- An AR model is a linear model that belongs to the Box and Jenkins (1970) class of models. This model characterizes the time series assuming a linear relationship between observations
- AR model of order p is denoted as AR(p) and can be represented using the following form:

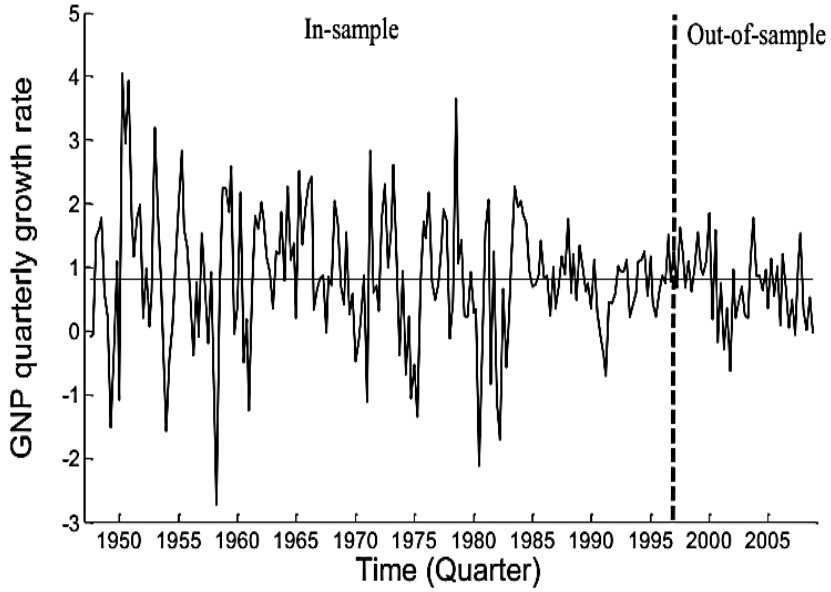
$$y_t = \sum_{i=1}^{p} \alpha_i \, y_{t-i} + \varepsilon_t, where \;\; \varepsilon_t \sim \text{NID}(0, \sigma^2)$$

- where y_t is an observation in the time series at time t
- α_i denotes AR model parameters for $i = 1, 2, \ldots, p$
- ε_t is an Normally and Independently Distributed (NID) process with mean zero and variance σ^2
- p corresponds to the number of lagged variables in this model

Autoregressive models

- The model coefficients (α) can be estimated using ordinary least squares (OLS).
- The model order (p) can be selected using IC, such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), etc.
- $y_t \sim$AR(p): y_t depends linearly on previous p values
- Simplest example: AR(1) process, $y_t = \alpha y_{t-1} + \varepsilon_t$
- If $\alpha = 1$, $y_t = y_{t-1} + \varepsilon_t$. Forecast of future periods is the most recent observations $\hat{y}_{t+k} = y_t$ (persistence/random walk benchmark)

Practical examples using US GNP time series growth rate



US GNP quarterly growth rate. In-sample data: 1947Q2-1996Q4 (199 observations), Out-of-sample data: 1997Q1-2008Q3 (47 observations).

Estimate models on the log of quarterly growth rate (first differences) of the GNP, i.e. $x_t = 100 \log(y_t/y_{t-1})$, where y_t is the actual GNP value at time instant t. This transformation is based on the assumption that raw GNP quarterly observations are generated from a non-stationary process. Multiply the quarterly returns by 100, in order to obtain percentage returns

AR model estimation

- **Model**: $y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i \, y_{t-i} + \varepsilon_t \, , where \;\; \varepsilon_t \sim \text{NID}(0, \sigma^2)$
- Select p using Akaike Information Criterion (AIC), and α using Ordinary Least Squares (OLS)

PARAMETER ESTIMATES FOR THE AR(4) MODEL AND CORRESPONDING STANDARD ERRORS (IN PARENTHESIS) FOR US GNP

Parameter	*Estimate (Standard Error)*
α_0	0.6499 (0.1132)
α_1	0.3120 (0.0719)
α_2	0.1284 (0.0751)
α_3	-0.0832 (0.0751)
α_4	-0.1259 (0.0716)
σ	0.9624
Total Parameters	5

Matlab command: *regress*

Moving average (MA)

- A moving average (MA) model of order q is denoted as MA(q) and can be represented using the following form:

$$y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q}, where \;\; \varepsilon_t \sim \text{NID}(0, \sigma^2)$$

β_i denotes MA model parameters for $i = 1, 2, \ldots, q$

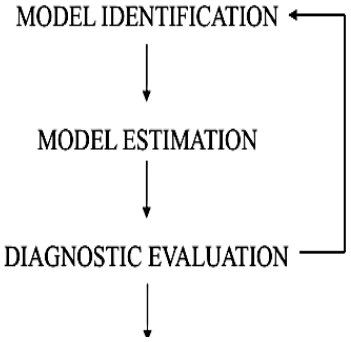
- Basically, ε_t are pure stochastic terms affecting y_t
- Small commodity market receives news about crops. News will have immediate effect and discounted effect as market assimilates importance
- Simplest example: MA(1) process, $y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1}$

Autoregressive moving average (ARMA)

- An ARMA(p, q) model consists of two polynomials, an AR(p) part and a MA(q) part, and is represented as:

$$y_t = \alpha_1 \, y_{t-1} + \alpha_2 \, y_{t-2} + \ldots + \alpha_p \, y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q}$$

- These models rely on there being some regularity in the underlying data generating process to make predictions
- We are assuming presence of an *autocorrelated* structure
- Use of autocorrelation structure relies on series being *stationary*
- Box and Jenkins methodology to select best model



Correlation, ACF, and PACF

- Given two random variables, $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$ and $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$, the (Pearson) correlation coefficient is given as

$$r_{x,y} = \frac{\sum_{t=1}^{N}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{N}(x_t - \bar{x})^2 (y_t - \bar{y})^2}} = \frac{cov(x,y)}{std(x)std(y)} \in [-1,1]$$

- Other useful measures to quantify the similarity between two variables (Spearman's ρ, Kendall's τ, Normalized mutual information)

- The Autocorrelation Function (ACF, ρ_l) measures the correlation between y_t and y_{t-l}, for lag l

$$\rho_l = \frac{\sum_{t=l+1}^{N}(y_t - \bar{y})(y_{t-l} - \bar{y})}{\sum_{t=1}^{N}(y_t - \bar{y})^2} = \frac{cov(y_t, y_{t-l})}{var(y_t)}$$

- The Partial Autocorrelation Function ($PACF_l$) measures the correlation between y_t and y_{t-l}, minus the part explained by intervening lags $\boldsymbol{y}^* = \{y_{t-1}, y_{t-2}, \ldots, y_{t-(l-1)}\}$

$PACF_l = corr(y_t - E^*(y_t|\boldsymbol{y}^*), y_{t-l})$, where $E^*(y_t|\boldsymbol{y}^*)$ is the minimum MSE predictor of y_t by $\boldsymbol{y}^*$