\ Y_{2t} &= Y_{2t-1} + u_{2t}, \\ u_{1t}, u_{2t} \sim N(0, \sigma^2_{u_t}) &\quad , \quad Cov(u_{1t}, u_{2t}) = 0 \end{aligned}$$

The problem of *spurious regressions* was first addressed by Granger and Newbold (1974) (G&N). The intuition is relatively simple: because integrated series have a tendency to "wander", it is often the case that a regression of one on the other will appear to yield significant results, even if the two series are completely independent. G&N's study was purely a simulation; subsequently, Phillips (1986) showed that there is an analytic basis for this result as well: under very general conditions for the error terms, sample moments of the Y variables converge not to constants, but rather to functions of Brownian motion. This means that standard distributional results for OLS fall completely apart:

- Conventional t-statistics (e.g., $\frac{\hat{\beta}}{s.e.(\hat{\beta})}$) do not have limiting distributions, but instead diverge as $T \to \infty$,
- this means that there are *no* asymptotically correct critical values for these tests, and
- the rejection rate will (in general) increase with the sample size used, consistent with G&N.
- In contrast, R^2 does have a limiting distribution, and that the value of the Durbin-Watson statistic d goes to zero as $T \to \infty$.

The last two points are also consistent with G&N, who note that their Monte Carlo studies produced regressions with low-to-moderate R^2 statistics, and very low D-W statistics.

Nor surprisingly, the driving force behind the spurious regression phenomenon is the error term e_t. In particular, its pretty easy to see that, since the error is itself a combination of I(1) processes, it too will (generally) be integrated:

$$\begin{aligned} e_t &= Y_{1t} - \hat{\beta}_0 - \hat{\beta}_1 Y_{2t} \\ &= -\hat{\beta}_0 - \sum u_{1t} - \hat{\beta}_1 \sum u_{2t} \end{aligned} \tag{6}$$

This means that we can "solve" the problem of spurious regressions by simply including a lagged Y_1 on the right-hand side of the equation (or, equivalently, by differencing the equation):

$$Y_{1t} = \beta_0 + \beta_1 Y_{2t} + \beta_2 Y_{1t-1} + e_t \tag{7}$$

This model eliminates the integration in the es, and allows for "normal" OLS-based estimation and testing.

2.2 Spurious Regression: An Example

Here's an example, using made-up data, in Stata 6.0:

```
.  set obs 500
obs was 0, now 500

.  gen t=_n

.  gen y1=0

.  gen y2=0

.  gen u1=invnorm(uniform())

.  gen u2=invnorm(uniform())

.  replace y1=y1[_n-1]+u1 if y1[_n-1] =.
(499 real changes made)

.  replace y2=y2[_n-1]+u2 if y2[_n-1] =.
(499 real changes made)
```

Regressing Y_1 on Y_2 yields the following results, for different lengths of T:

N	β_0	β_1	t-statistic for β_1	R^2	F	D-W statistic
100	1.50	-0.41	-4.75	0.19	22.6	0.29
250	1.80	-0.27	-9.10	0.25	82.8	0.10
500	3.18	-0.20	-7.63	0.10	58.3	0.06

Despite the fact that the two series were created independently, the fact that each is a random walk induces a correlation in them. While the us for the 500 observations are correlated at only 0.04 ($p > .20$), the two series correlate at -0.32 ($p < .001$).

That the problem can be solved by including a lagged Y_1 is also easily shown by estimating the model in (7):

N	β_1	t-statistic for β_1	β_2	t-statistic for β_2	R^2	D-W statistic
100	-0.02	-0.45	0.89	18.7	0.82	2.24
250	0.004	0.047	0.98	51.8	0.94	2.09
500	-0.002	-0.35	0.97	88.7	0.95	2.08

2.3 Wrap-up

The fact of spurious regressions is the major reason why, in many instances, analysts automatically difference variables they believe to be I(1). In fact, however, there is a class of multivariate models where differencing I(1) variables is *not* recommended. If the regression of two I(1) variables yields errors which are not I(1) (that is, stationary), then the series are said to be **cointegrated**; in that case, differencing is NOT the thing to do. We'll talk about this more in the future.