POLS571 - Longitudinal Data Analysis
September 20, 2001

Unit Roots, and Integration, Part II

1 Variants on the Dickey-Fuller Test

The Dickey-Fuller test requires that the us be uncorrelated. But suppose we have a model like the following, where the first difference of Y is a stationary AR(p) process:

$$\Delta Y_t = \sum_{i=1}^{p} d_i \Delta Y_{t-i} + u_t \tag{1}$$

This model yields a model for Y_t that is:

$$Y_t = Y_{t-1} + \sum_{i=1}^{p} d_i \Delta Y_{t-i} + u_t \tag{2}$$

If this is really what's going on in our series, and we estimate a standard D.F. test:

$$Y_t = \hat{\rho} Y_{t-1} + u_t \tag{3}$$

the term $\sum_{i=1}^{p} d_i \Delta Y_{t-i}$ gets lumped into the errors u_t. This induces an AR(p) structure in the us, and the standard D.F. test statistics will be *wrong*. There are two ways of dealing with this problem:

- Change the model (known as the *augmented Dickey-Fuller test*), or
- Change the test statistic (the *Phillips-Perron test*).

1.1 The Augmented Dickey-Fuller Test

1.1.1 Statistics

Rather than estimating the model in (3), we can instead estimate:

$$\Delta Y_t = Y_{t-1} + \sum_{i=1}^{p} d_i \Delta Y_{t-i} + u_t \tag{4}$$

and test whether or not $\hat{\rho} = 0$. This is the *Augmented Dickey-Fuller test*. As with the D-F test, we can include a constant/trend term to differentiate between a series with a unit root and one with a deterministic trend:

$$\Delta Y_t = \alpha + \beta t + Y_{t-1} + \sum_{i=1}^{p} d_i \Delta Y_{t-i} + u_t \tag{5}$$

The purpose of the lags of ΔY_{t-i} is to ensure that the us are white noise. This means that in choosing p (the number of lagged ΔY_{t-i} terms to include), we have to consider two things:

1. Too few lags will leave autocorrelation in the errors, while
2. too many lags will reduce the power of the test statistic.

This suggests, as a practical matter, a couple different ways to go about determining the value of p:

1. Start with a large value of p, and reduce it if the values of $\hat{d}_i$ are insignificant at long lags – This is generally a pretty good approach.
2. Start with a small value of p, and increase it if values of $\hat{d}_i$ are significant. This is a less-good approach...
3. Estimate models with a range of values for p, and use an AIC/BIC/F-test to determine which is the best option. This is probably the best option of all...

A sidenote: AIC and BIC tests:

The Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) are general tests for model specification. They can be applied across a range of different areas, and are like F-tests in that they allow for the testing of the relative power of nested models. Each, however, does so by penalizing models which are overspecified (i.e., those with "too many" parameters). The AIC statistic is:

$$AIC(p) = ln\hat{\sigma}_p^2 + \frac{2p}{N} \tag{6}$$

where N is the number of observations in the regression, p is the number of parameters in the model (including ρ and α), and $\hat{\sigma}_p^2$ is the estimated σ^2 for

the regression including p total parameters. Similarly, the BIC statistic is calculated as:

$$BIC(p) = ln\hat{\sigma}_p^2 + \frac{p\ lnN}{N} \tag{7}$$

The idea is to calculate these statistics for a range of different values of p, and then choose the model in which the statistic is the *lowest*. Note that the BIC statistic imposes a greater"penalty" for larger numbers of parameters; this means that the model "selected" using the BIC statistic will always be at least as parsimonious as that chosen using AIC.

1.1.2 An Example

Consider the data we discussed on Tuesday: The (logged) number of bills passed by the U.S. Congress, 1789-1990. While its possible to do ADF tests by "brute force", the canned Stata routines are much easier to use. `-dickey-`, `-dfuller-` and `-unitroot-` will estimate ADF tests for unit roots:

- For `-dickey-`, the user must specify the number of lags of ΔY to include, as well as whether a trend term is to be included or not. Also, specifying `-f-` (which stands for `-findlag-`) will report the number of lags "chosen" by the RMSE, AIC and BIC (labeled SIC) statistics. In addition, if you want a report of the RMSE/AIC/SIC statistics, use the `-detail-` option. These options make `-dickey-` a useful command.

- For `-dfuller-`, we are required to specify the number of lags of ΔY we wish to include, as well as whether or not we wish to include a trend or exclude the constant.

- `-unitroot-` is essentially the same as `-dickey-`, but without the `-findlag-` option. However, it also estimates Phillips-Perron test statistics (see below).

Examining ADF tests for unit roots on the "bills" data, we find the following:

Lags	No Constant (using `-dfuller-`)	With Constant (using `-dickey-`)	With Trend (using `-dickey-`)
1	0.31	-2.53	-2.69
2	0.31	-2.27	-2.35
3	0.39	-2.43	-2.40
4	0.44	-2.06	-1.72
5	0.54	-2.19	-1.62
6	0.63	-2.31	-1.50
7	0.81	-2.37	-1.03
8	0.75	-2.44	-1.11
Optimal Lags:			
RMSE		7	7
AIC		0	7
BIC		0	7

These results suggest a few things. First, if we buy the tests, we should use a standard D-F test (i.e., one with zero lags of ΔY on the right-hand side) for the model with no trend, and one with seven lags of ΔY for the model with a trend. Examining these statistics, and comparing them to the critical values given the other day, we can see that under no circumstances can we reject the null of a unit root in the data.

1.2 Phillips-Perron Tests

1.2.1 Intuition

The Phillips - Perron test for a unit root adopts a little different strategy. Rather than changing the model estimated, the P-P test, we stick with the model we talked about before:

$$\Delta Y_t = \alpha + \rho Y_{t-1} + u_t \tag{8}$$

The requirement that the us be white noise comes from the fact that the limiting distributions of the test statistics depend on the correlation of the residuals. In particular, the shape of the distributions depend on the ratio $\frac{\sigma^2}{\sigma_e^2}$, where σ^2 is just the variance of the innovations (the us) and $\sigma_e^2 = \lim_{T\to\infty} T^{-1} \sum_{j=1}^{T} E[(\sum_{i=1}^{t} u_i)_j^2]$. This latter term is a measure of the temporal covariance of the residuals u. In a nutshell, so long as $\sigma^2 = \sigma_e^2$, then the test statistics converge to the Dickey-Fuller distribution we discussed Tues-

day. If they are not equal, then the (asymptotic) shape of the distribution changes; the larger the difference between σ^2 and σ_e^2, the farther away from a "true" D-F distribution the test statistic will be.

The idea behind the Phillips-Perron tests is to use an empirical estimate of σ^2 and σ_e^2 to adjust the statistic itself, so that it more closely conforms to the "standard" D-F distribution. The calculation of the statistics is complicated, and differs depending on whether the model includes a constant term and/or a trend (see Maddala and Kim or Hamilton (pp. 507-515) for good discussions) but there are essentially two statistics, Z_ρ and Z_t (sometimes called Z_τ). The former test statistic follows the same limiting distribution as the $T(\rho - 1)$ D-F statistic (which we really didn't talk about) while the latter uses the same critical values as the D-F $\hat{\rho}$ statistic.

As with so many of these other tests, the researcher is required to specify the number of lags which s/he believes to be important in estimating $\hat{\sigma_e}$. Generally, small numbers of lags are preferred on both empirical and theoretical grounds; and sensitivity checking is usually a good idea.

1.2.2 Implementation

Hamilton (1994, pp. 511-512) has a good discussion of how to calculate the P-P test "by hand" from regular regression output (its not all that hard). A better option is to use either the `-unitroot-` or `-pperron-` commands in Stata. Given the choice, I prefer `-pperron-`, since it reports test statistics as well as critical values. Estimating this test for the Congressional bills data yields the following results:

```
. pperron lnbills, lags(1)
...
. pperron lnbills, lags(1) trend
...
```

Lags	No Trend $Z(\rho)$	No Trend $Z(t)$	With Trend $Z(\rho)$	With Trend $Z(t)$
1	-10.45	-2.58	-20.72	-3.30
2	-9.89	-2.53	-20.86	-3.32
3	-9.79	-2.53	-21.72	-3.38
4	-9.24	-2.48	-21.56	-3.37
5	-9.17	-2.47	-22.19	-3.41
6	-9.09	-2.46	-22.71	-3.45
7	-8.82	-2.44	-22.79	-3.46
8	-9.05	-2.46	-23.79	-3.53
$p < .05$ Critical Values	-13.70	-3.89	-20.70	-3.45

Note how stable the $Z(t)$ statistics are to changes in the number of autocovariance lags; this would be a good thing. The problem is that, while the results are unambiguous for the model without a trend term, the same is not true for the trending model. This would be a tough call; taken together with the other results, however, I'd tend to think that there's still a unit root in the series.

2 Other Unit Root Tests

2.1 The KPSS test

One potential problem with all the unit root tests so far described is that they take a unit root as the null hypothesis. Kwiatkowski et. al. (1992) provide an alternative test (which has come to be known as the *KPSS test*) for testing the null of stationarity against the alternative of a unit root. This method considers models with constant terms, and either with (their ν_τ statistic) or without (ν_{mu}) a deterministic trend term. Thus, the KPSS test tests the null of a level- or trend-stationary process against the alternative of a unit root.

Formally, the KPSS test is equal to:

$$LM = \sum_{t=1}^{T} \frac{S_t^2}{\hat{\sigma}_\epsilon^2} \tag{9}$$

where $S_t^2 = \sum_{i=1}^{t} \hat{u}_i$ is the running partial sum of the residuals and $\hat{\sigma}_\epsilon^2$ is the

estimated error variance from the regression:

$$Y_t = \alpha + \epsilon_t \tag{10}$$

or:

$$Y_t = \alpha + \beta t + \epsilon_t \tag{11}$$

for the model with a trend.

The practical advantages to the KPSS test are twofold. First, they provide an alternative to the DF/ADF/PP tests in which the null hypothesis is stationarity. They are thus good "complements" for the tests we've focused on so far. A common strategy is to present results of both ADF/PP and KPSS tests, and show that the results are consistent (e.g., that the former reject the null while the latter fails to do so, or vice-versa). In cases where the two tests diverge (e.g., both fail to reject the null), the possibility of "fractional integration" should be considered (e.g. Baillie 1989; Box-Steffensmeier and Smith 1996, 1998).

The other practical advantage to the KPSS test is that there is a user-written Stata routine for estimating it (hooray!). The `-kpss-` command requires that one specify the lag length (or "bandwidth") since the denominator of the formula is an empirical estimate of the long-run variance of the time-series, as calculated by the estimated autocovariance function. Estimating a KPSS test on our now-ubiquitous Congressional bills data yields the following:

```
. kpss lnbills, notrend maxlab(8)
. kpss lnbills, maxlag(8)
```

Lags	No Trend	With Trend
0	6.23	1.24
1	3.33	0.72
2	2.31	0.52
3	1.79	0.42
4	1.48	0.36
5	1.26	0.32
6	1.11	0.29
7	0.99	0.26
8	0.90	0.24
$p < .05$ Critical Values	0.463	0.146

These results are broadly consistent with those for the DF/ADF/PP tests; we soundly reject the null of stationarity in the model without a trend term, but do so only inconsistently when a trend is included. Overall, these results make it somewhat difficult to determine whether the "bills" series either (a) has a unit root, or (b) is more-or-less stationary around a deterministic, upward-sloping trend.

2.2 Variance-Ratio Tests

A quick review of variance-ratio tests:

- Based on the idea that, if a series is stationary, the variance of the series is not oincreasing over time; while a series with a unit root has increasing variance.
- Intuition: Compare the variance of a subset of the data "early" in the series with a similarly-sized subset "later" in the process. In the limit, for a stationary series, these two values should be the same, while they will be different for an I(1) series. Thus, the null hypothesis is stationarity, as for the KPSS test.
- There's a good, brief discussion of these tests in Hamilton (p. 531-32). Other cites are Cochrane (1988), Lo and McKinlay (1988), and Cecchetti and Lam (1991).

- There currently isn't a Stata routine for estimating these (though I once wrote a RATS proc for doing so). Think of them as just another possible alternative for testing for a unit root.

3 General Issues in Unit Root Testing

The Sims (1988) article I assigned is to point out an issue with unit root econometrics in general: that classicists and Bayesians have very different ideas about the value of knife-edge unit root tests like the ones here.[1]

Unlike classical statisticians, Bayesians regard ρ (the "true value of the autocorrelation parameter) as a random variable, and the goal to describe the distribution of this variable, making use of the information contained in the data. One result of this is that, unlike the classical approach (where the distribution of $\hat{\rho}$ is skewed), the Bayesian perspective allows testing using standard t distributions. For more on why this is, see the discussion in Hamilton.

Another issue has to do with lag lengths. As in the case of ARIMA models, choosing different lag lengths (e.g. in the ADF, PP and KPSS tests) can lead to different conclusions. This is an element of subjectivity that one needs to be aware of, and sensitivity testing across numerous different lags is almost always a good idea.

Finally, the whole reason we do unit root tests will become clearer when we talk about cointegration in a few weeks.

[1]The discussion here relies on both Sims (1988) and the overview of these arguments in Hamilton (pp. 532-34).