Lecture 2: Univariate Forecasting Models

UCSD, January 18 2017

Allan Timmermann[1]

[1]UC San Diego

Introduction: ARMA models

- When building a forecasting model for an economic or financial variable, the variable's own past time series is often the first thing that comes to mind
  - Many time series are persistent
  - Effect of past and current shocks takes time to evolve
- **Auto Regressive Moving Average (ARMA)** models
  - Work horse of forecast profession since Box and Jenkins (1970)
  - Remain the centerpiece of many applied forecasting courses
  - Used extensively commercially

Why are ARMA models so popular?

1. Minimalist demand on forecaster's **information set**: Need only past history of the variable $\mathcal{I}_T = \{y_1, y_2, ..., y_{T-1}, y_T\}$
   - "Reduced form": No need to derive fully specified model for y
   - By excluding other variables, ARMA forecasts show how useful the past of a time series is for predicting its future
2. **Empirical success**: ARMA forecasts often provide a good 'benchmark' and have proven surprisingly difficult to beat in empirical work
3. ARMA models underpinned by **theoretical arguments**
   - **Wold Representation Theorem**: Covariance stationary processes can be represented as a (possibly infinite order) moving average process
   - ARMA models have certain optimality properties among linear projections of a variable on its own past and past shocks to the series
   - ARMA models are *not* optimal in a global sense - it may be optimal to use nonlinear transformations of past values of the series or to condition on a wider information set ("other variables")

Covariance Stationarity: Definition

A time series, or stochastic process, $\{y_t\}_{t=-\infty}^{\infty}$, is covariance stationary if

- The mean of y_t, $\mu_t = E[y_t]$, is the same for all values of t: $\mu_t = \mu$
  - without loss of generality we set $\mu_t = 0$ for all t [de-meaning]
- The **autocovariance** exists and does not depend on t, but only on the "distance", j, i.e., $E[y_t y_{t-j}] \equiv \gamma(j, t) = \gamma(j)$ for all t
- Autocovariance measures how strong the covariation is between current and past values of a time series
- If y_t is independently distributed over time, then $E[y_t y_{t-j}] = 0$ for all $j \neq 0$

Covariance Stationarity: Interpretation

- **History repeats**: if the series changed fundamentally over time, the past would not be useful for predicting the future of the series. To rule out this situation, we have to assume a certain degree of stability of the series. This is known as covariance stationarity
- Covariance stationarity rules out shifting patterns such as
  - trends in the mean of a series
  - breaks in the mean, variance, or autocovariance of a series
- Covariance stationarity allows us to use historical information to construct a forecasting model and predict the future
- Under covariance stationarity $Cov(y_{2016}, y_{2015}) = Cov(y_{2017}, y_{2016})$. This allows us to predict y_{2017} from y_{2016}

White noise

- Covariance stationary processes can be built from white noise:

Definition

A stochastic process, ε_t, is called white noise if it has zero mean, constant variance, and is serially uncorrelated:

$$\begin{aligned} E[\varepsilon_t] &= 0 \\ Var(\varepsilon_t) &= \sigma^2 \\ E[\varepsilon_t \varepsilon_s] &= 0, \text{ for all } t \neq s \end{aligned}$$

Wold Representation Theorem

Any covariance stationary process can be written as an infinite order MA model, $MA(\infty)$, with coefficients θ_i that are independent of t :

Theorem

Wold's Representation Theorem: Any covariance stationary stochastic process $\{y_t\}$ can be represented as a linear combination of serially uncorrelated lagged white noise terms ε_t and a linearly deterministic component, μ_t:

$$y_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} + \mu_t$$

where $\{\theta_i\}$ are independent of time and $\sum_{j=0}^{\infty} \theta_j^2 < \infty$.

Wold Representation Theorem: Discussion

- Since $E[\varepsilon_t] = 0$, $E[\varepsilon_t^2] = \sigma^2 \geq 0$, $E[\varepsilon_t \varepsilon_s] = 0$, for all $t \neq s$, ε_t is not predictable using linear models of past data
- Practical concern: *MA* order is potentially infinite
  - Since $\sum_{j=0}^{\infty} \theta_j^2 < \infty$, the parameters are likely to die off over time - a finite approximation to the infinite MA process could be appropriate
  - In practice we need to construct ε_t from data (filtering)
- *MA* representation holds apart from a possible deterministic term, μ_t, which is perfectly predictable infinitely far into the future
  - e.g., constant, linear time trend, seasonal pattern, or sinusoid with known periodicity

Estimation of Autocovariances

- Autocovariances and autocorrelations can be estimated from sample data (sample $t = 1, ..., T$):

$$
\begin{aligned}
\widehat{Cov}(Y_t, Y_{t-j}) &= \frac{1}{T-j-1} \sum_{t=j+1}^{T} (y_t - \bar{y})(y_{t-j} - \bar{y}) \\
\hat{\rho}_j &= \frac{\widehat{cov}(y_t, y_{t-j})}{\widehat{var}(y_t)}
\end{aligned}
$$

  where $\bar{y} = (1/T) \sum_{t=1}^{T} y_t$ is the sample mean of Y

- Testing for autocorrelation: $Q-$stat can be used to test for serial correlation of order $1, ..., m$:

$$
Q = T \sum_{j=1}^{m} \hat{\rho}_j^2 \sim \chi_m^2
$$

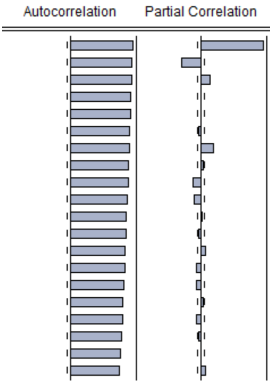
  Small p-values (below 0.05) suggest significant serial correlation

Autocovariances in matlab

- *autocorr*: computes sample autocorrelation
- *parcorr*: computes sample partial autocorrelation
- *lbqtest*: computes Ljung-Box Q-test for residual autocorrelation

Sample autocorrelation for US T-bill rate

Date: 06/10/14 Time: 08:26
Sample: 1927M01 2012M12
Included observations: 1032



	AC	PAC	Q-Stat	Prob
1	0.992	0.992	1018.5	0.000
2	0.979	-0.297	2012.1	0.000
3	0.968	0.148	2982.9	0.000
4	0.957	-0.007	3933.1	0.000
5	0.946	0.002	4863.6	0.000
6	0.935	-0.032	5773.7	0.000
7	0.927	0.189	6668.4	0.000
8	0.921	0.029	7552.2	0.000
9	0.913	-0.133	8422.6	0.000
10	0.903	-0.102	9274.3	0.000
11	0.892	0.018	10106.	0.000
12	0.881	-0.027	10917.	0.000
13	0.871	0.078	11711.	0.000
14	0.860	-0.076	12486.	0.000
15	0.848	-0.076	13239.	0.000
16	0.836	0.039	13973.	0.000
17	0.825	-0.075	14688.	0.000
18	0.812	-0.036	15382.	0.000
19	0.799	-0.006	16054.	0.000
20	0.786	0.076	16706.	0.000

Sample autocorrelation for US stock returns

Date: 06/10/14 Time: 08:33
Sample: 1960M01 2012M12
Included observations: 636

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.048	0.048	1.4615	0.227
		2	-0.036	-0.038	2.2889	0.318
		3	0.039	0.042	3.2405	0.356
		4	0.025	0.020	3.6483	0.456
		5	0.074	0.075	7.1434	0.210
		6	-0.065	-0.074	9.9043	0.129
		7	-0.025	-0.014	10.320	0.171
		8	-0.008	-0.019	10.365	0.240
		9	-0.013	-0.011	10.478	0.313
		10	-0.000	-0.001	10.478	0.400
		11	0.004	0.016	10.491	0.487
		12	0.045	0.044	11.790	0.463
		13	-0.021	-0.026	12.085	0.521
		14	-0.075	-0.072	15.750	0.329
		15	0.008	0.007	15.793	0.396
		16	0.003	-0.006	15.798	0.467
		17	0.026	0.029	16.231	0.508
		18	-0.006	0.003	16.255	0.575
		19	-0.010	0.002	16.327	0.635
		20	-0.023	-0.037	16.690	0.673

Autocorrelations and predictability

- The more strongly autocorrelated a variable is, the easier it is to predict its mean
  - strong serial correlation means the series is slowly mean reverting and so the past is useful for predicting the future
  - strongly serially correlated variables include
    - interest rates (in levels)
    - level of inflation rate (year on year)
  - weakly serially correlated or uncorrelated variables include
    - stock returns
    - *changes* in inflation
    - growth rate in corporate dividends

Lag Operator and Lag Polynomials

- The lag operator, L, when applied to any variable simply lags the variable by one period:

$$\begin{aligned} Ly_t &= y_{t-1} \\ L^p y_t &= y_{t-p} \end{aligned}$$

- Lag polynomials such as $\phi(L)$ take the form

$$\phi(L) = \sum_{i=0}^{p} \phi_i L^i$$

For example, if $p = 2$ and $\phi(L) = 1 - \phi_1 L - \phi_2 L^2$, then

$$\begin{aligned} \phi(L)y_t &= 1 \times y_t - \phi_1 L y_t - \phi_2 L^2 y_t \\ &= y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} \end{aligned}$$

ARMA Models

- Autoregressive models specify y as a function of its own lags
- Moving average models specify y as a weighted average of past shocks (innovations) to the series
- $ARMA(p, q)$ specification for a stationary variable y_t :

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}$$

- In lag polynomial notation

$$\begin{aligned} \phi(L) y_t &= \theta(L) \varepsilon_t \\ \phi(L) &= 1 - \sum_{j=0}^{p} \phi_i L^i \\ \theta(L) &= \sum_{i=0}^{q} \theta_i L^i = 1 + \theta_1 L + ... + \theta_q L^q \end{aligned}$$

AR(1) Model

- $ARMA(1,0)$ or $AR(1)$ model takes the form:

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \varepsilon_t \\ (1 - \phi_1 L) y_t &= \varepsilon_t, \theta(L) = 1 \end{aligned}$$

- By recursive backward substitution,

$$y_t = \phi_1 \underbrace{(\phi_1 y_{t-2} + \varepsilon_{t-1})}_{y_{t-1}} + \varepsilon_t = \phi_1^2 y_{t-2} + \varepsilon_t + \phi_1 \varepsilon_{t-1}$$

- Iterating further backwards, we have, for $h \geq 1$,

$$\begin{aligned} y_t &= \phi_1^h y_{t-h} + \sum_{s=0}^{h-1} \phi_1^s \varepsilon_{t-s} \\ &= \phi_1^h y_{t-h} + \theta(L) \varepsilon_t, \quad \text{where} \\ \theta(L) &: \theta_i = \phi_1^i \quad (\text{for } i = 1, .., h-1) \end{aligned}$$

AR(1) Model

- $AR(1)$ model is equivalent to an $MA(\infty)$ model as long as $\phi_1^h y_{t-h}$ becomes "small" in a mean square sense:

$$E\left[y_t - \sum_{s=0}^{h-1} \phi_1^s \varepsilon_{t-s}\right]^2 = E\left[\phi_1^h y_{t-h}\right]^2 \leq \phi_1^{2h} \gamma_y(0) \to 0$$

  as $h \to \infty$, provided that $\phi_1^{2h} \to 0$, i.e., $|\phi_1| < 1$
- Stationary $AR(1)$ process has an equivalent $MA(\infty)$ representation
- The root of the polynomial $\phi(z) = 1 - \phi_1 L = 0$ is $L^* = 1/\phi_1$, so $|\phi_1| < 1$ means that the root exceeds one. This is a necessary and sufficient condition for stationarity of an $AR(1)$ process
- Stationarity of an $AR(p)$ model requires that all roots of the equation $\phi(z) = 0$ exceed one (fall outside the unit circle)

MA(1) Model

- $ARMA(0,1)$ or $MA(1)$ model:

$$\begin{aligned} y_t &= \varepsilon_t + \theta_1 \varepsilon_{t-1}, \quad \text{i.e.,} \\ \phi(L) &= 1, \theta(L) = 1 + \theta_1 L \end{aligned}$$

  Backwards substitution yields

$$\varepsilon_t = \frac{y_t}{1+\theta_1 L} = \sum_{s=0}^{h} (-\theta_1)^s y_{t-s} + (-\theta_1)^h \varepsilon_{t-h}$$

- ε_t is equivalent to an $AR(h)$ process with coefficients $\phi_s = (-\theta_1)^s$ provided that $E[(-\theta_1)^h \varepsilon_{t-h}]$ gets small as h increases, i.e., $|\theta_1| < 1$
- $MA(q)$ is **invertible** if the roots of $\theta(z)$ exceed one
- Invertible MA process can be written as an infinite order AR process
- A stationary and invertible $ARMA(p,q)$ process can be written as either an AR model or as an MA model, typically of infinite order

$$y_t = \phi(L)^{-1}\theta(L)\varepsilon_t \quad \text{or } \theta(L)^{-1}\phi(L)y_t = \varepsilon_t$$

ARIMA representation for nonstationary processes

- Suppose that d of the roots of $\phi(L)$ equal unity (one), while the remaining roots of $\tilde{\phi}(L)$ fall outside the unit circle. Factorization:

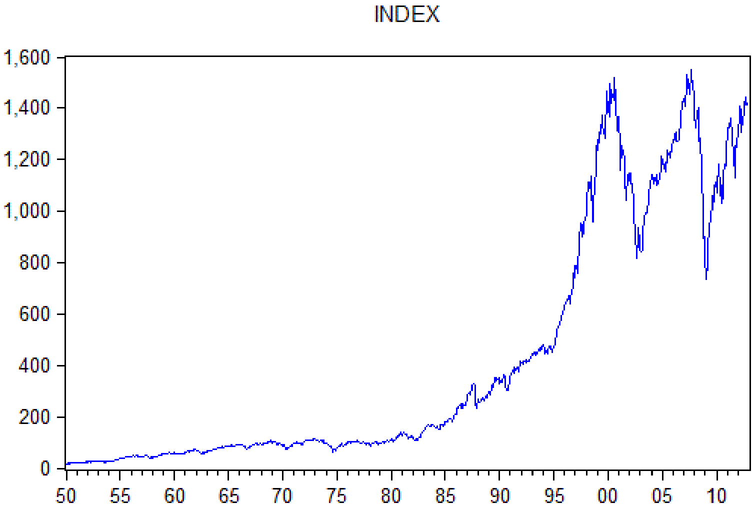
$$\phi(L) = \tilde{\phi}(L)(1-L)^d$$

- Applying $(1-L)$ to a series is called **differencing**
- Let $\tilde{y}_t = (1-L)^d y_t$ be the d^{th} difference of y_t. Then

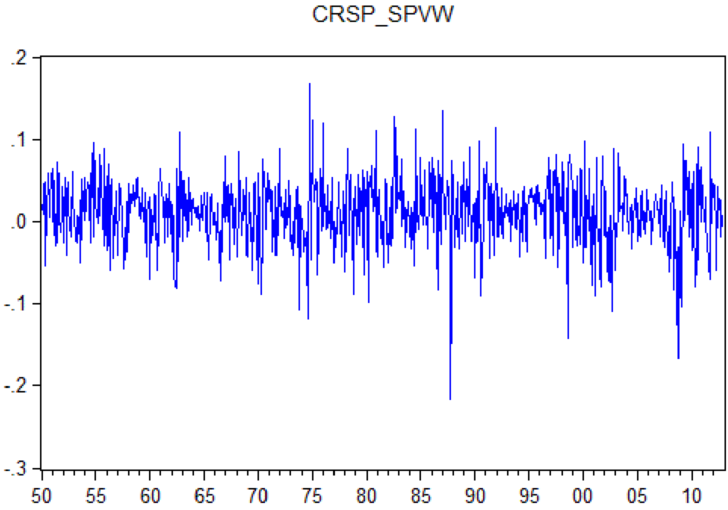
$$\tilde{\phi}(L)\tilde{y}_t = \theta(L)\varepsilon_t$$

- By assumption, the roots of $\tilde{\phi}(L)$ lie outside the unit circle so the differenced process, $\tilde{y}_t$, is stationary and can be studied instead of y_t
- Processes with $d \neq 0$ need to be differenced to achieve stationarity and are called $ARIMA(p, d, q)$

US stock index



Monthly US stock returns (first-differenced prices)



Forecasting with AR models

- Prediction is straightforward for $AR(p)$ models

$$y_{T+1} = \phi_1 y_T + ... + \phi_p y_{T-p+1} + \varepsilon_{T+1}, \quad \varepsilon_{T+1} \sim WN(0, \sigma^2)$$

- Treat parameters as known and ignore estimation error
- Using that $E[\varepsilon_{T+1} | \mathcal{I}_T] = 0$ and $\{y_{T-p+1}, ..., y_T\} \in \mathcal{I}_T$, the forecast of y_{T+1} given $\mathcal{I}_T$ becomes

$$f_{T+1|T} = \phi_1 y_T + ... + \phi_p y_{T-p+1}$$

- $f_{T+1|T}$ means the forecast of y_{T+1} given information at time T
- $x \in \mathcal{I}_T$ means "x is known at time T, i.e., belongs to the information set at time T"

Forecasting with AR models: The Chain Rule

- When generating forecasts multiple steps ahead, unknown values of y_{T+h} $(h \geq 1)$ can be replaced with their forecasts, $f_{T+h|T}$, setting up a recursive system of forecasts:

$$
\begin{aligned}
f_{T+2|T} &= \phi_1 f_{T+1|T} + \phi_2 y_T + \ldots + \phi_p y_{T-p+2} \\
f_{T+3|T} &= \phi_1 f_{T+2|T} + \phi_2 f_{T+1|T} + \phi_3 y_T + \ldots + \phi_p y_{T-p+3} \\
&\vdots \\
f_{T+p+1|T} &= \phi_1 f_{T+p|T} + \phi_2 f_{T+p-1|T} + \phi_3 f_{T+p-2|T} + \ldots + \phi_p f_{T+1|T}
\end{aligned}
$$

- 'Chain rule' is equivalent to recursively expressing unknown future values y_{T+i} as a function of y_T and its past
- Known values of y affect the forecasts of an $AR(p)$ model up to horizon $T+p$, while forecasts further ahead only depend on past forecasts themselves

Forecasting with MA models

- Consider the $MA(q)$ model

$$y_{T+1} = \varepsilon_{T+1} + \theta_1 \varepsilon_T + ... + \theta_q \varepsilon_{T-q+1}$$

One-step-ahead forecast:

$$f_{T+1|T} = \theta_1 \varepsilon_T + ... + \theta_q \varepsilon_{T-q+1}$$

Sequence of shocks $\{\varepsilon_t\}$ are not directly observable but can be computed recursively (estimated) given a set of assumptions on the initial values for ε_t, $t = 0, ..., q-1$

- For the $MA(1)$ model, we can set $\varepsilon_0 = 0$ and use the recursion

$$\begin{aligned}
\varepsilon_1 &= y_1 \\
\varepsilon_2 &= y_2 - \theta_1 \varepsilon_1 = y_2 - \theta_1 y_1 \\
\varepsilon_3 &= y_3 - \theta_1 \varepsilon_2 = y_3 - \theta_1 (y_2 - \theta y_1)
\end{aligned}$$

- Unobserved shocks can be written as a function of the parameter value θ_1 and current and past values of y

Forecasting with MA models (cont.)

- Simple recursions using past forecasts can also be employed to update the forecasts. For the $MA(1)$ model we have

$$f_{t+1|t} = \theta_1 \varepsilon_t = \theta_1 (y_t - f_{t|t-1})$$

- MA processes of infinite order: y_{T+h} for $h \geq 1$ is

$$\begin{aligned} y_{T+h} &= \theta(L)\varepsilon_{T+h} \\ &= (\underbrace{\varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + \ldots + \theta_{h-1}\varepsilon_{T+1}}_{unpredictable} + \underbrace{\theta_h \varepsilon_T + \theta_{h+1}\varepsilon_{T-1} + \ldots}_{predictable} \end{aligned}$$

Hence, if ε_T were observed, the forecast would be

$$\begin{aligned} f_{T+h|T} &= \theta_h \varepsilon_T + \theta_{h+1}\varepsilon_{T-1} + \ldots \\ &= \sum_{j=h}^{\infty} \theta_j \varepsilon_{T+h-j} \end{aligned}$$

$MA(q)$ model has limited memory: values of an $MA(q)$ process more than q periods into the future are not predictable

Forecasting with mixed ARMA models

- Consider a mixed $ARMA(p, q)$ model

$$y_{T+1} = \phi_1 y_T + \phi_2 y_{T-1} + ... + \phi_p y_{T-p+1} + \varepsilon_{T+1} + \theta_1 \varepsilon_T + ... + \theta_q \varepsilon_{T-q+1}$$

- Separate AR and MA prediction steps can be combined by recursively replacing future values of y_{T+i} with their predicted values and setting $E[\varepsilon_{T+j}|\mathcal{I}_T] = 0$ for $j \geq 1$:

$$\begin{aligned}
f_{T+1|T} &= \phi_1 y_T + \phi_2 y_{T-1} + ... + \phi_p y_{T-p+1} + \theta_1 \varepsilon_T + ... + \theta_q \varepsilon_{T-q+1} \\
f_{T+2|T} &= \phi_1 f_{T+1|T} + \phi_2 y_T + ... + \phi_p y_{T-p+2} + \theta_2 \varepsilon_T + ... + \theta_q \varepsilon_{T-q+2} \\
&\vdots \\
f_{T+h|T} &= \phi_1 f_{T+h-1|T} + \phi_2 f_{T+h-2|T} + ... + \phi_p f_{T-p+h|T} + \theta_h \varepsilon_T + ... + \theta_q \varepsilon_{T-q+h}
\end{aligned}$$

- Note: $f_{T-j+h|T} = y_{T-j+h}$ if $j \geq h$, and we assumed $q \geq h$

Mean Square Forecast Errors

- By the Wold Representation Theorem, all stationary ARMA processes can be written as an MA process with associated forecast error

$$y_{T+h} - f_{T+h|T} = \varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + ... + \theta_{h-1} \varepsilon_{T+1}$$

- Mean square forecast error:

$$\begin{aligned} E\left[(y_{T+h} - f_{T+h|T})^2\right] &= E[(\varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + ... + \theta_{h-1} \varepsilon_{T+1})^2] \\ &= \sigma^2 (1 + \theta_1^2 + ... + \theta_{h-1}^2) \end{aligned}$$

- For the $AR(1)$ model, $\theta_i = \phi_1^i$ and so the MSE becomes

$$\begin{aligned} E[(y_{T+h} - f_{T+h|T})^2] &= \sigma^2 (1 + \phi_1^2 + ... + \phi_1^{2(h-1)}) \\ &= \frac{\sigma^2 (1 - \phi_1^{2h})}{1 - \phi_1^2} \end{aligned}$$

Direct vs. Iterated multi-period forecasts

- Two ways to generate multi-period forecasts ($h > 1$):
  - **Iterated approach**: forecasting model is estimated at the highest frequency and iterated upon to obtain forecasts at longer horizons
  - **Direct approach**: forecasting model is matched with the desired forecast horizon: One model for each horizon, h. The dependent variable is y_{t+h} while all predictor variables are dated period t
- Example: AR(1) model $y_t = \phi_1 y_{t-1} + \varepsilon_t$
  - Iterated approach: use the estimated value, $\hat{\phi}_1$, to obtain a forecast $f_{T+h|T} = \hat{\phi}_1^h y_T$
  - Direct approach: Estimate $h-$period lag relationship:

$$y_{t+h} = \underbrace{\phi_1^h}_{\tilde{\phi}_{1h}} y_t + \underbrace{\sum_{s=0}^{h-1} \phi_1^s \varepsilon_{t-s}}_{\tilde{\varepsilon}_{t+h}}$$

Direct vs. Iterated multi-period forecasts: Trade-offs

- When the autoregressive model is correctly specified, the iterated approach makes more **efficient** use of the data and so tends to produce better forecasts
- Conversely, by virtue of being a linear projection, the direct approach tends to be more **robust** towards misspecification
    - When the model is grossly misspecified, iteration on the misspecified model can exacerbate biases and may result in a larger MSE
- Which approach performs best depends on the true DGP, the degree of model misspecification (both unknown), and the sample size
- Empirical evidence in Marcellino et al. (2006) suggests that the iterated approach works best on average for macro variables

Estimation of ARIMA models

- *ARIMA* models can be estimated by maximum likelihood methods
- *ARIMA* models are based on linear projections (regressions) which provide reasonable forecasts of linear processes under MSE loss
- There may be nonlinear models of past data that provide better predictors:
    - Under MSE loss the best predictor is the conditional mean, which need not be a linear function of the past

Estimation (continued)

- $AR(p)$ models with known $p > 0$ can be estimated by ordinary least squares by regressing y_T on $y_{T-1}, y_T, .., .y_{T-p}$
- Assuming the data are covariance stationary, OLS estimates of the coefficients $\phi_1, .., \phi_p$ are consistent and asymptotically normal
- If the AR model is correctly specified, such estimates are also asymptotically efficient
    - Least squares estimates are not optimal in finite samples and will be biased
    - For the $AR(1)$ model, $\hat{\phi}_1$ has a downward bias of $(1 + 3\phi_1)/T$
    - For higher order models, the biases are complicated and can go in either direction

Estimation and forecasting with ARMA models in matlab

- *regARIMA*: creates regression model with ARIMA time series errors
- *estimate*: estimates parameters of regression models with ARIMA errors
- Pure AR models: can be estimated by OLS
- *forecast*: forecast ARIMA models

Lag length selection

- In most situations, forecasters do not know the true or optimal lag orders, p and q
  - Judgmental approaches based on examining the autocorrelations and partial autocorrelations of the data
  - Model selection criteria: Different choices of (p, q) result in a set of models $\{M_k\}_{k=1}^{K}$, where M_k represents model k and the search is conducted over K different combinations of p and q
  - Information criteria trade off fit versus parsimony

Information criteria

- Information criteria (IC) for linear ARMA specifications:

$$IC_k = \ln \hat{\sigma}_k^2 + n_k g(T)$$

- ICs trade off fit (gets better with more parameters) against parsimony (fewer parameters is better). Choose k to minimize IC
- $\hat{\sigma}_k^2$: sum of squared residuals of model k. Lower $\hat{\sigma}_k^2 \Leftrightarrow$ better fit
- $n_k = p_k + q_k + 1$: number of estimated parameters for model k
- $g(T)$: penalty term that depends on the sample size, T:

Criterion	$g(T)$
AIC (Akaike (1974))	$2T^{-1}$
BIC (Schwartz (1978))	$\ln(T)/T$

In matlab: *aicbic*

Marcellino, Stock and Watson (2006)

Table 3
Relative MSFEs of each univariate forecast method, relative to iterated AR(4), and the fraction of times each forecast method is best

Forecast horizon	Summary statistic	Iterated					Direct				
		AR(4)	AR(12)	BIC	AIC	Sum	AR(4)	AR(12)	BIC	AIC	Sum
(A) *All series*											
3	Mean	1.00	0.99	1.01	0.99		0.99	0.99	0.99	0.99	
	Median	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00	
	Fraction best	0.15	0.22	0.21	0.12	0.70	0.06	0.14	0.06	0.08	0.33
6	Mean	1.00	0.97	1.00	0.97		0.99	0.98	0.98	0.98	
	Median	1.00	1.00	1.00	1.00		1.00	1.01	1.01	1.00	
	Fraction best	0.15	0.25	0.15	0.19	0.75	0.05	0.14	0.05	0.06	0.31
12	Mean	1.00	0.98	1.00	0.97		1.00	1.01	1.00	1.00	
	Median	1.00	1.01	1.01	1.00		1.01	1.03	1.02	1.02	
	Fraction best	0.25	0.23	0.14	0.17	0.79	0.07	0.09	0.05	0.05	0.25
24	Mean	1.00	1.01	1.00	1.00		1.05	1.10	1.05	1.08	
	Median	1.00	1.01	1.00	1.00		1.05	1.09	1.04	1.08	
	Fraction best	0.22	0.22	0.16	0.21	0.81	0.09	0.05	0.05	0.04	0.22

Random walk model

- The random walk model is an AR(1) with $\phi_1 = 1$:

$$y_t = y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2)$$

This model implies that the change in y_t is unpredictable:

$$\Delta y_t = y_t - y_{t-1} = \varepsilon_t$$

- For example, the level of stock prices is easy to predict, but not its change (rate of return if using logarithm of stock index)
- Shocks to the random walk have permanent effects: A one unit shock moves the series by one unit forever. This is in sharp contrast to a mean-reverting process

Random walk model (cont)

- The variance of a random walk increases over time so the distribution of y_t changes over time. Suppose that y_t started at zero, $y_0 = 0$:

$$\begin{aligned} y_1 &= y_0 + \varepsilon_1 = \varepsilon_1 \\ y_2 &= y_1 + \varepsilon_2 = \varepsilon_1 + \varepsilon_2 \\ &\vdots \\ y_t &= \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_{t-1} + \varepsilon_t \end{aligned}$$

From this we have

$$\begin{aligned} E[y_t] &= 0 \\ var(y_t) &= t\sigma^2, \quad \lim_{t\to\infty} var(y_t) = \infty \end{aligned}$$

- The variance of y grows proportionally with time
- A random walk does not revert to the mean but wanders up and down at random

Forecasts from random walk model

- Recall that forecasts from the AR(1) process $y_t = \phi_1 y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim WN(0, \sigma^2)$ are simply

$$f_{t+h|t} = \phi_1^h y_t$$

- For the random walk model $\phi_1 = 1$, so for all forecast horizons, h, the forecast is simply the current value:

$$f_{t+h|t} = y_t$$

- The basic random walk model says that the value of the series next period (given the history of the series) equals its current value plus an unpredictable change:

$$\text{Forecast of tomorrow} = \text{today's value}$$

- Random steps, ε_t, makes y_t a "random walk"

Random walk with a drift

- Introduce a non-zero drift term, δ :

$$y_t = \delta + y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2)$$

- This is a popular model for the logarithm of stock prices
- The drift term, δ, plays the same role as a time trend. Assuming again that the series started at y_0, we have

$$y_t = \delta t + y_0 + \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{t-1} + \varepsilon_t$$

Similarly,

$$\begin{aligned} E[y_t] &= y_0 + \delta t \\ var(y_t) &= t\sigma^2 \\ \lim_{t \to \infty} var(y_t) &= \infty \end{aligned}$$

Summary of properties of random walk

- Changes in random walk are unpredictable
- Shocks have permanent effects
- Variance grows in proportion with the forecast horizon
- These points are important for forecasting:
  - point forecasts never revert to a mean
  - since the variance goes to infinity, the width of interval forecasts increases without bound as the forecast horizon grows
  - Uncertainty grows without bounds

Logs, levels and growth rates

- Certain transformations of economic variables such as their logarithm are often easier to forecast than the "raw" data
- If the standard deviation of a time series is approximately proportional to its level, then the standard deviation of the change in the logarithm of the series is approximately constant:

$$\begin{aligned} Y_{t+1} &= Y_t \exp(\varepsilon_{t+1}), \quad \varepsilon_{t+1} \sim (0, \sigma^2) \Leftrightarrow \\ \ln(Y_{t+1}) - \ln(Y_t) &= \varepsilon_{t+1} \end{aligned}$$

- Example: US GDP follows an upward trend. Instead of studying the level of US GDP, we can study its growth rate which is not trending
- The first difference of the log of Y_t is $\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$
- The percentage change in Y_t between $t-1$ and t is approximately $100\Delta \ln(Y_t)$. This can be interpreted as a growth rate

Unit root processes

- Random walk is a special case of a unit root process which has a unit root in the AR polynomial, i.e.,

$$(1-L)\tilde{\phi}(L)y_t = \theta(L)\varepsilon_t$$

where the roots of $\tilde{\phi}(L)$ lie outside the unit circle

- We can test for a unit root using an Augmented Dickey Fuller (ADF) test:

$$\Delta y_t = \alpha + \beta y_{t-1} + \sum_{i=1}^{p} \Delta y_{t-i} + \varepsilon_t$$

- In matlab: *adftest*
- Under the null of a unit root, $\beta = 0$. Under the alternative of stationarity, $\beta < 0$
- Test is based on the t-stat of β. Test statistic follows a non-standard distribution

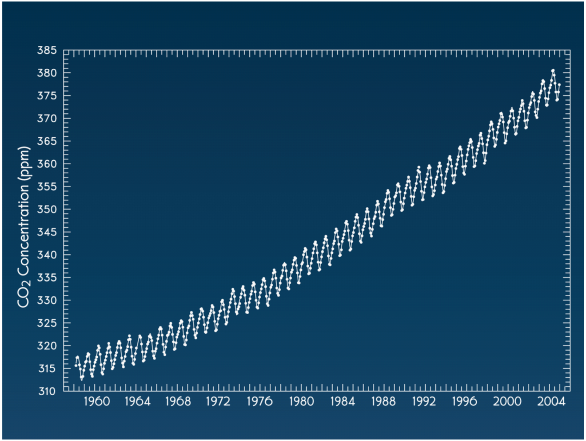
Critical values for Dickey-Fuller test

Critical values for Dickey–Fuller *t*-distribution.				
	Without trend		With trend	
Sample size	1%	5%	1%	5%
T = 25	−3.75	−3.00	−4.38	−3.60
T = 50	−3.58	−2.93	−4.15	−3.50
T = 100	−3.51	−2.89	−4.04	−3.45
T = 250	−3.46	−2.88	−3.99	−3.43
T = 500	−3.44	−2.87	−3.98	−3.42
T = ∞	−3.43	−2.86	−3.96	−3.41

Classical decomposition of time series into three components

- **Cycles** (stochastic) - captured using ARMA models
- **Trend**
  - trend captures the slow, long-run evolution in the outcome
  - for many series in levels, this is the most important component for long-run predictions
- **Seasonals**
  - regular (deterministic) patterns related to time of the year (day), public holidays, etc.

Example: CO2 concentration (ppm) - Dave Keeling, Scripps, 1957-2005



Seasonality

- Sources of seasonality: technology, preferences and institutions are linked to the calendar
  - weather (agriculture, construction)
  - holidays, religious events
- Many economic time series display seasonal variations:
  - home sales
  - unemployment figures
  - stock prices (?)
  - commodity prices?

Handling seasonalities

- One strategy is to remove the seasonal component and work with seasonally adjusted series
- Problem: We might be interested in forecasting the actual (non-adjusted) series, not just the seasonally adjusted part

Seasonal components

- Seasonal patterns can be deterministic or stochastic
- Stochastic modeling approach uses differencing to incorporate seasonal components - e.g., year-on-year changes
- Box and Jenkins (1970) considered seasonal ARIMA, or SARIMA, models of the form

$$\phi(L)(1 - L^S)y_t = \theta(L)\varepsilon_t.$$

  - $(1 - L^S)y_t = y_t - y_{T-S}$: computes year-on-year changes

Modeling seasonality

- Seasonality can be modeled through seasonal dummies. Let S be the number of seasons per year.
  - $S = 4$ (quarterly data)
  - $S = 12$ (monthly data)
  - $S = 52$ (weekly data)
- For example, the following set of dummies would be used to model quarterly variation in the mean:

$$
\begin{array}{ll}
D_{1t} = & \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \\
D_{2t} = & \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \\
D_{3t} = & \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \\
D_{4t} = & \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}
\end{array}
$$

- D_1 picks up mean effects in the first quarter. D_2 picks up mean effects in the second quarter, etc. At any point in time only one of the quarterly dummies is activated

Pure seasonal dummy model

- The pure seasonal dummy model is

$$y_t = \sum_{s=1}^{S} \delta_s D_{st} + \varepsilon_t$$

- We only regress y_t on intercept terms (seasonal dummies) that vary across seasons. δ_s summarizes the seasonal pattern over the year
- Alternatively, we can include an intercept and $S - 1$ seasonal dummies.
  - Now the intercept captures the mean of the omitted season and the remaining seasonal dummies give the seasonal increase/decrease relative to the omitted season
- Never include both a full set of S seasonal dummies and an intercept term - perfect collinearity

General seasonal effects

- Holiday variation (*HDV*) variables capture dates of holidays which may change over time (Easter, Thanksgiving) - v_1 of these:

$$y_t = \sum_{s=1}^{S} \delta_s D_{st} + \sum_{i=1}^{v_1} \delta_i^{HDV} HDV_{it} + \varepsilon_t$$

Seasonals

- ARMA model with seasonal dummies takes the form

$$\phi(L)y_t = \sum_{s=1}^{S} \delta_s D_{st} + \theta(L)\varepsilon_t$$

- Application of seasonal dummies can sometimes yield large improvements in predictive accuracy
- Example: day of the week, seasonal, and holiday dummies:

$$\mu_t = \sum_{day=1}^{7} \beta_{day} D_{day,t} + \sum_{holiday=1}^{H} \beta_{holiday} D_{holiday,t} + \sum_{month=1}^{12} \beta_{month} D_{month,t}$$

- Adding deterministic seasonal terms to the ARMA component, the value of y at time $T+h$ can be predicted as follows:

$$\begin{aligned} y_{T+h} = \sum_{day=1}^{7} \beta_{day} D_{day,T+h} + \sum_{holiday=1}^{H} \beta_{holiday} D_{holiday,T+h} & + \sum_{month=1}^{12} \beta_{month} D_{month,T+h} + \tilde{y}_{T+h}, \\ \phi(L)\tilde{y}_{T+h} & = \theta(L)\varepsilon_{T+h} \end{aligned}$$

Deterministic trends

- Let $Time_t$ be a deterministic time trend so that

$$Time_t = t, \quad t = 1, ..., T$$

- This time trend is perfectly predictable (deterministic)
- Linear trend model:

$$Trend_t = \beta_0 + \beta_1 Time_t$$

- β_0 is the intercept (value at time zero)
- β_1 is the slope which is positive if the trend is increasing or negative if the trend is decreasing

Examples of trended variables

- US stock price index
- Number of residents in Beijing, China
- US labor participation rate for women (up) or men (down)
- Exchange rates over long periods (?)
- Interest rates (?)
- Global mean temperature (?)

Quadratic trend

- Sometimes the trend is nonlinear (curved) as when the variable increases at an increasing or decreasing rate
- For such cases we can use a quadratic trend:

$$Trend_t = \beta_0 + \beta_1 Time_t + \beta_2 Time_t^2$$

- Caution: quadratic trends are mostly considered adequate local approximations and can give rise to a variety of unrealistic shapes for the trend if the forecast horizon is long

Log-linear trend

- log-linear trends are used to describe time series that grow at a constant exponential rate:

$$Trend_t = \beta_0 \exp(\beta_1 Time_t)$$

- Although the trend is non-linear in levels, it is linear in logs:

$$\ln(Trend_t) = \ln(\beta_0) + \beta_1 Time_t$$

Deterministic Time Trends: summary

- Three common time trend specifications:

$$\begin{aligned} \textit{Linear} &: \quad \mu_t = \mu_0 + \beta_0 t \\ \textit{Quadratic} &: \quad \mu_t = \mu_0 + \beta_0 t + \beta_1 t^2 \\ \textit{Exponential} &: \quad \mu_t = \exp(\mu_0 + \beta_0 t) \end{aligned}$$

- These global trends are unlikely to provide accurate descriptions of the future value of most time series at long forecast horizons

Estimating trend models

- Assuming MSE loss, we can estimate the trend parameters by solving

$$\hat{\theta} = \arg\min_{\theta} \left\{ \sum_{t=1}^{T} (y_t - Trend_t(\theta))^2 \right\}$$

- Example: with a linear trend model we have

$$\begin{aligned} Trend_t(\theta) &= \beta_0 + \beta_1 Time_t \\ \theta &= \{\beta_0, \beta_1\} \end{aligned}$$

and we can estimate β_0, β_1 by OLS

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \left\{ \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 Time_t)^2 \right\}$$

Forecasting Trend

- Suppose a time series is generated by the linear trend model

$$y_t = \beta_0 + \beta_1 Time_t + \varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma^2)$$

Future values of ε_t are unpredicable given current information, $\mathcal{I}_t$:

$$E[\varepsilon_{t+h} | \mathcal{I}_t] = 0$$

- Suppose we want to predict the series at time $T + h$ given information $\mathcal{I}_T$:

$$y_{T+h} = \beta_0 + \beta_1 Time_{T+h} + \varepsilon_{T+h}$$

Since $Time_{T+h} = T + h$ is perfectly predictable while ε_{T+h} is unpredictable, our best forecast (under MSE loss) becomes

$$f_{T+h|T} = \hat{\beta}_0 + \hat{\beta}_1 Time_{T+h}$$