Regression with random walks and Cointegration

Spurious Regression

• Generally speaking (an exception will follow) it’s a really really bad idea to regress one random walk process on another.

• That is, if \( y_t \) and \( x_t \) both follow a random walk looks like it could be a problem:

\[
y_t = \beta_0 + \beta_1 x_t + \epsilon_t
\]

\[
\hat{\beta}_1 = \frac{\text{cov}(x, y)}{\sigma_x^2} = \frac{\sigma_y \rho_{xy}}{\sigma_x^2} = \frac{\sigma_y \rho_{xy}}{\sigma_x} \rightarrow \infty \rho_{xy}
\]

• Remember the variance (and standard deviation) of a random walk don’t exist – they are infinite!
Simulations

• Simulate from two independent random walk models and regress $y_t$ on $x_t$.
• We can show that the estimated coefficient and the t-statistic are just random draws from some distribution!
• Even if the random walk models are totally unrelated you will get a slope estimate that doesn’t go to zero as the sample size gets large and will usually appear to be statistically significant!

It does make sense to regress mean reverting stuff on mean reverting stuff

• The usual solution is to make the series stationary by modeling returns or growth rates instead of price levels.
• That is if the two series $x$ and $y$ follow a random walk (like log prices), then take the first difference and model the differences in the series.
\[
\begin{bmatrix}
\Delta y_t \\
\Delta x_t
\end{bmatrix} = \Delta y_t = \beta_0 + \sum_{j=1}^{p} \beta_j \Delta y_t + \nu_t
\]
• Now we are regressing mean reverting series on mean reverting series and we no longer have the infinite variance problem.
Cointegration

• Cointegration is a special relationship that two non-mean reverting series can exhibit.
• Sometimes a pair of series might each follow a random walk, but over the long run their paths are connected.
• While they wander around over time, the two series can’t get far apart.

• Bid/Ask example.
• Tomato example.
• The $/Euro exchange rate and the price of converting from $/Yen and then from Yen/$.
Price of tomatoes per pound

Bid and Ask prices
**Formal Definition of Cointegration**

The series $y_t$ and $x_t$ are said to be cointegrated if both $y_t$ and $x_t$ follow a random walk but there exists a linear combination

$$z_t = y_t - \gamma x_t$$

where $z_t$ is stationary (mean reverting).

$\gamma$ describes the cointegrating relationship, often it is 1 (like in the preceding examples).

**Testing for Cointegration. CASE 1, KNOWN COINTEGRATING RELATIONSHIP.**

- First, test to see that both $y_t$ and $x_t$ have a unit root.
- Create sequence of $z_t = y_t - \gamma x_t$ (known value)
- Test the $z_t$ series using a standard random walk test (Dickey Fuller) test.
- If $z_t$ is mean reverting then x and y are cointegrated.
CASE 2: UNKNOWN COINTEGRATING RELATIONSHIP

- First, test to see that both $y_t$ and $x_t$ follow a random walk.
- Second regress $y_t$ on $x_t$ and estimate $\gamma$. Create the residual series

$$y_t = \hat{\gamma}_t x_t + z_t \iff z_t = y_t - \hat{\gamma}_t x_t$$

- These estimates of $\gamma$ turn out to not suffer from the spurious regression problem when $x$ and $y$ are cointegrated.
- When $y_t$ on $x_t$ are cointegrated, this is the one time its OK to run this regression.

- Test to see if $z_t$ has a unit root.
- We can’t use the standard Dickey-Fuller because the fact we have to estimate $\gamma$ (as opposed to plugging in the known value messes things up.
- We must use another special distribution.
- Eviews will perform this test for us.
Bid ask prices

Case 1: Create \( z_t = y_t - x_t \) series
Perform the Augmented Dickey Fuller Test

- Null is that \( z \) has a unit root which means that there is no cointegration.
- **We reject the null that \( z \) has a unit root and conclude that the series \( x \) and \( y \) are cointegrated.**

What about unknown cointegrating relationship?

- Estimate from data the value of \( \gamma \) by least squares regression.
- Perform “special” Dickey Fuller.
There are four ways that the test can be done. Each one has a p-value I use the first p-values here.

Error Correction Model (ECM)

- If $y_t$ and $x_t$ are cointegrated, then we CANNOT model changes in $y_t$ and $x_t$ by a vector autoregression.
- Instead, the model for changes in $y_t$ and $x_t$ follows what is called the Error Correction Model ECM.
- The model is specified as a VAR in changes in $y_t$ and $x_t$, but it includes a special term on the right hand side.
Error Correction Model (ECM)

• If the $x_t$ and $y_t$ are cointegrated then you should fit an error correction model.

• Since $x$ and $y$ follow a random walk, take the first difference of $x$ and $y$.

\[
\begin{bmatrix}
\Delta y_t \\
\Delta x_t
\end{bmatrix} = \Delta y_t = \beta_0 + \sum_{j=1}^{p} \beta_j \Delta y_{t-j} + \alpha z_{t-1} + v_t
\]

This is the usual VAR($p$) part.

• Where $z_t = y_t - \gamma x_t$ is the error correcting term that is new in the VAR specification and $\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$. (recall that bold $y$ is a vector, unbold $y$ is not)

• So this is just a regular VAR model for the changes in $x$ and $y$ but it has the “error correction term”

\[
\alpha z_{t-1} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} (y_t - \gamma x_t)
\]

• Often $\gamma$ is 1, so this term tells us how to update changes in $y$ and $x$ as a function of the difference between $y$ and $x$ at time $t$. 
• Step 1) estimate cointegrating relationship via regression of $y_t$ on $x_t$
• Step 2) create series $z_t$
• Step 3) estimate error correction model, equation by equation. $z_t$ would be an additional right hand side variable (exogenous).

• If $x$ and $y$ are cointegrated then the correct model is the error correction model.
• If we simply fit a VAR on the changes, the model will be misspecified since we need both *changes* and *levels* on the right hand side of the equation.
• The fact that the VAR depends on $z_{t-1}$ allows the series to correct for misalignments – hence the term error correction.
• A simple error correction model for logarithmic bid and ask prices is given by:

\[
\begin{bmatrix}
\Delta \ln(\text{ask}_t) \\
\Delta \ln(\text{bid}_t)
\end{bmatrix} =
\begin{bmatrix}
\beta_0^a & \beta_1^a & \beta_2^a \\
\beta_0^b & \beta_1^b & \beta_2^b
\end{bmatrix}
\begin{bmatrix}
\Delta \ln(\text{ask}_{t-1}) \\
\Delta \ln(\text{bid}_{t-1})
\end{bmatrix} +
\begin{bmatrix}
\alpha^a \\
\alpha^b
\end{bmatrix} \text{Spd}_{t-1} +
\begin{bmatrix}
\epsilon^a_t \\
\epsilon^b_t
\end{bmatrix}
\]

recall \( \text{Spd}_t = z_t = \ln(\text{ask}_t) - \ln(\text{bid}_t) \)

• So \( \alpha \) determine how the bid and ask prices change as a function of the spread.

• Market forces should force wide spreads to narrow.

• We should expect that a wide spread should lead to an increase in the bid and a decrease in the ask.

• This is true empirically with \( \alpha_3 \) around \(-.15\) and \( \beta_3 \) around \(.15\). (we get a little larger values in our sample).
Additional “x” variables can be included in the Error Correction Model.

\[
\begin{align*}
\Delta \ln(ask_t) &= \beta_0^a + \beta_1^a \Delta \ln(ask_{t-1}) + \beta_2^a \Delta \ln(bid_{t-1}) + \alpha^a \Spd_{t-1} + \theta^a x^a_{t-1} + \epsilon_t^a \\
\Delta \ln(bid_t) &= \beta_0^b + \beta_1^b \Delta \ln(ask_{t-1}) + \beta_2^b \Delta \ln(bid_{t-1}) + \alpha^b \Spd_{t-1} + \theta^b x^b_{t-1} + \epsilon_t^b
\end{align*}
\]

- A by sell indicator (+1 for buy, -1 for sell)
- The (signed for buy or sell) size of the previous trade.
- Multiple variables can be included.
• Buys tend to raise both the bid and the ask.
• Sells tend to decrease both the bid and the ask.
• However, the ask tends to raise by more than the bid following a buy and the bid tends to fall by more than the ask following a sell.
• Trade size matters. Larger trades have a larger price impact. The effect increases at a decreasing rate.