Introduction to Time Series Regression and Forecasting

606EC Econometria - Mod. II

2023-2023

Martin Magris

Overview I

Overview II

8. Nonstationarity I: Trends
   - Problems Caused by Stochastic Trends
   - Detecting Stochastic Trends: Testing for a Unit AR Root
   - Avoiding the Problems Caused by Stochastic Trends
9. Nonstationarity II: Breaks
   - Testing for Breaks
   - Detecting Breaks Using Pseudo Out-of-Sample Forecasts
   - Avoiding the Problems Caused by Breaks

Introduction to Time Series Data and Serial Correlation I

Time series data: data collected for a single entity at multiple points in time.

Used to address questions like:

- what is the causal effect on a variable of interest, Y, of a change in another variable, X, over time?
- what is the dynamic causal effect on Y of a change in X?

Lags, First Differences, Logarithms, and Growth Rates I

- The observation on the time series variable Y made at date t is denoted Y_t
- the total number of observations is denoted T
- The interval between observations, the period of time between observation t and observation $t+1$, is some unit of time such as weeks, months, quarters , or years
- A set of T observations on a time series variable Y is denoted as $Y_1, \ldots, Y_T$ or

$$\{Y_t\}, \quad t = 1, \ldots, T$$

- $t = 1$ corresponds to the first date, T to the last date in the dataset

Lags, First Differences, Logarithms, and Growth Rates II

Definition (First difference)

The change in the value of Y between period $t-1$ and period t is $Y_t - Y_{t-1}$; this change is called the first difference in the variable Y_t. In time series data, "Δ" is used to represent the first difference, so $\Delta Y_t = Y_t - Y_{t-1}$.

Definition (Lagged value)

The value of Y in the previous period (relative to the current period, t) is called its **first lagged value** and is denoted by Y_{t-1}.

Its j-**th lagged value** (or, more simply, its j-th lag) is its value j periods ago, which is $Y_t - Y_{t-j}$. Similarly, Y_{t+1} denotes the value of Y one period into the future.

Lags, First Differences, Logarithms, and Growth Rates III

Economic time series are often analysed after computing their logarithms or the changes in their logarithms:

- many economic series exhibit growth that is approximately exponential; that is, over the long run, the series tends to grow by a certain percentage per year on average.

$\rightarrow$ the logarithm of the series grows approximately linearly.

- the standard deviation of many economic time series is approximately proportional to its level.

$\rightarrow$ the standard deviation of the logarithm of the series is approximately constant

- Changes in the transformed series ($\Delta \log(Y_t)$) are proportional. changes in the original series ($\Delta Y_t / Y_{t-1}$).

Lags, First Differences, Logarithms, and Growth Rates IV

In fact, the change of the logarithm of a variable is approximately equal to the proportional change of that variable:

$$\log(X + a) - \log(X) \approx \frac{a}{X}$$

for $a/X \approx 0$ (small).

In the time-series case,

$$\begin{aligned}\Delta \log(Y_t) &= \log(Y_t) - \log(Y_{t-1}) \\ &= \log(Y_{t-1} + \Delta Y_t) - \log(Y_{t-1}) \approx \frac{\Delta Y_t}{Y_{t-1}}\end{aligned}$$

Note that $100\frac{\Delta Y_t}{Y_{t-1}}$ is the percentage change the series Y: this can be approximated with $100\Delta \log(Y_t)$.

Lags, First Differences, Logarithms, and Growth Rates

- The first lag of a time series Y_t is Y_{t-1}; its j^{th} lag is Y_{t-j}.
- The first difference of a series, ΔY_t, is its change between periods $t-1$ and t; that is, $\Delta Y_t = Y_t - Y_{t-1}$.
- The first difference of the logarithm of Y_t is $\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$.
- The percentage change of a time series Y_t between periods $t-1$ and t is approximately $100\Delta \ln(Y_t)$, where the approximation is most accurate when the percentage change is small.

Autocorrelation I

In time series data, the value of Y in one period typically is correlated with its value in the next period.

Definition (Autocorrelation)

The correlation of a series with its own lagged values is called **autocorrelation** or **serial correlation**

- The first autocorrelation (or **autocorrelation coefficient**) is the correlation between Y_t and Y_{t-1}, i.e. the correlation of Y between any two adjacent dates.
- The second autocorrelation is the correlation between Yt and Y_{t-2}.
- The j-th autocorrelation is the correlation between Y_t and Y_{t-j}.
- Similarly, the j-th **autocovariance** is the covariance between Yt and Y_{t-j}.

Autocorrelation II

Autocorrelation (Serial Correlation) and Autocovariance

The j^{th} autocovariance of a series Y_t is the covariance between Y_t and its j^{th} lag, Y_{t-j}, and the j^{th} autocorrelation coefficient is the correlation between Y_t and Y_{t-j}. That is,

$$j^{\text{th}} \text{ autocovariance} = \text{cov}(Y_t, Y_{t-j}) \tag{15.2}$$

$$j^{\text{th}} \text{ autocorrelation} = \rho_j = \text{corr}(Y_t, Y_{t-j}) = \frac{\text{cov}(Y_t, Y_{t-j})}{\sqrt{\text{var}(Y_t)\text{var}(Y_{t-j})}}. \tag{15.3}$$

The j^{th} autocorrelation coefficient is sometimes called the j^{th} serial correlation coefficient.

Autocorrelation III

The j-th population autocovariances and autocorrelations can be estimated by the j-th sample autocovariances and autocorrelations:

$$\hat{\mathbb{C}\text{ov}}(Y_t, Y_{t-j}) = \frac{1}{T}\sum_{t=j+1}^{T}(Y_t - \bar{Y}_{j+1:T})(Y_j - \bar{Y}_{1:T-j})$$

$$\hat{\rho}_j = \hat{\mathbb{C}\text{orr}}(Y_t, Y_{t-j}) = \frac{\hat{\mathbb{C}\text{ov}}(Y_t, Y_{t-j})}{\hat{\mathbb{V}}(Y_t)}$$

where $\bar{Y}_{j+1:T}$ denotes the sample average of Y_t compute using the observations $t = j+1, \ldots, T$, thus $\bar{Y}_{j+1:T} = \frac{1}{T-j}\sum_{t=j+1}^{T} Y_t$ and $\hat{\mathbb{V}}(Y_t)$ is the sample variance of Y_t.

Note that in $\hat{\rho}_j$ we are implicitly assuming that $\mathbb{V}(Y_t) = \mathbb{V}(Y_{t-j})$ (stationarity).

Stationarity I

Main idea: Time series forecasts use data on the past to forecast the future. Doing so presumes that the future is similar to the past in the sense that the correlations, and more generally the distributions, of the data in the future will be like they were in the past. If the future differs fundamentally from the past, then historical relationships might not be reliable guides to the future.

In the context of regression with time series data this is formalized by the concept of **stationarity.**

$\rightarrow$ Under the assumption of stationarity, regression models estimated using past data can be used to forecast future values.

Stationarity II

Stationarity

A time series Y_t is *stationary* if its probability distribution does not change over time—that is, if the joint distribution of $(Y_{s+1}, Y_{s+2}, \ldots, Y_{s+T})$ does not depend on s, regardless of the value of T; otherwise, Y_t is said to be *nonstationary*. A pair of time series, X_t and Y_t, are said to be *jointly stationary* if the joint distribution of $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2}, \ldots, X_{s+T}, Y_{s+T})$ does not depend on s, regardless of the value of T. Stationarity requires the future to be like the past, at least in a probabilistic sense.

Stationarity III

Stationarity can fail to hold for multiple reasons, in which case the time series is said to be **nonstationary**, e.g.:

- the (unconditional) mean of a time series might have a trend.
- the population regression coefficients change at a given point in time.

...for now, we assume that the time series is stationary.

Forecasts and Forecast Errors I

Consider the problem of forecasting the value of a time series variable Y in the period immediately following the end of the available data - that is, of forecasting Y_{T+1} using data through date T (**one-step ahead forecast**).

Let $\hat{Y}_{T+1|T}$ denote a candidate one-step ahead forecast of Y_{T+1}:

- subscript $T+1|T$ indicates that the forecast is of the value of Y at time $T+1$, made using data through time T
- the "hat" indicates that the forecast is based on an estimated model
- thus $\hat{Y}_{T+1|T}$ is specific for the estimated model

Forecasts and Forecast Errors II

Definition (Forecast)

A **forecast** refers to a prediction made for a future date that is not in the data set used to make the forecast - that is, the forecast is for an out-of-sample future observation.

Definition

The forecast error is the mistake made by the forecast, which is realized only after time has elapsed and the actual value of Y_{T+1} is observed.

$$\text{Forecast error} = Y_{T+1} - \hat{Y}_{T+1|T}$$

The Mean Squared Forecast Error I

Because forecast errors are inevitable, the aim of the forecaster is not to eliminate errors but rather to make them as small as possible - that is, to make the forecasts as accurate as possible.

$\rightarrow$ we need a quantitative measure of what it means for a forecast error to be small.

Definition

The **mean squared forecast error** (MSFE), which is the expected value of the square of the forecast error:

$$MSFE = \mathbb{E}\left[\left(Y_{T+1} - \hat{Y}_{T+1|T}\right)^2\right]$$

Note that in MSFE large errors receive a much greater penalty than small ones (due to the square).

The Mean Squared Forecast Error II

The **root mean squared forecast error** (RMSFE) is the square root of the MSFE.

- it has the same units as Y (easy to interpret)
- if the forecast is unbiased, forecast errors have mean zero mean $\rightarrow$ the RMSFE is the standard deviation of the out-of-sample forecast made using a given model
  (recall that, for a r.v. X, $\mathbb{V}(X) = \mathbb{E}(X_t^2) - \mathbb{E}(X_t)^2$).

The Mean Squared Forecast Error III

The MSFE incorporates **two sources of randomness [PROOF]**:

- The first is the randomness of the future value, Y_{T+1}
- The second is the randomness arising from **estimating** a forecasting model
- (Actually, there is also the model specification involved)

Note that: The MSFE is an unknown population expectation, so to use it in practice it must be estimated using data.

The Mean Squared Forecast Error IV

From the perspective of the MSFE, the best-possible prediction (minimizes the MSFE) is the conditional mean given the in-sample observations

$$\hat{Y}_{T+1} = \mathbb{E}(Y_{T+1}|Y_1, \ldots, Y_T) \tag{1}$$

also called the **oracle forecast**.

- The oracle forecast is infeasible because the conditional mean is unknown in practice
- The oracle forecast minimizes the MSFE: is a conceptual benchmark against which to assess an actual forecast

The First-Order Autoregressive Model I

An autoregression expresses the conditional mean of a time series variable Y_t as a linear function of its own lagged values.

first-order autoregression uses only one lag of Y in this conditional expectation:

$$\mathbb{E}(Y_{t+1}|Y_1, \ldots, Y_t) = \beta_0 + \beta_1 Y_t$$

The first-order autoregression AR(1) model can be written in the familiar form of a regression model as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t \tag{2}$$

where u_t is the error term.

The first-order autoregression in eq.(2) is a population autoregression with two unknown coefficients with β_0 and β_1.

The First-Order Autoregressive Model II

- The unknown population coefficients β_0 and β_1 in eq.(2) can be estimated by OLS.
- Equation (2) has the form of a standard regression model, with X being the first lag of Y.
- to estimate β_0 and β_1, one must create a new variable, the first lag of Y, and then use that as the regressor.

Forests and forecasts errors with the AR(1) model I

If the population coefficients in eq.(2) were known, then the one-step ahead forecast of Y_{T+1}, made using data through date T, would be $\beta_0 + \beta_1 Y_T$.

Although β_0 and β_1 are unknown, the forecaster can use their OLS estimates instead:

$$\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated using historical data through time T. The forecast error is $Y_{T+1|T} - \hat{Y}_{T+1|T}$.

The p-th -Order Autoregressive Model I

The AR(1) model uses Y_{t-1} to forecast Y_t, but doing so ignores potentially useful information in the more distant past.

The p^{th}-order autoregressive model, AR(p), represents Y_t as a linear function of p of its lagged values; that is, in the AR(p) model, the regressors are $Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p}$, plus an intercept.

The p-th -Order Autoregressive Model II

Autoregressions

The p^{th}-order autoregressive [AR(p)] model represents the conditional expectation of Y_t as a linear function of p of its lagged values:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + u_t, \qquad (15.12)$$

where $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots) = 0$. The number of lags p is called the order, or the lag length, of the autoregression.

Properties of the forecast and error term in the AR(p) model I

The assumption that the conditional expectation of u_t is 0 given past values of Y_t, i.e. $\mathbb{E}(u_t|Y_{t-1}, Y_{t-2}, ...)$ has two important implications.

1. The best forecast of $Y_{T+1|T}$ based on its entire history depends on only the most recent p past values.

$$\begin{aligned} Y_{T+1|T} &= \mathbb{E}(Y_{T+1}|Y_T, Y_{T-1}, \ldots) \\ &= \beta_0 + \beta_1 Y_T + \beta_2 Y_{T-1} + \cdots + \beta_p Y_{T-p+1} \end{aligned}$$

2. The errors u_t are serially uncorrelated **[PROOF]**

Autoregressive Distributed Lag Model I

Economic theory often suggests other variables that could help forecast a variable of interest. These other variables, or predictors, can be added to an autoregression to produce a time series regression model with multiple predictors.

Autoregressive distributed lag (ADL) model:

i *autoregressive* because lagged values of the dependent variable are included as regressors, as in an autoregression.

ii *distributed lag because* the regression also includes multiple lags of an additional predictor.

Autoregressive Distributed Lag Model II

The Autoregressive Distributed Lag Model

The autoregressive distributed lag model with p lags of Y_t and q lags of X_t, denoted ADL(p, q), is

$$\begin{aligned} Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \\ + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \cdots + \delta_q X_{t-q} + u_t, \end{aligned} \tag{15.17}$$

where $\beta_0, \beta_1, \ldots, \beta_p, \delta_1, \ldots, \delta_q$, are unknown coefficients and u_t is the error term with $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots) = 0$.

- The assumption that the errors in the ADL model have a conditional mean of 0 given all past values of Y implies that no additional lags of either Y or X belong in the ADL model.
- The lag lengths p and q are the true lag lengths, and the coefficients on additional lags are all 0.

The Least Squares Assumptions for Forecasting with Multiple Predictors I

The Least Squares Assumptions for Forecasting with Time Series Data

The general time series regression model allows for k additional predictors, where q_1 lags of the first predictor are included, q_2 lags of the second predictor are included, and so forth:

$$\begin{aligned} Y_t = {} & \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \\ & + \delta_{11} X_{1t-1} + \delta_{12} X_{1t-2} + \cdots + \delta_{1q_1} X_{1t-q_1} \\ & + \cdots + \delta_{k1} X_{kt-1} + \delta_{k2} X_{kt-2} + \cdots + \delta_{kq_k} X_{kt-q_k} + u_t, \end{aligned} \tag{15.18}$$

where

1. $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots, X_{1t-1}, X_{1t-2}, \ldots, X_{kt-1}, X_{kt-2}, \ldots) = 0$;
2. (a) The random variables $(Y_t, X_{1t}, \ldots, X_{kt})$ have a stationary distribution, and (b) $(Y_t, X_{1t}, \ldots, X_{kt})$ and $(Y_{t-j}, X_{1t-j}, \ldots, X_{kt-j})$ become independent as j gets large;
3. Large outliers are unlikely: $X_{1t}, \ldots, X_{kt}$ and Y_t have nonzero, finite fourth moments; and
4. There is no perfect multicollinearity.

The Least Squares Assumptions for Forecasting with Multiple Predictors II

A1 u_t has conditional mean 0 given the history of all the regressors. This extends the assumption used in the AR and implies that the oracle forecast of Y_t using all past values of Y and the X's is given by the regression in Equation (15.18).

A2 The second assumption for time series regression replaces the i.i.d. assumption in cross-sectional data

$$(X_{1i}, \ldots, X_{ki}, Y_i), \ i = 1, \ldots, n \qquad \sim \text{i.i.d}$$

by a more appropriate one with two parts (see later).

A3 same as for cross-sectional data

A4 same as for cross-sectional data

The Least Squares Assumptions for Forecasting with Multiple Predictors III

A2 - part a: that the data are drawn from a stationary distribution:

- This is a time series version of the *identically distributed* part of the i.i.d. assumption: the joint distribution of the variables, including lags, not change over time
- The assumption of stationarity implies that the conditional mean for the data used to estimate the model is also the conditional mean for the out-of-sample observation of interest
- Thus the assumption of stationarity is also an assumption about external validity

The Least Squares Assumptions for Forecasting with Multiple Predictors IV

A2 - part b: the random variables become independently distributed when the amount of time separating them becomes large:

- this is a time series version of the *independent* part of the i.i.d. assumption: the variables are independently distributed when they are separated by long periods of time
- This assumption is sometimes referred to as **weak dependence** (ensures that in large samples there is sufficient randomness in the data for the law of large numbers and the central limit theorem to hold)

Estimation of the MSFE I

The MSFE, is an expected value that depends on the distribution of Y and on the forecasting model. Because it is an expectation, its value is not known and must be estimated from the data.

Replacing the expectation with an average over out-of-sample observations is unfeasible as the out-of-sample data are not observed.

Estimation of the MSFE II

Three methods:

M1: focuses only on future uncertainty and ignores uncertainty associated with estimation of the regression coefficients.

M2: incorporates future uncertainty and estimation error, under the assumption of stationarity so that the conditional expectation estimated by the model applies to the out-of-sample forecast.

M3: incorporates uncertainty and estimation error and in addition allows for the possibility that the conditional expectation might change over the course of the sample.

Estimation of the MSFE III

M1 and M2 are derived based on an expression for the MSFE derived from eq.(1) and the assumption of stationarity.

For the AR(p) this is,

$$MSFE = \sigma_u^2 + \mathbb{V}\left(\hat{\beta}_0 + \hat{\beta}_1 \hat{\mathrm{Y}}_T + \cdots + \hat{\beta}_p \hat{\mathrm{Y}}_{T-p+1}\right) \tag{3}$$

and it immediately extends to models with additional predictors (ADL).

- The first term in eq.(3) is the variance of Y_{T+1} around its conditional mean. This is the variance of the oracle forecast.
- The second term in eq.(3) arises because the coefficients of the autoregression are unknown and must be estimated.

[PROOF] eq.(3) for the AR(1) model.

Estimation of the MSFE IV

Method 1: Estimating the MSFE by the standard error of the regression (SER)

- Because the variance of the OLS estimator is proportional to $1/T$, the second term in eq.(3) is proportional to $1/T$.
- (the first term is fixed, is the variance of u_t)
- if the number of observations T is large relative to the number of autoregressive lags p, the contribution of the second term is small relative to the first term

That is, if T is large relative to p:

$$\hat{MSFE}_{SER} = s^2_{\hat{u}} = \frac{SSR}{T - P - 1},$$

with the statistics $s^2_{\hat{u}}$ being the (usual) square of the standard error of the regression.

Estimation of the MSFE V

Method 2: Estimating the MSFE by the final prediction error
If T is not large relative to p, the sampling error of the estimated autoregression coefficients can be sufficiently large that the second term in Equation (15.19) should not be ignored, as in Method 1.

The **final prediction error (FPE)** is an estimate of the MSFE that incorporates both terms in eq.(3), under the additional assumption that the errors are homoskedastic.

- with homoskedastic errors

$$\mathbb{V}\left(\hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_T + \cdots + \hat{\beta}_p \hat{Y}_{T-p+1}\right) \approx \sigma_u^2 \frac{p+1}{T}$$

- By substituting the above in eq.(3):

$$MSFE = \sigma_u^2 + \sigma_u^2 \frac{p+1}{T} = \sigma_u^2 \left[1 + \frac{p+1}{T}\right]$$

Estimation of the MSFE VI

The FPE uses this expression along with the $s^2_{\hat{u}}$ estimator for σ^2_u,

$$\begin{aligned} \hat{MSFE}_{FPE} &= s^2_{\hat{u}} \frac{T+p+1}{T} = \frac{SSR}{T-p-1} \frac{T+p+1}{T} \\ &= \frac{T+p+1}{T-p-1} \frac{SSR}{T} \end{aligned} \tag{4}$$

The FPE estimator improves upon the squared SER in Method 1 by adjusting for the sampling uncertainty in estimating the autoregression coefficients.

Estimation of the MSFE VII

Method 3: Estimating the MSFE by pseudo out-of-sample forecasting This method uses the data to simulate out-of-sample forecasting and estimate the MSFE.

Divide the data in two parts

1. An initial sample of $T - P$ observations: used to estimate the model
2. A reserved sample of (the last) P observations:
   - 2.0 (i) Use the estimated model to forecast the *first* observation in the reserved sample.
   - 2.1 (i) the estimation sample is augmented by the *first* observation in the reserved sample, (ii) the model is reestimated, and (iii) used to forecast the second observation in the reserved sample.
   - 2.2 (i) the estimation sample is augmented by the *first two* observation in the reserved sample, (ii) the model is reestimated, and (iii) used to forecast the *third* observation in the reserved sample.
   - ... Repeat

Estimation of the MSFE VIII

2.j (i) the estimation sample is augmented by the *first* j-*th* observation in the reserved sample, (ii) the model is reestimated, and (iii) used to forecast the $(j+1)$-*th* observation in the reserved sample.

... Repeat

2.P-1 (i) the estimation sample is augmented by the *first* $P-1$ observations in the reserved sample, (ii) the model is reestimated, and (iii) used to forecast the P-*th* observation in the reserved sample.

This procedure produces P forecasts and thus P forecast errors, used to estimate the MSFE.

Estimation of the MSFE IX

This method of estimating a model on a subsample of the data and then using that model to forecast on a reserved sample is called **pseudo out-of-sample forecasting**:

- *out-of-sample* because the observations being forecasted were not used for model estimation.
- *pseudo* because the reserved data are not truly out-of-sample observations.

Estimation of the MSFE X

Pseudo Out-of-Sample Forecasts

Pseudo out-of-sample forecasts are computed using the following steps:

1. Choose a number of observations, P, for which you will generate pseudo out-of-sample forecasts; for example, P might be 10% or 20% of the sample size. Let $s = T - P$.
2. Estimate the forecasting regression using the estimation sample—that is, using observations $t = 1, \ldots, s$.
3. Compute the forecast for the first period beyond this shortened sample, $s + 1$; call this $\widetilde{Y}_{s+1|s}$.
4. Compute the forecast error, $\widetilde{u}_{s+1} = Y_{s+1} - \widetilde{Y}_{s+1|s}$.
5. Repeat steps 2 through 4 for the remaining periods, $s = T - P + 1$ to $T - 1$ (reestimate the regression for each period). The pseudo out-of-sample forecasts are $\widetilde{Y}_{s+1|s}, s = T - P, \ldots, T - 1$, and the pseudo out-of-sample forecast errors are $\widetilde{u}_{s+1}, s = T - P, \ldots, T - 1$.

Estimation of the MSFE XI

Denoting with $\tilde{u}_s$, $s = T - P + 1, \ldots, T$ the pseudo out-of-sample forecast errors, the pseudo out-of-sample (POOS) estimate of the MSFE is

$$\hat{MSFE}_{POOS} = \frac{1}{P} \sum_{s=T-P+1}^{P} \tilde{u}_s^2$$

Estimation of the MSFE XII

Pros:

- does not rely on the assumption of stationarity: conditional mean might differ between the estimation and the reserved samples.

$\rightarrow$ e.g. if the AR coefficients are not the same in the two samples, the POOS error shall not have mean 0: this bias can be captured by $\hat{MSFE}_{POOS}$ (but not by the other two methods)

Cons:

1. If Y is stationarity, $\hat{MSFE}_{POOS}$ has higher variance than the estimators from Method 1 and 2.
2. It is more difficult to compute.
3. Requires choosing P (10%-20% of T, is a good trade-off between the precision of the coefficient estimates and the number of observations available for estimating the MSFE).

Forecast intervals I

One measure of the uncertainty of a forecast is its root mean squared forecast error (RMSFE). Under the additional assumption that the errors u_t are normally distributed, the estimates of the RMSFE can be used to construct a forecast interval.

Definition

A forecast interval is like a confidence interval except that it pertains to a forecast. For example, a 95% forecast interval is an interval that contains the future value of the variable being forecasted in 95% of repeated applications.

Forecast intervals II

CI The usual confidence interval form

$$CI = \text{estimator} \pm 1.96SE$$

is *justified by the central limit theorem* and therefore holds for a wide range of distributions of the error term.

FI Note that the forecast error includes the future value of the error u_{T+1}: for computing a Forecast Interval (FI),

- Either the distribution of the error is somehow estimated
- Or assumptions on the distribution of the error are required.

$\rightarrow$ In practice it is convenient to assume that u_{T+1} is normally distributed

Forecast intervals III

If u_{T+1} is normally distributed, a 95% forecast interval is obtained by

$$\hat{Y}_{T+1|T} \pm 1.96 \hat{RMSE}_{\text{Method}}$$

with $\hat{RMSE}_{\text{Method}}$ being estimated with one of the three discussed methods.

Proof: The forecast error is the sum of (i) u_{T+1} and (ii) the errors reflecting the estimation error of the regression coefficients (normally distributed in large samples). Thus, if u_{T+1} is normally distributed the forecast error is normal as well and has variance equal to the MSFE.

Note: so far, all the discussion assumed that σ_u^2 is homoskedastic. If not, one needs to develop a model of the heteroskedasticity so that the term σ_u^2 in eq.(3) can be estimated given the most recent values of Y (and X)

Proof: MSFE decomposition I

To show: the MSFE involves two sources of randomness.

A very simple model: Y_{T+1} is forecasted as its historical mean value μ_Y.

- μ_Y is unknown: $\hat{\mu}_Y$ denotes its estimate
- thus $\hat{Y}_{T+1} = \hat{\mu}_Y$
- the forecast error is $Y_{T+1} - \hat{Y}_{T+1|T} = Y_{T+1} - \hat{\mu}_Y$
- work on the MSFE, with Y_{T+1} begin uncorrelated with μ_Y:

$$
\begin{aligned}
MSFE &= \mathbb{E}\left[(Y_{T+1} - \hat{\mu}_Y)^2\right] \\
&= \mathbb{E}\left[(Y_{T+1} - \hat{\mu}_Y + \mu_Y - \mu_Y)^2\right] \\
&= \mathbb{E}\left[(Y_{T+1} - \mu_Y)^2\right] + \mathbb{E}\left[(\hat{\mu}_Y - \mu_Y)^2\right] + 0
\end{aligned}
$$

Proof: MSFE decomposition II

1. $\mathbb{E}\left[(Y_{T+1} - \mu_y)^2\right]$ is the error the forecaster would make if the population mean were known:

   $\rightarrow$ This term captures the random future (out-of-sample) fluctuations in Y_{T+1} around the population mean

2. The second term in this expression is the additional error made because the population mean is unknown.

Determining the Order of an Autoregression I

How many lags should you include in a time series regression?

- if the order of an estimated autoregression is too low, you will omit potentially valuable information contained in the more distant lagged values
- if it is too high, you will be estimating more coefficients than necessary, which in turn introduces additional estimation error into your forecasts.

There are three approaches: (1) F-statistics, (2) AIC, (3) BIC.

Thr F-Statistics approach I

One approach to choosing p is to start with a model with many lags and to perform hypothesis tests on the final lag.

Example: start by estimating an AR(6) and test whether the coefficient on the sixth lag is significant at the 5% level; if not, drop it and estimate an AR(5), test the coefficient on the fifth lag, and so forth...

$\rightarrow$ The method tends to produce large models: if the true value of p is five, this method will estimate p to be six 5% of the time. In fact, 5% test using the t-statistic will incorrectly reject this null hypothesis 5% of the time just by chance.
$\rightarrow$ If the number of models being compared is not small, then this F-statistic method is not practical to use.

The BIC approach I

estimate p by minimizing an information the **Bayes information criterion** (BIC), also called the Schwarz information criterion (SIC),

$$BIC(p) = \ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{\ln(T)}{T} \tag{5}$$

where $SSR(p)$ is the sum of squared residuals of the estimated AR(p).

The BIC estimator of p, $\hat{p}$, is the value that minimizes BIC(p) among the possible choices $p = 0, 1, \ldots, p_{max}$, where p_{max} is the largest value of p considered and $p = 0$ corresponds to the model that contains only an intercept.

The BIC approach II

Interpretation:

- Because the coefficients are estimated via OLS, the first term in eq.(5) decreases when you add a lag
- The second term increases when you add a lag and thus provides a penalty for including another lag.
- The BIC trades off these two forces: the number of lags that minimizes the BIC $\hat{p}$ is a **consistent** estimator of the true lag length.

The AIC approach I

Another information criterion is the **Akaike information criterion** (AIC):

$$AIC(p) = \ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{2}{T} \tag{6}$$

The difference w.r.t. eq.(5) is that for eq.(6) the second terms is smaller.
$\rightarrow$ Thus a smaller decrease in the SSR is needed in the AIC to justify including another lag.

- In large samples, it corresponds to choosing p that minimizes the $MSFE_{\text{FPE}}$ **[PROOF]**.
- However, even in large samples, the AIC estimator of p is **not** consistent.
- In large samples the AIC **overestimates** p with non-zero probability

The AIC approach II

Notes:

- For the AIC and BIC to decide between competing regressions with different numbers of lags, those regressions must be estimated using the **same** observations
- Both the AIC and the BIC are widely used in practice (the BIC is more conservative on p)
- If you are concerned that the BIC might yield a model with too few lags, the AIC provides a reasonable alternative

The AIC approach III

- When there are multiple predictors (k besides the lagged values of Y) (ADL), this approach is computationally demanding (one should consider all combinations of lag parameters on $y, q_1, \ldots, q_k$). A convenient shortcut is to require all the regressors to have the same number of lags

$$p = q_1 = \ldots q_k$$

so that only $p_{max} + 1$ models need to be compared, corresponding to $p = 0, 1, \ldots, p_{max}$.

Proof: AIC I

To show: in large samples, minimizing the AIC corresponds to minimizing the $MSFE_{\text{FPE}}$.

Start from eq.(4):

$$\hat{MSFE}_{FPE} = \frac{T+p+1}{T-p-1}\frac{SSR}{T}$$

this is, in general, a positive quantity (certainly non-negative, and not-zero except in very particular cases), then we can consider its log:

Proof: AIC II

$$
\begin{aligned}
\log \hat{MSFE}_{FPE} &= \log\left[\frac{T+p+1}{T-p-1}\frac{SSR}{T}\right] \\
&= \log\left[\frac{T+p+1}{T-p-1}\right] + \log\left[\frac{SSR}{T}\right] \\
&= \log\left[\frac{T+p+1\,\frac{1}{T}}{T-p-1\,\frac{1}{T}}\right] + \log\left[\frac{SSR}{T}\right] \\
&= \log\left[\frac{\frac{T+p+1}{T}}{\frac{T-p-1}{T}}\right] + \log\left[\frac{SSR}{T}\right] \\
&= \log\left[\frac{1+\frac{p+1}{T}}{1-\frac{p+1}{T}}\right] + \log\left[\frac{SSR}{T}\right] \\
&= \log\left[1+\frac{p+1}{T}\right] - \log\left[1-\frac{p+1}{T}\right] + \log\left[\frac{SSR}{T}\right]
\end{aligned}
$$

Now recall that, as a general result, for $x \approx 0$ (small values of x),

Proof: AIC III

$$\log(1+x) \approx x$$

and thus $\log(1-x) = \log(1+(-x)) \approx -x$. Therefore, with $x = (p+1)/T$,

$$\underbrace{\log\left[1+\frac{p+1}{T}\right]}_{\approx \frac{p+1}{T}}, \qquad \underbrace{\log\left[1-\frac{p+1}{T}\right]}_{-\frac{p+1}{T}}$$

and the approximation is good, indeed if the sample size T is large relative to the number of parameters, $(p+1)/T$ is almost zero. Therefore we

Proof: AIC IV

conclude that,

$$\begin{aligned}\log \hat{MSFE}_{FPE} &= \frac{p+1}{T} - \left(-\frac{p+1}{T}\right) + \log\left[\frac{SSR}{T}\right] \\ &= 2\frac{p+1}{T} + \log\left[\frac{SSR}{T}\right] \\ &\overset{(6)}{=} AIC(p) \end{aligned} \tag{7}$$

Thus we conclude that, in large samples

1a $AIC(p) \approx \log \hat{MSFE}_{FPE}$ (eq. (7)).

1b minimizing the function $AIC(p)$ corresponds to minimizing $\log \hat{MSFE}_{FPE}$,

$\rightarrow$ the optimal lag choice of p^* determined with the AIC criterion is also the lag choice that minimizes $\log \hat{MSFE}_{FPE}$.

2 Since the log is a monotonic function, p^* also minimizes $\hat{MSFE}_{FPE}$.

Proof: AIC V

$\rightarrow$ In fact, the value p^* that minimizes the log of a function of p is also the value that minimizes the original function. In our case, as p^* minimizes $\log \hat{MSFE}_{FPE}$, it also minimizes $\hat{MSFE}_{FPE}$ [a basic fact that you should recall from your first course in math].

Nonstationarity

So far, it was assumed that the dependent variable and the regressors are stationary. If this is not the case, then conventional hypothesis tests, confidence intervals, and forecasts can be unreliable.

We examine two types of nonstationarity that are frequently encountered in economic time series:

1. Trends
2. Breaks

What Is a Trend? I

A **trend** is a persistent long-term movement of a variable over time: a time series variable fluctuates around its trend.

There are two types of trends in time series data:

1. A **deterministic trend** is a nonrandom function of time
2. A **stochastic trend** is random and varies over time

What Is a Trend? II

"We [SW] think it is more appropriate to model economic time series as having stochastic rather than deterministic trends. It is hard to reconcile the predictability implied by a deterministic trend with the complications and surprises faced year after year by workers, businesses, and governments."

$\rightarrow$ treatment of trends in economic time series focuses on stochastic rather than deterministic trends, and when we refer to "trends" in time series data, we mean stochastic trends unless we explicitly say otherwise

The random walk model of a trend I

Definition (Random walk)

A time series Y_t is said to follow a **random walk** if the change in $Y_t$is i.i.d., that is, if

$$Y_t = Y_{t-1} + u_t$$

where u_t is i.i.d.

Definition (Martingale)

A time series Y_t such that

$$Y_t = Y_{t-1} + u_t$$

where u_t has conditional mean zero, $\mathbb{E}(u_t|Y_{t-1}, Y_{t-2}, \ldots) = 0$, is called **martingale**. Analogously, a time series for which $\mathbb{E}(\Delta Y_t|Y_{t-1}, Y_{t-2}, \ldots) = 0$ is a martingale.

The random walk model of a trend II

- for a random walk, the value of tomorrow is today's plus an unpredictable change
- The definition of martingale is more relaxed than that of the random walk: a random walk is a martingale, but not the other way around.
- In time series, by "random walk" we actually mean a martingale.
- The conditional mean of Y_t based on data through time $t-1$ is Y_{t-1},

$$\mathbb{E}(u_t|Y_{t-1}, Y_{t-2}, \ldots) = 0 \quad \rightarrow \quad \mathbb{E}(Y_t|Y_{t-1}, Y_{t-2}, \ldots) = Y_{t-1},$$

  then the best forecast of tomorrow's value is its value today.
- If Y_t follows a random walk, its variance increases over time **[PROOF]**: a random walk is **nonstationary**.

The random walk model of a trend III

Definition (Random walk with drift)

A time series Y_t is a random walk with drift if

$$Y_t = \beta_0 + Y_{t-1} + u_t \tag{8}$$

where $\mathbb{E}(u_t|Y_{t-1}, Y_{t-2}, \ldots) = 0$ and β_0 is a constant.

Comments:

- If β_0 is positive, then Y_t increases on average
- the best forecast of the series tomorrow is the value of the series today plus the drift β_0
- The random walk model (with drift, as appropriate) is the primary model for trends.

Stochastic trends, autoregressive models, and a unit root I

The random walk model is a special case of the AR(1) model, where $\beta_1 = 1$. Thus if Y_t follows an AR(1) with $\beta_1 = 1$ is non-stationary.

However (it can be proved that) if,

$$|\beta_1| < 1$$

and u_t is stationary, Y_t is stationary. I.e., then the joint distribution of Yt and its lags does not depend on t.

Stochastic trends, autoregressive models, and a unit root II

For an AR(p), the corresponding condition is that the roots of the polynomial

$$1 - \beta_1 z - \beta_2 z^2 - \cdots - \beta_p z^p$$

must be all greater than 1 in absolute value.

An AR(p) that has a root equal to 1, is said to have a **unit root**.

- If Y_t has a unit root, then it contains a stochastic trend.
- If Y_t is stationary, it does not have a unit root and it does not contain a stochastic trend.
- Thus, "stochastic trend" and "unit root" are used *interchangeably*

Problems Caused by Stochastic Trends I

If a regressor has a stochastic trend, we identify two major issues:

1. inferences made using the OLS estimator of the autoregressive coefficient can be misleading
2. two series that are independent but have stochastic trends will, with high probability, misleadingly appear to be related

Downward bias and nonnormal distributions of the OLS estimator and t-statistic I

Problem: If a regressor has a stochastic trend, then its usual OLS t-statistic can have a nonnormal distribution under the null hypothesis, even in large samples, and the estimate of the autoregressive coefficient is biased toward 0.

- This nonnormal distribution means that conventional confidence intervals are inapplicable
- Hypothesis testing cannot be conducted as usual
- For and AR(1) with coefficient 1, the OLS estimate will tend to be less than 1: oracle forecasts are now all biased
- This downward bias is however not detectable: as the distribution of the t-statistics is not normal standard inference is not effective.

Spurious regression I

Stochastic trends can lead two time series to appear related when they are not, a problem called **spurious regression**

`https://www.tylervigen.com/spurious-correlations`

- The series do correlate, because both have a stochastic trend. However, there is no compelling economic or political reason to think that the trends in these two series are related: the regressions are spurious.
- Any causal interpretation on a possibly estimated coefficient is obviously a nonsense.

Spurious regression II

Note:
One special case in which certain regression-based methods are reliable is when the trend component of the two series is the same—that is, when the series contain a common stochastic trend; in such a case, the series are said to be **cointegrated**.

The Dickey–Fuller test in the AR(1) model I

The starting point for detecting a trend in a time series is inspecting its time series plot.

$\rightarrow$ If the series looks like it might have a trend, the hypothesis that it has a stochastic trend can be tested using a **Dickey–Fuller test**.

The Dickey–Fuller test in the AR(1) model II

The random walk in eq.(8) is a special case of the AR(1) model with $\beta_1 = 1$. Thus, when Y_t follows an AR(1), the hypothesis that Y_t has a stochastic trend corresponds to

$$H_0 : \beta_1 = 1 \quad \text{vs} \quad H_1 : \beta_1 < 1 \tag{9}$$

where

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

The null hypothesis in eq.(9) is that the AR(1) has a unit root, and the one-sided alternative is that it is stationary.

The Dickey–Fuller test in the AR(1) model III

This test is most easily implemented by estimating a modified version of eq.(9), obtained by subtracting Y_{t-1} from both sides:

$$\Delta Y_t = Y_t - Y_{t-1} = \beta_0 + (\beta_1 - 1) Y_{t-1} + u_t = \beta_0 + \delta Y_{t-1} + u_t$$

where $\delta = \beta_1 - 1$.
Now, testing $\beta = 1$ ($\beta < 1$) is equivalent to testing $\delta = \beta_1 - 1 < 0$ ($\delta < 0$). Thus the initial hypotheses in eq.(9) read:

$$H_0 : \delta = 0 \quad \text{vs} \quad H_1 : \delta < 0 \tag{10}$$

where

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + u_t$$

$\rightarrow$ The OLS t-statistic testing $\delta = 0$ in eq.(10) is called the **Dickey–Fuller statistic**.

The Dickey–Fuller test in the AR(1) model IV

Notes:

- Under the null hypothesis of a unit root, the usual nonrobust standard errors produce a t-statistic that is, robust to heteroskedasticity: The Dickey-Fuller statistic is computed using nonrobust standard errors.
- Under the null hypothesis of a unit root, the Dickey–Fuller statistic does not have a normal distribution, even in large samples.
- Because its distribution is nonnormal, a different set of critical values is required.
- Critical values are tabulated. The critical values substantially larger (more negative) than the one-sided critical values of -1.28 (at the 10% level) and -1.64 (at the 5% level) from the standard normal distribution.
- Because the alternative hypothesis of stationarity implies that $\delta < 0$ in eq.(10), the ADF test is one-sided.

The Dickey–Fuller test in the AR(p) model I

The **Augmented Dickey-Fuller test**.
The extension of the Dickey–Fuller test to the AR(p) model entails including $p-1$ lags of ΔY_t as additional regressors, eq.(10) becomes:

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta Y_{t-2} + \cdots + \gamma_{p-1} \Delta Y_{t-p+1} + u_t$$

Under the null hypothesis that $\delta = 0$, Y_t has a stochastic trend; under the alternative hypothesis that $\delta < 0$, Y_t is stationary.

The Dickey–Fuller test in the AR(p) model II

Notes:

- The t-statistic testing the hypothesis that $\delta = 0$ in eq.(10) is called the **augmented Dickey–Fuller (ADF) statistic**.
- the lag length p is unknown, but it can be estimated using an information criterion in regressions eq.(10), or various values of p.
- Studies suggest that it is better to have too many lags than too few: it is recommended to use the AIC instead of the BIC.

Avoiding the Problems Caused by Stochastic Trends I

The most reliable way to handle a trend in a series is to transform the series so that it does not have the trend.

If Y_t follows a random walk (with or without drift), then

$$\Delta Y_t = (\beta_0) + u_t$$

is stationary.

$\rightarrow$ first differences eliminate random walk trends in a series.

Avoiding the Problems Caused by Stochastic Trends II

Notes:

- The failure to reject the null hypothesis does not necessarily mean that the null hypothesis is true; rather, it simply means that you have insufficient evidence to conclude that it is false.
- Failure to reject the null hypothesis of a unit root using the ADF statistic does not mean that the series actually has a unit root.
- Failure to reject the null hypothesis: it still can be reasonable to approximate the true autoregressive root as equalling 1 and therefore to use differences of the series rather than its levels.

What Is a Break? I

A second type of nonstationarity arises when the population regression function changes over the course of the sample.

If such changes, or **breaks**, occur, then a regression model that neglects those changes can provide a misleading basis for inference and forecasting.

Breaks can arise either:

1. from a discrete change in the population regression coefficients at a distinct date (e.g., a major change in macroeconomic policy)
2. from a gradual evolution of the coefficients over a longer period of time (e.g., gradual changes in macroeconomic policy, or ongoing changes in the structure of the economy).

$\rightarrow$ Thus there are two kinds of tests.

What Is a Break? II

If a break occurs in the population regression function during the sample,

- then the OLS regression estimates over the full sample will estimate a relationship that holds on average, in the sense that the estimate combines the two different periods.
- the "average" regression function can be quite different from the true regression function at the endpoints of the sample,
- this leads to poor forecasts.

Testing for a break at a known date I

In some applications, you might suspect that there is **a break at a known date**. If the date of the hypothesized break in the coefficients is known, then the null hypothesis of no break can be tested using a binary variable interaction regression.

Consider an ADL(1,1):
Let τ denote the hypothetized break date, and let $D_t(\tau)$ be a binary variable that equals 0 before the break date and 1 after:

$$D_t(\tau) = \begin{cases} 0 & t \leq \tau \\ 1 & t > \tau \end{cases}$$

Testing for a break at a known date II

Then the (testing) regression including the binary break indicator and all interaction terms is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \gamma_0 D_t(\tau) + \gamma_1 [D_t(\tau) Y_{t-1}] + \gamma_2 [D_t(\tau) X_{t-1}] + u_t \tag{11}$$

Testing in the above equation $H_0 : \gamma_0 = \gamma_1 = \gamma_2 = 0$ with the F-statistics is known as the **Chow test** for a break at a known break date. I.e., if there is a break in *any* of the parameters at least one of the γ's will be non-zero, resulting in a large F-statistics.

Testing for a break at a known date III

The meaning of the above eq. (11) is the following:

Before the break, all the $D_t(\tau)$'s are zero, and the regression equation is:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} \tag{12}$$

After the break, $D_t(\tau)$'s are one, and the regression equation is:

$$Y_t = (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1) Y_{t-1} + (\beta_2 + \gamma_2) X_{t-1} + u_t \tag{13}$$

The eq. (12) and eq. (13) are jointly written together under a unique regression model that can be estimated as eq. (11). If there is a break in the coefficients at time τ then they do change, and, e.g., β_1 shifts to $\beta_1 + \gamma_1$. Testing that any of the γ's is statistically different from zero unveils if indeed there was such a change after τ.

Testing for a break at a known date IV

Interpretation:

- If there is not a break, then the population regression function is the same over both parts of the sample, so the terms involving the break binary variable $D_t(\tau)$ do not enter in eq. (11).
- Under the null hypothesis of no break, $\gamma_0 = \gamma_1 = \gamma_2 = 0$, testable with the F-statistics.
- Under the alternative, a population parameter is different before and after the break date τ: at least one of the γ's is non-zero.

Testing for a break at a known date V

Extensions:

- If there are multiple predictors or more: include binary variable interaction variables for all the regressors and test the hypothesis that all the coefficients on terms involving $D_t(\tau)$ are 0.
- Test a subset of coefficients: including only the binary variable interactions for the subset of regressors of interest. E.g., for testing that there is a break involving the intercept only, modify (11) as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \gamma_0 D_t(\tau) + u_t$$

(here you may use a t-test for checking $H_0 : \gamma_0 = 0$, as a particular case).

Testing for a break at an unknown date I

Often the date of a possible break is unknown or known only within a range.

Suppose, for example, you suspect that a break occurred sometime between two dates, τ_0 and τ_1.

$\rightarrow$ Quandt likelihood ratio (QLR) test

Main idea:
testing with the Chow test for breaks at all possible dates τ between τ_0 and τ_1 and then using the largest of the resulting F-statistics to test for a break at an unknown date.

Testing for a break at an unknown date II

- The QLR statistics is the largest of many F-statistics: its distribution is not the same as an individual F-statistic
- The critical values for the QLR statistic must be obtained from a special distribution
- This distribution depends:
  - on the number of restrictions being tested, q, the number of coefficients (incl. the intercept) that are being allowed to break or change under the alternative hypothesis
  - on τ_0/T and τ_1/T, the endpoints of of the subsample over which the F-statistics are computed, expressed as a fraction of the total sample size

Testing for a break at an unknown date III

About the subsample:

- The subsample endpoints, τ_0 and τ_1, cannot be too close to the beginning or the end of the sample (for the large-sample approximation to the distribution of the QLR statistic to be good)
- A common choice is to use 15% trimming: set $\tau_0 = 0.15\,T$ and $\tau_1 = 0.85\,T$
- With 15% trimming, the F-statistic is computed for break dates in the central 70% of the sample.

$\rightarrow$ See table 15.5

Testing for a break at an unknown date IV

By examining F-statistics at many possible break dates, the QLR statistic has many opportunities to reject the null hypothesis, leading to QLR critical values that are larger than the individual F-statistic critical values.

A If there is a **discrete break** at a date within the range tested, the date at which the constituent F-statistic is at its maximum, $\hat{\tau}$, is an estimate of the break date τ.

B The QLR statistic also rejects the null hypothesis with high probability in large samples when there are multiple discrete breaks or when the break comes in the form of a slow evolution of the regression function. This means that the QLR statistic detects forms of instability **other than a single discrete break**

Testing for a break at an unknown date V

$\rightarrow$ If the QLR rejects the null hypothesis, it can mean that

1. there is a single discrete break.
2. there are multiple discrete breaks.
3. or that there is slow evolution of the regression function.

Testing for a break at an unknown date VI

The QLR test can be used to test for a break in only **some** of the regression coefficients by using interactions between the date binary indicators and only the variables in question, and then computing the largest of the resulting F-statistics: same critical values as the general case.

Testing for a break at an unknown date VII

The QLR Test for Coefficient Stability

Let $F(\tau)$ denote the F-statistic testing the hypothesis of a break in the regression coefficients at date τ; in the regression in Equation (15.35), for example, this is the F-statistic testing the null hypothesis that $\gamma_0 = \gamma_1 = \gamma_2 = 0$. The QLR (or sup-Wald) test statistic is the largest of the F-statistics in the range $\tau_0 \leq \tau \leq \tau_1$:

$$\text{QLR} = \max[F(\tau_0), F(\tau_0 + 1), \ldots, F(\tau_1)]. \qquad (15.36)$$

1. Like the F-statistic, the QLR statistic can be used to test for a break in all or just some of the regression coefficients.
2. In large samples, the distribution of the QLR statistic under the null hypothesis depends on the number of restrictions being tested, q, and on the endpoints τ_0 and τ_1 as a fraction of T. Critical values are given in Table 15.5 for 15% trimming ($\tau_0 = 0.15T$ and $\tau_1 = 0.85T$, rounded to the nearest integer).
3. The QLR test can detect a single discrete break, multiple discrete breaks, and/or slow evolution of the regression function.
4. If there is a distinct break in the regression function, the date at which the largest Chow statistic occurs is an estimator of the break date.

Detecting Breaks Using Pseudo Out-of-Sample Forecasts I

The ultimate test of a forecasting model is its out-of-sample performance -that is, its forecasting performance in "real time", after the model has been estimated.

Pseudo-out-of-sample forecasting simulates the real-time performance of a forecasting model and can be used to detect breaks **near the end of the sample**.

Detecting Breaks Using Pseudo Out-of-Sample Forecasts II

Two approaches:

- Use a time series plot of the in-sample predicted values, the pseudo-out-of-sample forecasts, and the actual values of the series. A visible deterioration of the forecasts in the pseudo-out-of-sample period is a red flag warning of a possible breakdown of the forecasting model.
- Compare $M\hat{S}FE_{POOS}$ with $M\hat{S}FE_{FPE}$ computes on the first $T - P$ observations (same sample as $M\hat{S}FE_{POOS}$). If the series is stationary the two measures should be close: a value of $M\hat{S}FE_{POOS}$ that is much higher than $M\hat{S}FE_{FPE}$ suggests some violation of stationarity, *possibly* a breakdown of the forecasting equation.

Avoiding the Problems Caused by Breaks I

- If a distinct break occurs at **a specific date**, that break will be detected with high probability by the QLR statistic, and the break date can be estimated.

  $\rightarrow$ The regression function can then be reestimated using a binary variable indicating the two subsamples associated with this break and including interactions with the other regressors as appropriate
  - If **all** the coefficients break, then this simplifies to reestimating the regression using the post-break data
  - then subsequent inference on the regression coefficients can proceed as usual
  - forecasts can be produced using the regression function estimated using the post-break model
- If the break is **not distinct** but rather arises from a slow, ongoing change in the parameters, the remedy is more difficult...