preamble

Time Series Analysis - Tutorial 2.a

Roberto Baragona
Sapienza University of Rome, Italy

International Conference on Pattern Recognition
and Machine Intelligence (PReMI–2011)
Moscow, Russia, June 28 - June 30, 2011
WWW.PREMI11.HSE.RU

Univariate time series in practice

- A data set whose elements are ordered with time is called a time series and in this case ’variable’ is used as synonymous if no ambiguity arises.
- The univariate time series analysis is useful to check the properties of the time series in a set one at a time so that relationships between them may be given their proper meaning.
- As an example we will examine the personal income in the United States from 1-st quarter 1954 to 4-th quarter 1994, values increase with a rate which is approximately constant.

Autocorrelation

- The possible existence of a correlation between the observations of a time series is a peculiar property that in general other data sets don't own.
- For example,many macroeconomic time series change slowly with time so that an observation at time t (the current variable) tends tobe similar to that at time $t-1$ (the lagged variable).
- This property is called 'autocorrelation', it is much less evident in the growth rate time series.

The linear correlation coefficient

- Let X and Y denote a pair of statistical variables for which n observations are available.
- Let r denote the linear correlation coefficient between $x = (x_1, x_2, \ldots, x_n)'$ and $y = (y_1, y_2, \ldots, y_n)'$.
- The defining formula is as follows:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Properties of the linear correlation coefficient

- r is always included in the interval $(-1, 1)$.
- If $r > 0$ then there is positive correlation between X and Y, while if $r < 0$ there is negative correlation and $r = 0$ means that X and Y are uncorrelated.
- $r = 1$ implies perfect positive correlation, while $r = -1$ perfect negative correlation.
- The correlation between X and Y is the same as the correlation between Y and X.
- The correlation of a statistical variable with itself is always equal to 1.

Some comments on the linear correlation coefficient

- Correlation does not imply causation though causality often produces a non zero correlation coefficient.
- The value of r alone is inadequate to explain the relationships between two variables and knowledge about the application field helps to appreciate the estimated value.
- r is a 'nonsense correlation' between X and Y when the correlation is due to a third variable Z which is correlated with both X and Y.
- The correlation between X and Y reflects the linear relationships only.
- While independent variables are uncorrelated, in general uncorrelated variables are not independent.

The autocorrelation function for the personal income in the United States from 1954 to 1994 (quarterly data)

The autocorrelation function

- The estimated autocorrelation function is often used as a first approach to the analysis of a time series.
- If n observations of X_t are available, the estimated autocorrelation $\hat{r}(1)$ between X_t and X_{t-1} i.e. the correlation of a time series with its lagged version is the linear correlation coefficient between the two vectors $(x_1, x_2, \ldots, x_{n-1})'$ and $(x_2, x_3, \ldots, x_n)'$.
- Likewise $\hat{r}(h)$ is the 'order h' autocorrelation between X_t and X_{t-h}, i.e. the linear correlation coefficient between the two vectors $(x_1, x_2, \ldots, x_{n-h})'$ and $(x_{h+1}, x_{h+2}, \ldots, x_n)'$.
- As a practical rule we may select a maximum value for h, $h_{\max}$ say, and take the data from $t = h + 1$ on into account.

The personal income in the United States

- Let us consider the quarterly time series data of the personal income in the United States from 1954, first quarter, to 1994, fourth quarter (plot in Fig. 10).
- The original variable is million dollars but it is useful to transform the data by taking the their natural logarithm.
- The time series ∇x_t represents the rate of growth of the original variable between times $t-1$ and t (plot in Fig. 11).
- The personal income increases approximately by 1% per quarter but the variance of this growth rate varies considerably with time.
- Table 1 reports for comparison the autocorrelation functions of Y and ∇Y

The personal income in the United States

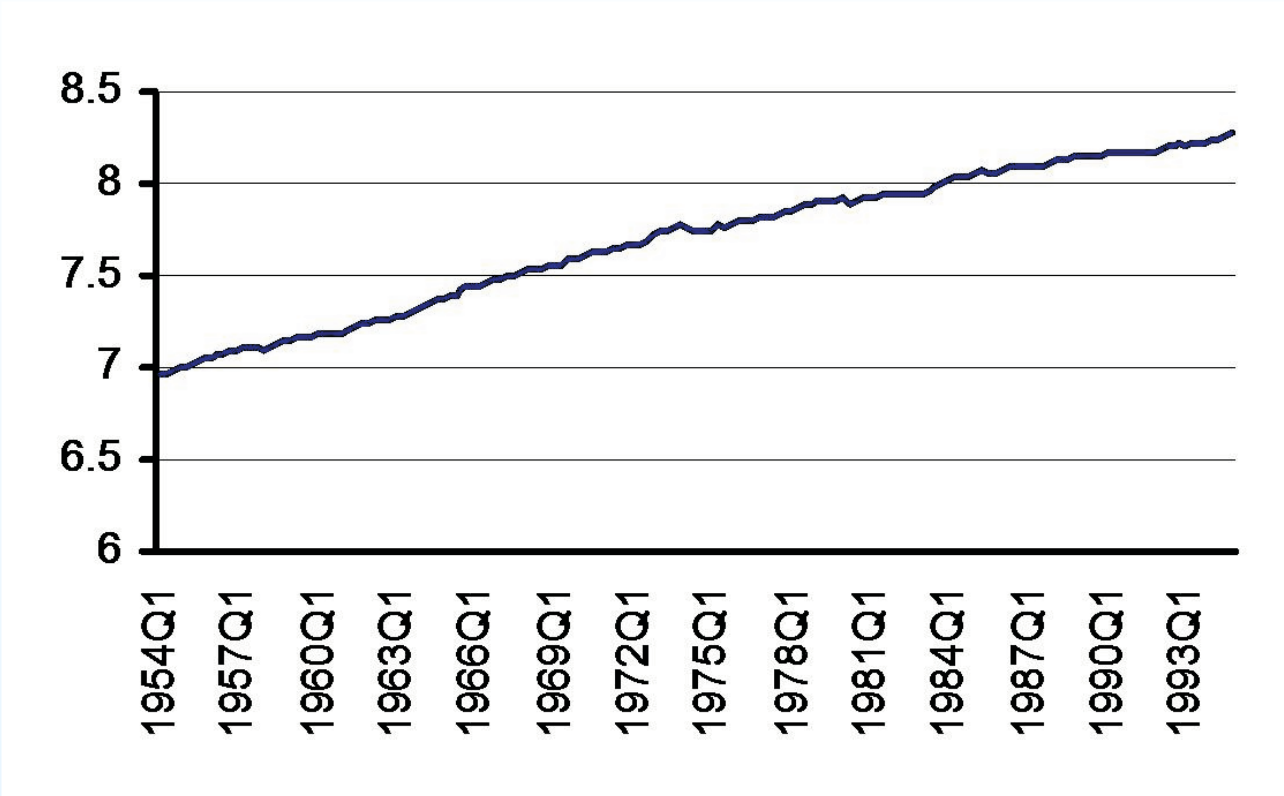

Figure 1: Natural logarithm of the personal income time series from first quarter 1954 to fourth quarter 1994

Variation of the personal income in the United States

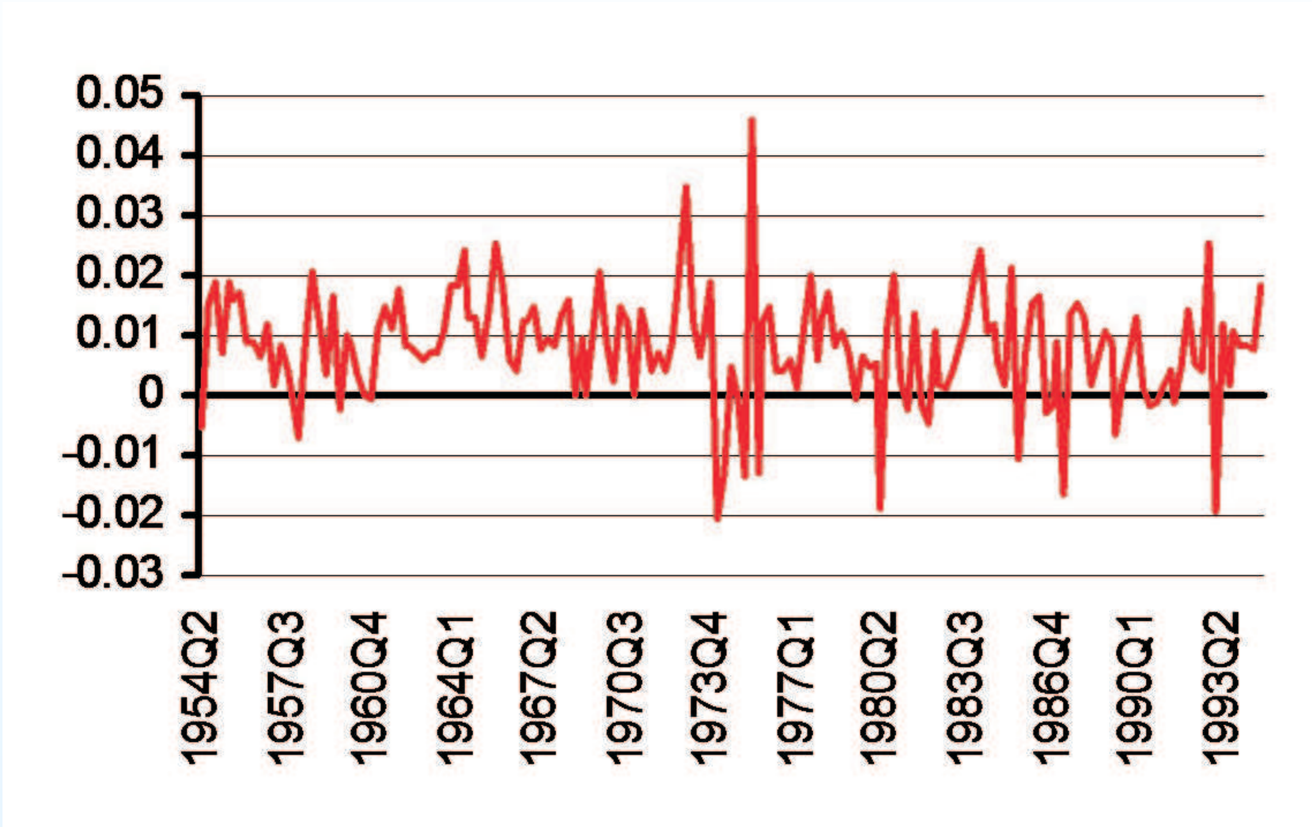

Figure 2: Variation of natural logarithm of the personal income time series from first quarter 1954 to fourth quarter 1994

The autocorrelation function of Y and ∇Y

Table 1: Autocorrelation function at lags $1-12$ of the personal income Y and the variations in the level $\nabla Y = Y_t - Y_{t-1}$

lag τ	personal income $r(\tau)$	variation in the level $r(\tau)$
1	0.9997	-0.01
2	0.9993	0.0121
3	0.9990	0.1341
4	0.9986	0.0082
5	0.9983	-0.1562
6	0.9980	0.0611
7	0.9978	-0.035
8	0.9975	-0.0655
9	0.9974	0.0745
10	0.9972	0.1488
11	0.9969	0.033
12	0.9966	0.0363

Some comments on the personal income autocorrelation results

- Levels Y are highly autocorrelated up to three years ($\tau = 12$) while ∇Y shows little or no autocorrelations.
- The personal income in past quarters is a good predictor of the current personal income level while the variations in the levels are essentially useless for prediction.
- The persistence of high autocorrelations for large lags suggests that Y has 'long memory' while the variations in the levels don't own such property.
- We may think of Y as a non stationary time series and of ∇Y as a stationary one.
- The two autocorrelation functions (see Fig. 14 and Fig. 15 as well) are typical of the two cases of stationarity and non stationarity.

Autocorrelation function of the personal income in the United States

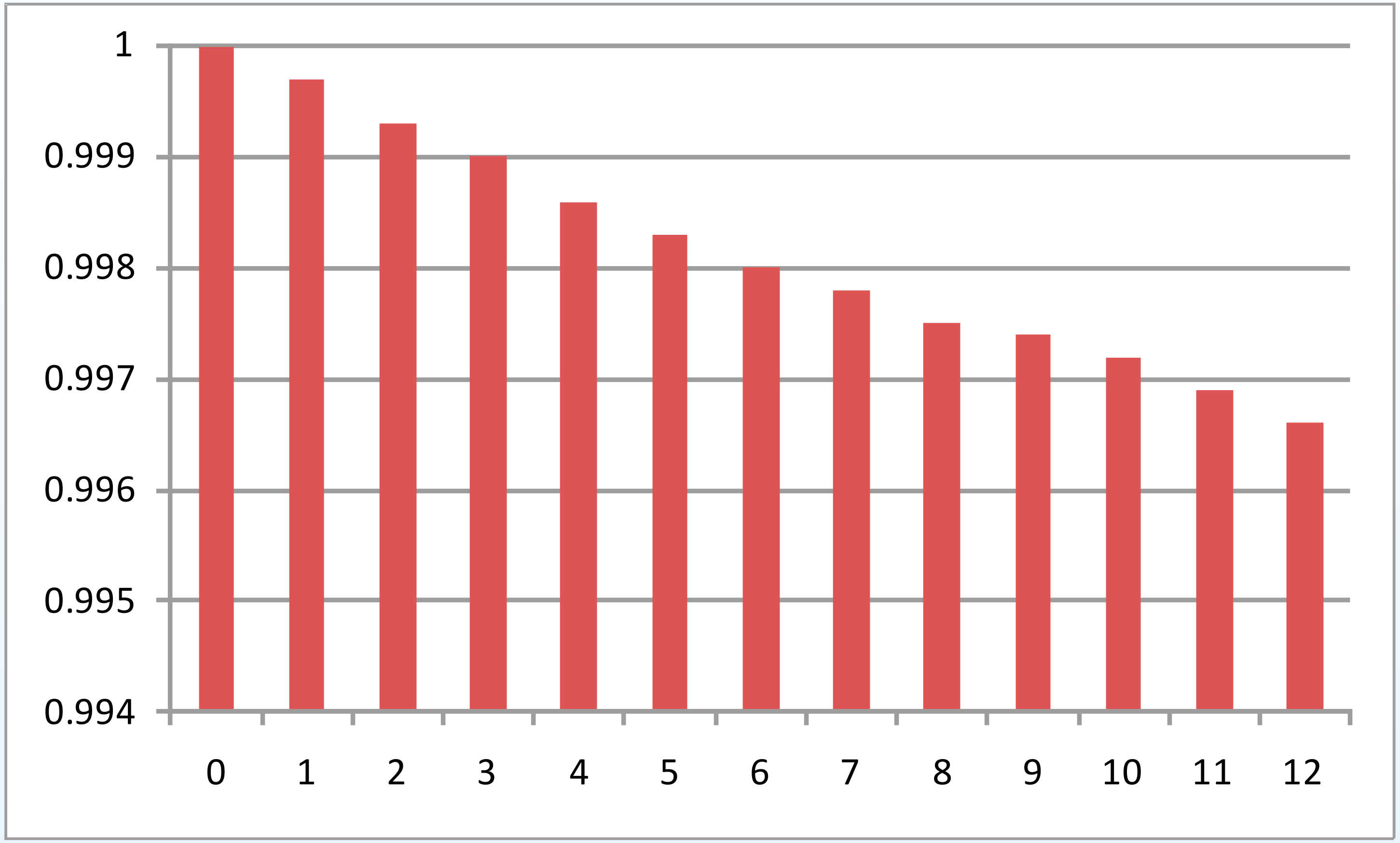

Figure 3: Autocorrelation functions of Y for lags $\tau = 1, \ldots, 12$.

Autocorrelation of the personal income increase in the United States

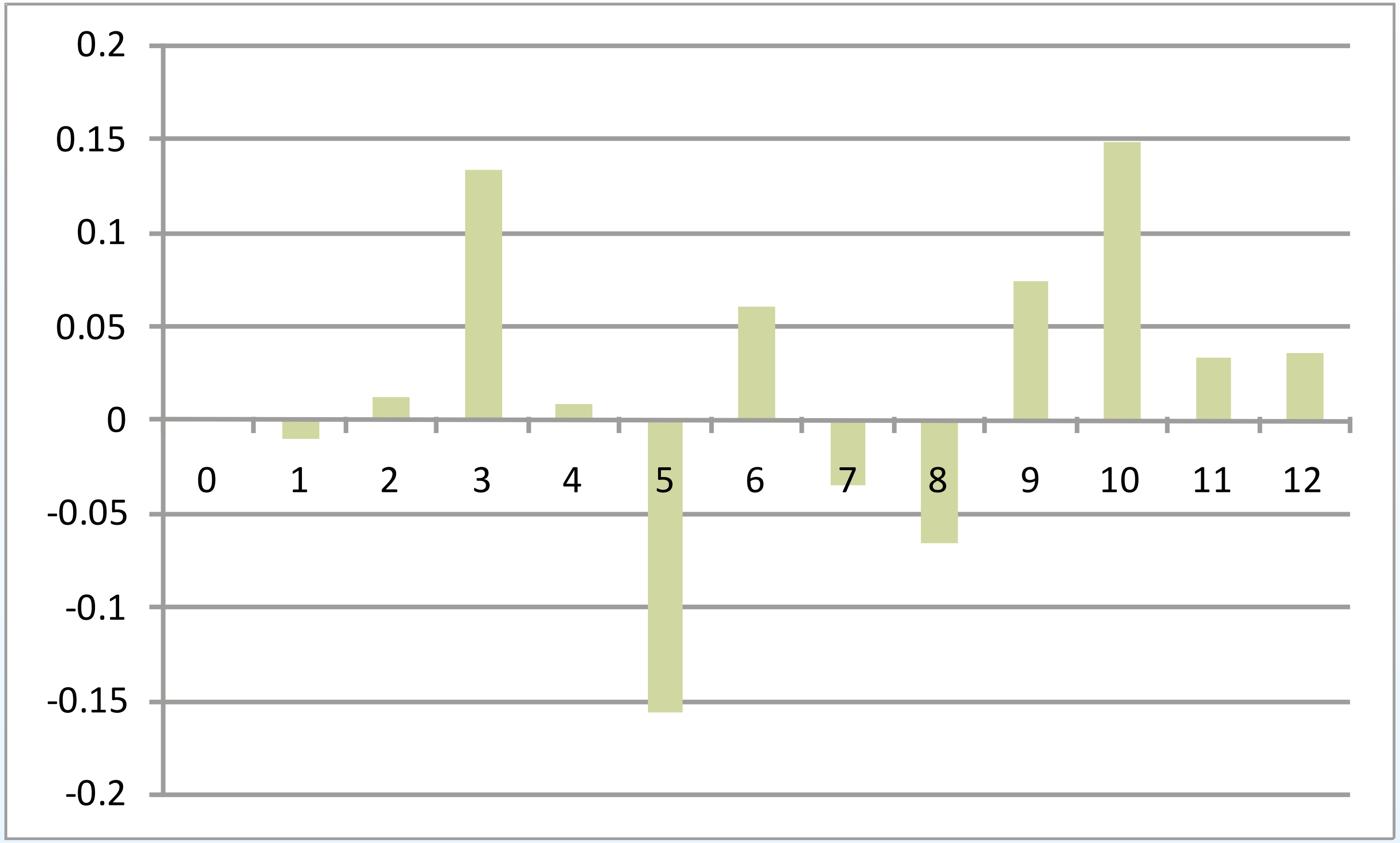

Figure 4: Autocorrelation functions of ∇Y for lags $\tau = 1, \ldots, 12$.

The autoregressive model

Autoregressive model for univariate time series

- The autoregressive (AR) model may be viewed as a regression model where the input variables are the lagged values of the response variable.
- The simplest AR model is the AR model of order 1, or AR(1), where the response variable is y_t and the input variable is y_{t-1}:

$$y_t = \alpha + \phi y_{t-1} + \epsilon_t.$$

The behavior of a AR(1) depends on the coefficient ϕ

- Simulated time series from an AR(1) with $\phi = 0$, $\phi = 1$ and $\phi = 0.8$, and fixed α, are very different each other:
  - $\phi = 0$, casual fluctuations similar to the 'level variations' which is often observed for differenced time series $\nabla y_t = y_t - y_{t-1}$.
  - $\phi = 1$, a trend as in most 'level' time series.
  - $\phi = 0.8$, a behavior which is half way between the previous two cases.

- We may distinguish however the case $\phi = 1$, typical of non stationary time series, from the other cases ($\phi = 0$ and $\phi = 0.8$) which characterize stationary time series.

Stationarity of an AR(1) time series

- A time series generated by an AR(1) is stationary if $|\phi| < 1$
  - is stationary if $|\phi| < 1$,
  - non stationary if $|\phi| = 1$.
- The case $|\phi| > 1$ in general is not to be considered as it implies the ’explosive’ behavior of the time series, this is of little relevance in practical applications.
- Often non stationary is used to mean every situation where the stationarity hypothesis is untenable.

A non stationarity case: the unit root

- If a time series is generated by an AR(1) with $\phi = 1$ then it is said to have an 'unit root'.
- If y_t has an unit root, then the autocorrelations are close to 1 and are slowly decreasing as the lag increases.
- If y_t has an unit root, then y_t is a 'long memory' random process.
- If y_t has an unit root a trend is present in the time series behavior and this is specially apparent when α is non zero.
- If y_t has an unit root then ∇y_t is a stationary random process and y_t is called 'difference stationary'.

A difference time series model

- Let y_t be generated from an AR(1) model.
- Then ∇y_t follows the model, where $\rho = \phi - 1$,

$$\nabla y_t = \alpha + \rho y_{t-1} + \epsilon_t.$$

- For this model it is possible to test the hypothesis $\rho = 0$ that the time series y_t has an unit root.
- Notice that the stationarity condition $-1 < \phi < 1$ is equivalent to $-2 < \rho < 0$.
- If $\rho = 0$, or, which it is the same, $\phi = 1$, and, in addition, $\alpha = 0$, then the AR(1) model is said to be a 'random walk' $y_t = y_{t-1} + \epsilon_t$.

Extension of the AR(1) to the AR(p) model

- The AR(p) model is obtained by assuming as input variables beside y_{t-1} the lagged variables $y_{t-2}, \ldots, y_{t-p}$.

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t.$$

- The properties of the AR(p) model though similar are more general than those of the AR(1).
- An useful property consists in that the AR(p), $p > 1$, may be used for modeling not only a trend but the (pseudo) periodic fluctuations of the time series as well.

The unit roots in the AR(p) model

- The AR(p) model may be written in the equivalent form:

$$\nabla y_t = \alpha + \rho y_{t-1} + \gamma_1 \nabla y_{t-1} + \ldots + \gamma_{p-1} \nabla y_{t-p+1} + \epsilon_t.$$

- The coefficients $\rho, \gamma_1, \ldots, \gamma_{p-1}$ may be calculated from $\phi_1, \ldots, \phi_p$.
- If $\rho = 0$ the time series y_t has a unit root while it is stationary if $-2 < \rho < 0$.
- If $\rho = 0$ the time series y_t is difference stationary, i.e. ∇y_t follows an AR(p) model.

Unit roots and the stochastic trend

- We could see that the time series data have to be ’differenced’ to obtain a stationary time series in the presence of a unit root.
- This simple statement of fact is of importance in practice because time series may be jointly modeled only if they are stationary.
  - An important exception is the case of ’cointegration’, i.e. a linear combination of non stationary time series exists which is found to be stationary.
- ...

Unit roots and the stochastic trend (cont'd)

- Unit root time series contain a stochastic trend.
  - To see this, consider for example an AR(1) with $\phi = 1$ and assume that the time series is known at an initial time $t = 1$ in the past.
  - Then we may write recursively:

$$
\begin{aligned}
&y_t = y_{t-1} + \epsilon_t, \\
&y_{t-1} = y_{t-2} + \epsilon_{t-1} \Rightarrow y_t = y_{t-2} + \epsilon_t + \epsilon_{t-1}, \\
&y_{t-2} = y_{t-3} + \epsilon_{t-2} \Rightarrow y_t = y_{t-3} + \epsilon_t + \epsilon_{t-1} + \epsilon_{t-2}, \\
&\vdots \\
&y_t = y_1 + \textstyle\sum_{i=0}^{t-2} \epsilon_{t-i}.
\end{aligned}
$$

- The summation $\sum_{i=0}^{t-2} \epsilon_{t-i}$ is said to be the stochastic trend.

Deterministic trend

- The unit root AR models have a trend, but time series with trend exist that are not unit root AR.
- An example of a 'deterministic trend' in an AR model is:

$$y_t = \alpha + \phi y_{t-1} + \delta t + \epsilon_t.$$

- If in a deterministic trend model the random process is modeled by a stationary AR(1) ($|\phi| < 1$), the time series is said to be 'around trend stationary'.

The AR(p) model with deterministic trend

- Often the following model proved useful in many applications:

$$\nabla y_t = \alpha + \rho y_{t-1} + \gamma_1 \nabla y_{t-1} + \ldots + \gamma_{p-1} \nabla y_{t-p+1} + \delta t + \epsilon_t.$$

- We have to check if a unit root is present: it suffices to test the hypothesis $\rho = 0$.
- In general $y_{t-1}, \nabla y_{t-2}, \ldots, \nabla y_{t-p+1}$ result uncorrelated variables unlike $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$.
- For time series with seasonal periodicities trend is a complicated polynomial function of time.

The AR(p) model with deterministic trend and seasonal dummy variables

- If a time series includes a seasonal component with period s the dummy variables (or 'dummies') $D_1, D_2, \ldots, D_{s-1}$ may be defined that equal 1 if there is an observation in this period while equal zero otherwise.

- The observations which are not accompanied by a dummy are intended to refer to the period s.

- All dummies and trend coefficients are to be included in the linear regression and may be estimated by the ordinary least square (OLS) method.

The AR(p) model with deterministic trend and seasonal dummy variables (cont'd)

- If the time series are quarterly data then $s = 4$ and we may write the model

$$\nabla y_t = \alpha + \rho y_{t-1} + \gamma_1 \nabla y_{t-1} + \gamma_2 \nabla y_{t-2} + \gamma_3 \nabla y_{t-3}$$

$$+ \delta_1 D_{1t} + \delta_2 D_{2t} + \delta_3 D_{3t} + \delta t + \epsilon_t .$$

- The usual test for checking if the coefficients are equal to zero apply to the coefficients of the dummies and of time.

Exercises

- Simulate 100 observation from the model $\nabla y_t = \alpha + \rho y_{t-1} + \delta t + \epsilon_t$ where $\{\epsilon_t\}$ is a sequence of independent identically distributed normal random variables with zero mean and variance $\sigma_\epsilon = 7.85 \times 10^{-5}$, for the following values of α, ρ, δ:

  1. $0.0078, -0.95, 1.45 \times 10^{-5}$
  2. $0.0087, -0.2155, 1.9 \times 10^{-5}$
  3. $0.004, -0.1758, 0.0016 \times 10^{-5}$
  4. $0.018, -0.78, 0.0097$

- Perform the unit root test for each of the 4 time series.

...

[FIGURES: Immagine1.eps, Immagine2.eps, Immagine31.eps, Immagine32.eps]
[Data files and electronic sheets: * Example: FIG95.xls, FIG96.xls, FIG97.xls, FIG98.xls * Personal income, time series: income.xls * Personal income, correlations: CORRELATIONS.xls]

References

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., 1994. Time Series Analysis: Forecasting and Control (3rd edition). Prentice Hall, Englewood Cliffs NJ.

Dickey, D. A., Fuller, W. A., 1981. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49(4).

Engle, R. F., Granger, 1987. Co-integration and error correction: representation, estimation and testing. Econometrica 55(2) 251–276.

Reinsel, G. C., 1993. Elements of Multivariate Time Series Analysis. Springer. New York.