607

Persistence: some practical observations and suggestions

Jon Faust

http://e105.org/e607

November 30, 2015

▶ **Where we at on persistence?**

- We've had a slog through lots of unit root stuff

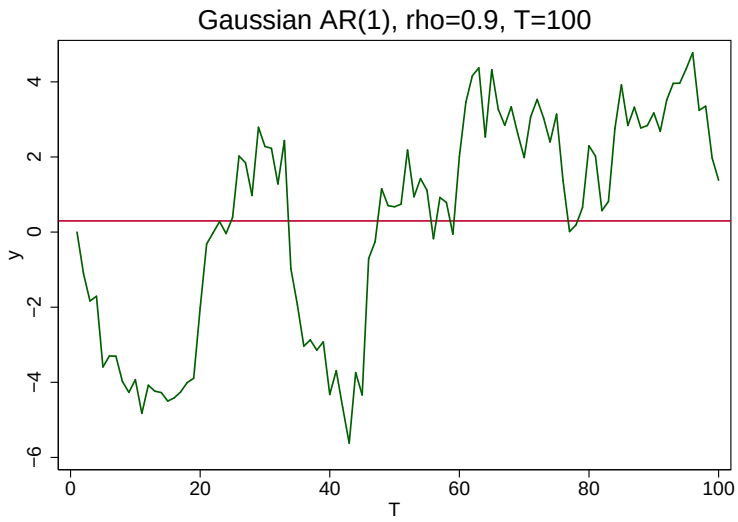
  and I've mainly said that it isn't very helpful in practice.

- In this section, I'll offer some observations and suggestions about how to proceed.

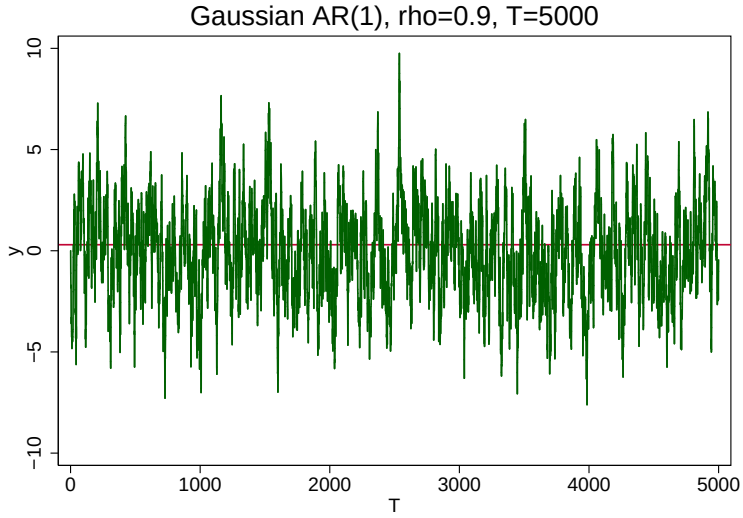
▶ **Persistence: a practical definition**

- A variable shows persistence on a given sample if, say, having risen above (or below) its sample mean it tends to stay there for a *considerable period*.
- *Considerable period* should be thought of as a significant or nontrivial share of the sample period at hand.
- This is a degreed notion: a variable can show a lesser or greater degree of persistence.
- Consider a Gaussian AR(1) with $\rho = 0.9$.

▶ **A Gaussian AR(1),** $\rho = 0.9$**,** $T = 100$



- **A Gaussian AR(1),** $\rho = 0.9$**,** $T = 5000$



- **Lesson: Same process** ($\rho = 0.9$)**,**
  - Small sample: wanders away from the mean and stays for an appreciable share of the sample span.
  - Larger sample, same process, wanders away from mean, but only for trivial part of sample.
- **Persistence that is relevant in practice is a function of the sample size and not simply a feature of the process driving the data**
- **Persistence that lasts a nontrivial share of the sample span will often cause CAN framework approximations to be poor**

▶ **Observations and suggestions**

- These will all be couched in the context of linear regression models.
- Suitable adaptation to, e.g., GMM context is not too difficult conceptually

▶ **1. Make sure your residuals are not persistent**

- Long ago (before HAC standard errors) folks usually assumed that residuals were not serially correlated.
- Good practice involved checking for serial correlation and re-working the specification until that assumption seemed justified.
- Nowadays, sometimes folks just hope that any serial correlation in residuals can be accommodated by using HAC standard errors.
- This may work ok so long as the residual serial correlation is modest, but when residuals look very persistent, this is probably a bad approach

▶ **Suggestion**

- Consider any regression results to be highly suspicious if the residuals look persistent

  you should probably respecify the model or re-think things more generally in this case.

▶ **Aside:: Spurious regression**

- In the spurious regression case, the residuals will look persistent
- Thus, under this suggestion, nobody would ever misinterpret such results as meaningful.

▶ **A useful observation**

- In practice, if you include at least one lag of every included persistent variable the residuals should not be persistent.

  (so long as the variables are at most I(1))

- Loosely speaking, the regression can difference any variable it needs to difference in order to reach stationary residuals

▶ **An observation about inference regarding features other than persistence**

- Suppose you have a specification in which the residual is not persistent.

  Thus, you've adhered to the first suggestion

- If your coefficient of interest can be written as being on a nonpersistent variable, the distribution of the coefficient and standard test statistics about that coefficient should not be affected by persistence problems.
- Formally, coefficients on I(0) variables follow standard CAN results asymptotically.
- Key theory results here come from Sims, Stock, Watson, 1990.

  Very nice explanation is in Watson's Handbook of Econometrics chapter go

  http://www.sciencedirect.com/science/article/pii/S1573441205800169

▶ **Aside:: Intuition**

- Problems arise when small alteration of the value of a coefficient may dramatically alter the persistence of the residual
- The extreme case of this is estimating an AR(1) on random walk data.
- With $\rho = 1$ the residual has finite variance; with any $\rho < 1$, it has infinite variance.
- The spurious regression case shows that once the residual has infinite variance all sorts of mischief can happen involving interactions with other infinite variance variables—even unrelated ones.
- When you are estimating a regression, however, so long as the regression can reach a nonpersistent residual, coefficients on nonpersistent variables should be fine.
- Suppose truth is:

$$y_t = \beta^{*\prime} x_t + \gamma^{*\prime} z_t + \varepsilon_t$$

  where x_t and y_t are persistent, but z_t and ε_t are not.
- If you choose $\beta = \tilde{\beta} \neq \beta^*$:

$$y_t = \tilde{\beta}' x_t + \gamma^{*\prime} z_t + [\varepsilon_t + (\beta^* - \tilde{\beta})' x_t]$$

  the new residual, in [], now has a persistent component.
- If you choose $\gamma = \tilde{\gamma} \neq \gamma^*$:

$$y_t = \beta^{*\prime} x_t + \tilde{\gamma}' z_t + [\varepsilon_t + (\gamma^* - \tilde{\gamma})' x_t]$$

  since z is not persistent, neither is the new residual
- It is mainly when changing coefficients dramatically alters persistence that unconventional statistical properties emerge.

▶ **More generally**

- We can state this a bit more generally.
- Suppose you have a general maintained model and the null hypothesis that involves restrictions on that model
- Suppose that the residual looks I(0) and not very persistent under the general model, and looks similarly persistent under the restricted model.
- In this case, the hypothesis is not much intertwined with persistence issues and test statistics are likely to behave in a standard manner.

▶ **Where are we**

- Make sure your residuals are not persistent
- And then if your hypotheses don't involve persistence properties, you should be ok.

▶ **But what if you want to do inference about persistence properties?**

▶ **Two additional topics**

- What do we do when the questions of intrerest inherently involve persistence properties?
- Where do impulse response inferences fall in this discussion?

▶ **Inference about phenomena lasting a significant share of the total sample span.**

▶ **Inference about persistent phenomena**

- First observation is that inference about persistent phenomena is fraught with difficulty.
  (With only slight exaggeration: it is not clear that anyone has ever generated a valuable applied result in this area.)
- Very often, an honest appraisal is that the answer folks are seeking is simply not in the data

▶ **Ok, so you want to go on despite the near futility**

▶ **If you are drawing inferences about phenomena that last a significant share of the sample span, you probably should be guided by a theory designed specifically for this enterprise**

▶ **Motivating example: variance ratio tests**

- One test statistic for determining the 'random-walk-like' properties of a series is

$$V_T(k) = \frac{\hat{\text{var}}((1-L^k)y_t)}{k\hat{\text{var}}((1-L)y_t)}$$

  where $\hat{\text{var}}$ is the sample variance

- In population, $V(k)$ is 1 when y_t is a random walk

  variance of $(1 - L^k)y_t$ increases linearly for a random walk

- In the old days, we tested $H_0 : V_T(k) = 1$ with a main alt. of interest $V_T(k) < 1$

  When $V_T(k) < 1$, shocks to the series die out (at least a bit) over k periods, rather than remaining eternally in the series at full value as in a random walk

▶ **Asymptotics: Two versions**

- Conventional asymptotics: Under the random walk null (and sensible regularity conditions):

$$\sqrt{T}(V_T(k) - 1) \to N(0, \nu^2(k))$$

  where T is growning, k is fixed, and ν^2 is a function of k.

▶ **Monte Carlo**

Rejection rate, VR(k), crit. val. 1.96

	T					
k	50	100	150	200	500	1000
10	0.020	0.032	0.040	0.043	0.046	0.050
20	0.005	0.019	0.024	0.031	0.043	0.047
30	.	0.013	0.020	0.023	0.038	0.045
40	.	0.005	0.016	0.021	0.034	0.042
100	.	.	.	.	0.020	0.029
200	.	.	.	.	0.004	0.020

Note: see `vrkMC.m` saved with this lecture.

- The conventional asymptotics become progressively less accurate when k/T gets larger.
- We are focussing on whether dependence dies out a bit over k periods.
- When this question involves a nontrivial share of the sample span the CAN framework approximations are not so good.
- They essentially assume T is large enough that nothing lasts a nontrivial share of the sample span.

▶ **An alternative asymptotic device**

- Suppose our sample is of size $T = 60$ and our chosen $k = 12$, so the stat. is

$$V_{60}(12)$$

- In this case $k/T = 1/5$. That is, we are asking about stuff that lasts about $1/5^{th}$ of the sample size.
- How about we consider the distribution of

$$V_T(T/5)$$

as T grows.

▶ **Aside:: Huh?**

- Suppose that the exact distribution is

$$V_T(T/5) \sim G_{T,k}$$

- CAN framework gave that as T grows,

$$G_{T,k} \to N(0, \nu^2(k))$$

- However, we could also consider the sequence of distributions given by

$$G_{T,T\kappa}$$

for fixed κ.

- Whenever $T\kappa$ and k correspond, these two are the same, but we are considering different limiting sequences

▶ **Same Monte Carlo again**

Rejection rate, VR(k), crit. val. 1.96

	T					
k	50	100	150	200	500	1000
10	0.020	0.032	0.040	0.043	0.046	0.050
20	0.005	0.019	0.024	0.031	0.043	0.047
30	.	0.013	0.020	0.023	0.038	0.045
40	.	0.005	0.016	0.021	0.034	0.042
100	.	.	.	.	0.020	0.029
200	.	.	.	.	0.004	0.020

Note: see `vrkMC.m` saved with this lecture.

- Along the diagonal, $\kappa = k/T$ is constant: 1/5
- Suggests a consistent limiting tail probability so long as $\kappa = k/T$ is constant.

- As it turns out, viewing the statistic as $V_T(T\kappa)$ for fixed κ leads to a asymptotic distribution that is a function of Brownian motions as discussed in the unit root asymptotics lecture.
- See Richardson and Stock, 'Drawing inferences from ...,' J. Financial Ecmet 1989, v1 p323
  go
  http://dx.doi.org/10.1016/0304-405X(89)90086-X

▶ **Aside:: The normalization**

- The discussion above was not strictly correct.
- Conventional asymptotics delivered

$$\sqrt{T}(VR(k) - 1) \rightarrow_d N(.,.)$$

- As is generally the case, the alternative asymptotics have a normalization by a different power of T, T^0 in fact:

$$VR(T\kappa) \rightarrow_d FBM(\kappa)$$

  where FBW means function of Brownian motion, and the function depends on κ.
- Thus, in the above discussion the $G_{T,T\kappa}$ distribution is under a different normalization.
- Not sure I should present it that way, but I did so to emphasize the essence

▶ **In short,**

- It pays to think about persistence and about our statistics relative to the sample size at hand.
- And there is often an asymptotic theory to help us formalize this idea.

▶ **Example 2: Roots near the unit circle, again.**

▶ **Unit roots**

- We don't know or care whether any process has a unit root.
- But if we are doing inference about issues such as the size of roots near the unit circle, persistence issues will affect the distribution our statistics.

▶ **Aside:: Unit root testing**

- In many courses, there would now ensue a long discussion of unit root tests
- But, we can't know and don't care.
- There are two reasons to do a small digression on unit root testing, however.
  - So you can converse with the profession

- Some stuff that might be useful is derived from the unit root testing framework.

- Thus, let me do a quick summary

▶ Dickey-Fuller and augmented Dickey-Fuller tests

- If you run an AR(1), perhaps with a constant or constant and time trend (or some other deterministic trend), you can base a unit root test either on the estimated ρ coefficient or the associated t statistic for testing $\rho = 1$.
- The distribution under the null, as noted in the first lecture, will differ depending on what deterministic elements are included.
- These are the Dickey-Fuller unit root tests and the distributions are sometimes called the Dickey-Fuller distribution
- Note: You can equivalenly run,

$$y_t = \rho y_{t-1} + \varepsilon_t$$

and test $\rho = 1$ or run

$$\Delta y_t = \alpha y_{t-1} + \varepsilon_t$$

and test $\alpha = 0$.

- We just described the simple AR(1) case. If there is residual correlation, this will affect the test.
- One 'solution' is to include enough lagged Δys to soak up this correlation.
- If you do so, the resulting test is called the augmented Dickey-Fuller test and the distributions are as in the simple case.
- (Phillips and Perron derived a different way of accouting for the residual correlation that in practice seems to work less well and in any case is not much used.)

▶ Optimal unit root tests (in a particular sense)

- There is no uniformly most powerful test of the unit root null against all natural alternatives.
- But folks have derived tests that are optimal under some more general criterion
- In particular, you form a point optimal test following the Neyman-Pearson lemma.

  this test will differ by which alternative value of the root you presume

- And then from among the point optimal tests create or choose a single test that performs well against a broad range of alternatives.
- Elliot, Rothenberg, and Stock have done this for the unit root case.

  (Specifically, they start with tests that have optimal power against point hypotheses of the local-to-unity variety discussed below.)

- Cite:

  Elliott, Graham, Rothenberg, Thomas J & Stock, James H, 1996. Efficient Tests for an Autoregressive Unit Root, Econometrica, Econometric Society, vol. 64(4), pages 813-36, July.

- Great use of theory. Practical value: can't know, don't care.
- O.K., enough of that

▶ Practical inference

- Suppose we want to know how quickly shocks die out in the limit or some similar question that will be intermingled with the size of roots near the unit circle.
- As an example, we'll consider forming a confidence interval for the inverse of the root nearest the unit circle.

  That is, we want ρ where the root nearest the unit circle is the root of $(1 - \rho L)$

- We are going to need a sample size dependent notion of a large ρ.

▶ Nearly integrated or local to unity asymptotics

- Define the process

$$(1 - \rho_T L)A(L)y_t = \varepsilon_t$$

  where $E\varepsilon_t = 0$ and ε is, say, a Martingale difference sequence, and

$$\rho_T = 1 - c/T$$

  for some fixed c.

  For any fixed c, as T grows, ρ approaches 1 and the process becomes more persistent.

- Thus, we'll call the asymptotic results, 'local-to-unity' asymptotics
- Clearly, we cannot observe such processes in practice: every time a new observation comes in, the entire process changes.
- But we can still derive limiting distributions of statistics under this changing DGP

▶ Spell it out a bit

- We have $Y \sim P_{\theta(T)}, \theta(T) \in \Theta$

  Where the distribution is as defined above.

- In particular, $\theta(T) = \{c, T, \psi\}$ where ψ is any parameters of the εs.
- As usual, we want to know the distribution of some statistic $\phi(Y)$
- For concreteness, you can suppose $\phi(Y)$ is the t statistic for testing that ρ equals the true ρ_T.

- Each $\theta(T)$ will induce an exact distribution for ϕ in a sample of size T, call it:

$$G_{\theta(T),T}(.)$$

- This is like our usual case except θ changes with T.
- We nonetheless have a well-defined sequence of distributions.

  And we can ask whether the sequence of distributions converges.

- As it turns out, in this case we have that,

$$G_{\theta,T}(z) \to G^{\infty}(z;c)$$

  where G^{∞} is (as usual) a FBM

▶ **More generally ...**

- I've been referring to the t statistic but we can derive an analogous distribution for lots of other statistics.
- Nice discussion: Stock J. Confidence Intervals for the Largest Autoregressive Root in Macroeconomic Time Series. Journal of Monetary Economics. 1991;28:435-459.

  go

  http://scholar.harvard.edu/files/stock/files/confidence%5Fintervals%5Ffor%5Fthe%5Flargest%5Fautoregressive%5Froot%5Fin%5Fmacroeconomic%5

▶ **A sketch of how to use this result**

- Suppose we want a confidence interval for the largest ρ of a process.
- We could run the augmented Dickey-Fuller-style regression

$$y_t = \alpha + \rho y_{t-1} + \sum \gamma_j \delta y_{t-j} + \varepsilon_t$$

- Ignoring persistence issues, we could compute, say, a size 5 percent t test of $H_0 : \rho = \rho_0$ for all possible $|\rho| < 1$.

  We could do this taking the t statistic to be asymptotically $N(0,1)$.

- Then we could take the nonrejected ρs as a 95 percent confidence interval.
- As you know, this would lead to the standard $\hat{\rho} \pm 2\hat{\sigma}$ style confidence interval.

▶ **The alternative approach**

- Run the same regression

- Take the same set of t statistics
- But the statistics for testing $H_0 : \rho = \rho_0$ is instead viewed as a test of

$$H_0 : c = c_0$$

where c_0 solves

$$\rho_0 = 1 - c_0/T$$

and T is the sample size at hand.

- Having re-conceptualized the test, we then take the critical value from the distribution $G^{\infty}(.; c_0)$ described above.
- Because c affects the distribution of the statistic, and because this is not just a shift, the confidence interval will not be just '$\hat{\rho}$ plus and minus *something*'.

  But the idea of collecting the non-rejected c and, hence, the implied non-rejected ρs is straightforward.

- Stock shows how to create the confidence interval

▶ **Elaborations**

▶ **Elaboration: Adding a bootstrap wrinkle**

- Taking as given that we should think in terms of local-to-unity asymptotics, it is still true that we might want to use a bootstrap to help capture additional features of the case at hand.
- Thus, rather than simply using $G^{\infty}(.; c_0)$ we might want some bootstrap distribution depending on c_0.
- Thus, instead of using $G^{\infty}(; c_0)$ as the basis on which to reject $H_0 : c = c_0$, we could run a bootstrap for each c_0.
- Since c is continuous, we actually would choose to run a bootstrap for some grid of c_0s.
- An then we would take, say, the outer envelope of implied ρs for nonrejected c_0 our confidence interval.
- This is Hansen's grid bootstrap

  Bruce E. Hansen, The grid bootstrap and the autoregressive model, Review of Economics and Statistics (1999) go

  http://www.ssc.wisc.edu/ bhansen/papers/restat%5F99.html

▶ **Elaboration: Using more powerful tests**

- We have described inverting the acceptance region for the t test (under the nonstandard asymptotic approximation) to form a confidence interval

- We could use any valid test here.
- And indeed we could use a different test for each c_0.

  So long as each is a valid, say, 5 percent test.

- This might make sense if there is no uniformly most powerful test; that is, if the most powerful test varies with c_0 and perhaps with the alternative.
- Elliot and Stock discuss how to use point optimal tests in this context.

  Elliott, G. and Stock, J. H. (2001) Confidence intervals for autoregressive coefficients near one. J. Ecmet., 103, 155–81. go

  http://dx.doi.org/10.1016/S0304-4076(01)00042-2

▶ **Takeaways?**

▶ **Takeaways**

- The key insight here is that we are doing inference about phenomena spanning a nontrivial share of the sample.
- We probably want to be guided by theory in which as the sample size grows, we continue to be performing inferences about features the span a nontrivial share of the sample.
- The variance ratio test and local to unity examples show how one might do this.

▶ **More generally**

- Ulrich Mueller has, with a number of co-authors, developed a body of inference techniques regarding persistence features
- These explicitly recognize that there is very little information about these features and are focussed on accurate inference in light of the limited information.
- If you hope to do meaningful inference about persistence features, you should go learn this stuff.
- Nicely summarized in Mueller and Watson

  Low Frequency Econometrics go

  http://www.princeton.edu/ umueller/ULFE.pdf

- This was going to be presented in our macro seminar this term, but was cancelled due to scheduling issues. Hopefully, will be presented soon.

▶ **And finally, a note about impulse response inferences**

▶ **Impulse responses**

- In a simple sense, the response of variable k to a shock ℓ periods earlier involves a coefficient on a not serially correlated variable: the shock.
- Thus, we should be able to do inference on impulse responses without worrying about persistence.
- But, in general, we estimate AR models and invert them to form our impluse responses.
- Thus, the impulse responses may inherit an ill-behavior in the estimates of the AR parameters due to persistence.

▶ **AR(1) case**

- The univariate AR(1) case makes this clear.
- We estimate ρ, and the impulse responses are then $1, \hat{\rho}, \hat{\rho}^2, \ldots$.
- Any problem estimating ρ will obviously affect the estimated impulse responses.
- On the other hand, at large lags, ρ^j for large j, the problem may not be too serious.
- More generally, when we estimate a VAR we estimate $\hat{A}(L)$ in

$$A(L)y_t = \varepsilon_t$$

  And the impulse responses are the coefficients of $\hat{A}(L)^{-1}$
- Just how persistence issues affect the estimate of $\hat{A}(L)$ and then how this affects inferences after performing the highly nonlinear inverse transformation is a bit opaque.
- If we had a sense of the number of problem roots, say, N, we could think about using an N-dimensional version of Hansen's grid bootstrap.

  In some cases, this may be becoming computationally feasible, but few seem to go this direction.
- Best practice at present seems to be some approach such as the Kilian bootstrap approach already described.
- In light of what we've just been discussing, the Kilian method takes account of the distribution of the estimate of the largest root by bias adjusting all coefficients of the bootstrap DGP

  But then does not bias adjust on each bootstrap draw.
- In practice, this seems to work ok.
- Best practice in this area is still evolving.