The RATSletter

Estima

Volume 30, No. 2 June, 2018

ARCH/GARCH and Volatility Models E-Course, 2nd Edition

The popular *ARCH/GARCH and Volatility Models* e-course has been updated to a 2nd edition. The *State-Space and DSGE Models*, *Structural Breaks and Switching Models* and *VAR* courses are also on their 2nd editions. The other subjects for e-courses are *Panel Data,* and *Bayesian Econometrics.*

The e-courses go into considerably greater detail on specific subjects than we can provide in the main documentation. "Diagnostics in Large Data Sets" below is based upon a new section in the GARCH e-course and more information on the updated content is on page 2. For details on *all* the courses see

https://estima.com/courses_completed.shtml

Diagnostics in Large Data Sets

If you do quite a bit of work with large data sets (typically financial models), you may find it frustrating to get a model which "passes" a standard set of diagnostic tests for model adequacy. This, however, is a well-known (though rarely discussed) problem when dealing with real-world data. In Berkson(1938), the author points out that tests for Normality will almost always reject in very large data sets with *p*-values "beyond any usual limit of significance." Now, the key is that we're talking about "real-world" data. There have been dozens, if not hundreds, of papers published in the literature which propose a test, demonstrate that it has asymptotically the correct size when simulated under the null, demonstrate it has power against simulated alternatives, and ... that's the end of the paper. What's missing in many of these papers is any attempt to demonstrate how it works with real data, not simulated data. If the new test doesn't produce qualitatively different results from an existing (probably simpler) test, is there any real point to it?

A very readable discussion of what's happening with tests in large data sets is provided in Leamer(1974), Chapter 4. Effectively no model is "correct" when applied to real world data. Consider, for instance, Berkson's example of a test for Normality. In practice, actual data is likely to be bounded—and the Normal distribution isn't. Various Central Limit Theorems show that the sums of many relatively small independent (or weakly correlated) components approach the Normal, but in practice, some

(continued on page 3)

Toda-Yamamoto Causality Test: A Cautionary Tale

It's been known for quite some time that the standard form "Granger causality" lag exclusion test has non-standard (and basically uncomputable) asymptotics if the variables involved are *I(1)*. This is one of the implications of Sims, Stock and Watson(1990)—in a linear regression involving (possibly) non-stationary series, a test has "standard" asymptotics (that is, Wald tests have the correct distribution asymptotically) only if the test can be arranged to be on stationary variables. For the Granger test regression, if we are testing for x causing y, the lag polynomial on x using p+1 lags is:

$$\sum_{i=1}^{p+1}\beta_i x_{t-i}$$

and the Granger test would be a joint exclusion test on all the β's. By repeatedly substituting

$$x_{t-i} = \Delta x_{t-i} + x_{t-(i+1)}$$

we can replace this with a sum in the form

$$\beta^*_{p+1} x_{t-(p+1)} + \sum_{i=1}^{p}\beta^*_i \Delta x_{t-i}$$

where the β^* are linear combinations of the original β. This is *almost* a linear combination of the stationary differences, but while you can do a different sequence of substitutions to shift the one lagged level term to the beginning rather than the end (or even can put it somewhere in the middle), no matter what you do, you will end up with a non-stationary term.

The idea behind Toda-Yamamoto(1995) (and independently Dolado-Lutkepohl(1996)) is to run an "augmented" regression with at least one extra lag and test all the lags *except the final one*. Because this is equivalent to a test only on differences, this has standard asymptotics whether x is *I(1)* or *I(0)*.

While this, at first glance, looks reasonable, it's, in fact, incorrect as a method for testing Granger causality, and examining why can help in understanding how non-stationarity can affect estimators.

First off, note that the result depends upon the "correct" lag length p being known, which will never be true in practice. But, even if p *is* known, the result is still flawed. The problem is that the

(continued on page 4)

Procedures and Paper Replication Links

Long-time RATS users have been familiar with the PDF files distributed with the software that list the "Procedures and Examples" and the "Paper Replication Programs". While somewhat useful, the problem is that these provide only a "snapshot" of those at the time the particular list was generated. We are now replacing these with pages in the on-line help, so they will always be up-to-date, and will also include links to more detailed descriptions where those are available (which is true for over 80% of the procedures and about 30% of the paper replications). The links are

https://estima.com/ratshelp/examples.html

https://estima.com/ratshelp/procedures.html

https://estima.com/ratshelp/paperreplications.html

ARCH/GARCH and Volatility E-Course, 2nd Edition

This is the 2nd edition of a course that was originally produced in 2012. We've significantly updated the course primarily to deal with issues that users have faced in analyzing volatility with difficult data sets (often from developing countries or emerging markets). Roughly half the course concerns the use of the existing **GARCH** instruction—determining the best specification, handling the estimation and doing tests of the adequacy of the model. The other half is on alternative methods of modeling volatility (from simple ones like rolling sample estimates to technically complex ones like stochastic volatility), less-standard GARCH specifications that require general likelihood maximization, and simulation methods applied to GARCH models for computing quantiles or doing inference.

It's important to note that all GARCH models are approximations. If they weren't, Tim Bollerslev wouldn't have a 44 page glossary of various flavors of GARCH on his web site:

http://public.econ.duke.edu/~boller/Papers/glossary_arch.pdf

Not only is it important to pick the most appropriate model for a task, but it's often even more important to pick the correct data series, data frequency and data range. The assumption underlying the estimation of a GARCH model is that the same process generates the data throughout the range. In many cases, it's not clear that that is the case. A thinly traded asset might have far too many days where nothing happens—switching to weekly returns may improve the ability to handle that. A market which has had a major change in regulation almost certainly will not admit a single model in a sample crossing the intervention point, and only rarely will some form of dummy shifts be able to fix that. A common question that we have to ask people who ask for technical support on GARCH models is "Have you looked at your data?" where simply graphing it shows that the assumption of a common model is dubious. Unfortunately, it's relatively rare for published papers to go into detail about why a particular set of data was chosen (which countries or assets, what sample range, etc.) which leaves the impression for readers that the modeling procedure is relatively painless. Sometimes it is; often it is far from it. The course gives enough theory to understand GARCH and related models, but the main point is to help a user decide upon a proper model, check how well it works and fix it if it needs fixing. The **"Diagnostics in Large Data Sets"** story in this newsletter is based upon an appendix in the course to answer a common question about problems getting a model which "passes the tests".

Some of the new material in the 2nd edition includes

- Greater coverage of non-trivial mean models, such as ARMA models and Vector Error Correction Models (VECM's).
- Tests for "spillover" in the mean and in the variance.
- Methods for detecting/handling structural breaks and extreme outliers.
- Generation of VECH representations for BEKK models.
- Use/abuse of "rolling window" estimates.
- Calculation of Variance Impulse Response Functions (VIRF) both closed form (for models that allow them) or through simulations (for models that don't).
- Computation and display of time-varying hedge ratios and portfolio weights.
- Cermeño-Grier-style "panel GARCH" models.

All examples have been updated to use newer features of the software such as the improvements to the **GARCH** instruction to simplify diagnostics and more complicated mean and variance models.

To order the GARCH e-course or any of the other e-courses, go to

https://estima.com/shopcart/courses.shtml

Diagnostics, contd from page 1

of the components probably will be a bit more dominant than the CLT's expect.

While Berkson was talking about using a chi-squared test, it's simpler to look at what happens with the commonly used Jarque-Bera test. Suppose we generate a sample from a t with 50 degrees of freedom (thus close to, but not quite, Normal). The following is a not atypical set of results from applying the Jarque-Bera test to samples of varying sizes:

N	Kurtosis	JB	Signif
100	0.243	0.448	0.799469
500	0.287	4.071	0.130635
1000	0.230	2.638	0.267441
2000	0.208	3.676	0.159127
5000	0.201	8.507	0.014212
10000	0.201	16.883	0.000216
100000	0.152	96.544	0.000000

The *N=100000* line would be an example of the "beyond any usual level of significance." The theoretical excess kurtosis for this is roughly .13—not a particularly heavy-tailed distribution, and one which is barely distinguishable in any meaningful way from the Normal. If we stick with the conventional significance level of .05, regardless of sample size, we are choosing to allow the probability of Type I errors to remain at .05 while driving the probability of Type II errors down effectively to zero. As we get more and more data, we should be able to push down the probabilities of *both* types of errors, and a testing process which doesn't is hard to justify.

The JB test has an asymptotic χ^2_2 distribution. That has a .05 critical value of 5.99, a .01 critical value of 9.21 and a .001 critical value of 13.81. In effect, the JB estimates two extra parameters (the skewness and excess kurtosis) and sees whether they are different from the 0 values that they would have if the distribution were Normal. If we look at how the SBC (or BIC) would deal with the decision about allowing for those extra 2 parameters, its "critical values" (the point at which we would choose the larger model) is 2 log T, which is 9.2 for T=100, 13.3 for T=1000, 18.4 for T=10000 to 23.0 for T=100000. Eventually (by T=100000), the data evidence in favor of the (correct) non-normal distribution is strong enough to cause us to choose it, at T=10000 the decision somewhat marginal (16.883 vs a critical value of 18.4), but at the smaller sample sizes, we would rather strongly favor the simpler model—the difference between the $t(50)$ and Normal just isn't apparent enough until we get a truly enormous amount of data.

Now JB has degrees of freedom fixed at 2. Let's look at a typical diagnostic test in time series analysis like the Ljung-Box Q test for serial correlation.

$$(1)\quad Q = T(T+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{(T-k)}$$

where $\hat{\rho}_k$ is the lag k autocorrelation (typically of residuals in some form). This is easier to examine more carefully if we take out some of the small-sample corrections and look at the simpler (asymptotically equivalent) Box-Pierce test

$$(2)\quad Q = T\sum_{k=1}^{h}\hat{\rho}_k^2$$

In practice, h is often fairly large (10 or more) and, in fact, the recommendation is that it increase (slowly!) with T. However, let's fix on h=25. Suppose we have T=2500 (a not uncommon size for a GARCH model). The .05 critical value for a χ^2_{25} is 37.7, the .01 is 44.3. Because of the T multiplier in (2), the size of correlations that triggers a "significant" result is quite small—if the typical autocorrelation is a mere .03, it would generate a Q of 56.3, which has a p-value of .0003. That's despite the fact that .03 autocorrelations are probably not correctable by any change you could make to the model: they're statistically significant, but *practically insignificant*. This is a common problem with high degrees of freedom tests. Now the h log T "critical value" suggested by the SBC (which would be 195.6 here) is a bit extravagant, since we're aren't really estimating a model with 25 extra parameters, but we should not be at all surprised to see a fairly reasonable-looking model on this size data set produce a Q statistic at least 2 or 3 times h, without the test suggesting *any* change to improve the model.

This is not to suggest that you simply brush aside results of the diagnostic tests. For instance, if residual autocorrelations are large on the small lags (1 and 2, particularly), that suggests that you need longer lags in the mean model—it's a "significant result" triggered by (relatively) large correlations at odd locations like 7 and 17 that is unlikely to be improved by a change to the model. One simple way to check whether there is anything to those is to compute the diagnostic on a split sample. What you would typically see is that the pattern of "large" autocorrelations is completely different on the subsamples and thus isn't a systematic failure of the model. Remember that the point of the diagnostics is to either suggest a better model or warn you of a serious problem with the one you're using. This type of "significant" test does neither.

Berkson, J.(1938), "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test," *Journal of American Statistical Association*, 33(303), 526–536.

Leamer, E.(1974), *Specification Searches*. New York: Wiley.

Causality Test, contd from page 1

connection between the TY test and Granger causality assumes that the extra, untested, lag is zero (since the true model has p lags, the coefficient on lag p+1 *must* be zero). However, that's where the flaw is. To see why in a simpler situation, suppose x is just a random walk:

$$x_t = x_{t-1} + u_t$$

For this model p is known to be 1. However, if you run a regression with lag 50 only (thus a 50 lag regression excluding lags 1 to 49), the coefficient on the lag 50 will not be zero; in fact, it will be near one. The problem with the TY logic is that if x is *I(1)*, even distant lags can proxy for the omitted ones. Note that this is not really dependent upon the process being *I(1)*. If the process has, for instance, a dominant root of .9, it's true that 49 lags may be enough to create an effective lack of correlation, a more typical 4 or 5 lags will leave the augmenting lag as a fairly strong proxy.

Now there have been Monte Carlo examinations of the TY test which seem to validate the approach. The fact that the "size" seems to be correct isn't a surprise, since that's what Sims, Stock and Watson would predict. And, it turns out that there are alternatives against which the TY procedure has power. However, there are others against which it does not.

To see where it "works", consider a case where y and x are *I(1)*, *but aren't* cointegrated. If that's the case, then the relationship between y and x can be written in differences. So if we look at the augmented polynomial for x:

$$\beta^*_{p+1} x_{t-(p+1)} + \sum_{i=1}^{p} \beta^*_i \Delta x_{t-i}$$

if some of the β^* on the differences are non-zero, the augmenting lagged level will be a *poor* proxy for those differences with non-zero coefficients—asymptotically, it has a zero correlation with what are (to it) future differences. So the restricted regression will end up being almost the same as a full exclusion of all lagged x. Where the TY test has a "blind spot" is where the series are cointegrated and there is "long-run" causality from x to y, that is, the coefficient on the cointegrating vector is non-zero. Because the causality comes in on a lagged *level* (rather than lagged differences), the augmenting lag can do a much better job of picking it up. For example, suppose we have

$$x_t = x_{t-1} + u_{xt}$$
$$y_t = y_{t-1} - .025\left(y_{t-1} - x_{t-1}\right) + u_{yt}$$

where the residuals are independent *N(0,1)*. *p=1* is known, so you don't have the uncertainty about the lag length that you would have in practice. The TY test runs a regression of y on two lags of x and y and tests the exclusion of lag one of x. And with 500 observations, it barely has any power, rejecting non-causality only about 10% of the time using a 5% significance level. By contrast, the standard Granger test (same regression but excluding *both* lags of x) rejects non-causality 80% of the time. Now that test has a non-standard distribution, but when the distribution is simulated, it isn't that far off from the usual F in this case, so the 80% rate is roughly correct.

Note that not only is this fundamentally flawed, but there have been papers written more recently which have "bootstrapped" the TY test. Since the whole (only) "point" of the TY test is that it has standard asymptotics, bootstrapping it makes no sense at all from a statistical standpoint.

The program used for doing the simulations is available on the RATS forum at

https://estima.com/forum/viewtopic.php?f=8&t=3136

Also, note that there is a rather lengthy topic in the RATS help on various issues with causality testing at

https://estima.com/ratshelp/causalitytesting.html

which includes links to a number of programs for testing for short- and long-run causality, causality in GARCH models, causality in ARDL models, causality in panel data and the use/abuse of rolling window causality tests.

References:

Dolado, J. J. and H. Lutkepohl(1996), "Making Wald Tests Work for Cointegrated Systems", *Econometric Reviews*, vol 15, pp 369-386.

Sims, C.A., J.H. Stock, and M.W. Watson (1990), "Inference in Linear Time Series Models with Some Unit Roots", *Econometrica*, vol. 58, No. 1, pp 113-144.

Toda, H.Y. and T. Yamamoto (1995), "Statistical inference in vector autoregression with possibly integrated processes", *Journal of Econometrics*, vol 66, pp 225-250.

The RATSletter

1560 Sherman Ave, Suite 1029
Evanston, IL 60201 USA
www.estima.com
sales@estima.com
874–864–8772