

Tests for stationarity and stability in time-series data

Christopher F Baum

Boston College and DIW Berlin

January 2009



When working with time-series data, we must be concerned with two attributes: *stationarity* and *stability*. The former property applies to a single time series: is the series covariance stationary, or does its autoregressive representation contain one or more *unit roots*?



When working with time-series data, we must be concerned with two attributes: *stationarity* and *stability*. The former property applies to a single time series: is the series covariance stationary, or does its autoregressive representation contain one or more *unit roots*?

The latter property refers to a bivariate or multivariate relationship: is the relationship *temporally stable*? The latter issue may relate to the conditional mean of a series, or indeed to its variance or autocorrelation function. In this lecture, we consider several aspects of these two time-series properties, and present software tools that may be used in their evaluation.



Unit root tests

The “first generation” unit root tests, such as the Dickey–Fuller, Augmented Dickey–Fuller and Phillips–Perron tests have been shown to have relatively low power to reject their null hypothesis: that the series is non-stationary ($I(1)$) rather than stationary ($I(0)$). In particular, any sort of structural break in the series is likely to cause a failure to reject, even if the series is stationary before and after the structural break.



To deal with the well-known low power of these tests, researchers have devised more powerful tests such as the DF-GLS test of Elliott, Rothenberg, Stock (*Econometrica*, 1996). The standard Dickey–Fuller test is essentially an OLS regression: in the simplest form, of the difference of the series (ΔX_t) on the lagged level of the series (X_{t-1}). The “Augmented” Dickey-Fuller or ADF test adds a number of lagged differences to the specification. The DF-GLS test makes use of generalized least squares (GLS) rather than OLS, and has been shown to have considerably higher power in many circumstances.



The `dfgls` command is now part of official Stata. Its original implementation was provided by Baum (*STB-57*, 2000) and Baum and Sperling (*STB-58*, 2000). `dfgls` performs the Elliott–Rothenberg–Stock (ERS) efficient test for an autoregressive unit root. This test is similar to an (augmented) Dickey-Fuller t test, as performed by `dfuller`, but has the best overall performance in terms of small-sample size and power, dominating the ordinary Dickey-Fuller test. The `dfgls` test “has substantially improved power when an unknown mean or trend is present” (ERS, p. 813).



The `dfgls varname` command applies a generalized least squares (GLS) detrending (demeaning) step to the *varname*:

$$y_t^d = y_t - \hat{\beta}' z_t$$

For detrending, $z_t = (1, t)'$ and $\hat{\beta}_0, \hat{\beta}_1$ are calculated by regressing $[y_1, (1 - \bar{\alpha}L) y_2, \dots, (1 - \bar{\alpha}L) y_T]$ onto $[z_1, (1 - \bar{\alpha}L) z_2, \dots, (1 - \bar{\alpha}L) z_T]$ where $\bar{\alpha} = 1 + \bar{c}/T$ with $\bar{c} = -13.5$, and L is the lag operator. For demeaning, $z_t = (1)'$ and the same regression is run with $\bar{c} = -7.0$.



The `dfgls varname` command applies a generalized least squares (GLS) detrending (demeaning) step to the *varname*:

$$y_t^d = y_t - \hat{\beta}' z_t$$

For detrending, $z_t = (1, t)'$ and $\hat{\beta}_0, \hat{\beta}_1$ are calculated by regressing $[y_1, (1 - \bar{\alpha}L)y_2, \dots, (1 - \bar{\alpha}L)y_T]$ onto $[z_1, (1 - \bar{\alpha}L)z_2, \dots, (1 - \bar{\alpha}L)z_T]$ where $\bar{\alpha} = 1 + \bar{c}/T$ with $\bar{c} = -13.5$, and L is the lag operator. For demeaning, $z_t = (1)'$ and the same regression is run with $\bar{c} = -7.0$.

The values of \bar{c} are chosen so that “the test achieves the power envelope against stationary alternatives (is asymptotically MPI (*most powerful invariant*)) at 50 percent power” (Stock, *Handbook of Econometrics*, 1994, p. 2769).



The augmented Dickey-Fuller regression is then computed using the y_t^d series:

$$\Delta y_t^d = \alpha + \gamma t + \rho y_{t-1}^d + \sum_{i=1}^m \delta_i \Delta y_{t-i}^d + \epsilon_t$$

where $m = \text{maxlag}$. The `notrend` option suppresses the time trend in this regression.



The augmented Dickey-Fuller regression is then computed using the y_t^d series:

$$\Delta y_t^d = \alpha + \gamma t + \rho y_{t-1}^d + \sum_{i=1}^m \delta_i \Delta y_{t-i}^d + \epsilon_t$$

where $m = \text{maxlag}$. The `notrend` option suppresses the time trend in this regression.

Approximate critical values for the GLS detrended test are taken from ERS, Table 1 (p. 825). Approximate critical values for the GLS demeaned test are identical to those applicable to the no-constant, no-trend Dickey-Fuller test, and are computed using the `dfuller` code.



The `maxlag(p)` option specifies the maximum lag order to be considered. The test statistics will be calculated for each lag up to the maximum lag order (which may be zero). If not specified, the maximum lag order for the test is by default calculated from the sample size using a rule provided by Schwert (*JBES*, 1989) using $c=12$ and $d=4$ in his terminology. Whether the maximum lag is explicitly specified or computed by default, the sample size is held constant over lags at the maximum available sample.



The `maxlag(p)` option specifies the maximum lag order to be considered. The test statistics will be calculated for each lag up to the maximum lag order (which may be zero). If not specified, the maximum lag order for the test is by default calculated from the sample size using a rule provided by Schwert (*JBES*, 1989) using $c=12$ and $d=4$ in his terminology. Whether the maximum lag is explicitly specified or computed by default, the sample size is held constant over lags at the maximum available sample.

The `dfgls` routine includes a very powerful lag selection criterion, the “modified AIC” (MAIC) criterion proposed by Ng and Perron (*Econometrica*, 2000). They have established that use of this MAIC criterion may provide “huge size improvements” (2000, abstract) in the `dfgls` test. The criterion, indicating the appropriate lag order, is printed on `dfgls`’ output, and may be used to select the test statistic from which inference is to be drawn.



It should be noted that all of the lag length criteria employed by `dfgls` (the sequential t test of Ng and Perron, the Schwarz Criterion (SC), and the MAIC) are calculated, for various lags, by holding the sample size fixed at that defined for the longest lag. These criteria cannot be meaningfully compared over lag lengths if the underlying sample is altered to use all available observations. That said, if the optimal lag length (by whatever criterion) is found to be much less than that picked by the Schwert criterion, it would be advisable to rerun the test with the `maxlag` option specifying that optimal lag length, especially when using samples of modest size.



The KPSS test

As an alternative to the Dickey–Fuller style tests for stationarity, we may consider the *KPSS* test of Kwiatkowski, Phillips, Schmidt and Shin (*J. Econometrics*, 1992). This test (and those derived from it) have the more “natural” null hypothesis of stationarity ($I(0)$), where a rejection indicates non-stationarity ($I(1)$ or $I(d)$). The KPSS test may be used to confirm the findings of a DF-GLS test; their verdicts will not necessarily agree, but if they do, that is strong evidence in favor of (non-)stationarity.



The `kpss` command (`findit kpss` to install) performs the KPSS test for stationarity of a time series. The test may be conducted under the null of either trend stationarity (the default) or level stationarity. Inference from this test is complementary to that derived from those based on the Dickey–Fuller distribution (such as `dfgls`, `dfuller` and `pperron`). The KPSS test is often used in conjunction with those tests to investigate the possibility that a series is fractionally integrated (that is, neither $I(1)$ nor $I(0)$): see Lee and Schmidt (*J. Econometrics*, 1996).



The series is detrended (demeaned) by regressing y on $z_t = (1, t)'$ ($z_t = (1)'$), yielding residuals e_t . Let the partial sum series of e_t be s_t . Then the zero-order KPSS statistic is

$$k_0 = \frac{T^{-2} \sum_{t=1}^T s_t^2}{T^{-1} \sum_{t=1}^T e_t^2}$$

For $\text{maxlag} > 0$, the denominator is computed as the Newey-West estimate of the long run variance of the series; see [R] newey. Approximate critical values for the KPSS test are taken from KPSS (1992).



The `kpss` routine has been enhanced to add two options recommended by the work of Hobijn et al. (Econometric Institute WP, Rotterdam, 1998). An automatic bandwidth selection routine has been added, rendering it unnecessary to evaluate a range of test statistics for various lags. An option to weight the empirical autocovariance function by the Quadratic Spectral kernel, rather than the Bartlett kernel employed by KPSS, has also been introduced.



The `kpss` routine has been enhanced to add two options recommended by the work of Hobijn et al. (Econometric Institute WP, Rotterdam, 1998). An automatic bandwidth selection routine has been added, rendering it unnecessary to evaluate a range of test statistics for various lags. An option to weight the empirical autocovariance function by the Quadratic Spectral kernel, rather than the Bartlett kernel employed by KPSS, has also been introduced.

These options may be used separately or in conjunction. It is in conjunction that Hobijn et al. found the greatest improvement in the test: “Our Monte Carlo simulations show that the best small sample results of the test in case the process exhibits a high degree of persistence are obtained using both the automatic bandwidth selection procedure and the Quadratic Spectral kernel.” (1998, p.14)



Covariate-augmented unit root tests

Returning to the DF-GLS unit root test, we now consider an improved version of that test proposed by Elliott and Jansson (*J. Econometrics*, 2003) that adds stationary covariates to gain additional power.



Covariate-augmented unit root tests

Returning to the DF-GLS unit root test, we now consider an improved version of that test proposed by Elliott and Jansson (*J. Econometrics*, 2003) that adds stationary covariates to gain additional power.

As is well known in the applied economics literature, even a test with DF-GLS's favorable characteristics may still lack power to distinguish between the null hypothesis of nonstationary behavior ($I(1)$) and the stationary alternative ($I(0)$). In many applications, using a longer time series is not feasible due to known structural breaks, institutional changes, and the like.



Covariate-augmented unit root tests

Returning to the DF-GLS unit root test, we now consider an improved version of that test proposed by Elliott and Jansson (*J. Econometrics*, 2003) that adds stationary covariates to gain additional power.

As is well known in the applied economics literature, even a test with DF-GLS's favorable characteristics may still lack power to distinguish between the null hypothesis of nonstationary behavior ($I(1)$) and the stationary alternative ($I(0)$). In many applications, using a longer time series is not feasible due to known structural breaks, institutional changes, and the like.

Another potential alternative, the panel unit root test, brings its own set of complications (do we assume all series are $I(1)$? or that all are $I(0)$?).



Elliott and Jansson addressed this issue by considering a model in which there is one potentially nonstationary ($I(1)$) series, y , which potentially covaries with some available stationary variables, x . This idea was first put forth by Bruce Hansen in 1995 who proposed a covariate augmented D–F test, or CADF test and showed that this test had greater power than those which ignored the covariates.



Elliott and Jansson addressed this issue by considering a model in which there is one potentially nonstationary ($I(1)$) series, y , which potentially covaries with some available stationary variables, x . This idea was first put forth by Bruce Hansen in 1995 who proposed a covariate augmented D–F test, or CADF test and showed that this test had greater power than those which ignored the covariates.

The authors extended Hansen's results to show that such a test could be conducted in the presence of unknown nuisance parameters, and with constants and trends in the model. Their proposed test may be readily calculated by estimating a vector autoregression (var) in the $\{y, x\}$ variables and performing a sequence of matrix manipulations.



The model considered is:

$$z_t = \beta_0 + \beta_1 t + u_t, \quad t = 1, \dots, T$$
$$A(L) \begin{pmatrix} (1 - \rho L)u_{y,t} \\ u_{x,t} \end{pmatrix} = e_t$$

with $z_t = \{y_t, x_t'\}'$, x_t an $m \times 1$ vector,

$\beta_0 = \{\beta_{y0}, \beta'_{x0}\}'$, $\beta_1 = \{\beta_{y1}, \beta'_{x1}\}'$ and $u_t = \{u_{y,t}, u'_{x,t}\}'$.

$A(L)$ is a stable matrix polynomial of finite order k in the lag operator L .



The model considered is:

$$z_t = \beta_0 + \beta_1 t + u_t, \quad t = 1, \dots, T$$
$$A(L) \begin{pmatrix} (1 - \rho L)u_{y,t} \\ u_{x,t} \end{pmatrix} = e_t$$

with $z_t = \{y_t, x_t'\}'$, x_t an $m \times 1$ vector,

$\beta_0 = \{\beta_{y0}, \beta'_{x0}\}'$, $\beta_1 = \{\beta_{y1}, \beta'_{x1}\}'$ and $u_t = \{u_{y,t}, u'_{x,t}\}'$.

$A(L)$ is a stable matrix polynomial of finite order k in the lag operator L .

This is a vector autoregression (VAR) in the model of x and the quasi-difference of y . The relevant test is that the parameter ρ is equal to unity, implying that y has a unit root, against alternatives that ρ is less than one.



The potential gain in this test depends on the R^2 between y and the set of x covariates. As Elliott and Pesavento (*J. Money, Credit, Banking* 2006) point out, the relevant issue is the ability of a unit root test to have power to distinguish between $I(1)$ and a local alternative. The local alternative is in terms of $c = T(\rho - 1)$ where ρ is the largest root in the AR representation of y .



The potential gain in this test depends on the R^2 between y and the set of x covariates. As Elliott and Pesavento (*J. Money, Credit, Banking* 2006) point out, the relevant issue is the ability of a unit root test to have power to distinguish between $I(1)$ and a local alternative. The local alternative is in terms of $c = T(\rho - 1)$ where ρ is the largest root in the AR representation of y .

How far below unity must ρ fall to give a unit root test the ability to discern stationary, mean-reverting behavior (albeit with strong persistence, with $\rho > 0.9$) from nonstationary, unit root behavior? The various tests in this literature differ in their power against relevant local alternatives.



The c parameter can be expressed in terms of the half-life (k) of a shock, where a unit root implies an infinite half-life:
 $k = \log(0.5) / \log(\rho)$. When related to the local alternative, $k/T = \log(0.5) / c$. For about 120 observations (30 years of quarterly data), $c = -5$ corresponds to a half-life of 16.8 time periods (over four years at a quarterly frequency). From Elliott, Rothenberg, Stock (ERS, *Econometrica*, 1996), the standard Dickey-Fuller test (df_{fuller}) has 12% power to reject the alternative. The ERS DF-GLS test (df_{gls}) has 32% power.



The c parameter can be expressed in terms of the half-life (k) of a shock, where a unit root implies an infinite half-life: $k = \log(0.5)/\log(\rho)$. When related to the local alternative, $k/T = \log(0.5)/c$. For about 120 observations (30 years of quarterly data), $c = -5$ corresponds to a half-life of 16.8 time periods (over four years at a quarterly frequency). From Elliott, Rothenberg, Stock (ERS, *Econometrica*, 1996), the standard Dickey-Fuller test (df_{fuller}) has 12% power to reject the alternative. The ERS DF-GLS test (df_{gls}) has 32% power.

In contrast, with an $R^2 = 0.2$, the Elliott–Jansson (EJ) test has power of 42%. The power rises to 53% (69%) for $R^2 = 0.4$ (0.6). For higher absolute values of c (shorter half-lives), the gains are smaller. For $c = -10$, or a half-life of 8.4 periods, the power of the D-F (DF-GLS) test is 31% (75%). The EJ test has power of 88%, 94% and 99% for $R^2 = 0.2, 0.4, 0.6$. One clear conclusion: DF-GLS always has superior power compared to df_{fuller} .



Like other unit root tests, you must specify the deterministic model assumed for y . As in `dfgls`, you can specify that the model contains no deterministic terms, a constant, or constant and trend. But as the Elliott–Jansson model contains one or more x variables as well, you may also specify that those variables' deterministic model contains no deterministic terms, a constant, or constant and trend.

Five cases are defined:

1. No constant nor trend in model



Like other unit root tests, you must specify the deterministic model assumed for y . As in `dfgls`, you can specify that the model contains no deterministic terms, a constant, or constant and trend. But as the Elliott–Jansson model contains one or more x variables as well, you may also specify that those variables' deterministic model contains no deterministic terms, a constant, or constant and trend.

Five cases are defined:

1. No constant nor trend in model
2. Constant in y only



Like other unit root tests, you must specify the deterministic model assumed for y . As in `dfgls`, you can specify that the model contains no deterministic terms, a constant, or constant and trend. But as the Elliott–Jansson model contains one or more x variables as well, you may also specify that those variables' deterministic model contains no deterministic terms, a constant, or constant and trend.

Five cases are defined:

1. No constant nor trend in model
2. Constant in y only
3. Constants in both $\{y, x\}$



Like other unit root tests, you must specify the deterministic model assumed for y . As in `dfgls`, you can specify that the model contains no deterministic terms, a constant, or constant and trend. But as the Elliott–Jansson model contains one or more x variables as well, you may also specify that those variables' deterministic model contains no deterministic terms, a constant, or constant and trend.

Five cases are defined:

1. No constant nor trend in model
2. Constant in y only
3. Constants in both $\{y, x\}$
4. Constant and trend in y , constant in x



Like other unit root tests, you must specify the deterministic model assumed for y . As in `dfgls`, you can specify that the model contains no deterministic terms, a constant, or constant and trend. But as the Elliott–Jansson model contains one or more x variables as well, you may also specify that those variables' deterministic model contains no deterministic terms, a constant, or constant and trend.

Five cases are defined:

1. No constant nor trend in model
2. Constant in y only
3. Constants in both $\{y, x\}$
4. Constant and trend in y , constant in x
5. No restrictions



The `urcovar` Stata command implements the EJ test using Mata to perform a complicated sequence of matrix manipulations that produce the test statistic. EJ's Table 1 of asymptotic critical values is stored in the program and used to produce a critical value corresponding to the R^2 for your data. The command syntax:

```
urcovar depvar varlist [if exp] [in range] [ , maxlag(#) case(#)  
firstobs ]
```

where the `case` option specifies the deterministic model, with default of case 1. The `maxlag` option specifies the number of lags to be used in computing the VAR (default 1). The `firstobs` option specifies that the first observation of `depvar` should be used to define the first quasi-difference (rather than zero). The `urcovar` command is available for Stata 9.2 or Stata 10 via `findit urcovar`.



As an illustration of `urcovar` use, we consider an experiment similar to that tested in EJ, who in turn refer to a Blanchard–Quah model. The variable of interest is U.S. personal income (`PINCOME`). The single stationary covariate to be considered is the U.S. unemployment rate (`UNRATE`). Both are acquired from the FRED database with the `freduse` command (Drukker, *Stata Journal*, 2006) and have been converted to the common quarterly frequency for 1950Q2–1987Q4 using `tscollap` (Baum, *STB-57*, 2000).

We first present a line plot of these two series, then the output from a conventional DF-GLS test, followed by the output from `urcovar`, cases 3 and 5. The `maxlag` considered in both the `dfgls` and `urcovar` tests is set to eight quarters.





```
. dfglS PINCOME, maxlag(8) trend
```

```
DF-GLS for PINCOME
```

```
Number of obs = 142
```

[lags]	DF-GLS tau Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
8	-0.830	-3.519	-2.875	-2.593
7	-0.585	-3.519	-2.890	-2.607
6	-0.173	-3.519	-2.905	-2.620
5	-0.071	-3.519	-2.918	-2.633
4	-0.272	-3.519	-2.932	-2.645
3	-0.334	-3.519	-2.944	-2.656
2	0.332	-3.519	-2.955	-2.666
1	0.655	-3.519	-2.966	-2.676

```
Opt Lag (Ng-Perron seq t) = 8 with RMSE 11.44531
```

```
Min SC = 5.103851 at lag 3 with RMSE 11.96667
```

```
Min MAIC = 5.000489 at lag 7 with RMSE 11.56349
```



```
. urcovar PINCOME UNRATE, maxlag(8) case(3)
Elliott-Jansson unit root test for PINCOME      1950q2 - 1987q4
Number of obs:      143
Stationary covariates: UNRATE
Deterministic model: Case 3
Maximum lag order:  8
Estimated R-squared: 0.9950
H0: rho = 1 [ PINCOME is I(1) ]
H1: rho < 1 [ PINCOME is I(0) ]
Reject H0 if Lambda < critical value
Lambda:              11.8307
5% critical value:  17.9900
```

```
. urcovar PINCOME UNRATE, maxlag(8) case(5)
Elliott-Jansson unit root test for PINCOME      1950q2 - 1987q4
Number of obs:      143
Stationary covariates: UNRATE
Deterministic model: Case 5
Maximum lag order:  8
Estimated R-squared: 0.8407
H0: rho = 1 [ PINCOME is I(1) ]
H1: rho < 1 [ PINCOME is I(0) ]
Reject H0 if Lambda < critical value
Lambda:              17.1558
5% critical value:  29.1170
```



The DF-GLS test is unable to reject its null of $I(1)$ at any reasonable level of significance. When we augment the test with the stationary covariate in the EJ test, quite different results are forthcoming. Case 3 allows for constant terms (but no trends) in both the quasi-difference of $PINCOME$ and $UNRATE$. The R^2 in this system is over 0.99. Case 5 allows constant terms and trends in both equations of the VAR, with an R^2 of 0.84.



The DF-GLS test is unable to reject its null of $I(1)$ at any reasonable level of significance. When we augment the test with the stationary covariate in the EJ test, quite different results are forthcoming. Case 3 allows for constant terms (but no trends) in both the quasi-difference of $PINCOME$ and $UNRATE$. The R^2 in this system is over 0.99. Case 5 allows constant terms and trends in both equations of the VAR, with an R^2 of 0.84.

Like the DF-GLS test, the EJ test has a null hypothesis of nonstationarity ($I(1)$). The EJ test statistic, λ , must be compared with the interpolated 5% critical value. A value of λ **smaller** than the tabulated value leads to a rejection, and *vice versa*. In both cases, we may reject the null hypothesis at the 95% level of confidence in favor of the alternative hypothesis of stationarity.



Unit root tests allowing for structural breaks

Unit root tests are particularly susceptible to breaks in the structure of a relationship. For instance, the hypothesis of purchasing power parity (PPP) in international trade implies that real exchange rates should be stationary stochastic processes. A vast literature contains numerous instances lacking support for this hypothesis (e.g., Baum et al., *J. Intl. Money and Fin.*, 2001).



Unit root tests allowing for structural breaks

Unit root tests are particularly susceptible to breaks in the structure of a relationship. For instance, the hypothesis of purchasing power parity (PPP) in international trade implies that real exchange rates should be stationary stochastic processes. A vast literature contains numerous instances lacking support for this hypothesis (e.g., Baum et al., *J. Intl. Money and Fin.*, 2001).

One rationale that has been put forth for rejection of the PPP hypothesis is the existence of structural breaks. In this section, we discuss how standard unit root tests may be modified to allow for structural breaks.



Perron and Vogelsang (*JBES*, 1992), building on work by Perron (*JBES*, 1990), demonstrate that nonrejection of the unit-root hypothesis may be “associated with an apparent permanent change in the level of the series” (1992, p. 302). As Perron demonstrated with a simulation experiment, “...if the magnitude of the change is significant, one could hardly reject the unit-root hypothesis even if the series would consist of *i.i.d.* disturbances around a deterministic component (albeit one with a shift in mean)...The problem is one of model misspecification.” (1990, p.155)



Perron and Vogelsang (*JBES*, 1992), building on work by Perron (*JBES*, 1990), demonstrate that nonrejection of the unit-root hypothesis may be “associated with an apparent permanent change in the level of the series” (1992, p. 302). As Perron demonstrated with a simulation experiment, “...if the magnitude of the change is significant, one could hardly reject the unit-root hypothesis even if the series would consist of *i.i.d.* disturbances around a deterministic component (albeit one with a shift in mean)...The problem is one of model misspecification.” (1990, p.155)

To deal with this source of bias in unit-root tests, Perron and Vogelsang propose a class of test statistics which allow for two alternative forms of change: the *additive outlier* (AO) model, capturing a sudden change, and the *innovational outlier* (IO) model, appropriate for modeling a gradual shift in the mean of the series.



The test statistics do not require a priori knowledge of the breakpoint, as their computation involves search over the sample for a single break date. The breakpoint, should it occur, is denoted by T_b , $1 < T_b < T$, where T is the sample size. The AO model considers the dynamics of y_t to be given by

$$y_t = \delta DT_{bt} + y_{t-1} + w_t, \quad t = 2, \dots, T \quad (1)$$

with $DT_{bt} = 1$ for $t = T_b + 1$, and 0 otherwise, under the null hypothesis of a unit root. Under the alternative hypothesis,

$$y_t = c + \delta DU_t + v_t, \quad t = 2, \dots, T \quad (2)$$

where $DU_t = 1$ for $t > T_b$, and 0 otherwise.



This more general specification nests the null hypothesis (1) in the case that the distribution of v_t may be factored into a unit root and a stationary ARMA process. The test strategy is then to estimate the regression

$$y_t = \mu + \delta DU_t + \tilde{y}_t \quad (3)$$

the residuals of which (\tilde{y}_t) are regressed on their lagged values, lagged differences, and a set of dummy variables, the latter needed to ensure that the distribution of the test statistic will be manageable:

$$\tilde{y}_t = \sum_{i=0}^k \omega_i DT_{bt-i} + \alpha \tilde{y}_{t-1} + \sum_{i=1}^k \theta_i \Delta \tilde{y}_{t-i} + e_t, \quad t = k+2, \dots, T \quad (4)$$



This regression, similar in nature to the common Augmented Dickey–Fuller (ADF) model, yields an estimate of α which will be significantly less than one in the presence of stationarity. Perron and Vogelsang provide critical values and describe the method by which they may be simulated.



This regression, similar in nature to the common Augmented Dickey–Fuller (ADF) model, yields an estimate of α which will be significantly less than one in the presence of stationarity. Perron and Vogelsang provide critical values and describe the method by which they may be simulated.

The equivalent process for the innovational outlier (IO) model expresses the shock (for instance, the effect of δ in (1) above) as having the same effect on y_t as any other shock, so that the dynamic effects of DT_b have the same ARMA representation as do other shocks to the model.



This formulation, when transformed, generates the finite AR model

$$y_t = \mu + \delta DU_t + \vartheta DT_{bt} + \alpha y_{t-1} + \sum_{i=1}^k \theta_i \Delta y_{t-i} + e_t, \quad t = k+2, \dots, T \quad (5)$$

which again yields a test of α differing from one in the presence of stationarity. In both the AO and the IO models, the appropriate values of T_b (the breakpoint) and k (the autoregressive order) are unknown. This is resolved for T_b by estimating the model for each feasible breakpoint, and following one of several proposed rules to identify the optimal single breakpoint. In our application, we search for the minimum t-statistic on δ . Conditional on that T_b , the autoregressive order k is chosen, as Perron (1990) suggests, by a sequence of pairs of F-tests for the significance of lags, starting from an appropriately large maximum order.



The unit-root test statistics forthcoming from the AO and IO models will account for one-time level shifts which might otherwise be identified as departures from stationarity. However, the behavior of real exchange rate series over our sample period may not be adequately characterized by a single shift; as Lothian (*JIMF*, 1998) has noted, US dollar-based real exchange rates appear to have exhibited two shifts in mean over the 1980-1987 period, approximately reverting to their pre-1980 level after 1987. In these circumstances, allowing for a single level shift will not suffice.



The Perron–Vogelsang methodology has been extended to double mean shifts by Clemente et al. (*Econ.Letters*, 1988), who demonstrate that a two-dimensional grid search for breakpoints (T_{b1} and T_{b2}) may be used for either the AO or IO models, and provide critical values for the tests. In this context, the AO model involves the estimation of:

$$y_t = \mu + \delta_1 DU_{1t} + \delta_2 DU_{2t} + \tilde{y}_t \quad (6)$$

and subsequently searching for the minimal t -ratio for the hypothesis $\alpha = 1$ in the model:

$$\tilde{y}_t = \sum_{i=0}^k \omega_i DT_{b1,t-i} \sum_{i=0}^k \omega_i DT_{b2,t-i} + \alpha \tilde{y}_{t-1} + \sum_{i=1}^k \theta_i \Delta \tilde{y}_{t-i} + e_t, \quad (7)$$

$$t = k + 2, \dots, T$$



For the IO model, the modified equation to be estimated becomes:

$$y_t = \mu + \delta_1 DU_{1t} + \delta_2 DU_{2t} + \vartheta_1 DT_{b1,t} + \vartheta_2 DT_{b2,t} + \alpha y_{t-1} + \sum_{i=1}^k \theta_i \Delta y_{t-i} + \mathbf{e}_t, \quad t = k + 2, \dots, T \quad (8)$$

with a search for the minimal t -ratio for the hypothesis $\alpha = 1$. These tests customarily are applied to a trimmed sample; we trimmed 5% of the sample from each end when searching for the breakpoints.



In Baum et al. (*JIFMIM*, 1999), the results from these two-mean-break models are quite consistent over 17 countries and both CPI and WPI price series. In none of the 58 cases considered do the unit-root test statistics surpass their approximate 5% critical values, although the t-statistics for δ_1 and δ_2 generally indicate the presence of meaningful level shifts in almost every instance.



In Baum et al. (*JIFMIM*, 1999), the results from these two-mean-break models are quite consistent over 17 countries and both CPI and WPI price series. In none of the 58 cases considered do the unit-root test statistics surpass their approximate 5% critical values, although the t-statistics for δ_1 and δ_2 generally indicate the presence of meaningful level shifts in almost every instance.

Even with structural breaks taken into account, the evidence in favor of nonstationarity is overwhelmingly strong and consistent across countries for both CPI-based and WPI-based real exchange rate series.



Therefore, we may conclude that the inability to reject the unit-root hypothesis for the post-Bretton Woods era using standard univariate unit-root tests is not likely to be overturned by allowing for one or two mean breaks in the series. Such instability is quite apparent in a first-order Markov model of the real exchange rate, but even when unit-root tests are adjusted for its presence, the null hypothesis of nonstationarity cannot be rejected in favor of mean reversion.



Therefore, we may conclude that the inability to reject the unit-root hypothesis for the post-Bretton Woods era using standard univariate unit-root tests is not likely to be overturned by allowing for one or two mean breaks in the series. Such instability is quite apparent in a first-order Markov model of the real exchange rate, but even when unit-root tests are adjusted for its presence, the null hypothesis of nonstationarity cannot be rejected in favor of mean reversion.

These routines for unit root tests in the presence of structural breaks are available as Stata commands `clemao1`, `clemao2`, `clemio1`, `clemio2`. To install, findit `clemao`.



A general test for structural stability: qll

Elliott and Müller's 2006 paper in *Review of Economic Studies* (EM) addresses the large literature on testing a time series model for structural stability. They consider “tests of the null hypothesis of a stable linear model

$$y_t = X_t' \bar{\beta} + Z_t' \gamma + \epsilon$$

against the alternative of a partially unstable model

$$y_t = X_t' \beta_t + Z_t' \gamma + \epsilon$$

where the variation in β_t is of the strong form” (p. 907), or nontrivial.



Consideration of this alternative has led to a huge literature based on the “diversity of possible ways $\{\beta_t\}$ can be non-constant.” EM point out that optimal tests and their asymptotic distributions have not been derived for many particular models of the alternative.



Consideration of this alternative has led to a huge literature based on the “diversity of possible ways $\{\beta_t\}$ can be non-constant.” EM point out that optimal tests and their asymptotic distributions have not been derived for many particular models of the alternative.

Their approach develops a single unified framework, noting that the “seemingly different approaches of ‘structural breaks’ and ‘random coefficients’ are in fact equivalent.” (p.908) EM unify the approaches that describe a breaking process with a number of non-random parameters with tests that specify stochastic processes for $\{\beta_t\}$ without requiring to specify its exact evolution.



The processes considered include breaks that occur in a random fashion, serial correlation in the changes of the coefficients, a clustering of break dates, and so on. Under a normality assumption on the disturbances, “small sample efficient tests in this broad set are asymptotically equivalent” and “leaving the exact breaking process unspecified (apart from a scaling parameter) does not result in a loss of power in large samples.” (p. 908)



The consequences of this approach to the problem of structural stability are profound. “The equivalence of power over many models means that there is little point in deriving further optimal tests for particular processes in our set” (p. 908) and the researcher can carry out (almost) efficient inference without specifying the exact path of the breaking process.



The consequences of this approach to the problem of structural stability are profound. “The equivalence of power over many models means that there is little point in deriving further optimal tests for particular processes in our set” (p. 908) and the researcher can carry out (almost) efficient inference without specifying the exact path of the breaking process.

Furthermore, the computation of EM’s Quasi-Local Level (q_{LL}) test statistic is straightforward, and it remains valid for very general specifications of the error term and covariates. The computation requires no more than $(k + 1)$ OLS regressions for a model with k covariates, in contrast to many approaches which require T or T^2 regressions. No arbitrary trimming of the data is required.



In the *structural break* literature, a fixed number of N breaks at τ_1, \dots, τ_N are assumed. Much of the literature addresses $N = 1$: e.g. the “Chow test”, cusums tests of Brown–Durbin–Evans, Bai and Perron, Andrews and Ploberger, etc.



In the *structural break* literature, a fixed number of N breaks at τ_1, \dots, τ_N are assumed. Much of the literature addresses $N = 1$: e.g. the “Chow test”, cusums tests of Brown–Durbin–Evans, Bai and Perron, Andrews and Ploberger, etc.

In contrast, the *time-varying parameter* literature considers a random process generating β_t : often considered as a random walk process. The approaches of Leybourne and McCabe, Nyblom, and Saikkonen and Luukonen are based on classical statistics, while Koop and Potter and Giordani et al. consider a Bayesian approach. All of these approaches are very analytically challenging.



In the *structural break* literature, a fixed number of N breaks at τ_1, \dots, τ_N are assumed. Much of the literature addresses $N = 1$: e.g. the “Chow test”, cusums tests of Brown–Durbin–Evans, Bai and Perron, Andrews and Ploberger, etc.

In contrast, the *time-varying parameter* literature considers a random process generating β_t : often considered as a random walk process. The approaches of Leybourne and McCabe, Nyblom, and Saikkonen and Luukonen are based on classical statistics, while Koop and Potter and Giordani et al. consider a Bayesian approach. All of these approaches are very analytically challenging.

EM argue that tests for one of these phenomena will have power against the other, and *vice versa*. Therefore a single approach will suffice.



EM raise the interesting question: why do we test for parameter constancy? They consider three motivations:

1. Stability relates to theoretical constructs such as the Lucas critique of economic policymaking



EM raise the interesting question: why do we test for parameter constancy? They consider three motivations:

1. Stability relates to theoretical constructs such as the Lucas critique of economic policymaking
2. Forecasting will depend crucially on a stable relationship



EM raise the interesting question: why do we test for parameter constancy? They consider three motivations:

1. Stability relates to theoretical constructs such as the Lucas critique of economic policymaking
2. Forecasting will depend crucially on a stable relationship
3. Standard inference on $\bar{\beta}$ will be useless if $\{\beta_t\}$ varies in a permanent fashion; persistent changes will render a fixed model misleading



EM raise the interesting question: why do we test for parameter constancy? They consider three motivations:

1. Stability relates to theoretical constructs such as the Lucas critique of economic policymaking
2. Forecasting will depend crucially on a stable relationship
3. Standard inference on $\bar{\beta}$ will be useless if $\{\beta_t\}$ varies in a permanent fashion; persistent changes will render a fixed model misleading



EM raise the interesting question: why do we test for parameter constancy? They consider three motivations:

1. Stability relates to theoretical constructs such as the Lucas critique of economic policymaking
2. Forecasting will depend crucially on a stable relationship
3. Standard inference on $\bar{\beta}$ will be useless if $\{\beta_t\}$ varies in a permanent fashion; persistent changes will render a fixed model misleading

“The more pervasive these three motivations are, the more persistent the changes in $\{\beta\}$.” (p. 912) Therefore, EM propose that a useful test should maximize its power against persistent changes in $\{\beta_t\}$.



The conditions underlying the EM test allow for diverse breaking models, from relatively rare (including a single break) to very frequent small breaks (such as breaks every period with probability p). Breaks can also occur with a regular pattern, such as every 16 quarters following U.S. presidential elections.



The conditions underlying the EM test allow for diverse breaking models, from relatively rare (including a single break) to very frequent small breaks (such as breaks every period with probability p). Breaks can also occur with a regular pattern, such as every 16 quarters following U.S. presidential elections.

Computation of the q_{LL} test statistic is straightforward, relying only on OLS regressions and construction of an estimate of the long-run covariance matrix of $\{X_t\epsilon_t\}$. For uncorrelated ϵ_t , a robust covariance matrix will suffice. For possibly autocorrelated ϵ_t , a HAC (Newey–West) covariance matrix is appropriate.



The conditions underlying the EM test allow for diverse breaking models, from relatively rare (including a single break) to very frequent small breaks (such as breaks every period with probability p). Breaks can also occur with a regular pattern, such as every 16 quarters following U.S. presidential elections.

Computation of the q_{LL} test statistic is straightforward, relying only on OLS regressions and construction of an estimate of the long-run covariance matrix of $\{X_t\epsilon_t\}$. For uncorrelated ϵ_t , a robust covariance matrix will suffice. For possibly autocorrelated ϵ_t , a HAC (Newey–West) covariance matrix is appropriate.

The null hypothesis of parameter stability is rejected for **small** values of \widehat{q}_{LL} : that is, values more negative than the critical values. Asymptotic critical values are provided by EM for $k = 1, \dots, 10$ and are independent of the dimension of Z_t (the set of covariates assumed to have stable coefficients).



The `qll` Stata command implements the EM q_{LL} test using Mata to produce the test statistic. EM's Table 1 of asymptotic critical values is stored in the program and used to produce 10%, 5% and 1% critical values corresponding to number of regressors with potentially unstable parameters. The command syntax:

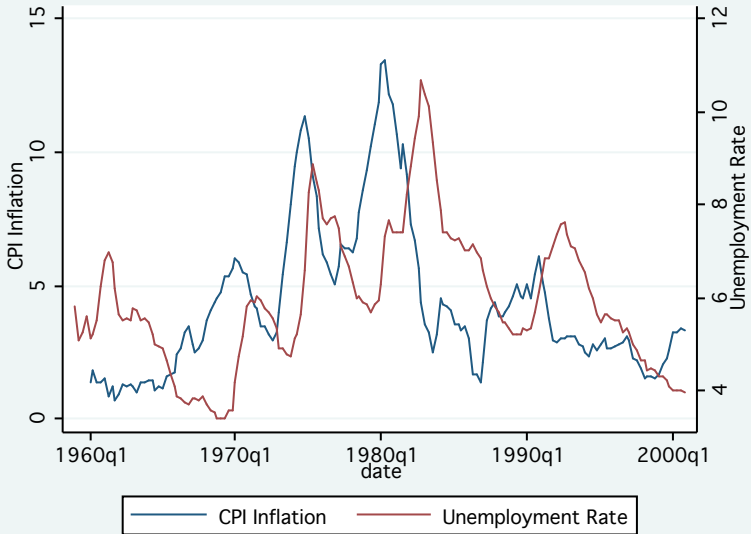
```
qll depvar varlist [if exp] [in range] [ , (zvarlist) rlag(#) ]
```

where the parenthesized `zvarlist` optionally specifies the list of covariates assumed to have stable coefficients (none are required). The `rlag` option specifies the number of lags to be used in computing the long-run covariance matrix of $\{X_t \epsilon_t\}$. If a negative value is given, the optimal lag order is chosen by the BIC criterion. The `qll` command is available for Stata 9.2 or Stata 10; `findit qll`.



We consider a regression of inflation on the lagged unemployment rate, the Treasury bill rate and the Treasury bond rate. We assume the latter two coefficients are stable over the period. We test over the full sample and a 1990–2000 subsample.





```

. qll inf L.UR (TBILL TBON), rlag(8)
Elliott--Müller qLL test statistic for time varying coefficients
in the model of inf, 1960q1 - 2000q4
Allowing for time variation in 1 regressors
H0: all regression coefficients fixed over the sample period (N = 164)
Test stat.    1% Crit.Val.    5% Crit.Val.    10% Crit.Val.
      -2.260          -11.05          -8.36          -7.14
Long-run variance computed with 8 lags.

. qll inf L.UR (TBILL TBON) if tin(1990q1,), rlag(8)
Elliott-Müller qLL test statistic for time varying coefficients
in the model of inf, 1990q1 - 2000q4
Allowing for time variation in 1 regressors
H0: all regression coefficients fixed over the sample period (N = 44)
Test stat.    1% Crit.Val.    5% Crit.Val.    10% Crit.Val.
      -6.647          -11.05          -8.36          -7.14
Long-run variance computed with 8 lags.

```



In both samples, using eight lags to calculate the long-run covariance matrix, the null hypothesis that the coefficients on the lagged unemployment rate ($L.UR$) are stable cannot be rejected at the 10% level of confidence. The Elliott–Müller q_{LL} test indicates that the stability of this regression model, allowing for instability in the coefficient of the unemployment rate only, cannot be rejected by the data.

