Abstract

A number of authors have sought to improve univariate forecasting of trending variables through the use of pre-tests for a unit root. The unconditional forecast distribution becomes then a mixture between a fully estimated model and one that imposes a unit root. This idea defines a very wide class of forecast model estimators — for any single unit root test choice over size means an infinite number of estimators per unit root test. There are also many different unit root tests that could be chosen. This paper examines more closely the qualities one would desire in a test for a unit root when it is to be used for constructing forecasts in this manner. We also examine the choice of the size of the test. We show that the deviation of the last observation in the sample from its trend component is an important determinant of the difference between the two possible models, and derive tests that have power properties that distinguish the models more closely along the lines required to produce good forecasts. This along with the choice of size yeilds a method that almost completely dominates OLS estimation of the root in a model that includes a time trend.

*Corresponding author.
1 Introduction

Most macroeconomic and financial series exhibit a great deal of trending, so past data provide rich data for prediction. However at the same time it is well understood that ordinary least squares (OLS) estimators for the unknown parameters are heavily biased despite being consistent. It is less well appreciated that when the model contains a unit root (or near unit root) that the OLS estimator for the autoregressive component does not have any optimality properties, even asymptotically. This is quite different from the stationary case, where the OLS estimator is asymptotically efficient. The implication is that it is theoretically reasonable to expect that better methods than OLS for estimation of forecasting models exist.

This has led to the employment of strategies other than simply estimating autoregressive models by OLS. Often researchers simply impose a unit root on the model, and estimate the remaining model parameters from the differenced model. Even when this causes over-differencing, it may well have better properties than using the biased OLS estimator. A second approach, which follows the Box and Jenkins (1981) type of method, is to pre-test for a unit root and condition the forecast model on the outcome of the pre-test.

Diebold and Kilian (2000) examined use of the Dickey and Fuller (1979) test for a unit root as a pre-test in a model where the general model contained a time trend. They found that pre-testing improved forecast mean square error (MSE) in a series of Monte Carlo experiments. They found that the effect was better the smaller the sample and the longer the forecast horizon. Small samples mean large biases in the OLS estimator, however pre-tests tend to fail to reject so the unit root model is imposed. When forecasting over a longer horizon larger powers are taken of the OLS estimator for the root, increasing the small sample bias, and making the tradeoff starker.

In their conclusion Diebold and Kilian (2000) conjecture that more powerful pre-tests should be even more useful. The intuition is obvious — better pre-tests should enable the more correct model to be used for forecasting more often. However Fomby (unpublished presentation) presented Monte Carlo evidence that this effect was small.

This paper seeks to understand the problem through considering properties of pre-tests that might be desirable in forecasting. It is not directly obvious that higher powered tests would provide forecasts that outperform those based on lower powered tests. Indeed, a number of issues arise. We show that it is not simply power of the test that matters,
but having power at points where the OLS model outperforms imposing a unit root as a forecasting method. One direction in which to differentiate forecast methods is the difference in model performance when the last data point (which we will call the end point of the sample) is typical or unusual. The point can be demonstrated in an example.

In Figure 1 we plot the departures of the log of US GDP from a fitted linear trend. By definition of the trend most of the observations on the log of income are close to the trend. However there are well known episodes where this is not true — the log of real GDP departs from the trend. These include periods such as the late 60’s, early 70’s where income was above trend, and also periods where income was below trend. The point from which one forecasts impacts the cost and hence decision as to whether or not one would want to impose a unit root on the model. In the early 1970’s, where income was above trend, a unit root assumption would mean employing forecasts that do not revert to the trend whereas for the estimated model the forecasts would to some extent revert back towards the trend. The differences between the forecasts obtained by each method, by virtue of being made when the current level of income is far from the trend, will be large.

We can see this effect in Figure 2. Here the point from which we forecast, which is a few quarters before the first oil shock, is above trend. A forecasting model imposing a unit root will yield forecasts such that the forecast is the last value (i.e. log(GDP) in the fourth
Figure 2: Large Deviation - Forecast vs Actual. Forecasts from the unit root model given by heavy dashes, from estimated AR(1) model by short dashes. The actual log of GDP is given by the solid line.

quarter of 1974) plus the average drift over the sample times the number of periods being forecast. This means that the forecast displays no tendency to mean revert, and will remain above the trend line (since we are starting from a point that we know ex post is above the trend line). The forecast that does not assume a unit root but instead uses OLS estimates of the AR(1) model does have mean reverting behavior. This mean reverting behavior — small but nonetheless there — is the curving set of forecasts given in the short-dash line of Figure 2. The end result is that in this instance it is greatly preferable to have estimated the stochastic behavior here, the difference in the forecasts two years out is just over 2% in GDP.

Alternatively, if the point from which we are forecasting is close to the trend then either the forecast based on assuming a unit root and estimating the root are not likely to be so far apart and also the likely error is smaller. The reason, which will be shown more explicitly in the next section, is that even though the AR(1) coefficient estimate may be different from one in the forecast, in the construction of the forecast it is being multiplied by a small deviation when the last observation is close to trend. The result is that there is not much difference between the forecasts based on either the estimated model or the model with a unit root imposed. This can be seen in Figure 3. Figure 3 examines forecasting with the
two models when we can see ex post that the point from which we are forecasting is close to the trend. Both forecasts are relatively similar, differing after eight quarters out by less than one percent.

In the next section we provide some analytical and Monte Carlo results to examine the importance of the 'end point' or observation on which the forecast is based. We establish that the power of standard unit root tests is decreasing in the distance of the final observation from its trend component. We then in Section 3 examine the construction of a test for a unit root that directs power away from rejecting when the null is false and the end point is close to trend towards rejecting when the null is false and the end point is far from trend. The forecast performance of these tests is then examined and compared to previously suggested pre-test forecast models. A final section concludes. All proofs are contained in an appendix.

2 Unit root pre-tests and univariate forecasts

In this paper, we consider the univariate model

$$y_t = X_t \beta + w_t \quad t = 1, \ldots, T$$

$$w_t = \rho w_{t-1} + v_t \quad t = 2, \ldots, T$$
where $X_t$ is a vector of deterministic terms that includes a constant, $\beta$ are unknown parameters and \{\nu_t\} are mean zero disturbances. We will assume that when $\rho < 1$ that $w_1$ is drawn from the unconditional distribution for $w_t$, so that the data are stationary when the root is not equal to one. There is no need to make an assumption on $w_1$ when $\rho = 1$ as it is not separately identified from the deterministic component (see Mueller and Elliott (2003)).

In the context of this family of models we are interested in forecasting future values for $y_{T+h}$ given the data $\{y_t\}_{t=1}^T$ where lack of knowledge of the parameters means that we are uncertain as to the correct model to choose. Further, we are considering a forecasting situation that occurs often in economics where it is known that $\rho$ is in the region of unity. This 'knowledge' comes in the form of Granger (1966)'s typical spectral shape or Nelson and Plosser (1982)'s failure to reject unit roots. In both cases it is suggested that the root $\rho$ is close to one statistically. Since asymptotic power for hypothesis tests testing $H_0 : \rho = 1$ versus $H_a : \rho < 1$ are nontrivial only in a neighborhood of the null hypothesis of $\rho = 1$ of the form where $\rho = 1 - \gamma/T$, where $\gamma > 0$ is fixed, we will consider this neighborhood. This is the local-to-unity framework developed by Phillips (1987) and Chan and Wei (1987).

This problem has similarities and differences to the more usual forecasting problem where the data is sufficiently stationary so that large sample asymptotic approximations hold well for available sample sizes. In such stationary cases the contribution of parameter uncertainty is to add a term to the mean square error of the forecast that dissappears at rate $T$ (i.e. the square of the size of the error induced by estimating parameters that are consistent for their population values at rate $T^{1/2}$) and is in expectation equal to $k/T$ where $k$ is the number of parameters to be estimated. In these situations the effect of parameter uncertainty does not in large enough samples depend on the parameter values of the coefficients to be estimated. It is well known in these situations however that for $k$ large enough methods that trade off bias for variance reduction are available and so there is no 'optimal' procedure for estimating the unknown parameters.

The problem at hand is similar to the stationary problem in that the effect of parameter uncertainty on the mean square error of the forecast continues to dissapear at the same rate $T$ even though the variables may or may not be stationary. This result arises even though parameter estimates may be converging at a rate faster than $T^{1/2}$ — the reason is that the error component here involves not only the parameter estimate but also the covariate and in precisely those cases where parameter estimates converge faster as the sample size increases.
the covariate has a variance that grows with the sample size offsetting exactly the faster convergence rate of the estimate. To see this, we will make the following assumption on the error process \( v_t \).

**Condition 1** The stationary sequence \( \{v_t\} \) has a strictly positive spectral density function \( \nu_\nu(\lambda) \); it has a moving average representation \( v_t = \sum_{s=0}^{\infty} \delta_s \varepsilon_{t-s} \) where the \( \varepsilon_t \sim (0, 1) \) are independent random variables and \( \sum_{s=0}^{\infty} s|\delta_s| < \infty \).

Under Condition 1 partial sums of \( v_t \) satisfy a functional central limit theorem when suitably scaled, i.e.

\[
T^{-1/2}w[Ts] \Rightarrow \begin{cases} \omega W(s) & \text{for } \gamma = 0 \\ \omega e^{-\gamma s} (2\gamma)^{-1/2} + \omega \int_0^s e^{-\gamma(s-\lambda)} dW(\lambda) & \text{else} \end{cases} \equiv \omega M(s)
\]  

(2)

where \( \Rightarrow \) denotes weak convergence of the underlying probability measures, \( W(s) \) is a standard Brownian motion, \( \lfloor \cdot \rfloor \) indicates the greatest lesser integer function and \( \omega^2 = 2\pi \nu_\nu(0) \).

In the case of the model in (1) and \( v_t \) serially uncorrelated the \( h \) step ahead forecast can be written

\[
y_{T+h} - y_T = \sum_{i=1}^{h} \rho^{h-i} \varepsilon_{T+i} + (\rho^h - 1)(y_T - \beta'X_T) + \beta'(X_{T+h} - X_T)
\]

The term \( \sum_{i=1}^{h} \rho^{h-i} \varepsilon_{T+i} \) represents an unforecastable component common to any method of forecasting and uncorrelated with the remaining terms. The forecast is then

\[
y_{T+h|T} - y_T = (\hat{\rho}^h - 1)(y_T - \hat{\beta}'X_T) + \hat{\beta}'(X_{T+h} - X_T)
\]

where different estimation methods provide different estimators for \( \{\hat{\rho}, \hat{\beta}\} \) that in turn provide different forecast losses. For any estimation technique the resulting forecast error is

\[
y_{T+h} - y_{T+h|T} = \sum_{i=1}^{h} \rho^{h-i} \varepsilon_{T+i} - (\hat{\rho}^h - \rho^h)(y_T - \beta'X_T) - (\hat{\beta} - \beta)'(X_{T+h} - X_T) + (\rho^h - 1)(\hat{\beta} - \beta)'X_T.
\]

In this form the above points can be seen. First, in any forecast procedure the unforecastable term \( \sum_{i=1}^{h} \rho^{h-i} \varepsilon_{T+i} \) is \( O_p(1) \) for the forecast horizon \( h \) fixed. This will be a common first order term in the evaluation of forecast errors, and because it is common to all methods we will
ignore it. Each of the remaining terms is at best of order $O_p(T^{1/2})$, just as in the stationary case. Reasonable estimators for $\rho$ are such that $T(\hat{\rho}^h - \rho^h)$ is $O_p(1)$, but notice that so is $T^{-1/2}(y_T - \beta'X_T) = T^{-1/2}w_T$. In the case where $X_t = 1$, then the term $T(\hat{\rho}^h - 1)[T^{-1/2}(\hat{\beta} - \beta)]$ is also $O_p(1)$. For more complicated deterministic terms this remains true. Thus just as in the stationary case the mean square forecast error has a second order term that is of order $T$. Hence we can rewrite the above forecast error as

$$y_{T+h} - y_{T+h|T} = \sum_{i=1}^{h} \rho^{h-i} e_{T+i} + T^{-1/2}[-T(\hat{\rho}^h - \rho^h)T^{-1/2}(y_T - \beta'X_T)$$

$$-T^{1/2}(\hat{\beta} - \beta)'(X_{T+h} - X_T) + T(\hat{\rho}^h - 1)T^{-1/2}(\hat{\beta} - \beta)'X_T]$$

where the second term is the contribution of the estimation error. In terms of mean square error of the forecast the difference in performance of different methods will amount to a term dissapearing at rate $T$ equal to

$$E \left[ - (\hat{\rho}^h - \rho^h)(y_T - \beta'X_T) - (\hat{\beta} - \beta)'(X_{T+h} - X_T) + (\hat{\rho}^h - 1)(\hat{\beta} - \beta)'X_T \right]^2 \quad (3)$$

As this is the difference between methods we will focus on this term in all calculations and comparisons1.

There exist a number of approaches to forecasting when confronted with this model. One method is to simply impose a unit root on the forecast model, so $\hat{\rho}^h = 1$ and then estimate $\hat{\beta}$ from the differenced model (so in the case of $X_t = 1$ there are no parameters to estimate, in the model with a time trend the coefficient on the time trend is the mean of the differenced data). The expression in (3) can be evaluated analytically. For the time trend case we have that the term is in large samples approximately

$$h^2 \left[ \frac{1 + (1 + \gamma)^2 - 2(1 + \gamma)e^{-\gamma}}{2\gamma} \right]$$

This is increasing in $\gamma$ with a slope that rapidly becomes equal to $h^2/2$. This is obviously unbounded as $\gamma$ gets large, hence successful use of this approach to forecasting relies on the model being close to one with a unit root.

A second popular approach is to simply estimate the unknown parameters by OLS. In this case an analytic expression for the second order term is difficult to derive. The problem

---

1It is more common to look at the entire forecast MSE, however such a statistic clearly depends on the number of observations making such comparisons difficult.
Figure 4: Second order term as a function of $\gamma$ for $h = 1$. The upward sloping solid line is for imposing $\gamma = 0$, short-long dashed line is for OLS. Upper long dashes and short dashes are for DF and $Q_T(10, 1)$ pretests respectively.

is that the terms $T(\hat{\rho}^h - \rho^h)$ and $T^{-1/2}(y_T - \beta'X_T)$ etc. have nonstandard asymptotic distributions that are functions of $M(\lambda)$. In the expression these are multiplied together, forming quantities that are difficult to analyse analytically. However the term can be evaluated in a Monte Carlo simulation where the number of observations is chosen to be large enough to approximate the asymptotic results. The size of these second order terms for various values for $\gamma$ are given in Figures (4), (5) and (6) for forecast horizons $h$ equal to 1, 4 and 8 periods respectively in the model where $X_t = (1, t)$.

The form of the tradeoff is as expected — for $\rho$ near one imposing the parameter to one results in forecasts that outperform forecasts based on the estimated values but as $\rho$ departs from one there is a point at which estimation is preferred. The Figure does give empirical content to this tradeoff\(^2\), in the time trend model the switch occurs at around $\gamma = 10$. Such a value for $\gamma$, according to $\rho = 0.9$ when a hundred observations are available, is difficult to distinguish from a unit root in practice.

As is clear from the figure, neither method dominates the other. Further, our typical uncertainty over the size of $\gamma$ results in no clear understanding as to which method to use.

\(^2\)These tradeoffs have been analysed in Turner (2004)?.
Figure 5: As for Figure 4 for $h = 4$

Figure 6: As for Figure 4 for $h = 8$
This has led to approaches that attempt to obtain the best of both models, using pre-tests for a unit root that obtain the gains from imposing the unit coefficient when $\gamma$ is small and estimation when it is large. Consider a generic test $\phi(y)$ which tests the null hypothesis of a unit root versus mean reversion, i.e. $H_0 : \rho = 1$ vs $H_1 : \rho < 1$. A very large number of tests are available, see Stock (1994) for a review. The pre-test estimator for the model now takes the form of

$$\hat{\rho}_{PT} = \hat{\rho}_{OLS}.1(\phi..rejects) + (1 - 1(\phi..rejects)) = 1 + (\hat{\rho}_{OLS} - 1)1(\phi..rejects)$$

where $1(\phi..rejects)$ is the event that the pretest rejects the unit root null hypothesis. A similar result follows for the estimation of $\hat{\beta}_{PT}$, which becomes a mixture between the estimator under the assumption that $\rho = 1$ and the estimator when this parameter is unknown.

Different tests for a unit root, through different rejection properties, provide different estimators for the forecast model parameters. Diebold and Kilian (2000) find that in the model with a time trend included that pre-testing using the Dickey and Fuller (1979) test provides gains over using OLS estimators for the parameters. They speculate that higher power tests may give even greater gains. Higher power will mean choosing the estimated model over the unit root model for $\gamma > 0$, which leads to the speculation that greater power will lead to better forecast performance. However this is only part of the story, since the second order term (3) is a complicated function of correlated terms.

As with the model estimated with OLS, we can examine this in a Monte Carlo experiment. The results for the size of the second order term in the MSE are also given in Figures (4), (5) and (6). In addition to the forecast based on the Dickey and Fuller (1979) test for a unit root we include the higher power $Q_T(10, 1)$ test from Elliott (1999) and Mueller and Elliott (2003). This latter test is an asymptotically optimal test for the null hypothesis of a unit root in this model when the errors are normally distributed. The results show that pre-testing helps greatly over OLS estimation when $\gamma$ is small. The reason of course is that imposing a unit coefficient here provides much better forecasts than estimating this parameter, and as these alternatives are close to the null power is low for the tests and so they tend to result in imposing the nearly true unit root. Similarly, for distant alternatives where power is large the methods select the estimated model.

The 'catch' comes for intermediate values for $\gamma$, where power is nontrivial but not equal to one, and the pre-test method is noticeably inferior to the estimated model. Tests with
higher power do naturally revert to the estimated model for a smaller value for $\gamma$, however the differences in the size of the second order term are not great. This is in a real sense disappointing as the power function for this test is very close to the power envelope, hence this is close to the limit in terms of finding pre-tests that have greater power in distinguishing between $\rho = 1$ and $\rho < 1$.

However it is not necessarily the case that the test with the highest power provides the best pre-test estimator. Indeed, given that pre-test estimators $\hat{\rho}_{PT}$ are a subclass of shrinkage estimators, it is not even the case that any pre-test estimator should provide the best forecast model estimator. We know from the normal case with independent data that there exists no estimator that dominates other estimators in terms of MSE when there are sufficient regressors, so there is no reason to expect a dominant method exists for the problem at hand.

A guide to avenues of providing pre-test estimators that do improve forecast performance lies in examining the second order term (3). This term depends in a complicated way on all of the pieces and their correlations. But consider the first term, i.e. $(\hat{\rho}^h - \rho^h)(y_T - \beta'X_T)$. This term tells us, at least for the contribution of this term, that the cost of estimation error in the autoregressive coefficient $\rho$ is small when $y_T$ is close to its deterministic component $\beta'X_T$ and large when it is far from the deterministic component. This distinction accords with intuition. Suppose that, such as in the late 1970’s, GNP is below trend. A mean reverting model suggests that GNP growth will pick up, as it reverts to trend. The unit root model on the other hand treats the shocks that led to GNP falling below trend as permanent, and will forecast that the lower level is here to stay. On the other hand, when GNP is close to trend then a constant growth model and a model that stays near the trend are giving the same prediction.

To evaluate this possibility, we can examine in Monte Carlo simulation the differences in MSE using forecast models that impose $\rho = 1$ versus those that estimate this parameter. Define $\xi = (y_T - \beta'X_T)$, which is of course unobserved when $\beta$ are unknown. The functional central limit theorem gives us that asymptotically $T^{-1/2}(y_T - \beta'X_T) \Rightarrow \omega M(1)$, and so we expect $\xi = (y_T - \beta'X_T)$ to be approximately zero mean with standard deviation $(2\gamma)^{-1/2}\omega T^{1/2}$. We will evaluate the MSE for a one step ahead forecast conditional on various values for $\alpha$ where $\alpha$ measures the number of standard deviations of the approximate distribution for the end point, (i.e. we set $\xi = \alpha(2\gamma)^{-1/2}\omega T^{1/2}$). Figures (7) and Figure (8) show the forecast
Figure 7: $\gamma = 10$, solid line is for $\rho = 1$ imposed and dashed line for $\rho$ estimated by OLS.

MSE for various $\alpha$ for $\gamma = 10$ and 15 respectively.

The main point is that for both values for $\gamma$ the difference in the MSE’s is zero when $\alpha = 0$ (so at the trend it does not matter which model is employed) and gets larger the larger is the value for $\alpha$. For $\alpha = 2$, there is a large difference in the MSE’s between the models. It is for these values that we are really concerned with model selection.

When $\gamma = 10$, we see that imposing $\rho = 1$ yeilds a better MSE than estimation regardless of the value for $\alpha$, although the curves are not far apart. They are however further apart with the same ordering for $\gamma = 5$ (not reported). However the ordering switches by $\gamma = 15$, and here the MSE from imposing the root is much worse for $\alpha > 0$ if $\rho = 1$ is imposed on the forecasting model.

This suggests that rather than focussing on the power of a test per se, perhaps a better pre-test estimator will arise if we focus on tests with good power properties when $(y_T - \beta'X_T)$ is large, rather than good power properties overall. We can by Monte Carlo examine the power of tests for a unit root conditional on this quantity. Figure (9) shows the power of both the Dickey and Fuller (1979) test and the $Q(10, 1)$ test where the data is drawn so that the end point is drawn conditional on various values of $\alpha$. Power is against the alternative that $\gamma = 10$. Clearly power drops precipitously as $\alpha$ gets large for both of the tests. The effect is less (the slope is flatter) for the Dickey and Fuller (1979) statistic, which may explain in part why forecasts based on the more powerful $Q(10, 1)$ test do not outperform this test as
we might have expected.

It is possible to develop tests for a unit root that have different power properties when \( \alpha \) is large. We develop such tests in the next section.

## 3 Unit Root Tests and the Final Observation

A practical issue in choosing between estimation and unit root imposition when the last observed value \( y_T \) is far from its trend \( \beta'X_T \) is that as the parameters \( \beta \) are unknown then so is the trend and hence so is the deviation \( y_T - \beta'X_T \). The reconsideration of the tradeoff is complicated because this is an unknown quantity — this conditioning does not amount to conditioning on \( y_T \). This is not to say that the data are not informative of this deviation, one can certainly estimate the trend and subtract it off the final observation. However this does not provide a consistent estimate, even asymptotically, since \( y_T - \hat{\beta}'X_T = \xi + (\hat{\beta} - \beta)'X_T \) and both terms are of the same order for all reasonable known (and feasible) estimators of the trend component.

One approach would be to construct estimators for \( \hat{\rho} \) that incorporate information in the data on the deviation \( \xi = y_T - \beta'X_T \). We have seen above that standard unit root pretest estimators have very low power for \( \xi \) far from zero when there is reversion to trend, resulting in a failure to reject and the imposition of a unit root on the forecasting model at a time when forecasts with mean reversion provide very different and much better forecasts. Given the difficulty in examining analytically the effect of different estimators on forecast mean square error, we take the approach of deriving a unit root pre-test that avoids the property of low power when \( \xi \) is large.

### 3.1 Small Sample Optimal Statistic

Consider the model in (1) but for the purposes of test construction assume that the distribution for \( v_t \) is known. In matrix form we can write the stochastic part of the model as \( \tilde{A}(\rho)\tilde{w} = \tilde{v} \) where \( \tilde{w} = (w_1, \ldots, w_T)' \), \( \tilde{v} = [v_1, \ldots, v_T]' \) and

\[
\tilde{A}(\rho) = \begin{pmatrix}
1 & 0 & 0 & 0 \\
-\rho & 1 & 0 & 0 \\
0 & \ddots & \ddots & 0 \\
0 & 0 & -\rho & 1
\end{pmatrix}
\]
Figure 8: $\gamma = 15$, solid line is for $\rho = 1$ imposed and dashed line for $\rho$ estimated by OLS.

Figure 9: Asymptotic Power of the DF (solid line) nad $Q_T(10,1)$ (dashed line) tests for a unit root against $\gamma = 10$. 
We will be interested in the distribution of quantities given $\xi = w_T = y_T - X_T' \beta$. First, notice that with the additional assumption of $\tilde{v} \sim N(0, \tilde{V}(\rho))$ we have

$$\tilde{w} \sim N\left(0, \tilde{A}(\rho)^{-1} \tilde{V}(\rho) \tilde{A}(\rho)^{-1}\right) = N(0, \tilde{W}(\rho))$$

Partition each of the above $T \times 1$ vectors into a $(T - 1) \times 1$ component and a $1 \times 1$ component (e.g. $\tilde{y} = (y', y_T')$). Partitioning the matrices after the $(T - 1)^{th}$ row and column we have that

$$\tilde{A}(\rho)^{-1} = \begin{pmatrix} A(\rho)^{-1} & 0 \\ v(\rho)' & 1 \end{pmatrix}$$

where $v(\rho)' = (\rho^{T-1}, \rho^{T-2}, ..., \rho)$. Consider the matrix

$$\tilde{V}(\rho) = \begin{pmatrix} V(\rho) & 0 \\ 0 & 1 \end{pmatrix}$$

so the variance covariance matrix is

$$\tilde{W}(\rho) = \begin{pmatrix} A(\rho)^{-1} V(\rho) A(\rho)^{-1} & A(\rho)^{-1} V(\rho) v(\rho) \\ v(\rho)' V(\rho) A(\rho)^{-1} & 1 + v(\rho)' V(\rho) v(\rho) \end{pmatrix}.$$ 

> From this we see that we have

$$w|\xi \sim N \left( \frac{A(\rho)^{-1} V(\rho) v(\rho)}{1 + v(\rho)' V(\rho) v(\rho)}, A(\rho)^{-1} V(\rho) A(\rho)^{-1} - \frac{A(\rho)^{-1} V(\rho) v(\rho) v(\rho)' V(\rho) A(\rho)^{-1}}{1 + v(\rho)' V(\rho) v(\rho)} \right)$$

$$= N(\beta_0(\xi), \Sigma(\rho))$$

The mean term in the conditional distribution depends on $\rho$ and adds what amounts to a deterministic arc to the data when $\rho < 1$. When $\rho = 1$ and we set $V(1) = I$ then since $v(1)$ is a $(T - 1) \times 1$ vector of ones we have $1 + v(1)' V(1) v(1) = T$ and so conditional mean is $(t/T) \xi$, i.e. it is a time trend joining the first observation and the last.

Given this conditional distribution, we can then follow Elliott, Rothenberg and Stock (1996), Mueller and Elliott (2003) and construct test statistics using the Neyman Pearson lemma. Typically we would like to obtain a test that is invariant with respect to the nuisance parameters $\{\beta, \xi\}$. Unfortunately, for the same reasons as detailed in Elliott and Mueller (2005) for the nuisance parameter introduced by more general assumptions on the initial condition, we cannot obtain invariance with respect to $\xi$. Instead we will examine tests that are invariant to the translation $\{y_t\}_{t=1}^T \rightarrow \{y_t + X_t'b\}_{t=1}^T \forall b, m$. Tests will hence be based on the maximal invariant statistic $\tilde{M}\tilde{y} = (I_T - \tilde{X}(\tilde{X}' \tilde{X})^{-1}\tilde{X}' \tilde{y})$ where $\tilde{X} = (X_1', X_2', ..., X_T')'$. 

16
The effect of invoking invariance is that in the case where \( X \) includes a time trend the conditional mean of the maximal invariant is zero under the null hypothesis, which is to say that \( \tilde{M} \tilde{R}(1) = 0 \) where \( \tilde{R}(\rho) = (R(\rho)', 1)' \). This removes the effect of \( \xi \) from the likelihood under the null hypothesis, simplifying the testing procedure. In the constant only case this is not true, and the likelihood for the maximal invariant depends on the nuisance parameter \( \xi \) under both the null and alternative hypotheses. This complicates the testing problem. For the remaining development we concentrate on the time trend case only.

The maximal sufficient statistic has the distribution

\[
\tilde{M} \tilde{y} \mid \xi \sim N(\tilde{M} \tilde{R}(\rho) \xi, \tilde{M} \tilde{\Sigma}(\rho) \tilde{M})
\]

where

\[
\tilde{\Sigma}(\rho) = \begin{pmatrix} \Sigma(\rho) & 0 \\ 0 & 0 \end{pmatrix}.
\]

since \( w_T = \xi \). Hence \( \tilde{M} \tilde{\Sigma}(\rho) \tilde{M} \) is singular.

The point optimal test for \( H_0 : \rho = 1 \) versus \( H_a : |\rho| < 1 \) conditional on a value \( \xi \) is given by the ratio

\[
LR_T = \exp \left\{ -\frac{1}{2} \left( \tilde{y} - \tilde{R}(\rho) \xi \right)' \tilde{M} \tilde{\Sigma}(\rho) \tilde{M} (\tilde{y} - \tilde{R}(\rho) \xi) - \tilde{M} (\tilde{y} - \tilde{R}(1) \xi) \right\}
\]

where \( \tilde{G}(\rho) = \tilde{\Sigma}(\rho)^{-1} - \tilde{\Sigma}(\rho)^{-1} \tilde{X} (\tilde{\Sigma}(\rho)^{-1} \tilde{X})^{-1} \tilde{X}' \tilde{\Sigma}(\rho)^{-} \) and we define

\[
\tilde{\Sigma}(\rho)^{-} = \begin{pmatrix} \Sigma(\rho)^{-1} & -\Sigma(\rho)^{-1} e' \\ -e' \Sigma(\rho)^{-1} & 1 + e' \Sigma(\rho)^{-1} e \end{pmatrix}.
\]

This follows Mueller and Elliott (2003), which also establishes for the constant and time trend cases that \( \tilde{M} \tilde{G}(\rho) \tilde{M} = \tilde{G}(\rho) \), a result that holds here.

The likelihood ratio statistic is (in the detrended case)

\[
LR_T = \exp \left\{ -\frac{1}{2\sigma^2} \left[ (\tilde{y} - \tilde{R}(\rho) \xi)' \tilde{G}(\rho)(\tilde{y} - \tilde{R}(\rho) \xi) - \tilde{y}' \tilde{G}(1) \tilde{y} \right] \right\}
\]

\[
= \exp \left\{ -\frac{1}{2\sigma^2} \tilde{y}'(\tilde{G}(\rho) - \tilde{G}(1)) \tilde{y} \right\} \exp \left\{ -\frac{1}{2\sigma^2} \left[ \xi^2 (\tilde{R}(\rho)' \tilde{G}(\rho) \tilde{R}(\rho))^2 - 2\xi \tilde{R}(\rho)' \tilde{G}(\rho) \tilde{y} \right] \right\}
\]
where the simplification came from using $\tilde{M}\tilde{R}(1) = 0$.

This test statistic depends on the unobserved quantity $\xi$. We could simply posit a value for this deviation, which would result in the test having maximal power against the alternative when the nuisance parameter takes this value. An alternative approach is to maximise power against a weighted average of these values, which is a more typical approach in econometrics. We maximise the weighted average power criterion

$$\int P(\varphi(y) \text{ rejects} | \rho = r, \xi = x) dF(x)$$

over all tests $\varphi(y)$ that satisfy the level constraint. Such a test has the largest average power under the alternative $\rho = r < 1$, where the averaging over possible initial conditions $\xi$ is according to $F$. Specifically, they employ a Gaussian weighting for different values of the initial condition $\xi$, so that $F$ in (4) is the cumulative distribution function of a $N(0, \lambda)$ random variable. The variance of this Gaussian weighting function determines the relative weight of small and large initial conditions: the larger the variance, the larger the relative weight of large $|\xi|$.

If we integrate over this statistic with respect to the weighting over $\xi$ we obtain the statistic

$$Q_f(r, \lambda) = \sigma^{-2} \tilde{y}'(\tilde{G}(r) - \tilde{G}(1))\tilde{y} - \frac{\sigma^{-2} \lambda(\tilde{R}(r)'\tilde{G}(r)\tilde{y})^2}{1 + \lambda \tilde{R}(r)'\tilde{G}(r)\tilde{R}(r)}.$$  

The statistic depends on both $\rho$ and $\lambda$ — i.e. the statistic will be different depending on the alternative point $r$ chosen as well as the weighting over the different end values $\lambda$. This means that there are a wide variety of possible tests with different properties. By choosing them carefully the researcher can direct power to the desired alternatives. A drawback to using this test directly is that it is derived under assumptions of no serial correlation, and will differ when there is serial correlation. In order to be able to apply the tests more generally we will construct in the next section a test that is asymptotically equivalent to this test when there is no serial correlation and also controls size for a wider variety of models of $v_t$.

### 3.2 Large Sample Approximate Statistic

We will derive a modified statistic based on the asymptotic distribution of $Q_f(r, \lambda)$. Given our interest in finding feasible, asymptotically equivalent tests, it is natural to set $r = 1 - g/T$ for some fixed $g$, in accordance with the asymptotic thought experiment concerning $\rho$ under local-to-unity asymptotics. For the purposes of testing or forecasting, it is only for $\rho$ such
that $\gamma = T(1 - \rho)$ is a moderate number between zero and around 50 that are difficult to distinguish from one. When $\gamma$ is larger than 50, $\rho$ is so far from unity that almost all tests will reject will probability very close to one. Under such asymptotics, it is reasonable to measure the deviation of $y_T$ from its deterministic component $|\xi|$ in multiples of the square root of the unconditional variance of a stationary process for $\gamma > 0$. We hence denote $a^2 = 2gT^{-1}\lambda$ so choosing $a$ reflects a choice of $\xi$ in terms of standard deviations when $\gamma = g$.

**Theorem 1** For the model in (1) with Condition 1 holding then

$$\frac{(\sigma/\omega)^2}{\gamma}Q_f(r, \lambda) \Rightarrow g^2 \int M^d(s)^2 ds - g + q_0 M^d(0)^2 + q_1 M^d(0) \int M^d(s) ds + q_2 M^d(0) \int s M^d(s) ds + q_3 \left( \int M^d(s) ds \right)^2 + q_4 \int M^d(s) ds \int s M^d(s) ds + q_5 \left( \int s M^d(s) ds \right)^2$$

where $M^d(s) = M(s) - M(1) + 6(s-1) \int M(r) dr - 12(s-1) \int r M(r) dr$ and $q_0 = (-12 - 12g + 4g^2 - 2a^2g^3 + 0.5a^2g^4)/(12 + 12g + 4g^2 + 0.5a^2g^3)$, $q_1 = 8g^2(-3 - 3g + 0.75a^2g - 0.5a^2g)/(12 + 12g + 4g^2 + 0.5a^2g^3)$, $q_2 = 12g^2(2 + 2g + 0.5a^2g^3)/(12 + 12g + 4g^2 + 0.5a^2g^3)$, $q_3 = 4g^3(-3 - 1.5a^2 + 1.5a^2g - 0.5a^2g^2)/(12 + 12g + 4g^2 + 0.5a^2g^3)$, $q_4 = 12g^4(2 - a^2 + 0.5a^2g)/(12 + 12g + 4g^2 + 0.5a^2g^3)$, and $q_5 = 12g^4(-1 - 0.5a^2g)/(12 + 12g + 4g^2 + 0.5a^2g^3)$.

Approximate percentiles of this distribution are given in Table 1. These values were constructed by replacing functions of $M^d(s)$ with sample analogs constructed from independent standard normal random variables, where 1500 steps were used in each approximation. The distribution was constructed from 5000 replications of the random variables.

We will employ an asymptotic approximation to the test (5) based on its asymptotic representation in Proposition 1. The test can be computed through the following steps.

1. Regress $y_t$ on $X_t$ using the full sample to obtain detrended data $y_t^d$

2. Estimate $\hat{\omega}^2$ using a consistent estimator of the spectral density of $(1 - \hat{\rho}L)y_t^d$ at frequency zero where $\hat{\rho}$ is the least squares estimate from regressing $y_t^d$ on its first lag.

3. Subtract the last observation from each datapoint, i.e. construct a new series $\hat{w}_t = y_t^d - y_T^d$. The last observation is dropped as it is zero.
4. Construct the statistic according to the formula

\[
\hat{Q}_f(g, \lambda) \Rightarrow g^2 \hat{\omega}^{2-2} T^{-2} \sum_{t=2}^{T} \hat{w}_{t-1}^2 - g + q_0 T^{-1} \hat{w}_1^2 + q_1 T^{-2} \hat{w}_1 \sum_{t=2}^{T} \hat{w}_{t-1} + q_2 T^{-3} \hat{w}_1 \sum_{t=2}^{T} t \hat{w}_{t-1}
\]

\[+ q_3 T^{-3} \left( \sum_{t=2}^{T} \hat{w}_{t-1} \right)^2 + q_4 T^{-4} \sum_{t=2}^{T} \hat{w}_{t-1} \sum_{t=2}^{T} t \hat{w}_{t-1} + q_5 T^{-5} \left( \sum_{t=2}^{T} t \hat{w}_{t-1} \right) \]

where the \( q_i \) weights are as in Theorem 1.

5. Reject for values below critical value from Table 1.

When \( v_t \) satisify Condition 1 the test has the asymptotic distribution given in Theorem 1.

For various choices of \( a \) we can evaluate the power profiles of the tests against various values for \( \alpha \). This is undertaken in Figure (10). This Figure shows the asymptotic power against the alternative that \( \gamma = 10 \). The solid line represents the test constructed with \( a = 1.5 \). The remaining downward sloping lines are for \( a = 0.5, 1 \) which are represented with short and long dashes respectively. The upward sloping lines are for tests based on \( a = 2 \) and 2.5 represented by long and short dashes respectively. The results show that the choice of \( a \) results in a tradeoff in power — larger values result in higher power for large \( \alpha \) but at the expense of lower power when \( \alpha \) is small. A full complement of tradeoffs is possible, very low power when \( \alpha \) is small or very low power when \( \alpha \) is large. The choices of \( a = 1.5 \) and \( \alpha = 2 \) give power curves that are relatively flat across values of that one might consider reasonable. The former has power falling over this range, the latter rising so could be seen to be 'different' choices in the types of power profiles without taking extreme stands on the final value.

We can also evaluate the effect of the use of these tests as pretests for forecasting the AR(1) model. For the same experiment as in Figures (4) and (5) we examine the second order term (3) for various \( \gamma \). This is undertaken in Figures (11) and (12) for \( h = 1 \) and \( h = 4 \) respectively where pretests based on \( a = 1.5 \) and \( a = 2 \) are employed for forecasting and compared against the previously examined pretest estimators. The higher power when \( \alpha > 0 \) leads to a different tradeoff than the usually applied unit root pretests. First, for \( a = 2 \), the tests do not reject as often. This is a property of redirecting power whilst maintaining size — the cost is lower power when \( \alpha \) is small. This impacts the performance of the forecasts. In both figures we see that the first order term in the MSE calculation is slightly larger.
Figure 10: Power for various $\alpha$ when $\gamma = 10$. Curves are for $a = 0.5, 1, 1.5, 2, 2.5$ where the curves at $\alpha = 0$ are in descending order.

when $\gamma$ is close to zero. The gain occurs for moderate $\gamma$. This is when the more typically applied pretests have estimation errors that are in excess of both imposing the unit root or estimating the root. Here the errors are still larger than those from using the OLS estimator always, however they are lower than those using DF or $Q_T(10, 1)$ pretests. Finally, for large $\gamma$ they achieve the same level as for estimating the parameters by OLS for values of $\gamma$ smaller than the usually applied pretests.

4 Choosing Size

Use of the test optimized for the size of the end point, whilst mitigating to some extent the 'hump' for moderate $\gamma$, does not completely remove this effect. So it would still be preferable to use the OLS estimator if the researcher was confident the root was not too close to the unit circle. Of course the tradeoff between imposing the unit root and using the OLS estimator for the forecast depends on size. The smaller the size chosen, the higher the likelihood that the test fails to reject and the greater the weight the imposed unit root receives in the forecast procedure.

Typically size is chosen to be relatively small, say 5%. This follows the usual approach
Figure 11: Second order term as a function of $\gamma$ when $h = 1$. The upward sloping solid line is for imposing $\gamma = 0$, short-long dashed line is for OLS. Upper long dashes and short dashes are for DF and $Q_T(10, 1)$ pretests. Lower short and long dashed lines are for $Q_f(10, 1.5)$ and $Q_f(10, 2)$ tests respectively.

Figure 12: As for Figure 11 with $h = 4$. 
in classical hypothesis testing, although such a choice has even less basis in constructing the forecast than it does in standard hypothesis tests. The effect of this choice is to ensure that when $\gamma$ is very close to zero that we use the imposed root very often. However consider the intermediate values for $\gamma$. For these values we have seen that the pre-test procedure does poorly compared to forecasts based on the OLS estimator. It would be preferable that the OLS estimator be chosen more often on average over these values, given that it does much better as a forecast estimator than imposing a unit root. By choosing size to be larger, this will occur. Such an improvement at these intermediate values for $\gamma$ must come at the expense of worse performance when $\gamma$ is closer to zero. As size goes to one, the method approaches only using the OLS estimator for forecast construction and hence approaches the tradeoff over various $\gamma$. This section will examine this tradeoff numerically.

Figures 13 and 14 repeat the experiments from the previous section for $h = 1$ and $h = 4$ for size equal to 50%. In order to be able to view the results clearly we have suppressed the results from imposing $\gamma = 0$ on the forecast method. Here the tradeoffs between the methods that do not have unbounded second order terms is clear. Using the OLS estimator is unsurprisingly worst when $\gamma$ is close to zero. For $h = 1$ and $\gamma$ near zero the effect of estimation is about $6 - 7$ standard deviations of the error term divided by the number of observations. Using the DF pre-test with 50% size achieves the same size errors but for $\gamma$ in the range of 5 – 10. Using the higher power $Q_T(10, 1)$ test we have that the peak loss is still at the same range for $\gamma$, but smaller. For the tests derived in section 3 the effects are smaller still. Indeed, for both choices of $a$ but particularly $a = 2$ the pre-test estimators based on $Q_f(10, a)$ almost dominate OLS everywhere. They do dominate the other pre-test estimators considered here.

For $h = 4$ similar results apply although the magnitudes of the effects differ.

5 Conclusion

In constructing a forecast univariate procedure for variables that may or may not have a unit root the forecaster faces the problem that OLS does not provide an optimal forecast model in any known sense, even when attempting to minimize mean square error of the forecast. This has led to a number of refinements of estimation. Here we examine unit root pre-test estimators and derive estimators that direct power towards realizations of the data where
Figure 13: Second order term as a function of $\gamma$ when $h = 1$. The short-long dashed line is for OLS. Upper long dashes and short dashes are for DF and $Q_T(10,1)$ pretests. Lower short and long dashed lines are for $Q_f(10,1.5)$ and $Q_f(10,2)$ tests respectively.

Figure 14: As for Figure 13 with $h = 4$. 

24
wrongly assuming a unit root is likely to have a greater cost.

Such realizations are situations where the observation from which the forecast is being made is far from its deterministic component. For example consider output, which moves slowly enough that recent observations are likely to better forecast the level in the near future than say using an estimated deterministic trend to capture the level. If output is close to the trend, forecasts based on the last observation or the trend would give similar results and the choice between the methods is not particularly interesting. However if output were either currently very low or high with respect to the trend, it might be useful in the forecast to use this information and construct forecasts that are more likely to be mean reverting towards its deterministic component.

In the presence of a time trend typical unit root tests have power profiles that go to zero when the final observation is far from its deterministic trend. This is to say that in precisely the times when the current level of output is far from trend, a unit root test is most likely to fail to reject a unit root even if the true model mean reverts. In turn then a forecasting model based on unit root pretests will then suggest that output is likely to stay away from trend. But this is precisely when such a forecast is likely to be particularly costly, as if there is mean reversion to the deterministic trend then it is in these situations where we have a greater amount of information as to the direction and likely size of the forecast.

This paper develops a family of tests for a unit root that allows different power profiles over the deviation of the last observation from trend. Through choice of a parameter of this test it is possible to choose different levels of robustness to these deviations. Employed in a pre-test gains are available over previously suggested pre-test methods.

6 Appendix

Lemma 1 Define the $T \times 1$ vectors $\mathbf{\hat{w}} = (\hat{w}_1, ..., \hat{w}_{T-1}), \tau = (1, 2, ..., T-1)'$ and $e = (1, 1, ..., 1)'$ where $\hat{w}_t$ is as defined in Step 2 of the test construction. Then for the model in (1) and Condition 1 holding

(i) \[ T(R(\rho) - e)'\Sigma(p)^{-1}(R(\rho) - e) \to g^2 \]

(ii) \[ T^{-1}(\tau - Te)'\Sigma(p)^{-1}(\tau - Te) \to 1 + g + g^2/3 \]

(iii) \[ (R(\rho) - e)'\Sigma(p)^{-1}(\tau - Te) \to g + g^2/2 \]
\[ (iv) \quad \sigma^{-1} T^{1/2} (\hat{w} - \hat{w}_T e)' \Sigma(\rho)^{-1} (R(\rho) - e) \Rightarrow - \left( g^2 \int M^d(r) dr + gM^d(0) \right) \]

\[ (v) \quad \sigma^{-1} T^{-1/2} (\hat{w} - \hat{w}_T e)' \Sigma(\rho)^{-1} (\tau - T e) \Rightarrow g^2 \left( \int rM^d(r) dr - \int M^d(r) dr \right) - (1 + g) M^d(0) \]

**Proof.** The variance covariance term \( \Sigma(r) \) can be simplified in a useful way — we have that

\[
\Sigma(r) = A(r)^{-1} V(r) A(r)^{-1} - \frac{A(r)^{-1} V(r) v(r) v(r)' V(r) A(r)^{-1}}{1 + v(r)' V(r) v(r)}
\]

\[
= A(r)^{-1} V(r)^{1/2} \left( I - \frac{V(r)^{1/2} v(r) v(r)' V(r)^{1/2}}{1 + v(r)' V(r) v(r)} \right) V(r)^{1/2} A(r)^{-1}
\]

\[
= A(r)^{-1} V(r)^{1/2} \left( I - \frac{v(r) v(r)'}{1 + v(r)' v(r)} \right) V(r)^{1/2} A(r)^{-1}
\]

where \( \bar{v}(r) = V(r)^{1/2} v(r) \). This is useful as we will require \( \Sigma(r)^{-1} \) where

\[
\Sigma(r)^{-1} = A(r)' V(r)^{-1/2} \left( I - \frac{\bar{v}(r) \bar{v}(r)'}{1 + \bar{v}(r)' \bar{v}(r)} \right)^{-1} V(r)^{-1/2} A(r)
\]

\[
= A(r)' V(r)^{-1/2} (I + \bar{v}(r) \bar{v}(r)') V(r)^{-1/2} A(r)
\]

\[
= A(r)' V(r)^{-1/2} (I + V(r)^{1/2} v(r) v(r)' V(r)^{1/2}) V(r)^{-1/2} A(r)
\]

\[
= A(r)' V(r)^{-1} A(r) + A(r)' v(r) v(r)' A(r)
\]

\[
= A(r)' V(r)^{-1} A(r) + r^2 e_{T-1} e_{T-1}'
\]

where \( e_j \) is defined as a \( T \times 1 \) vector with zeros everywhere except the \( j^{th} \) element.

Also note that we can write

\[
R(r)' \Sigma(r)^{-1} = (1 + v(r)' V(r) v(r))^{-1} v(r)' V(r)' A(r)^{-1} \Sigma^{-1}(r)
\]

\[
= (1 + v(r)' V(r) v(r))^{-1} (v(r)' V(r) A(r) + v(r)' V(r) v(r) v(r)' A(r))
\]

\[
= v(r)' V(r) A(r)
\]

so

\[
\sigma^{-1} T^{1/2} (\hat{w} - \hat{w}_T e)' \Sigma(\rho)^{-1} (R(\rho) - e) = \sigma^{-1} T^{1/2} v(r)' V(r) A(r)(\hat{w} - \hat{w}_T e)
\]

so

\[
R(r)' \Sigma(r)^{-1} e = v(r)' A(r) e = v(r)'((1 - r) v(r)' e + re_1) = r, \quad R(r)' \Sigma(r)^{-1} \tau = v(r)' A(r) \tau = (T - 1) r
\]

From direct calculation

\[
e' \Sigma(\rho)^{-1} e = e' [A(\rho)' V(\rho)^{-1} A(\rho) + A(\rho)' v(\rho) v(\rho)' A(\rho)] = e(T - 2)(1 - \rho)^2 + 1 \text{ when } \rho < 1.
\]

26
For the result in (i)

\[ T(R(r) - e)'\Sigma(p)^{-1}(R(r) - e) = T[R(r)'\Sigma(p)^{-1}R(r) + e'\Sigma(p)^{-1}e - 2R(r)'\Sigma(p)^{-1}e] \]
\[ = T[r^2 + (T - 2)(1 - r)^2 + 1 - 2r] \]
\[ = T(T - 1)(1 - r)^2 \]

hence \( T(R(\rho) - e)'\Sigma(p)^{-1}(R(\rho) - e) \to g^2 \).

For (ii)

\[ (R(\rho) - e)'\Sigma(p)^{-1}(\tau - Te) = R(\rho)'\Sigma(p)^{-1}\tau + Te'\Sigma(p)^{-1}e - TR(\rho)'\Sigma(p)^{-1}e - e'\Sigma(\rho)^{-1}\tau \]
\[ = (T - 1)\rho + T \left[ 1 + (T - 2)(1 - \rho)^2 - \rho \right] \]
\[ - T(1 - \rho) - (1 - \rho)^2 \sum_{t=2}^{T-2} t - \rho^2(T - 1) \]
\[ = \rho(T - 1)(1 - \rho) + (1 - \rho)^2 \sum_{t=2}^{T-2} t + T(T - 2)(1 - \rho)^2 \]
\[ \to g + g^2/2 \]

For (iii)

\[ T^{-1}(\tau - Te)'\Sigma(p)^{-1}(\tau - Te) = T^{-1} \left[ (e - Te_1) + T(1 - \rho)(T^{-1}\tau_{-1} - e_{-1}) \right] V(\rho)^{-1} \]
\[ \cdot \left[ (e - Te_1) + T(1 - \rho)(T^{-1}\tau_{-1} - e_{-1}) \right] + \rho^2/T \]
\[ = T^{-1}(T - 1)^2(1 - \rho^2) + (T - 1)/T + (T(1 - \rho))^2(T^{-3}\sum_{t=1}^{T-2} t) \]
\[ + (T - 1)/T + T^{-2}\sum_{t=1}^{T-2} t \]
\[ + 2T(1 - \rho)T^{-2}\sum_{t=1}^{T-2} t - 2T(1 - \rho)(T - 1)/T + \rho^2/T \]
\[ \to 1 + g + g^2/3 \]

For (iv) using the results above and rearranging yeilds

\[ (\hat{w} - \hat{w}_T e)'\Sigma(p)^{-1}(R(\rho) - e) = -\rho(1 - \rho)(\hat{w}_1 - \hat{w}_T) - (1 - \rho)^2 \sum_{t=1}^{T-1} (\hat{w}_t - \hat{w}_T) \]

and under condition 1 \( \omega^{-1/2}T^{-1/2}(\hat{w}_1 - \hat{w}_T) \Rightarrow M^d(0) \) and \( \omega^{-1/2}T^{-3/2} \sum_{t=1}^{T-1} (\hat{w}_t - \hat{w}_T) \Rightarrow \int M^d(s)ds. \)
For (v)
\[(\hat{w} - \hat{w}_T)e')\Sigma(\rho)^{-1}(\tau - Te) = (-\rho - T\rho(1 - \rho))(\hat{w}_1 - \hat{w}_T) + (1 - \rho)^2 \sum_{t=2}^{T} t(\hat{w}_{t-1} - \hat{w}_T)\]
\[-(T - 1)(1 - \rho)^2 \sum_{t=1}^{T-1} (\hat{w}_t - \hat{w}_T)\]
and under condition 1 \(\omega^{-1/2}T^{-5/2} \sum_{t=2}^{T} t(\hat{w}_t - \hat{w}_T) \rightarrow \int sM^d(s)ds\). □

**Proof.** Theorem 1.

Since \(a^2 = 2gT\lambda\) then
\[
\sigma^2 Q(\rho, \lambda) = \gamma'(\hat{G}(\rho) - \hat{G}(1))\gamma - \frac{a^2(T^{-2/2}\hat{R}(\rho)'\hat{G}(\rho)\gamma)^2}{2g + a^2T^{-1}\hat{R}(\rho)'\hat{G}(\rho)\hat{R}(\rho)}.
\]
Consider the pieces.

First
\[
\gamma'(\hat{G}(\rho) - \hat{G}(1))\gamma = (\hat{w} - \hat{w}_T)e' (\Sigma(\rho)^{-1} - \Sigma(1)^{-1})(\hat{w} - \hat{w}_T)e
\]
\[+ (\rho^2 - 1)(\hat{w}_{t-1} - \hat{w}_T)^2
\]
\[= \frac{[(\hat{w} - \hat{w}_T)e')\Sigma(\rho)^{-1}(\tau - Te)]^2}{(\tau - Te)'\Sigma(\rho)^{-1}(\tau - Te)}
\]
\[= (1 - \rho)^2 \sum_{t=2}^{T} (u_t^2 - u_T^2)^2 + (1 - \rho)\sum_{t=2}^{T} (\Delta(w_t - w_T))^2
\]
\[= (\rho^2 - 1 + (T^2 - T)^{-1}T^2) (w_1 - w_T)^2
\]
\[= \frac{[(\hat{w} - \hat{w}_T)e')\Sigma(\rho)^{-1}(\tau - Te)]^2}{(\tau - Te)'\Sigma(\rho)^{-1}(\tau - Te)}
\]
\[= (\rho^2 - 1)(\hat{w}_{t-1} - \hat{w}_T)^2
\]
We have
\[T^{-1}\hat{R}(\rho)'\hat{G}(\rho)\hat{R}(\rho) = T^{-1}\hat{R}(\rho)'\hat{G}(\rho)\hat{R}(\rho) - T^{-1}\hat{R}(\rho)'\hat{G}(\rho) - \hat{X}'\hat{G}(\rho) - \hat{X}'\hat{G}(\rho) - \hat{R}(\rho)
\]
where
\[\hat{R}(\rho)'\hat{G}(\rho) - \hat{R}(\rho) = [R(\rho)', 1] \left( \begin{array}{cc} \Sigma(\rho)^{-1} & -\Sigma(\rho)^{-1}e \\ -e'\Sigma(\rho)^{-1} & 1 + e'\Sigma(\rho)^{-1}e \end{array} \right) \left( \begin{array}{c} R(\rho) \\ 1 \end{array} \right)
\]
\[= \left( (R(\rho) - e)'\Sigma(\rho)^{-1} - (R(\rho) - e)'\Sigma(\rho)^{-1}e \right) \left( \begin{array}{c} R(\rho) \\ 1 \end{array} \right)
\]
\[= (R(\rho) - e)'\Sigma(\rho)^{-1}R(\rho) + 1 - (R(\rho) - e)'\Sigma(\rho)^{-1}e
\]
\[= 1 + (R(\rho) - e)'\Sigma(\rho)^{-1}(R(\rho) - e)
\]
and
\[
\tilde{R}(\rho)'\tilde{\Sigma}(\rho)\tilde{X} = [R(\rho)', 1] \begin{pmatrix}
\Sigma(p)^{-1} & -\Sigma(\rho)^{-1}e \\
-e'\Sigma(\rho)^{-1} & 1 + e'\Sigma(\rho)^{-1}e
\end{pmatrix} \begin{pmatrix}
e \\
1
\end{pmatrix}
= [R(\rho)', 1] \begin{pmatrix}
0 & \Sigma(p)^{-1}(\tau - Te) \\
1 & T - e'\Sigma(\rho)^{-1}(\tau - Te)
\end{pmatrix}
= \begin{pmatrix}
1 & (R(\rho)' - e)'\Sigma(\rho)^{-1}(\tau - Te) + T
\end{pmatrix}
\]

Also,
\[
\tilde{X}'\tilde{\Sigma}(\rho)\tilde{X} = \begin{pmatrix}
e' \\
\tau'
\end{pmatrix} \begin{pmatrix}
\Sigma(p)^{-1} & -\Sigma(\rho)^{-1}e \\
-e'\Sigma(\rho)^{-1} & 1 + e'\Sigma(\rho)^{-1}e
\end{pmatrix} \begin{pmatrix}
e \\
1
\end{pmatrix}
= \begin{pmatrix}
e' \\
\tau'
\end{pmatrix} \begin{pmatrix}
0 & \Sigma(p)^{-1}(\tau - Te) \\
1 & T - e'\Sigma(\rho)^{-1}(\tau - Te)
\end{pmatrix}
= \begin{pmatrix}
1 & T \\
T & (\tau - Te)'\Sigma(p)^{-1}(\tau - Te) + T^2
\end{pmatrix}
\]

Notice that the determinant of this is \((\tau - Te)'\Sigma(p)^{-1}(\tau - Te) + T^2 - T^2 = (\tau - Te)'\Sigma(p)^{-1}(\tau - Te)\). Hence
\[
(\tilde{X}'\tilde{\Sigma}(\rho)\tilde{X})^{-1} = \frac{1}{(\tau - Te)'\Sigma(p)^{-1}(\tau - Te)} \begin{pmatrix}
(\tau - Te)'\Sigma(p)^{-1}(\tau - Te) + T^2 & -T \\
-T & 1
\end{pmatrix}
\]

Putting these terms together we have
\[
TR(\rho)'\hat{G}(\rho)\tilde{R}(\rho) = TR(\rho)'\tilde{\Sigma}(\rho)\tilde{R}(\rho) - T\tilde{R}(\rho)'\tilde{\Sigma}(\rho)\tilde{X}(\tilde{X}'\tilde{\Sigma}(\rho)\tilde{X})^{-1}\tilde{X}'\tilde{\Sigma}(\rho)\tilde{R}(\rho)
= T + T(R(\rho) - e)'\Sigma(p)^{-1}(R(\rho) - e)
\]
\[
-\frac{[(\tau - Te)'\Sigma(p)^{-1}(\tau - Te) + [(R(\rho) - e)'\Sigma(p)^{-1}(\tau - Te)]^2}{(\tau - Te)'\Sigma(p)^{-1}(\tau - Te)}
\]
\[
= T(R(\rho) - e)'\Sigma(p)^{-1}(R(\rho) - e) - \frac{[(R(\rho) - e)'\Sigma(p)^{-1}(\tau - Te)]^2}{T^{-1}(\tau - Te)'\Sigma(p)^{-1}(\tau - Te)}
\]

Noticeing that we can replace \(\tilde{y}\) everywhere with its projection onto a constant and time trend, i.e. \((\tilde{w}', \tilde{w}_T)'\) then in the numerator of the final term we have
\[
\tilde{R}(\rho)'\tilde{G}(\rho)\tilde{y} = \tilde{R}(\rho)'\tilde{\Sigma}(\rho)\tilde{y} - \tilde{R}(\rho)'\tilde{\Sigma}(\rho)\tilde{X}(\tilde{X}'\tilde{\Sigma}(\rho)\tilde{X})^{-1}\tilde{X}'\tilde{\Sigma}(\rho)\tilde{y}
\]
where

$$\tilde{R}(\rho)'\tilde{\Sigma}(\rho)^{-}\tilde{y} = [R(\rho)', 1] \begin{pmatrix} \Sigma(p)^{-1} & -\Sigma(p)^{-1}e \\ -e'\Sigma(p)^{-1} & 1 + e'\Sigma(p)^{-1}e \end{pmatrix} \begin{pmatrix} \hat{w} \\ \hat{w}_T \end{pmatrix}$$

$$= (R(\rho) - e)'\Sigma(p)^{-1} 1 - (R(\rho) - e)'\Sigma(p)^{-1}e \begin{pmatrix} \hat{w} \\ \hat{w}_T \end{pmatrix}$$

$$= (R(\rho) - e)'\Sigma(p)^{-1}\hat{w} + \hat{w}_T - (R(\rho) - e)'\Sigma(p)^{-1}e\hat{w}_T$$

$$= \hat{w}_T + (R(\rho) - e)'\Sigma(p)^{-1}(\hat{w} - \hat{w}_Te)$$

and

$$\tilde{\Sigma}(\rho)^{-}\tilde{X} = [\hat{w}', \hat{w}_T] \begin{pmatrix} \Sigma(p)^{-1} & -\Sigma(p)^{-1}e \\ -e'\Sigma(p)^{-1} & 1 + e'\Sigma(p)^{-1}e \end{pmatrix} \begin{pmatrix} e \\ \tau \\ 1 \end{pmatrix}$$

$$= [\hat{w}', \hat{w}_T] \begin{pmatrix} 0 & \Sigma(p)^{-1}(\tau - Te) \\ 1 & T - e'\Sigma(p)^{-1}(\tau - Te) \end{pmatrix}$$

$$= \begin{pmatrix} \hat{w}_T (\hat{w} - \hat{w}_Te)'\Sigma(p)^{-1}(\tau - Te) + T\hat{w}_T \end{pmatrix}$$

The result is that

$$T^{-1/2}\tilde{R}(\rho)'\tilde{G}(\rho)\tilde{y} = T^{-1/2}\tilde{R}(\rho)'\tilde{\Sigma}(\rho)^{-}\tilde{y} - T^{-1/2}\tilde{R}(\rho)'\tilde{\Sigma}(\rho)^{-}\tilde{X}(\tilde{X}'\tilde{\Sigma}(\rho)^{-}\tilde{X})^{-1}\tilde{X}'\tilde{\Sigma}(\rho)^{-}\tilde{y}$$

$$= T^{-1/2}\hat{w}_T + T^{-1/2}(R(\rho) - e)'\Sigma(p)^{-1}(\hat{w} - e\hat{w}_T)$$

$$- T^{-1/2}[(R(\rho) - e)'\Sigma(p)^{-1}(\tau - Te)(\hat{w} - \hat{w}_Te)'\Sigma(p)^{-1}(\tau - Te)]$$

$$= T^{-1/2}(R(\rho) - e)'\Sigma(p)^{-1}(\hat{w} - e\hat{w}_T)$$

$$- T^{-1/2}[(R(\rho) - e)'\Sigma(p)^{-1}(\tau - Te)T^{-1/2}(\hat{w} - \hat{w}_Te)'\Sigma(p)^{-1}(\tau - Te)]$$

Putting these pieces together we have that

$$Q(g, a) = T^2(1 - \rho)^2T^{-2}\sum_{t=2}^{T}(w_t - w_T)^2 - T(1 - \rho)T^{-1}\sum_{t=2}^{T}(\Delta(w_t - w_T))^2$$

$$\quad + \rho T (1 - \rho) T^{-1}(w_1 - w_T)^2$$

$$- \frac{a^2(T^{-1/2}\tilde{R}(\rho)'\tilde{G}(\rho)\tilde{y})^2}{2g + a^2T^{-1}\tilde{R}(\rho)'\tilde{G}(\rho)\tilde{R}(\rho)} + o_p(1)$$

The results Lemma 1 and noting that

$$T^{-2}\sum_{t=2}^{T}(w_t - w_T)^2 \Rightarrow \omega^2 \int M^d(s)^2 ds, T^{-1}\sum_{t=2}^{T}(\Delta(w_t - w_T))^2 \Rightarrow \sigma^2$$

yield the asymptotic result where for the assumptions of the Theorem \(w = \sigma\). Rearranging and collecting terms gives the result in Theorem 1.
References


<table>
<thead>
<tr>
<th>test</th>
<th>critical value</th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{Q}_F(10, 0.5) )</td>
<td>-7.77</td>
<td>-6.90</td>
<td>-6.32</td>
<td>-1.78</td>
<td></td>
</tr>
<tr>
<td>( \hat{Q}_F(10, 1) )</td>
<td>-7.96</td>
<td>-7.19</td>
<td>-6.63</td>
<td>-2.98</td>
<td></td>
</tr>
<tr>
<td>( \hat{Q}_F(10, 1.5) )</td>
<td>-8.11</td>
<td>-7.51</td>
<td>-7.00</td>
<td>-3.88</td>
<td></td>
</tr>
<tr>
<td>( \hat{Q}_F(10, 2) )</td>
<td>-8.65</td>
<td>-7.91</td>
<td>-7.39</td>
<td>-4.41</td>
<td></td>
</tr>
<tr>
<td>( \hat{Q}_F(10, 2.5) )</td>
<td>-9.47</td>
<td>-8.24</td>
<td>-7.70</td>
<td>-4.73</td>
<td></td>
</tr>
</tbody>
</table>