gretl **working papers**

Response surfaces for DF-GLS p-values

Allin Cottrell[1]

working paper #6

January 27, 2021

[1]Department of Economics, Wake Forest University, North Carolina 27109, USA; email: cottrell@wfu.edu. I have benefited from discussions on this topic with Oleh Komashko. However, Oleh is not responsible for the results presented here, and in fact regards my rough and ready approach as somewhat disreputable! He is seeking a fresh theoretical approach to the issue and I await his results with interest.

1 Introduction

In a well known paper Elliott, Rothenberg and Scott (1996) (hereinafter, ERS) advocated a modification of the then-standard method of computing unit root tests of the Dickey–Fuller type. They argued that power could be increased by using a GLS procedure for removing the mean (and deterministic trend, if applicable) from the data to be tested; such tests are known as DF-GLS tests.

From the practitioner's point of view, one drawback with using these tests is that p-values are hard to come by. James MacKinnon (1996) produced an apparatus for calculating p-values for a wide variety of unit-root and cointegration tests, but his response surfaces have limited applicability to the DF-GLS tests: all they offer is asymptotic p-values for the case with a constant but no trend.[1] As regards the DF-GLS test with trend, the only generally available means of evaluating the test statistic is the table of critical values provided by ERS (their Table I, which gives simulated critical values at the 1%, 2.5%, 5% and 10% levels for sample sizes of 50, 100 and 200 as well as an asymptotic value).

This note applies MacKinnon's methodology—strictly speaking, a simplified version thereof—to DF-GLS, with the aim of providing reliable sample size-dependent p-values. Specifically, we aim to provide functions for computing p-values for the constant-only and constant-plus-trend cases of the test statistic of the Dickey–Fuller type, τ in MacKinnon's terminology; that is, the t statistic for a_0 in the regression

$$\Delta y_t^d = a_0 y_{t-1}^d + a_1 \Delta y_{t-1}^d + \cdots + a_p \Delta y_{t-p}^d + \text{error}$$

where y_t^d is the series of interest with mean (and possibly trend) removed via GLS.

2 Methodology

As in MacKinnon (1996), we carried out $N = 100$ trials for each of a set of sample sizes; after some experimentation we chose samples sizes of $T = 20, 25, 35, 50, 75, 100, 200, 500$ and 800. MacKinnon ran $M = 200,000$ replications per trial; taking advantage of Moore's Law, we raised this to $M = 10^6$.

At each replication, artificial data were generated in conformity with the given null: cumulated standard normal variates plus either a constant or a deterministic trend. Specifically, we used gretl version 1.10.90cvs of April 2015 to generate the pseudo-random normal variates. The PRNG is the DCMT implementation (Matsumoto and Nishimura (2000)) of the Mersenne Twister, with conversion to normality via the Marsaglia and Tsang (2000) "ziggurat" method, amended along the lines proposed by Doornik (2005).

For each of the N trials, a set of empirical quantiles was determined, based on the M replications, using α levels of 0.001, 0.0025, 0.005, 0.01, 0.02, 0.025, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 and 0.9 and 0.99. This α-grid is a good deal coarser than that used by MacKinnon, but as we argue below it seems to be fine enough for practical purposes.[2]

---

[1] As noted by ERS (p. 824), in this case "critical values and asymptotic power are those of the conventional Dickey–Fuller t statistic when there is no intercept."

[2] The α value of 0.025 was included specifically to allow comparison with the ERS table of critical values for DF-GLS with trend. In the constant-only case it was omitted.

Let $n_T = 9$ denote the number of sample sizes considered, and $n_\alpha = 26$ denote the number of α levels used. The output from the simulations then consists of an array of $N \times n_T = 900$ values of the τ statistic for each of the n_α quantiles. The next step is to estimate a response surface for each of the quantiles or significance levels.

MacKinnon's method was to average the critical τ values for each T across the N trials and then apply a GMM procedure, equivalent to weighted least squares, to these sample means: the sample means are regressed on a constant and up to three powers of $1/T$. A simpler procedure is to run a set of OLS regressions (of the same form) each one using all $N \times n_T$ data points for a given α. We didn't see compelling evidence of heteroskedasticity in the data, and decided to use OLS.[3] In our estimation we found that the cubic term in $1/T$ was strongly significant for all α, so our model for the trend case is

$$\tau_\alpha(T) = \beta_0 + \beta_1 T^{-1} + \beta_2 T^{-2} + \beta_3 T^{-3} + \varepsilon_\alpha \tag{1}$$

where eqn. (1) is estimated separately for each α. In the constant-only case, however, we found that T^{-4} was strongly significant at all α values, so we included that too.

3 Critical values: the trend case

A subset of results from this stage of the procedure is shown in Table 1, covering the significance levels that appear in ERS Table I for the trend case.

Table 1: Response surface estimates, trend case; $n = 900$ for each regression

α	1%	2.5%	5%	10%
β_0	−3.4069	−3.1029	−2.8471	−2.5593
	(0.00030)	(0.00022)	(0.00017)	(0.00014)
β_1	−20.025	−18.778	−18.184	−18.038
	(0.06458)	(0.04634)	(0.03687)	(0.02892)
β_2	22.426	50.000	74.291	104.71
	(3.2296)	(2.3175)	(1.8441)	(1.4463)
β_3	−942.71	−810.49	−750.39	−728.64
	(42.834)	(30.737)	(24.458)	(19.181)
SSR	0.0114	0.0058	0.0037	0.0023
R^2	0.9999	0.9999	0.9999	0.9999

Table 2 compares the critical values given by ERS with fitted values calculated using the coefficients in Table 1 (labeled "AC"), rounded to three significant figures for comparison. The values are mostly quite close, although note that the AC values resolve an anomaly in ERS, whereby in the 1% column the critical value for $T = 200$ is less extreme than that for infinite T.

A further perspective on the results is given by Figure 1, which shows OLS fitted curves based on Table 1, average critical values (per sample size) obtained in the present study,

---

[3]Our choice of OLS over WLS was not based solely on simplicity; we return to this point in section 3 below.

Table 2: Comparison of critical values

T	1% ERS	1% AC	2.5% ERS	2.5% AC	5% ERS	5% AC	10% ERS	10% AC
50	-3.77	-3.81	-3.46	-3.46	-3.19	-3.19	-2.89	-2.88
100	-3.58	-3.61	-3.29	-3.29	-3.03	-3.02	-2.74	-2.73
200	-3.46	-3.51	-3.18	-3.20	-2.93	-2.94	-2.64	-2.65
∞	-3.48	-3.41	-3.15	-3.10	-2.89	-2.85	-2.57	-2.56

and the critical values from ERS. These plots show two things: first, the very close fit to the AC sample average critical values inspires confidence in using the coefficients to compute critical values for intermediate sample sizes;[4] second, while the ERS values are mostly "in the ballpark" their 1 percent critical values in particular appear to be somewhat out of line (and there may be something awry with their values for $T = 200$).

The information in Figure 1 is relevant to our choice of OLS estimation as opposed to the WLS approach in MacKinnon. The general argument for WLS is that data-points with a higher error variance are less informative about the parameters of interest, so it makes sense to down-weight them. If we were quite confident that eqn. (1)—that is, a low-order polynomial in T^{-1}—is a correct specification of the dependence of DF-GLS critical values on T, and if it were also clear that the variance of critical values across trials is a function of T, this argument would have some force in the present context. As we show in section 5, there is some ground for saying that the variance of critical values is smaller for larger T, but the difference in variance is neither very great nor consistent across cases. And as the plots in Figure 1 show, the main point of interest is the curvature with respect to T at small to medium sample sizes. In this regard the results from trials with relatively small T may be *more* informative, allowing for the possibility of specification error; so it is not at all clear that it makes sense to down-weight them.

4   Critical values: constant only

Repeating the procedure described above for the constant-only case, we obtain a second set of response surface coefficients; the estimates for the most commonly used critical values are shown in Table 3.

As mentioned above, p-values obtained via MacKinnon's urcdist code (for the no-constant Dickey–Fuller case) are asymptotically valid for the DF-GLS case with no trend. It is of interest to see how useful these p-values are for finite samples. Table 4 is constructed as follows: the $\bar{\tau}$ values for each α and T are the sample means of empirical quantiles based, as before, on 100 trials of 10^6 replications each; and the $P_\infty(\bar{\tau})$ figures are p-values obtained using urcdist, passing $\bar{\tau}$ as parameter. It is hardly surprising to see that the asymptotic p-values are a poor approximation at small to moderate T.

---

[4]The points for $T = 800$ were omitted from the plots to avoid compressing the T axis unduly; however, these points fell almost exactly on the curves shown.

Table 3: Response surface estimates, constant-only case; $n = 900$

α	1%	5%	10%
β_0	-2.5666	-1.942	-1.6179
	(0.00039)	(0.00020)	(0.00015)
β_1	-20.186	-21.858	-23.817
	(0.13744)	(0.072541)	(0.054962)
β_2	272.12	378.58	449.01
	(12.294)	(6.4892)	(4.9166)
β_3	-4843.5	-6349.9	-7447.5
	(380.61)	(200.89)	(152.21)
β_4	$30613.$	$41599.$	$49745.$
	(3762.5)	(1985.9)	(1504.7)
SSR	0.0117	0.0033	0.0019
R^2	0.9998	0.9999	0.9999

Table 4: Asymptotic p-values, constant only

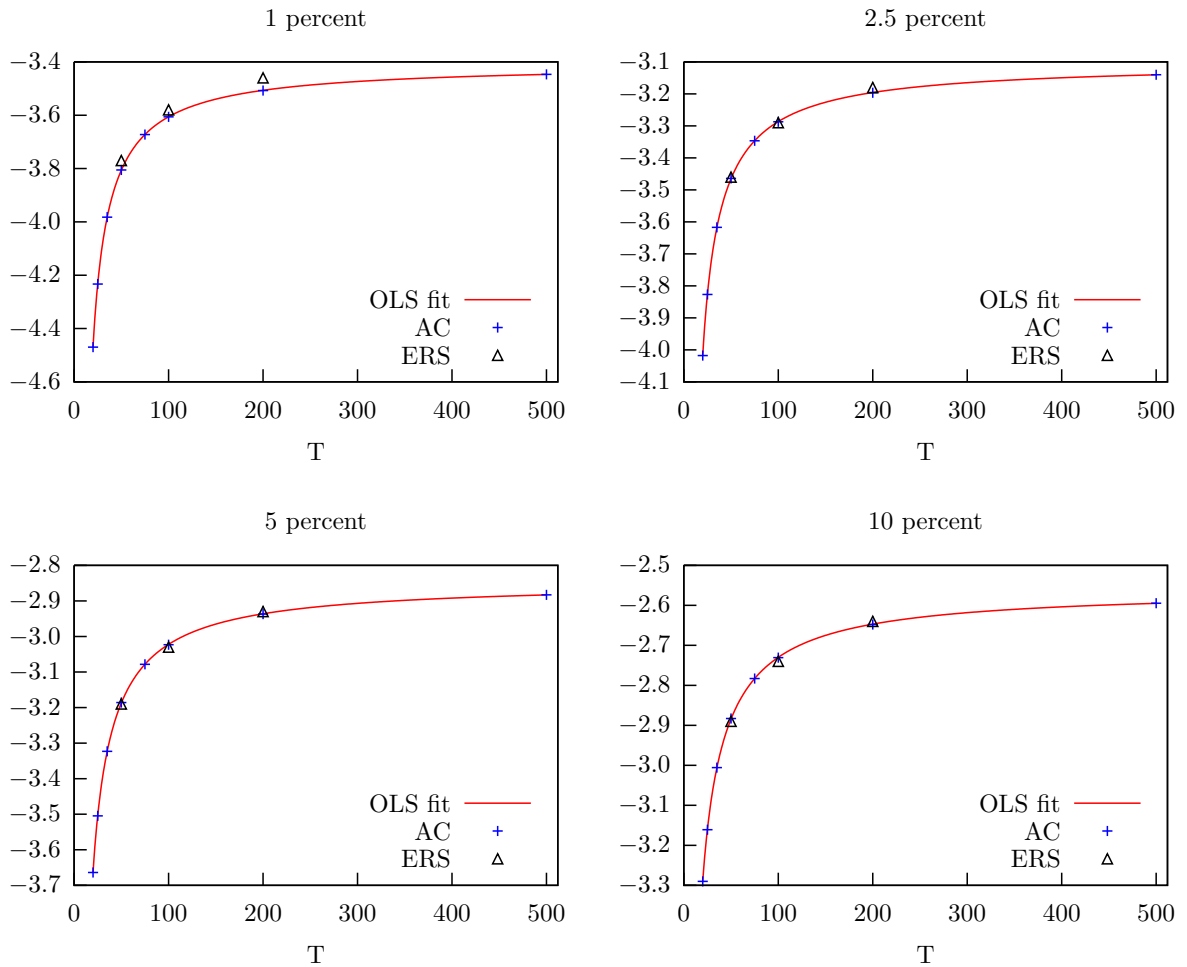| α	T	$\bar{\tau}$	$P_\infty(\bar{\tau})$
0.01	50	-2.8952	0.0037
	75	-2.7975	0.0050
	100	-2.7458	0.0059
	200	-2.6620	0.0075
	800	-2.5910	0.0093
0.05	50	-2.2718	0.0223
	75	-2.1795	0.0282
	100	-2.1285	0.0320
	200	-2.0432	0.0394
	800	-1.9683	0.0469
0.1	50	-1.9661	0.0472
	75	-1.8713	0.0585
	100	-1.8181	0.0658
	200	-1.7273	0.0798
	800	-1.6466	0.0942

Figure 1: Selected critical values as function of *T*

5 P-value functions

The response surfaces for critical values at a given set of significance levels are of some interest in their own right, but our goal is to create a means of producing reliable *p*-values for finite-sample DF-GLS tests. This requires implementing the second stage of MacKinnon's procedure. For a given test statistic, τ, obtained from a sample of size *T* the general idea is:

1. Calculate the critical values for all n_α cases covered by the analysis, at sample size *T*. This requires prior storage of the n_α sets of response-surface coefficients (a small subset of which we have shown above).

2. Determine which of these critical values is closest to the given τ, and in addition choose a suitable number of critical values in the "neighborhood" of τ.

3. Run a regression to gauge the shape of the relevant surface (α as a function of critical value) in the chosen vicinity.

4. Obtain an estimated *p*-value for the input τ based on a fitted value from this regres-

sion.

Theory does not dictate the form of the regression to be run at step 3, but MacKinnon suggests using a normal approximation:

$$\Phi^{-1}(\alpha) = \gamma_0 + \gamma_1 \hat{q}(\alpha) + \gamma_2 \hat{q}^2(\alpha) + \gamma_3 \hat{q}^3(\alpha) + \epsilon_\alpha \tag{2}$$

where the $\hat{q}$ values represent estimated quantiles and $\Phi^{-1}(\cdot)$ is the inverse of the standard normal CDF. This means that the dependent variable is the vector of significance levels associated with the selected critical values, run through $\Phi^{-1}(\cdot)$, and the independent variables are the levels, squares and (possibly) cubes of the selected critical values themselves.[5] Let $\hat{y}$ denote the fitted value obtained via (2), by applying the γ estimates to the input τ; then the estimated p-value is found as $\Phi(\hat{y})$.

As mentioned above, our grid of significance levels is much coarser than MacKinnon's. This means that we're not able to include as many local data points in estimating eqn. (2): we use 5 as against MacKinnon's 9.[6] Can we still produce p-values that are good enough for practical purposes?

There is no definitive metric for "good enough," but we have made this notion operational as follows. Taking our empirical quantiles as a basis, we try "round-tripping" by means of our p-value code. For instance, if $\bar{\tau}_{0.01}(50)$ is our empirical estimate of the 0.01 quantile of the τ distribution for $T = 50$ (the mean of 100 trials), we pass this $\bar{\tau}$ to our p-value code and see if it produces something close to 0.01. Allowing for sampling variation in the empirical quantiles, we also show the standard deviation of the quantiles across the trials, $\hat{\sigma}_\tau$, and compute p-values for $\bar{\tau} \pm \hat{\sigma}_\tau$; these are labeled $\hat{P}_T(\tau_-)$ and $\hat{P}_T(\tau_+)$.

The full set of results of this sort contains $n_\alpha \times n_T$ elements; we show a representative subset in Tables 5 (test with trend) and 6 (constant only). Table 7 is based on the same general idea, but uses a random selection of quantiles and sample sizes that were not included in the response-surface estimation.

It appears that there's an acceptable measure of agreement between the calculated p-values and the empirical quantiles. In all cases there is agreement to at least two significant figures, and in all but a few cases (of the trend test, for $\alpha = 0.9$) the $\hat{P}_T(\tau_-)$ and $\hat{P}_T(\tau_+)$ values (where shown) bracket the nominal α. We submit, therefore, that p-values produced in this way are in fact good enough for practical inference.

A further test of the generated p-values may be of interest: how self-consistent are they? This can be checked for a given test and sample size as follows.[7] First we fix on certain minimum and maximum values of τ—for example, the means of the simulated critical values associated with the smallest and largest values of α used in the analysis. We then run a loop from minimum to maximum, incrementing τ by a small step at each iteration. Computing the p-value at each iteration will give us an estimate of the CDF of τ; and theoretically this should, of course, be monotonically increasing.

However, the MacKinnon-type procedure carries no *a priori* guarantee of monotonicity. There's a potential discontinuity whenever the window of local points used in estimating

---

[5]MacKinnon drops the cubic term if its t-ratio turns out to be less than 2 in absolute value; we follow him in this.

[6]With only 1 degree of freedom (or 2 if the cubic term is dropped) this is a minimal piece of curve-fitting; the proof of the pudding is in the quality of the generated p-values. Note that by reducing the number of critical values, n_α, we get a basic results matrix of a size that can readily be stored within the code for generating p-values, rather than requiring that this matrix be read from file as in MacKinnon's urcdist.

[7]I have to thank Oleh Kamashko for suggesting this experiment.

Table 5: Interpolated p-values: trend case

α	T	$\bar{\tau}$	$\hat{\sigma}_\tau$	$\hat{P}_T(\bar{\tau})$	$\hat{P}_T(\tau_-)$	$\hat{P}_T(\tau_+)$
0.001	25	-5.2075	0.0134	0.0010	0.00097	0.00103
	50	-4.5819	0.0110	0.0010	0.00097	0.00104
	200	-4.1840	0.0077	0.0010	0.00097	0.00103
0.01	25	-4.2332	0.0042	0.0100	0.00989	0.01008
	50	-3.8053	0.0033	0.0100	0.00993	0.01011
	200	-3.5075	0.0031	0.0100	0.00988	0.01007
0.05	25	-3.5045	0.0021	0.0499	0.04968	0.05011
	50	-3.1859	0.0017	0.0501	0.04993	0.05034
	200	-2.9370	0.0018	0.0499	0.04967	0.05012
0.1	25	-3.1608	0.0015	0.0998	0.09952	0.10011
	50	-2.8829	0.0013	0.1002	0.09996	0.10051
	200	-2.6478	0.0014	0.0998	0.09949	0.10011
0.4	25	-2.3256	0.0011	0.3998	0.39922	0.40040
	50	-2.1229	0.0008	0.4004	0.39998	0.40084
	200	-1.8930	0.0008	0.3996	0.39909	0.40005
0.9	25	-1.3292	0.0012	0.8996	0.89929	0.89997
	50	-1.1945	0.0009	0.9000	0.89969	0.90024
	200	-0.9257	0.0010	0.9005	0.90021	0.90077

eqn. (2) moves a notch to the right (adding a rightward point and dropping a leftward one). If the approximation is insufficiently accurate, some τ values may have an estimated p-value less than that of their neighbor to the left. It is straightforward to write a program to detect such cases.

To establish a baseline we applied this check to MacKinnon's urcdist, in cases `nc` (no constant), `c` (constant only) and `ct` (constant plus trend) of the regular Dickey–Fuller test statistic, for an arbitrarily chosen sample size of 50. We ran the loop for $\hat{p}$ (estimated p-value) from 0.001 to 0.999 with a step size for τ of 0.0001: the results (anomalous cases only) are shown in Table 8. It turns out that there are a few instances of non-monotonicity: a negative $\Delta\hat{p}$ means that the estimated p-value decreased from $\tau - 0.0001$ to τ. While this is not nice from a theoretical point of view it is hardly a real concern for the practitioner: a handful of cases in tens of thousands of computations, with a maximum "reversal" of 0.000003. If we increase the step size for τ to 0.0005 the anomalous cases drop to 1 (in the `nc` model); and if we raise the step size to 0.0006 they disappear altogether. Moreover, 5 out of the 9 anomalous cases with the τ step of 0.0001 are in the range of $\hat{p} > 0.99$, which is of no interest for hypothesis testing.

Table 9 shows the results of the same experiment (with a τ-step of 0.0001) as applied to our variant of MacKinnon's procedure for the DF-GLS tests. There were just two anomalous cases in the contant-only test and none in the case with trend. The maximal reversal of

Table 6: Interpolated p-values: constant-only case

α	T	$\bar{\tau}$	$\hat{\sigma}_\tau$	$\hat{P}_T(\bar{\tau})$	$\hat{P}_T(\tau_-)$	$\hat{P}_T(\tau_+)$
0.001	25	-4.0505	0.0117	0.0010	0.00097	0.00103
	50	-3.6704	0.0101	0.0010	0.00097	0.00103
	200	-3.3892	0.0077	0.0010	0.00097	0.00103
0.01	25	-3.1701	0.0038	0.0100	0.00991	0.01010
	50	-2.8952	0.0036	0.0100	0.00990	0.01010
	200	-2.6620	0.0033	0.0100	0.00989	0.01008
0.05	25	-2.5103	0.0021	0.0500	0.04978	0.05027
	50	-2.2718	0.0018	0.0500	0.04980	0.05022
	200	-2.0432	0.0019	0.0499	0.04970	0.05015
0.1	25	-2.2012	0.0014	0.1000	0.09974	0.10035
	50	-1.9661	0.0014	0.1000	0.09973	0.10034
	200	-1.7273	0.0016	0.0999	0.09956	0.10019
0.4	25	-1.4639	0.0008	0.4004	0.39993	0.40090
	50	-1.2030	0.0009	0.4008	0.40034	0.40131
	200	-0.8855	0.0012	0.3997	0.39917	0.40031
0.9	25	-0.2749	0.0014	0.8998	0.89950	0.90006
	50	0.0936	0.0019	0.8997	0.89937	0.90001
	200	0.6543	0.0020	0.9002	0.89993	0.90056

the CDF is somewhat greater than with urcdist, though still too small to be of practical importance. Increasing the τ-step to 0.0004, the anomalies disappear.

Having computed CDFs, it may be of interest to show them: see Figures 2 and 3.

6 Augmented DF-GLS tests

The apparatus discussed above rests on simulations of Gaussian random walks with added deterministic terms. The practical question arises: how useful are p-values thus derived, in the common case where one or more lagged differences have to be added to the Dickey–Fuller regression to ensure a reasonable approximation to a white-noise error? There's no definitive, general answer to this question, but it may be worthwhile to consider a couple of examples.

We took two annual time series for the US—the log of real GDP and consumer price inflation—from 1947 to 2014 and conducted ADF-GLS tests, including a trend for log GDP and just a constant for inflation.[8] For the lag order, we tested down from a maximum of 10 using the modified AIC method of Ng and Perron (2001) as revised by Perron and Qu

---

[8]To be specific, the GDP series was the log of `GDPCA`, and inflation was the percentage change in the annual average of `CPIAUCSL`, both series taken from the `FRED` database at the St. Louis Fed.
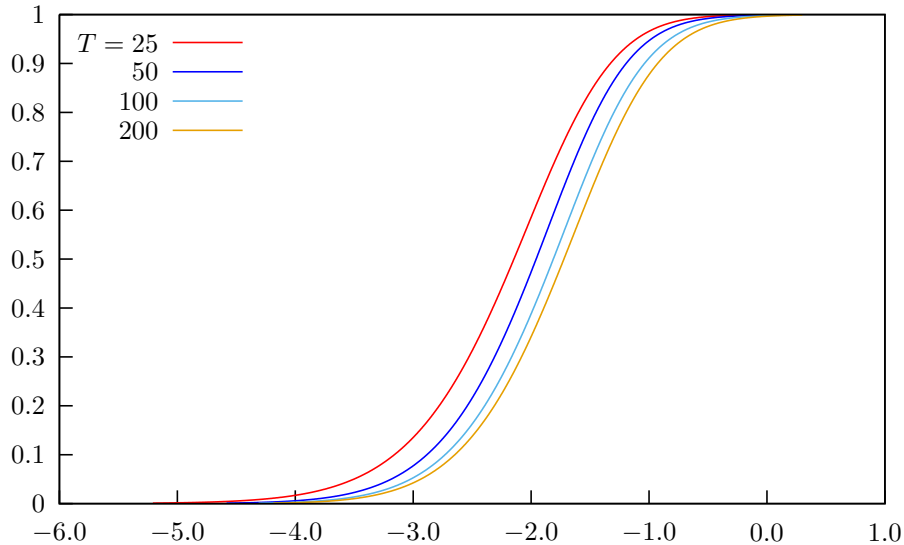
Figure 2: CDFs for τ, DF-GLS test with trend

(2007): this gave an order of 1 for GDP and 4 for inflation. We then simulated $I(1)$ data with the lag structure estimated for these two series via the Dickey–Fuller regression plus the deterministic components estimated via GLS, as follows:

	constant	trend	lagged differences
log GDP	7.60	0.033	0.179
inflation	5.62	–	0.174, −0.364, −0.116, −0.194

In each case we performed 100 trials with 250,000 replications per trial, for sample sizes of 50, 100 and 200, calculated the ADF-GLS test statistic (imposing the "known" lag order) on each replication, and computed quantiles for α = 0.01, 0.025, 0.05 and 0.1. Tables 10 and 11 show the means and standard deviations of critical values across the trials, $\bar{\tau}$ and $\hat{\sigma}_\tau$ respectively, along with estimated p-values calculated as described above, for both finite T and the limit distribution. For series with the characteristics described above, at any rate, the general picture is that one would over-reject the unit-root null by using asymptotic values, but somewhat under-reject by using finite-sample values based on the assumption of no autocorrelation of first differences. For $T = 100$ or less, one would be closer to correct inference by using finite-sample p-values.

7 Conclusion

We have presented a simplified variant of MacKinnon's (1996) approach to computing response surfaces for p-values for unit root tests, with application to the DF-GLS tests of Elliott *et al.* (1996). We have also provided evidence that the computed p-values are sufficiently accurate for practical purposes—that is, valid to 2 or 3 significant figures over a range that is of interest for hypothesis testing.

While we hope that this apparatus will be useful to practitioners, a word to the wise may be in order. Our claim of 2- or 3-digit accuracy does *not* extend to τ values in the far left
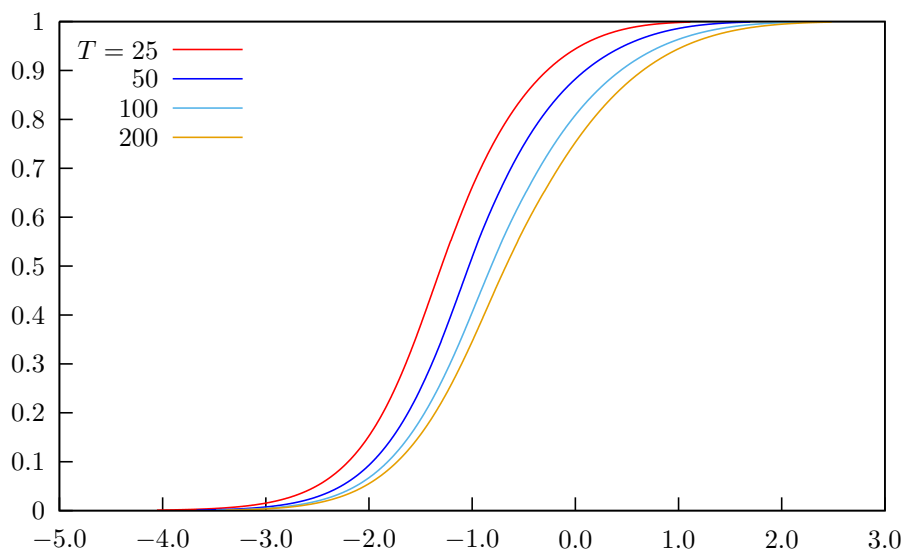
Figure 3: CDFs for τ, DF-GLS test with constant

tail. If you run a large sample of a perfectly stationary series through one of the DF-GLS tests you may get a test statistic on the order of -10 or -15. All we can really say about the probability of getting such a statistic under the null is that it's exceedingly small; the functions shown in the Appendix will print a numerical value, but in that range it may well be out by several orders of magnitude.

References

Doornik, J. A. (2005) 'An improved ziggurat method to generate normal random samples'. Discussion Paper, University of Oxford. URL `http://www.doornik.com/research/ziggurat.pdf`.

Elliott, G., T. J. Rothenberg and J. H. Stock (1996) 'Efficient tests for an autoregressive unit root', *Econometrica* 64: 813–836.

MacKinnon, J. G. (1996) 'Numerical distribution functions for unit root and cointegration tests', *Journal of Applied Econometrics* 11: 601–618.

Marsaglia, G. and W. W. Tsang (2000) 'The ziggurat method for generating random variables', *Journal of Statistical Software* 5: 3–30.

Matsumoto, M. and T. Nishimura (2000) 'Dynamic creation of pseudorandom number generators'. In *Monte-Carlo and Quasi-Monte Carlo Methods 1998*. Springer.

Ng, S. and P. Perron (2001) 'Lag length selection and the construction of unit root tests with good size and power', *Econometrica* 69(6): 1519–1554.

Perron, P. and Z. Qu (2007) 'A simple modification to improve the finite sample properties of Ng and Perron's unit root tests', *Economics Letters* 94(1): 12–19.

Table 7: Interpolated p-values: "out of sample" values

α	T	with trend			constant only		
		$\bar{\tau}$	$\hat{\sigma}_\tau$	$\hat{P}_T(\bar{\tau})$	$\bar{\tau}$	$\hat{\sigma}_\tau$	$\hat{P}_T(\bar{\tau})$
0.007	40	-4.0396	0.0044	0.0070	-3.0955	0.0044	0.0070
	60	-3.8620	0.0042	0.0070	-2.9724	0.0041	0.0070
	90	-3.7473	0.0043	0.0070	-2.8872	0.0041	0.0070
	120	-3.6895	0.0036	0.0070	-2.8410	0.0040	0.0070
0.034	40	-3.4274	0.0026	0.0341	-2.4947	0.0022	0.0340
	60	-3.2869	0.0022	0.0340	-2.3856	0.0024	0.0340
	90	-3.1921	0.0019	0.0339	-2.3055	0.0021	0.0340
	120	-3.1439	0.0019	0.0339	-2.2606	0.0021	0.0340
0.048	40	-3.2830	0.0020	0.0482	-2.3525	0.0018	0.0479
	60	-3.1492	0.0018	0.0481	-2.2444	0.0019	0.0480
	90	-3.0577	0.0018	0.0479	-2.1639	0.0019	0.0479
	120	-3.0111	0.0018	0.0479	-2.1189	0.0017	0.0480
0.083	40	-3.0413	0.0016	0.0832	-2.1149	0.0014	0.0829
	60	-2.9172	0.0015	0.0831	-2.0062	0.0014	0.0830
	90	-2.8303	0.0016	0.0829	-1.9236	0.0014	0.0830
	120	-2.7855	0.0015	0.0828	-1.8773	0.0015	0.0830
0.102	40	-2.9455	0.0014	0.1023	-2.0208	0.0012	0.1019
	60	-2.8246	0.0013	0.1021	-1.9113	0.0014	0.1020
	90	-2.7392	0.0014	0.1019	-1.8274	0.0014	0.1020
	120	-2.6949	0.0014	0.1018	-1.7800	0.0015	0.1020
0.351	40	-2.2720	0.0010	0.3513	-1.3666	0.0008	0.3513
	60	-2.1663	0.0009	0.3513	-1.2357	0.0009	0.3516
	90	-2.0849	0.0009	0.3508	-1.1308	0.0009	0.3512
	120	-2.0403	0.0009	0.3506	-1.0704	0.0010	0.3510

Table 8: Non-montonicity: Dickey–Fuller test using MacKinnon's urcdist code

	τ	$\hat{p}$	$\Delta\hat{p}$
nc	-2.77342	0.006482	$-6.78\mathrm{e}{-07}$
	2.40148	0.995524	$-4.45\mathrm{e}{-07}$
	2.63248	0.997545	$-3.14\mathrm{e}{-06}$
c	-4.06768	0.002452	$-3.52\mathrm{e}{-07}$
	-3.94858	0.003466	$-3.89\mathrm{e}{-07}$
	-3.49158	0.012265	$-8.91\mathrm{e}{-07}$
	0.72192	0.991517	$-9.57\mathrm{e}{-07}$
	1.16802	0.997546	$-2.84\mathrm{e}{-06}$
ct	0.21560	0.997547	$-1.09\mathrm{e}{-07}$

Note: evaluated from $p = 0.001$ to 0.999, with $T = 50$. Number of p-values: nc, 63777; c, 58320; ct, 54495.

Table 9: Non-montonicity: DF-GLS test using the author's code

	τ	$\hat{p}$	$\Delta\hat{p}$
c	-1.1177	0.44981	$-9.59\mathrm{e}{-05}$
	-0.7355	0.65226	$-2.74\mathrm{e}{-05}$

Note: evaluated from $p = 0.001$ to 0.999, with $T = 50$. Number of p-values: 53603.

Table 10: ADF-GLS p-values: mimicking US real GDP

α	T	$\bar{\tau}$	$\hat{\sigma}_\tau$	$\hat{P}_T(\bar{\tau})$	$\hat{P}_\infty(\bar{\tau})$
0.01	50	-3.8036	0.0067	0.0100	0.0026
	100	-3.6059	0.0057	0.0099	0.0053
	200	-3.5076	0.0057	0.0099	0.0071
0.025	50	-3.4616	0.0046	0.0306	0.0083
	100	-3.2874	0.0038	0.0300	0.0145
	200	-3.1969	0.0043	0.0299	0.0206
0.05	50	-3.1832	0.0033	0.0605	0.0217
	100	-3.0236	0.0032	0.0598	0.0412
	200	-2.9379	0.0032	0.0598	0.0507
0.1	50	-2.8798	0.0025	0.1109	0.0564
	100	-2.7317	0.0027	0.1096	0.0773
	200	-2.6492	0.0027	0.1095	0.0920

DGP: $y_0 = 7.60$; $y_t = y_{t-1} + 0.033 + 0.179\Delta y_{t-1} + u_t$

Table 11: ADF-GLS p-values: mimicking US inflation

α	T	$\bar{\tau}$	$\hat{\sigma}_\tau$	$\hat{P}_T(\bar{\tau})$	$\hat{P}_\infty(\bar{\tau})$
0.01	50	-2.7700	0.0072	0.0141	0.0055
	100	-2.6896	0.0069	0.0117	0.0071
	200	-2.6336	0.0069	0.0108	0.0083
0.025	50	-2.4347	0.0052	0.0336	0.0146
	100	-2.3572	0.0051	0.0285	0.0181
	200	-2.3011	0.0053	0.0266	0.0209
0.05	50	-2.1613	0.0035	0.0648	0.0299
	100	-2.0815	0.0041	0.0558	0.0364
	200	-2.0214	0.0032	0.0525	0.0420
0.1	50	-1.8633	0.0025	0.1243	0.0603
	100	-1.7757	0.0030	0.1092	0.0729
	200	-1.7086	0.0030	0.1038	0.0840

DGP: $y_0 = 5.62$; $y_t = y_{t-1} + \sum_{i=1}^{4} \rho_i \Delta y_{t-i} + u_t$

Appendix

The functions below are written in hansl, the scripting language of gretl, which is free, open-source software.[9] It should be straightforward to translate the code for, e.g., Matlab or R if anyone is so inclined.[10]

The first code block shows the main function, which is passed a τ value and sample size, T, along with two data matrices: beta ($n_\alpha \times 4$) holds the coefficients from estimation of eqn. (1) in the text and P ($1 \times n_\alpha$) holds the array of α values used.

```
# the main p-value function
function scalar dfgls_pval (scalar tau, int T,
                           const matrix *beta,
                           matrix P)
  scalar nalpha = rows(beta)
  scalar npow = cols(beta) - 1

  # compute all critical values for T and find
  # the index of the one closest to tau
  matrix C = beta * ((1/T) .^ seq(0, npow))'
  scalar imin = iminc(abs(C - tau))

  # get a suitable subset of points
  scalar npoints = 5
  scalar np2 = int(npoints/2)
  scalar i1 i2
  if imin > np2 && imin <= nalpha - np2
    i1 = imin - np2
  elif imin <= np2
    i1 = 1
  else
    i1 = nalpha - npoints + 1
  endif
  i2 = i1 + npoints - 1
  C = C[i1:i2]
  P = P[i1:i2]

  set svd on # hush collinearity warnings
  matrix V
  # run second-state regression via OLS
  matrix g = mols(qnorm(P)', C .^ seq(0, 3), null, &V)
  scalar tcube = g[4]/sqrt(V[4,4])

  if abs(tcube) < 2
    # drop the cubic term
    matrix g = mols(qnorm(P)', C .^ seq(0, 2))
    matrix tt = tau .^ seq(0, 2)
  else
    matrix tt = tau .^ seq(0, 3)
```

[9]See http://gretl.sourceforge.net/.

[10]Some of the idioms in the script are due to Jack Lucchetti, a master of the terse hansl statement. The expression "A .^ seq(0,3)" is a concise way of defining a matrix with as many rows as A, each of the form $\begin{pmatrix} 1 & a_i & a_i^2 & a_i^3 \end{pmatrix}$.

```
        endif

    return cnorm(tt*g)
end function
```

The next code block contains two "driver" functions, which just load the appropriate matrices (for the trend or constant-only case, as required) then call the main function. As stated in the text these data matrices are not very big and could easily be "inlined" rather then being read from file.

```
# the trend case
function scalar dfgls_ct_pvalue (scalar tau, int T)
  matrix beta = mread("beta-dfgls-ct.mat")
  matrix alpha = mread("alpha-ct.mat")
  return dfgls_pval(tau, T, &beta, alpha)
end function

# the constant only case
function scalar dfgls_c_pvalue (scalar tau, int T)
  matrix beta = mread("beta-dfgls-c.mat")
  matrix alpha = mread("alpha-c.mat")
  return dfgls_pval(tau, T, &beta, alpha)
end function
```