Chapter 2

Misspecification Analysis I: Coefficients' (In)Stability

Artur Silva Lopes, February 2022

ISEG–ULisboa

1 Introduction

Whatever is its purpose, any model should be used only if data validates all its assumptions. (Mis)Specification analysis consists precisely in this checking by means of statistical tests and that is the reason why D. F. Hendry proclaimed that the 3 golden rules of Econometrics are "*Test! Test! Test!*".

Examples of this checking are tests for heteroskedasticity or the RESET test for functional form.

Two of the assumptions of macroeconometric models are particularly sensitive:

a) the implicit assumption of parameter constancy or stability, in particular constancy or stability of their coefficients;

b) the assumption of no serial correlation of the errors.

2 Basic Tests

The coefficients' stability hypothesis is crucial but only implicit in

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t, \quad t = 1, 2, \ldots, T.$$

The absence of any index in the coefficients means that they are assumed to be the same for all the equations, i. e., the same through time.

Recall also that the crucial assumption for the model with predetermined regressors demands that the joint process $\{(y_t, \mathbf{x}_t)\}$ is stationary. For this to hold, it is required that the coefficients connecting y_t to the variables of $\mathbf{x_t}$ are constant, not varying with time, so that the process is stable through time.

Notice also that only tests assuming that stationarity holds will be considered. This is a restriction with some importance.

Some forecasting issues will be also handled. Notice that a model with unstable coefficients is useless to make both policy simulation and forecasting. Obviously, such a model is not trustworthy for any of those purposes. If a model has already shown instability issues inside the sample, how confident can we be on its forecasts? How can we trust that the same problems will not arise in the forecasting period?

Coefficient instability is also closely related to the well known “Lucas critique” to macroeconomic models.

Notice also that a misspecified model – e.g., with omitted regressors or a wrong functional form – may exhibit symptoms of coefficients’ instability. For instance, when the true relation is non-linear, e.g., convex, the linear model will likely present such erroneous symptons, hiding the true problem. Therefore, tests for coefficients’ stability may be also viewed as general specification tests or as *misspecification tests*.

2.1 Preliminary issues

The word *break* is used to denote a change of the regression function in the sample. Breaks can affect the coefficients, the error variance (heteroskedasticity) or both. Only the first class of problems will be studied but I recommend that you investigate what was the “Great Moderation”.

Considering only breaks in coefficients, there are 2 major types of breaks or forms of *structural change*:

a) discrete or abrupt or sudden changes or jumps in the regression coefficients, originating at least two distinct *regimes* in the sample;

b) smooth or gradual changes of the coefficients throughout the sample, which are often modeled using random walks:

$$\beta_{j,t} = \beta_{j,t-1} + \epsilon_{j,t},\ j = 1, 2, \ldots, k,\ t = 1, 2, \ldots, T.$$

The methods that we will study are designed mostly for the first type. However, they also have power to detect gradual change; i.e., statistical tests for discrete breaks often can detect gradual changes, and vice-versa. Implicitly, I am assuming that the null hypothesis refers to parameter constancy or stability in the sample period.

The main motivation to use the tests is usually associated to abrupt changes: a big change is known to have occured, e. g. a political one, and we wish to know whether economic relations changed as well, manifesting through changes in the regression coefficients.

The portuguese economy provides lots of these events: besides the big international shocks (as the oil shocks or the recent "Great Financial Crisis"), in the last 47 years we had the April Revolution (1974), the IMF "interventions" and more recently the one of the Troika, the adhesion to the EEC (EU) and to the European Monetary System.

2.2 Chow tests: known break date

The first Chow (1960) test or *"Chow breakpoint test"* is designed for those abrupt changes: one suspects that the sample may be broken into two, corresponding to different environments; e.g., different regimes of exchange rates or different economic policy regimes. Most importantly, the breakdate or break point must be known from the outset.

This is the most simple test: the sample is split in two (subsamples or subperiods). Any criticism to this feature is irrelevant because the analysis can be easily generalized to two or more breakdates, i.e., to partitions of the sample with more than two subperiods.

The real relevant limitation is the assumption on the *a priori* knowledge about the *time of the break*, assumed exogenous to data, T_b. Therefore the analysis is "pure", non-contaminated by any previous contact with data. Otherwise, power may be increased but at the cost of some size distortion, i.e., real size exceeding nominal size. Size inflation must be avoided.

The sample is split in two

$$y_t = \begin{cases} \mathbf{x}_t'\boldsymbol{\beta}_1 + u_t, & t = 1, 2, \ldots, T_b, \text{ or } t \leq T_b, \\ \mathbf{x}_t'\boldsymbol{\beta}_2 + u_t, & t = T_{b+1}, T_{b+2}, \ldots, T, \text{ or } t > T_b, \end{cases}$$

with T_1 and T_2 denoting the number of observations for each period, ($T_1 + T_2 = T$) and $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ the corresponding vectors of k coefficients. The convention to call T_b the date of the last observation of the first period is adopted but this is not universal (for some it is the first observation of the new regime).

And this is really a changing regime or *regime switching* (in T_b) model, or a *threshold* model, where the threshold variable is time and its threshold is T_b: when $t \leq T_b$ regime 1 rules and with $t > T_b$ regime 2 rules.

Hence, in matrix notation:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}.$$

It is also assumed that $\mathsf{E}(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \mathsf{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I}$, i.e., the errors must be serially uncorrelated and homoskedastic. Only the coefficients are allowed to vary, not the variances.

The purpose is to test

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta} \quad vs. \quad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2,$$

that is, to test the equality of the vectors of coefficients in the two subsamples conditional on the equality of variances.

2.2.1 First Chow test: $T_1 > k$ and $T_2 > k$

It is an ordinary test for linear restrictions which may be based on the usual F-statistic in terms of sums of squared residuals:

$$\frac{(SSR_R - SSR_{UR})/J}{SSR_{UR}/(T-k)},$$

exactly or asymptotically distributed as $F_{(J,T-k)}$ when H_0 is true (J denotes the number of linear restrictions *).

Restrictions here consist of imposing the same vector of coefficients to both subsamples ($\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}$). SSR_R may be denoted with SSR_0 or with $e_*' e_*$.

Free estimation, without restrictions, consists of applying OLS separately to each subsample, allowing that the data choose the best fitting estimates. Intuitively, SSR_{UR} equals the sum of separate SSRs for each subsample.

*Recall that the denominator is $\widehat{\sigma}^2_{UR}$.

Algebra confirms this intuition:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1'\mathbf{y}_1 \\ \mathbf{X}_2'\mathbf{y}_2 \end{bmatrix} \\
&= \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}_2'\mathbf{X}_2)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1'\mathbf{y}_1 \\ \mathbf{X}_2'\mathbf{y}_2 \end{bmatrix} \\
&= \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1 \\ (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}_2 \end{bmatrix},
\end{aligned}$$

Therefore

$$SSR_{UR} = \mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2.$$

Chow's statistic is simply the F's specialization to this case:

$$F_{CHOW} = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}_1'\mathbf{e}_1 - \mathbf{e}_2'\mathbf{e}_2)/k}{(\mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2)/(T-2k)} \sim F_{(k,T-2k)} \text{ under } H_0,$$

in the classical model. To calculate the statistic we need to run 3 regressions:

with the whole sample and with each of the sub-samples.

Very important in practical terms: these statistics are exactly algebraically equal to the F statistics to test the joint significance of the dummy variables coefficients introduced in the original model to allow its coefficients to change in the second subsample. For instance, when the model is

$$y_t = \beta_1 + \beta_2 \, x_t + \beta_3 \, z_t + \beta_4 \, w_t + u_t,$$

define the ordinary (step) dummy

$$D_t = \begin{cases} 0, & t = 1, 2, \ldots, T_b, \\ 1, & t = T_{b+1}, T_{b+2}, \ldots, T, \end{cases}$$

and respecify the model:

$$y_t = \beta_1 + \delta_1 \, D_t + \beta_2 \, x_t + \delta_2 \, D_t \times x_t + \beta_3 \, z_t + \delta_3 \, D_t \times z_t + \beta_4 \, w_t + \delta_4 \, D_t \times w_t + v_t.$$

It can be shown that the Chow statistic is numerically equal to the F statistic to test

$$H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = 0 \quad vs. \quad H_1 : \exists \delta_j \neq 0, \; j = 1, \ldots, 4,$$

i.e., the restrictions that impose that the coefficients in the two subperiods are equal.

The estimation of this model can be more informative than the Chow' statistic because it allows an informal analysis of the stability of the coefficients.

2.2.2 Second Chow test: $T_1 \leq k$ or $T_2 \leq k$

The null and the alternative hypotheses are the same. But the assumption of the normal distribution for the errors is now strictly necessary.

Note that if $T_1 \leq k$ or $T_2 \leq k$, then its corresponding SSR, $SSR_1 = 0$ or $SSR_2 = 0$. For instance, if T_1 (or T_2) $= 1$ and $k = 2$, there is an infinite number of OLS lines that minimize the SSR passing exactly on that single point, making the residual equal to zero and hence SSR_1 (or SSR_2) $= 0$. If T_1 (or T_2) $= 2$ and $k = 2$, there is a single line defined by those two points that makes each residual equal to zero and hence SSR_1 (or SSR_2) $= 0$. And in case we have only two observations to run a regression with $k = 3$ coefficients, we have again an infinite number of solutions that minimize the SSR annulling it: all the planes that pass through the 2 points annul the residuals for both observations. And we also have $SSR = 0$ when $T = 3$ and $k = 3$ because the surface defined by the 3 points annuls the residuals for the 3 observations, etc. .

Without any loss of generality, assume that the insufficient number of observations occurs in the second sub-period (as is usual): $T_2 \leq k$. Then, intuitively

$$F_{CHOW} = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}_1'\mathbf{e}_1)/T_2}{\mathbf{e}_1'\mathbf{e}_1/(T_1 - k)} \sim F_{(T_2, T_1 - k)} \text{ under } H_0,$$

because $\mathbf{e}_2'\mathbf{e}_2 = 0$, and degrees of freedom must be also adjusted accordingly. Only two regressions are now run: the one with all observations and the one with the data of the first sub-period only.

It is not surprising that this test is, in general, less powerful than the first Chow test.

Both Chow tests have special cases when only a subset of coefficients is scrutinized; for instance, only the slope coefficients or only the intercept. This poses no special problems but the dummy variable approach appears to be more attractive.

2.2.3 The forecast test

Actually it is only *one* of the several forecast tests, classical, called also post sample predictive test or *predictive failure* test.

Assume that some time has passed since the end of the sample and that g new observations became available. The purpose is to assess how good the model is to make forecasts comparing those new observations with the model forecasts.

Denote the new matrix and the new vector of observations with $\mathbf{X}_f$ and $\mathbf{y}_f$, respectively.

Forecasts are conditional in $\mathbf{X}_f$, they are denoted with $\widehat{\mathbf{y}}_f$ and are given by

$$\widehat{\mathbf{y}}_f = \widehat{\mathbf{y}}_f | \mathbf{X}_f = \mathbf{X}_f \widehat{\boldsymbol{\beta}}.$$

We seek a test statistic allowing the comparison between $\mathbf{y}_f$ and $\widehat{\mathbf{y}}_f$.

This is a crucial test to validate the model: successful forecasts are far more convincing than a good in-sample fit because the new observations are “fresh”, they were not previously used at the stages of model specification and estimation, when many attempts could have been made searching for the better fit, they were not used in a likely *data mining* process.

However, the previous condition is hard to comply with. Rather than waiting say a few years for the new observations, it is recommended that some observations at the end of the sample are reserved from the outset to perform the test.

Thus, the partition $T = T_1 + T_2$ now represents the splitting between observations for specification and estimation (T_1), and for forecast assesment ($T_2 = g$).

The restrictions we wish to test are now stochastic:

$$H_0 : \mathbf{y}_f = \mathbf{X}_f \boldsymbol{\beta} + \mathbf{u}_f \quad vs. \quad H_1 : \text{not } H_0,$$

The importance of parameter stability is clear: the new observations must be generated the same way as the old, particularly with the same coefficients.

However, all the usual assumptions are also included in H_0, at least implicitely: in particular, $\text{Var}(\mathbf{u}_f|\mathbf{X}_f) = \sigma^2 \mathbf{I}_f$, $\mathbf{u}_f|\mathbf{X} \sim \mathcal{N}$, etc. Therefore the test is sensitive to any departure from the basic assumptions: for instance, a possible increase in variance in the forecasting period is not a sufficient excuse to poor forecasts, far away from reality.

Denote with $\mathbf{f}$ the vector of forecast errors:

$$\begin{aligned}\mathbf{f} &= \mathbf{y}_f - \widehat{\mathbf{y}}_f = \mathbf{y}_f - \mathbf{X}_f \widehat{\boldsymbol{\beta}} \\ &= \mathbf{X}_f \boldsymbol{\beta} + \mathbf{u}_f - \mathbf{X}_f \widehat{\boldsymbol{\beta}} \\ &= -\mathbf{X}_f (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{u}_f \\ &= -\mathbf{X}_f (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u} + \mathbf{u}_f\end{aligned}$$

where H_0 was already imposed. Forecast errors have two components: a) the error in estimating $\boldsymbol{\beta}$, and b) the own model errors in the forecast period, both unavoidable.

Since forecast errors can be positive or negative and because they can be large possibly due to a large variance, the test statistic must be a quadratic form, and they should be weighted by their inverse covariance matrix:

$$\mathbf{f}' \left[\mathbf{Var}(\mathbf{f})\right]^{-1} \mathbf{f}.$$

To estimate Var($\mathbf{f}$) the assumption of no serial correlation between $\mathbf{u}$ and $\mathbf{u}_f$ is crucial. No serial correlation in the forecast period (inside $\mathbf{u}_f$) is also assumed.

Using also the normality assumption, it can be shown that

$$F = \frac{\mathbf{f}'[\mathbf{I}_g + \mathbf{X}_f(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_f']^{-1}\mathbf{f}/g}{\mathbf{e}_1'\mathbf{e}_1/(T_1 - k)} \sim F_{(g,T_1-k)} \text{ under } H_0,$$

where it is assumed that estimation is based only on the first T_1 observations.

It can be shown that the quadratic form in the numerator is the difference between SSRs for the whole sample and for the first subsample:

$$SSR_{T_1+g} - SSR_{T_1} = \mathbf{e}_*'\mathbf{e}_* - \mathbf{e}_1'\mathbf{e}_1.$$

Therefore, the test statistic really is the same of the second Chow test test (recall that $T_2 = g$). The only difference lies in the interpretation of the second subperiod, now the forecast assessment period.

However, contrarily to the Chow test, when $T_2 = g > k$ the test statistic does not change, it is computed the same way. That is, if $T_2 = g > k$ two distinct statistics can be computed: the one of the first Chow test and the one of the forecast test, each with its own interpretation (albeit this last one also has a possible stability interpretation).

But if $T_2 = g \leq k$ only one statistic can be computed and it has two distinct interpretations: one exclusively in terms of coefficients' stability and the other concerning the assessment of model's forecasts.

This test can still be deduced in a very different way. Define for each observation of the (pseudo) forecast period an impulse *dummy*, i.e., with 1 for that observation only; for instance:

$$D_{T_1+1} = \begin{cases} 1, \ t = T_1 + 1, \\ 0, \ t \neq T_1 + 1, \end{cases}$$

and do the same for the following observations, till the end of the sample.

Putting together all the observations, we have

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_f \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{X}_f & \mathbf{I}_g \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \gamma \end{bmatrix} + \begin{bmatrix} \mathbf{u} \\ \mathbf{u}_f \end{bmatrix}.$$

It can then be shown that:

a) the estimates of $\boldsymbol{\beta}$ from this model are exactly equal to the estimates obtained with the (T_1) sample data only $(\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u})^{\dagger}$;

b) the OLS estimate of the vector γ, $\widehat{\gamma}$, equals exactly the forecast errors that we get when the model is estimated with the first subsample (the first T_1 observations): $\widehat{\gamma} = \mathbf{f}$;

c) the statistic to test $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ $vs.$ $H_1 : \boldsymbol{\gamma} \neq \mathbf{0}$ is the same of the forecat

†Proposition: including impulse dummy variables for certain observations in the linear regression model amounts to remove those observations from the estimation process.

test. Now, under the strict exogeneity assumption, we have that

$$\mathsf{E}(\mathbf{f}|\,\mathbf{X}_f) = \mathsf{E}(\mathbf{X}_f\boldsymbol{\beta} + \gamma + \mathbf{u}_f - \mathbf{X}_f\widehat{\boldsymbol{\beta}}|\mathbf{X}_f) = \gamma,$$

which must be satisfied approximately when the regressors are only predetermined. Thus, the statistic is testing that mean forecast errors are zero, that is, that forecasts are unbiased.

2.3 The QLR test

The major weakness of Chow tests is the assumption about the knowledge of the breakdate, T_b. When T_b is unknown:

a) choosing T_b arbitrarily — e.g., the mid of the sample — produces tests with low power, i.e., even when there is instability it is hardly detected (we become too much dependent on luck).

b) But relying on a previous analysis of data to select T_b distorts inference because the real size of tests exceeds nominal size. This is similar to perform several tests searching for a rejection of the null — break “hunting” —, because the global size or the size of the joint procedure (largely) exceeds the size of an individual test [‡].

In both cases this selection of T_b is made endogenously to the data and the consequence is that $\alpha_R > \alpha_N$ (“R” denotes real and “N” nominal).

However, if there is previous information, *data-exogenous*, on a possible breakdate — for instance, due to an institutional change —, this information can and must be used at will. Provided it is exogenous, there is no size distortion, and the test should be powerful. It may occur, however, that data indicate a slightly different breakdate. One reason for this is that major changes may be non simultaneous with the date of the institutional change.

[‡]And it can be much larger when many tests are performed.

Otherwise, breakdate endogeneity must be acknowledged and incorporated in distribution theory, which is hard. In this case, T_b must be viewed as an additional parameter to estimate.

Richard Quandt was the first with the idea to endogeneize explicitly the estimation of T_b, 60 years ago: he suggested estimating 2 distinct regimes allowing all sample points and selecting the one that maximizes the likelihood function under the alternative hypothesis (of non-constancy). This entails using the likelihood ratio:

$$QLR = \frac{max(\mathcal{L}|H_0)}{max(\mathcal{L}|H_1)},$$

and searching for its minimum. In a sense, the maximization of the denominator must be double: maximimize the likelihood across all values for the coefficients and the variance but also across all possible breakdates. However, Quandt could not deduce the distribution of the test statistic and the problem had to wait more than 30 years to be solved by Andrews (1993), and in a rather general context, that of GMM estimation.

Represent T_b through the corresponding fraction of the sample, with π: $T_b = [\pi T]$, $[.]$ denoting the closest integer, or $\pi = T_b/T$; for instance, if the break occurs exactly in the mid of the sample, $\pi = 0.5$; as another example, if $\pi = 0.2$ the first regime covers the first 20% of the observations and the second the last 80%, etc.

Consider the subset Π from the $[0,1]$ interval, $\Pi \subset [0,1]$, of all possible break fractions and notice that the sample must be trimmed to ensure that the asymptotic approximation works. The most popular subset is $\Pi = (\pi_0; 1-\pi_0) = (0.15; 0.85)$, that is, $\pi_0 = 0.15$, as initially suggested by Andrews.

Andrews was able to derive the asymptotic distribution of the trinity of statistics

$$\sup_{\pi \in \Pi} W(\pi), \quad \sup_{\pi \in \Pi} LM(\pi), \quad \text{and} \quad \sup_{\pi \in \Pi} LR(\pi),$$

and particularly for the regression model with spherical errors, the statistic

$$\sup_{\pi \in \Pi} W = \max_{\pi \in \Pi} T \frac{\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}_1'\mathbf{e}_1 - \mathbf{e}_2'\mathbf{e}_2}{\mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2},$$

frequently known as QLR statistic or $\sup -Wald$ statistic. Notice that since

$$\sup W \propto \max Chow,$$

it is also known as $\max Chow$ (but it does not coincide with it).

The asymptotic distribution (under H_0) entails Wiener processes and requires Monte Carlo simulation. Corrected tables are available in Andrews (2003) and critical values depend on, besides α:

a) the number of parameters being tested, p;

b) π_0, the trimming parameter;

c) another parameter, λ, related with the asymmetry of the Π interval.

E.g., when $\pi_0 = 0.15$ and the interval is symmetric, the critical values are $8.68, 11.72, 14.13$ and 16.36, with $\alpha = 0.05$, for $p = 1, 2, 3$ e 4, respectively [Stock and Watson (2015), p. 611 have the critical values for the $\pi_0 = 0.15$ case for the $\max Chow$ statistic].

The case of known T_b located in the middle of the sample is also considered in these tables: $\pi_0 = 0.5$, $\Pi = \{0.5\}$ only and the distribution is obviously the chi-square.

An important by-product is an estimator(te) for T_b. Indeed, it makes sense to select

$$T_b = \arg\max QLR \Leftrightarrow T_b = \arg\min(\mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2),$$

because the expression inside brackets appears both in the numerator, subtracting $e_*'e_*$, and in the denominator. It is easy to see that this estimator is also a least squares one. Furthermore, it is also consistent; more precisely, what can be shown to be consistent is the estimator of π ($\widehat{\pi} \xrightarrow{p} \pi$).

As a small application example a data sample was generated with the DGP given by

$$y_t = \begin{cases} 1 + 2x_t + u_t, \ t = 1, 2, \ldots, 40, \\ 2 + 2x_t + u_t, \ t = 41, 42, \ldots, 100, \end{cases}$$

with $u \sim iid\mathcal{N}(0, 1)$ and $x_t \sim iid\mathcal{N}(0, 1)$ too.

Hence, there is a small break at $T_b = 40$. To detect the break the sample was swept at all points of $\Pi = (0.15; 0.85)$, producing, e.g., (TSP requires programming because it doesn't produce an automatic statistic):

T_b	16	17	...	38	39	40	41	...	83	84
W	3.92	3.76	...	26.11	28.40	26.98	23.89	...	7.55	7.36

Therefore $QLR = \sup W = 28.40$ and since with $\alpha = 0.05$ (asymptoticaly) the critical region is $RC = \{QLR : QLR > 11.72\}$, one clearly rejects stability. The test correctly detects a break, although not exactly its date, with $\widehat{T}_b = 39$; nevertheless, this estimate is very good.

In Eviews, after the estimation of the model, one can get the statistic with the options

`View => Stability Tests => Quandt-Andrews breakpoint test`.

By default, sample trimming is made with $\pi_0 = 0.15$ but it is possible to choose a different value. The EViews' version is max F (or max $Chow$) and it allows doing a test for partial stability, i. e., testing only the constancy of some of the coefficients.

To analyse the small sample behaviour of the test a modest MC study was performed, with only 5000 replicas for each case, the DGP given by $y_t = 1+1x_t+u_t$, with x_t and u_t both iid$\mathcal{N}(0,1)$. The results for different sample sizes, for 5% nominal size tests are the following:

Real size estimates of the QLR test, in %($\alpha_N = 0.05$)

T	30	50	100	200	400	800
$\widehat{\alpha}$	5.86	4.96	3.76	3.80	3.96	4.04

As expected since it is asymptotic, the test presents apparent size distortion when $T = 30$. However, this distortion or over-rejection is very light; so light indeed that it could be due to sampling error only.

When $T = 50$ the test shows an excellent behaviour and for samples with $T = 100$ and above it is conservative, rejecting the true null H_0 less frequently than the nominal 5%, approaching nominal size when $T \to \infty$ from below.

This is not however a free bonus. In such a case there must be situations where the power of the test is below what it should be, that is, cases where the test should reject a false null hypothesis and it doesn't do it; this is usually the price to pay for the size "discount".

Anyway, for this DGP the small sample size behaviour of this test is very good, particularly bearing in mind that the test is only asymptotic.