Forecasting & Predictive Analytics

Guillaume Chevillon (chevillon@essec.edu)
and Pierre Jacob (jacob@essec.edu)

October-December 2021
3rd set of slides
ARMA models

Overview

- Forecasting starts with a number of adjustments, transformations and decompositions on the original data.
- Forecasting can be done with deterministic methods such as exponential smoothing.
- Probabilistic approaches enable uncertainty estimates, and provide a coherent approach to estimation and model comparison.
- For stationary processes, we can hope to learn properties/parameters from time averages of observations.
- Today we explore the class of ARMA models for time series analysis and forecasting.

Playing with AR and MA models

```
library(shiny);
runGitHub(repo="shinyapps",ref="main",
username="pierrejacob",subdir="acfautoregressive/")
runGitHub(repo="shinyapps",ref="main",
username="pierrejacob",subdir="acfma/")
```

Linear processes

Definition

Noise terms (W_t), uncorrelated, mean zero, variance σ_W^2.

Linear processes:

$$Y_t = \mu + \sum_{j=-\infty}^{+\infty} \psi_j W_{t-j}.$$

Or restrict to a sum over $j \geq 0$:

$$Y_t = \mu + \sum_{j=0}^{+\infty} \psi_j W_{t-j}.$$

Weighted sum of all past terms, i.e. MA(∞) process.

Under some condition on (ψ_j), a linear process is stationary.

Wold representation theorem

If Y_t is stationary, then we can always write

$$Y_t = \sum_{j=0}^{\infty} \psi_j W_{t-j} + V_t,$$

where

- $\psi_0 = 1$, $\sum_{j=0}^{\infty} \psi_j^2 < \infty$,
- W_t is white noise with zero mean and variance σ_W^2,
- V_t is "predictable", satisfies conditions including $\mathbb{E}[V_t W_s] = 0$ for all s, t.

In other words, all weak stationary processes resemble a MA(∞) process plus a predictable process.

ARMA

Autoregressive process

Backshift or lag operator $B^k Y_t = Y_{t-k}$ for all $k \geq 1$.
AR(p) model:

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) Y_t = W_t.$$

Or we can include a mean parameter μ.

Autoregressive operator:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p,$$

leading to the concise notation:

$$\phi(B) Y_t = W_t.$$

From AR(1) to moving averages

- AR(1) with $|\phi| < 1$:

$$\begin{aligned} Y_t &= \phi Y_{t-1} + W_t \\ &= \phi^k Y_{t-k} + \phi^{k-1} W_{t-k+1} + \phi^{k-2} W_{t-k+2} + \ldots + \phi W_{t-1} + W_t \\ &= \sum_{j=0}^{\infty} \psi_j W_{t-j}, \end{aligned}$$

  where $\psi_j = \phi^j$. This is an MA(∞) representation.

- If we keep only q first terms, we obtain an MA(q) process:

$$Y_t = \sum_{j=0}^{q} \theta_j W_{t-j} = \theta(B) W_t,$$

  where $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$.

AR + MA = ARMA

A stochastic process is ARMA(p,q) if it is stationary and, for all t,

$$\phi(B) Y_t = \theta(B) W_t,$$

where

$$\begin{aligned}
\phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p, \\
\theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q,
\end{aligned}$$

with $\phi_p \neq 0$ and $\theta_q \neq 0$.
In other words,

$$\begin{aligned}
&Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \ldots - \phi_p Y_{t-p} \\
&= W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \ldots + \theta_q W_{t-q}.
\end{aligned}$$

So $\phi(B) Y_t$ follows an MA(q) process.

AR(p) to MA(∞)

- AR(p): $\phi(B)Y_t = W_t$. What are conditions on $\phi(B)$ for the process to be stationary and future-independent?

- If we can express $\phi(B)Y_t = W_t$ as

$$Y_t = \psi(B)W_t = \sum_{j=0}^{\infty} \psi_j W_{t-j},$$

  then we can use conditions on (ψ_j) such as $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

- Intuitively, we would like to consider the inverse $\phi^{-1}(B)$, such that

$$\phi^{-1}(B)\phi(B)Y_t = \phi^{-1}(B)W_t.$$

AR(p) to MA(∞)

- We can treat $\phi(B)$ as a polynomial, $\phi(z)$ for $z \in \mathbb{C}$.

- The condition for invertibility will be: the roots of ϕ are outside the unit circle, i.e.

$$\forall z \in \mathbb{C} \text{ s.t. } |z| < 1, \quad \phi(z) \neq 0.$$

- Allows to invert $1 - \varphi_1 z - \varphi_2 z^2 - \ldots - \varphi_p z^p$.

AR(p) to MA(∞)

- Find roots $\lambda_1, \ldots, \lambda_p$ such that

$$1 - \varphi_1 z - \varphi_2 z^2 - \ldots - \varphi_p z^p = \prod_{j=1}^{p} (1 - \lambda_j^{-1} z).$$

  Assume roots are distinct.

- Invert:

$$\varphi(z)^{-1} = \frac{1}{\prod_{j=1}^{p} (1 - \lambda_j^{-1} z)} = \sum_{j=1}^{p} \frac{a_j}{1 - \lambda_j^{-1} z}$$

  for some $a_1, \ldots, a_p$ that we can find.

- Then

$$\sum_{j=1}^{p} \frac{a_j}{1 - \lambda_j^{-1} z} = \sum_{k=0}^{\infty} (\sum_{j=1}^{p} a_j \lambda_j^{-k}) z^k,$$

  so that $\psi_k = (\sum_{j=1}^{p} a_j \lambda_j^{-k})$.

Restrictions for MA(q) processes

Consider both of these MA(1) processes, where (W_t) is WN$(0, 1)$,

- $Y_t = W_t + 5W_{t-1}$ ($\theta_1 = 5$, $\sigma_W = 1$),
- $\tilde{Y}_t = (5W_t) + 1/5 \times (5W_{t-1})$ ($\theta_1 = 1/5$, $\sigma_W = 5$).

The autocovariance satisfies

- $\gamma(0) = \sigma_W^2(1 + \theta_1^2) = 1^2 \times (1 + 5^2) = 5^2 \times (1 + 1/5^2)$,
- $\gamma(1) = \sigma_W^2\theta_1 = 1^2 \times 5 = 5^2 \times 1/5$,
- $\gamma(h) = 0$ for $|h| > 1$.

We restrict ourselves to MA(q) processes such that $\theta(B)$ is *invertible*, so that we have an AR(∞) representation:

$$Y_t = \theta(B)W_t \quad \Leftrightarrow \quad \theta(B)^{-1}Y_t = W_t.$$

Restrictions on ARMA(p,q) processes

Consider the equations $Y_t = W_t$, for all $t \in \mathbb{Z}$.

A process satisfying that, also satisfies for all α,

$$\alpha Y_{t-1} = \alpha W_{t-1}$$
$$Y_t - \alpha Y_{t-1} = W_t - \alpha W_{t-1}$$
$$(1 - \alpha B) Y_t = (1 - \alpha B) W_t$$

so that the white noise process is also an ARMA(1,1) process with $\varphi_1 = -\theta_1 = \alpha$, for any α.

To alleviate this redundancy, we impose that $\varphi(B)$ and $\theta(B)$ have no common root.

Summary: ARMA(p,q) processes

We will consider ARMA(p,q) processes

$$\phi(B)Y_t = \theta(B)W_t,$$

such that

- $\varphi(B)$ and $\theta(B)$ have their roots outside of the unit circle,
- $\varphi(B)$ and $\theta(B)$ have no common root,
- $\varphi_p \neq 0$ and $\theta_q \neq 0$.

Representations of ARMA(p,q) processes

For these ARMA(p,q), we can write an MA(∞) representation:

$$Y_t = \sum_{j=0}^{\infty} \psi_j W_{t-j},$$

where $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$, for all $|z| \leq 1$.

We can write an AR(∞) representation:

$$W_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j},$$

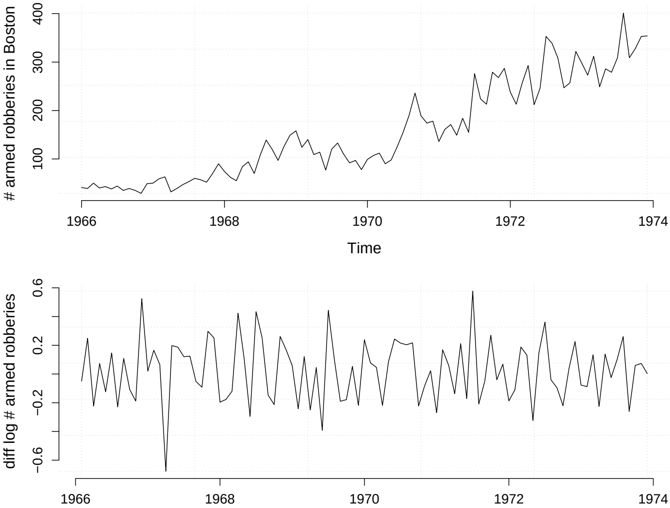
where $\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \phi(z)/\theta(z)$ for all $|z| \leq 1$.

Both representations are useful for prediction purposes. The coefficients (ψ_j) and (π_j) can be computed from φ and θ.

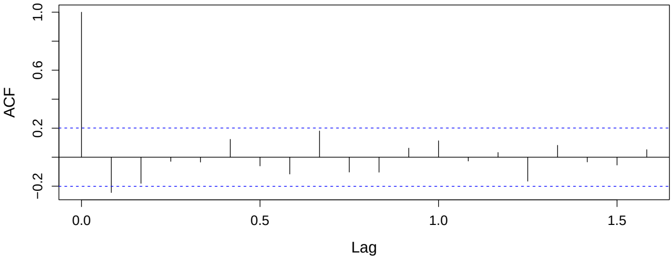
Forecasting with ARMA

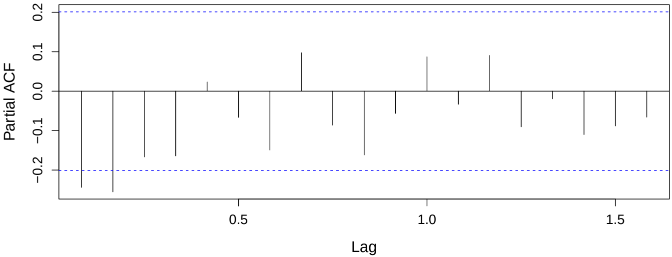
Armed robberies in Boston

Number of armed robberies reported each month in Boston, Massachusetts, from January 1966; see Deutsch & Alt 1975.

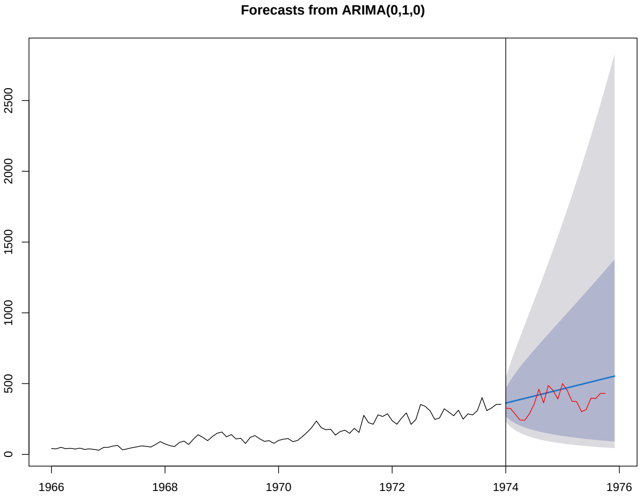


Armed robberies in Boston

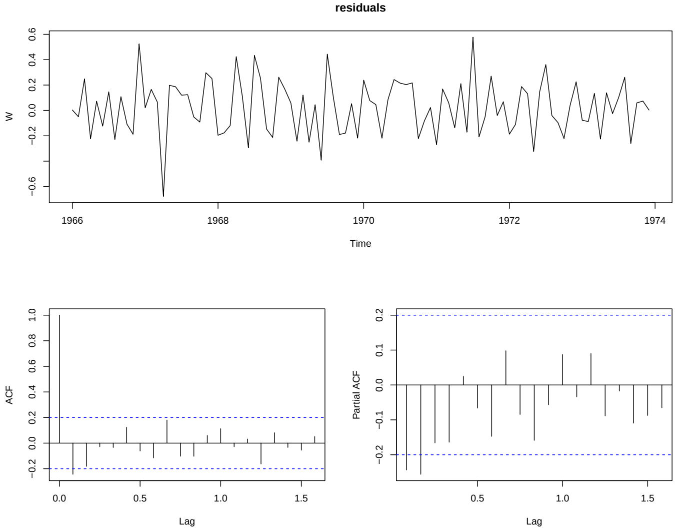




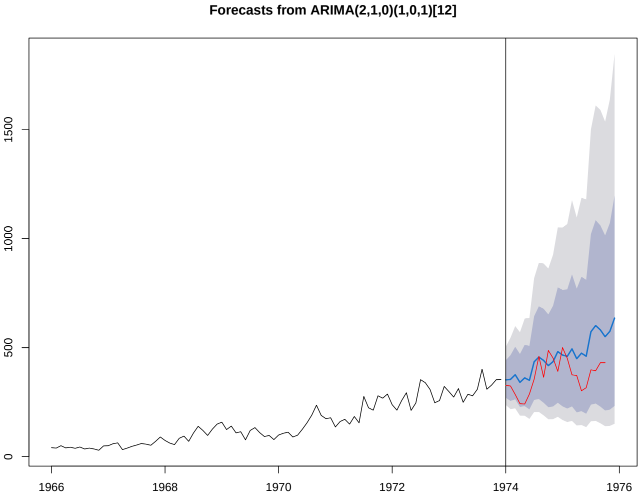
Armed robberies in Boston



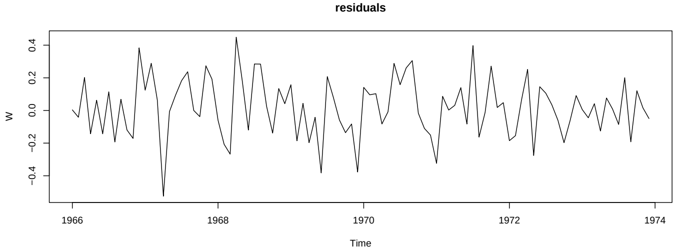
Armed robberies in Boston

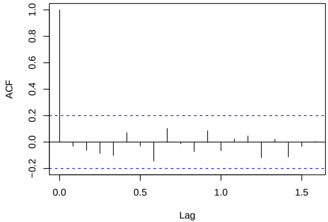


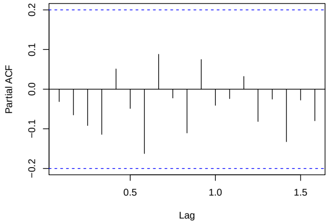
Armed robberies in Boston



Armed robberies in Boston







Forecast of ARMA(p,q) processes

- Consider the predictor $\tilde{Y}^t_{t+m} = \mathbb{E}[Y_{t+m} | Y_{-\infty:t}]$, using the infinite past.
- Consider the two equations:

$$Y_{t+m} = \sum_{j=0}^{\infty} \psi_j W_{t+m-j},$$

$$W_{t+m} = \sum_{j=0}^{\infty} \pi_j Y_{t+m-j}.$$

- Taking the conditional expectation $\mathbb{E}[\cdot | Y_{-\infty:t}]$ of the AR(∞) representation leads to

$$\tilde{Y}^t_{t+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{Y}^t_{t+m-j} - \sum_{j=m}^{\infty} \pi_j Y_{t+m-j}.$$

Requires knowing (π_j).

Forecast of ARMA(p,q) processes

Taking the conditional expectation of the MA(∞) representation leads to

$$\tilde{Y}_{t+m}^{t} = \sum_{j=m}^{\infty} \psi_j W_{t+m-j},$$

and in turn to

$$\mathbb{E}\left[\left(Y_{t+m} - \tilde{Y}_{t+m}^{t}\right)^2\right] = \sigma_W^2 \sum_{j=0}^{m-1} \psi_j^2$$

Requires knowing (ψ_j).

$\Rightarrow$ knowing (ψ_j) and (π_j) is useful for prediction!

Forecast of ARMA(p,q) processes

- In practice we condition on $Y_{1:t}$ instead of $Y_{-\infty:t}$; we can truncate the sum:

$$\tilde{Y}^t_{t+m} = -\sum_{j=1}^{m-1} \pi_j \tilde{Y}^t_{t+m-j} - \sum_{j=m}^{t+m-1} \pi_j Y_{t+m-j}.$$

- So, for models of the ARMA(p,q) form,

$$\varphi(B)Y_t = \theta(B)W_t,$$

we can estimate the parameters, and use the AR(∞) and MA(∞) representations to compute forecasts and prediction errors.

Identification

The Table

Table 3.1. Behavior of the ACF and PACF for ARMA Models

	AR(p)	MA(q)	ARMA(p,q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

From Shumway & Stoffer's Time Series Analysis book.

Identifying the order of ARMA models

- The ACF and PACF can be used to determine the order of an ARMA process (i.e. p and q), in the case of "pure AR" or "pure MA" processes.
- Cannot distinguish ARMA(p,q) models in general using ACF and PACF: they both "tail off" for most models.
- Other ways of selecting models (AIC, BIC, based on a bespoke loss function) are more principled and should be prefered.
- Given the order, we can estimate parameters, i.e. finding values for

  $$\sigma_W, \varphi_1, \dots, \varphi_p, \theta_1, \dots, \theta_q,$$

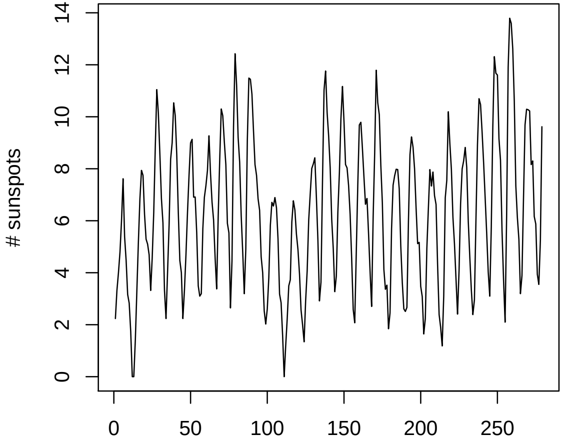
  and then we can perform prediction.

Identifying the order of ARMA models

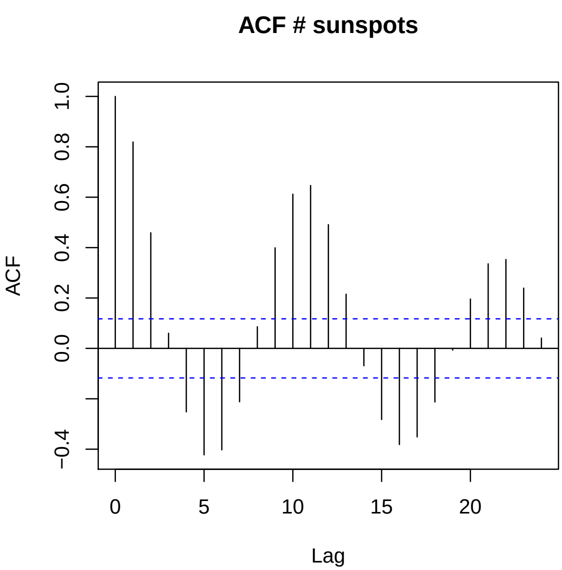
- We can try each model ARMA(p,q) for $p, q \geq 0$.
- Using ACF and PACF, we can focus the search for the best model to a subset of all ARMA(p,q) models.
- There is still some ambiguity left.

Example

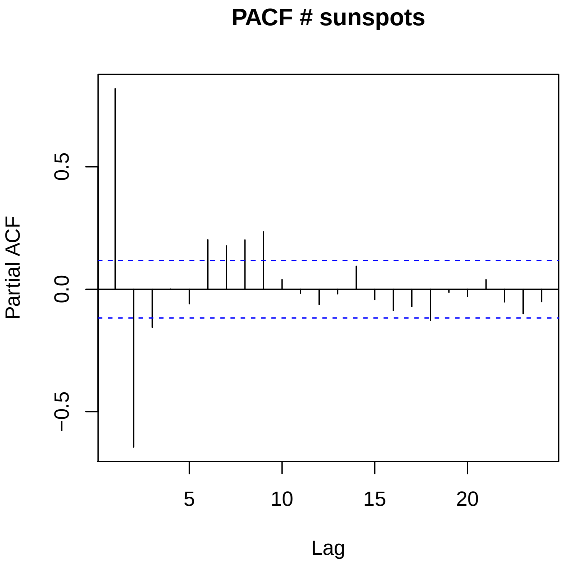
Sunspot data set, yearly observations since 1700.



Example



Example



Model choice

- Naively, we could fit an ARMA(p,q) model with large values of p and q. Why is this not a good idea?

- Occam's razor: "entities should not be multiplied unnecessarily."

  Pluralitas non est ponenda sine neccesitate.

- Isaac Newton stated it as: "we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances."

- Or, restated in our context, "if two competing models fit the data equally well, the simpler one is better."

Information Criteria

Let $k = p + q + 1$ be the number of parameters, and $\hat{\beta}$ the maximum likelihood estimator, that is $\hat{\varphi}, \hat{\theta}, \hat{\sigma}_W^2$.

$$\begin{aligned}
\text{AIC} &= -2 \log p(y_1, \ldots, y_n; \hat{\beta}) + 2k, \\
\text{BIC} &= -2 \log p(y_1, \ldots, y_n; \hat{\beta}) + k \log(n).
\end{aligned}$$

Bias-corrected AIC *for ARMA models* in Brockwell & Davis, Eq. (9.3.4),

$$\text{AICc} = -2 \log p(y_1, \ldots, y_n; \hat{\beta}) + \frac{2kn}{n - k - 1}.$$

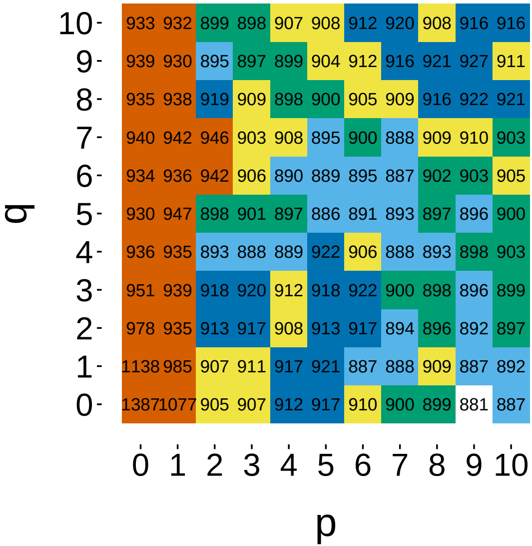
The smaller the value of AIC/BIC/AICc, the better.

AIC

Colours assigned according to values. Lower is better.

q \ p	0	1	2	3	4	5	6	7	8	9	10
10	890	885	849	844	848	847	847	852	836	840	836
9	899	887	848	846	845	845	850	851	852	854	835
8	898	898	876	862	847	845	847	848	851	853	849
7	907	906	906	860	861	844	845	830	847	845	834
6	905	904	906	866	846	842	844	832	844	841	839
5	905	918	865	864	857	842	843	842	842	838	838
4	914	910	864	855	852	883	862	841	842	843	845
3	933	917	893	891	880	882	882	857	850	846	844
2	963	917	891	891	879	880	880	854	853	845	847
1	1128	970	889	889	891	891	854	852	869	843	845
0	1380	1066	890	889	891	892	881	867	863	841	843

BIC


q
p
10 933 932 899 898 907 908 912 920 908 916 916
9 939 930 895 897 899 904 912 916 921 927 911
8 935 938 919 909 898 900 905 909 916 922 921
7 940 942 946 903 908 895 900 888 909 910 903
6 934 936 942 906 890 889 895 887 902 903 905
5 930 947 898 901 897 886 891 893 897 896 900
4 936 935 893 888 889 922 906 888 893 898 903
3 951 939 918 920 912 918 922 900 898 896 899
2 978 935 913 917 908 913 917 894 896 892 897
1 1138 985 907 911 917 921 887 888 909 887 892
0 1387 1077 905 907 912 917 910 900 899 881 887
0 1 2 3 4 5 6 7 8 9 10

Out-of-sample prediction

Split the dataset into $y_1, \ldots, y_{n-L}$ and $y_{n-L+1}, \ldots, y_n$.

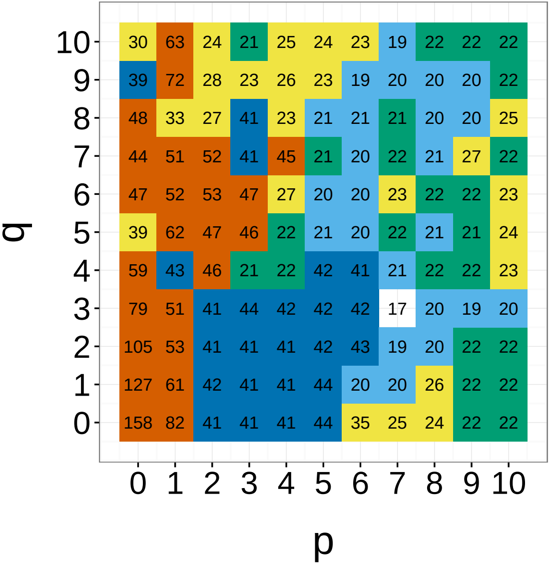
For each model,

- estimate the parameters using training data $y_1, \ldots, y_{n-L}$,
- predict test data, e.g. with $\hat{y}_{n-L+1}, \ldots, \hat{y}_n$,
- compute an assessment, for instance

$$\text{error} = \sum_{t=n-L+1}^{n} (y_t - \hat{y}_t)^2 .$$

Choose the model with lowest out-of-sample predictive error.

Out-of-sample prediction



Extensions: ARIMA, SARIMA

Differences

Let $(X_t)_{t \in \mathbb{Z}}$ be a stochastic process.

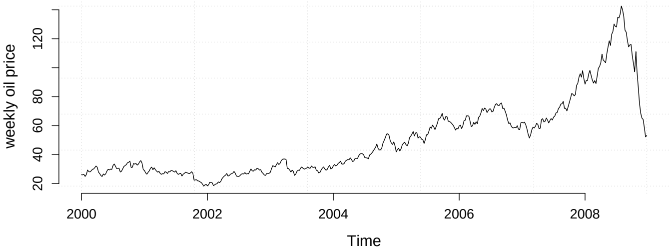
Let $Y_t = \nabla X_t = (1 - B)X_t = X_t - X_{t-1}$.

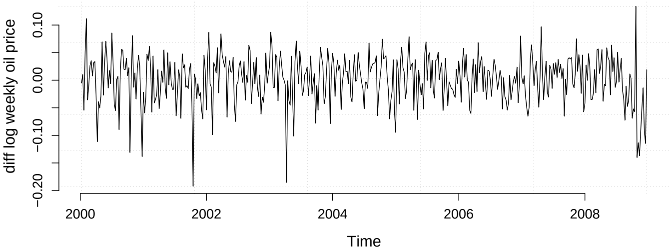
Likewise we can define $\nabla^d = (1 - B)^d$ for all $d \geq 1$.

If $Y_t = \nabla^d X_t$, follows an ARMA(p,q) model, we say that (X_t) follows an ARIMA(p,d,q) model.

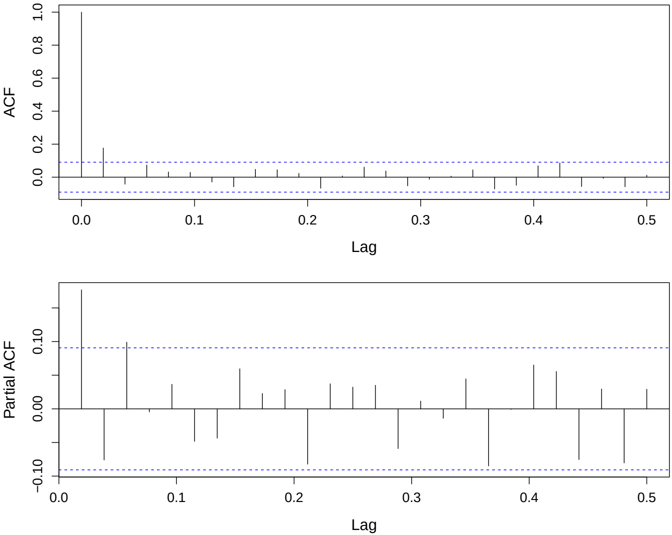
Oil prices

Crude oil, WTI spot price FOB (in dollars per barrel), weekly data

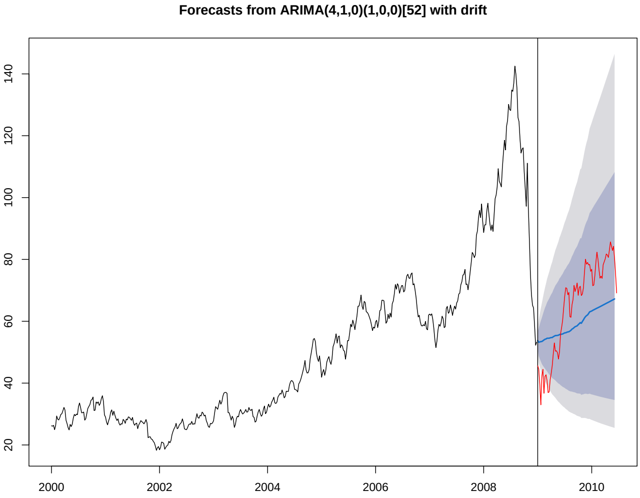




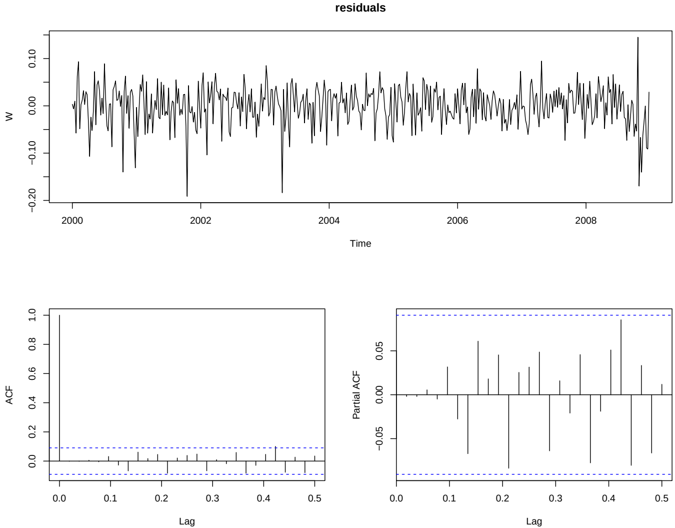
Oil prices


ACF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
Lag
Partial ACF
0.10
0.00
−0.10
0.0
0.1
0.2
0.3
0.4
0.5
Lag

Oil prices


Forecasts from ARIMA(4,1,0)(1,0,0)[52] with drift
20
40
60
80
100
120
140
2000
2002
2004
2006
2008
2010

Oil prices



residuals
W
0.10
0.00
−0.10
−0.20
2000
2002
2004
2006
2008
Time
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
Lag
Partial ACF
0.05
0.00
−0.05
0.0
0.1
0.2
0.3
0.4
0.5
Lag

Differences

Let $Y_t = \nabla X_t$.

- If (X_t) is stationary, then (Y_t) is also stationary.
- Warning: if (X_t) is white noise, then (Y_t) has non-zero autocorrelations.
- If (X_t) is a random walk with drift, then (Y_t) is stationary.
- If (X_t) is linear trend plus noise, then (Y_t) is stationary.

Testing integration: Dickey–Fuller test

- If Y_t is non-stationary but ∇Y_t is stationary, cannot be linearly related as in

$$\nabla Y_t = \phi Y_{t-1} + X_t \tag{1}$$

  where X_t is stationary. It must be the case that $\phi = 0$.

- AR(1) case: $Y_t = \rho Y_{t-1} + W_t$. Then

$$\nabla Y_t = (\rho - 1)\, Y_{t-1} + W_t$$

  The question is $\phi < 0$ or $\phi = 0$?

- Dickey–Fuller test: estimate (1) by OLS and perform a one-sided t-test for the null hypothesis $\mathrm{H}_0 : \phi = 0$ vs $\mathrm{H}_1 : \phi < 0$.
  t statistic has a nonstandard distribution

- Issues arise due to presence of a trend and X_t not being i.i.d.

Testing integration

A number of tests can be helpful to identify the order d of an ARIMA(p,d,q).

- Dickey–Fuller test: null hypothesis = one root of the AR polynomial is equal to 1.

  `library(tseries)`, `adf.test`.

- Also Philipps–Perron.

  `library(tseries)`, `pp.test`.

- Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test: null = stationarity, against alternative = unit root.

  `library(tseries)`, `kpss.test`.

http://faculty.washington.edu/ezivot/econ584/notes/unitroot.pdf.

Purely seasonal ARMA models

- Let Φ and Θ, polynomials of order P and Q:

$$\Phi(z) = 1 - \Phi_1 z - \Phi_2 z^2 - \ldots - \Phi_p z^P,$$
$$\Theta(z) = 1 + \Theta_1 z + \Theta_2 z^2 + \ldots + \Theta_q z^Q.$$

  then a purely seasonal ARMA(P,Q) of period 12 satisfies:

$$\Phi(B^s) Y_t = \Theta(B^s) W_t.$$

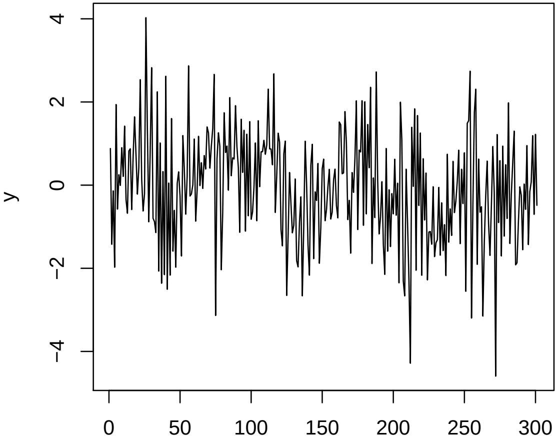
- For instance, purely seasonal ARMA(1,1) with period 12:

$$Y_t - \Phi_1 Y_{t-12} = W_t + \Theta_1 W_{t-12}.$$

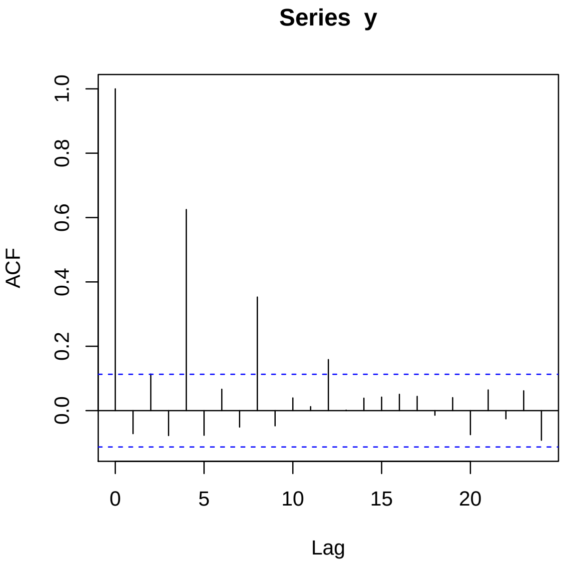
- More parcimonious than an ARMA(12,12).
- Same conditions as standard ARMA: Φ and Θ have roots outside unit circle, no common roots.

Purely seasonal ARMA models

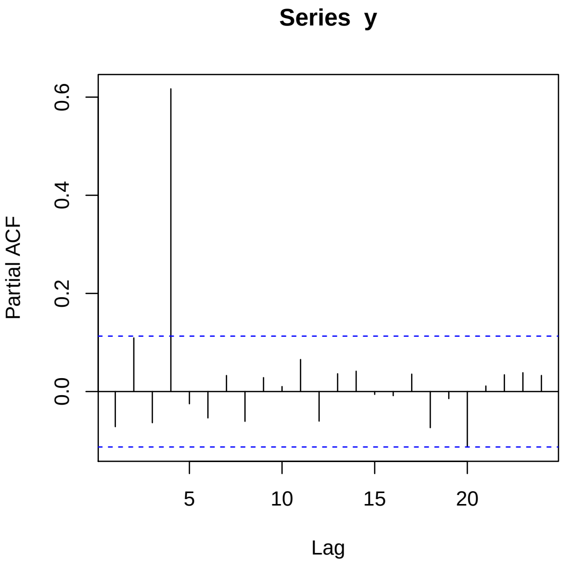
With $\Phi(B^4) = 1 - 0.65B^4$, $\sigma^2 = 1$, $T = 300$.



Purely seasonal ARMA models



Purely seasonal ARMA models



Mixed seasonal ARIMA models

- We can mix seasonal and non-seasonal components, e.g.

$$\nabla X_t = \Phi \nabla X_{t-12} + W_t + \theta W_{t-1},$$

or

$$Y_t = \phi Y_{t-1} + \Phi Y_{t-4} - \phi \Phi Y_{t-5} + W_t.$$

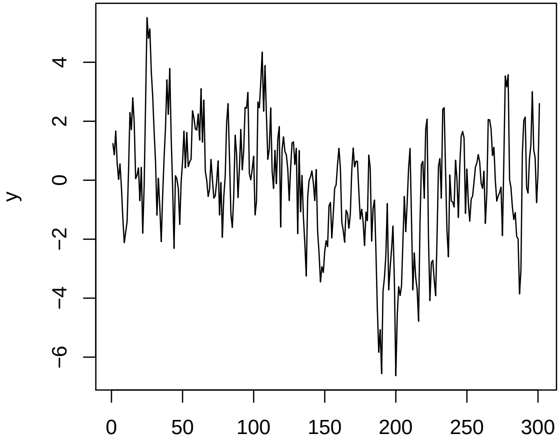
- A general SARIMA(p,d,q)$\times$(P,D,Q)$_s$ satisfies:

$$\Phi(B^s)\phi(B)\nabla^{Ds}\nabla^d Y_t = \Theta(B^s)\theta(B)W_t.$$

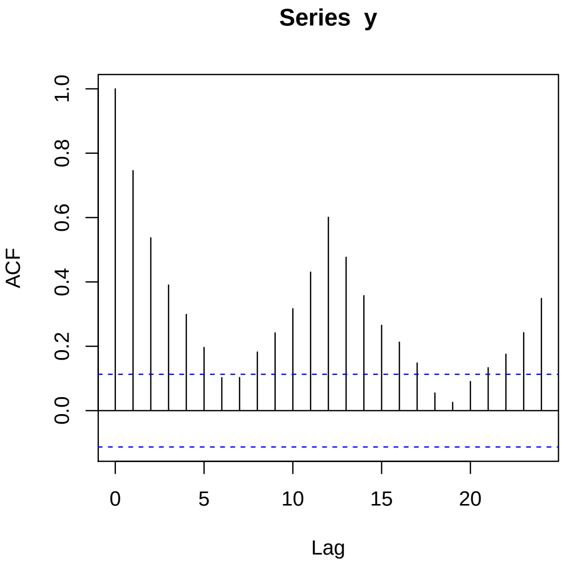
- Can generate process with complex dependencies, with relatively few parameters.

Mixed seasonal ARIMA models

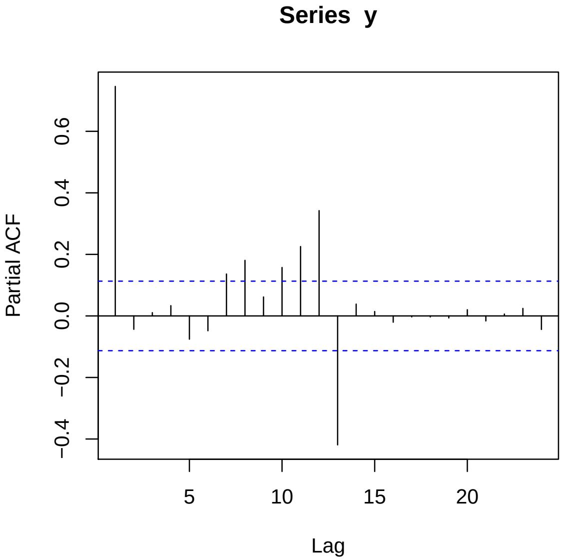
With $\Phi(B^{12}) = 1 - 0.65B^{12}$, $\phi(B) = 1 - 0.7B$, $\sigma^2 = 1$, $T = 300$.



Mixed seasonal ARIMA models



Mixed seasonal ARIMA models



ARIMA and exponential smoothing

ARIMA(0,1,1) model

Consider

$$Y_t = Y_{t-1} + W_t - \theta_1 W_{t-1}.$$

We assume $|\theta_1| < 1$, which allows to write $(1 - \theta_1 B)^{-1}$ as $\sum_{j=0}^{\infty} \theta_1^j B^j$.

We obtain

$$\left(\sum_{j=0}^{\infty} \theta_1^j B^j \right) (1 - B) Y_t = W_t$$

$$\text{or, equivalently, } \sum_{j=0}^{\infty} \theta_1^j B^j Y_t - \sum_{j=0}^{\infty} \theta_1^j B^{j+1} Y_t = W_t.$$

ARIMA(0,1,1) model

Note that the first sum is $Y_t + \sum_{j=1}^{\infty} \theta_1^j Y_{t-j}$, and the second sum is $\sum_{j=1}^{\infty} \theta_1^{j-1} Y_{t-j}$, so

$$Y_t = \sum_{j=1}^{\infty} (1-\theta_1)\theta_1^{j-1} Y_{t-j} + W_t.$$

From this expression, we can calculate the forecast of Y_{n+1} given $Y_{-\infty}, \ldots, Y_n$ using the conditional expectation:

$$\begin{aligned} Y_{n+1}^n &= \mathbb{E}[Y_{n+1} | Y_{-\infty}, \ldots, Y_n] \\ &= \sum_{j=1}^{\infty} (1-\theta_1)\theta_1^{j-1} Y_{n+1-j} \end{aligned}$$

Looks familiar! Benefits of the ARIMA view... parameter estimation, prediction intervals.

Exponential smoothing

```
m <- HoltWinters(co2)
p <- predict(m, 50, prediction.interval = TRUE)
plot(m, p)
```

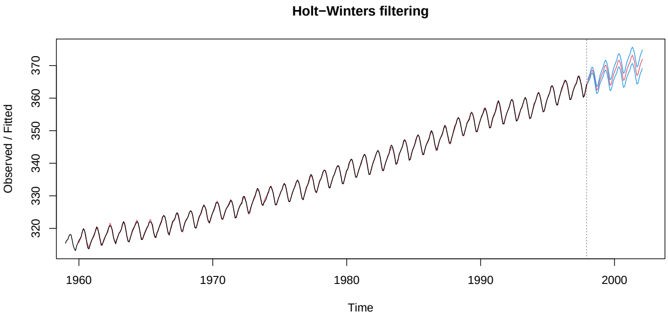


Figure: Atmospheric concentrations (monthly) of $CO2$ in Mauna Loa, expressed in parts per million (ppm), with predictions and intervals.

Prediction intervals obtained from ARIMA model.

State space representation

Linear Gaussian state space models

Observations (Y_t). The process (X_t) is "hidden" or "latent".

$$
\begin{aligned}
Y_t &= AX_t + V_t, \text{ with } V_t \sim \mathcal{N}(0, \Sigma_V), \\
X_t &= \Phi X_{t-1} + W_t \text{ with } W_t \sim \mathcal{N}(0, \Sigma_W).
\end{aligned}
$$

Parameters: A, Φ, Σ_V, Σ_W. We also need to specify X_0, e.g. $\mathcal{N}(m_0, C_0)$.

AR as state space models

Consider the state equation for X_t defined as

$$\begin{pmatrix} Y_{t-p+1} \\ Y_{t-p+2} \\ \vdots \\ Y_{t-1} \\ Y_t \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \varphi_p & \varphi_{p-1} & \varphi_{p-2} & \dots & \varphi_1 \end{pmatrix} \begin{pmatrix} Y_{t-p} \\ Y_{t-p+1} \\ \vdots \\ Y_{t-2} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} W_t.$$

We define the observation equation to be

$$Y_t = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \end{pmatrix} X_t + V_t,$$

where the noise term V_t is equal to zero.

MA as state space models

Next consider an MA(q) model:
$Y_t = W_t + \theta_1 W_{t-1} + \ldots + \theta_q W_{t-q}$. We can define the observation equation as

$$Y_t = \begin{pmatrix} \theta_q & \theta_{q-1} & \ldots & \theta_1 & 1 \end{pmatrix} X_t + V_t,$$

where again V_t is zero for all times t, and

$$X_t = \begin{pmatrix} W_{t-q} \\ W_{t-q+1} \\ \vdots \\ W_{t-1} \\ W_t \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \begin{pmatrix} W_{t-q-1} \\ W_{t-q} \\ \vdots \\ W_{t-2} \\ W_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} W_t.$$

In this representation, the latent process X_t contains all the noise terms that are used in the definition of Y_t.

ARMA as state space models

Define $r = \max(p, q+1)$, and extend φ or θ with zeros. Consider a latent process X_t made of r elements $(X_{t,1}, \ldots, X_{t,r})$ such that

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ \vdots \\ X_{t,r-1} \\ X_{t,r} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ \varphi_r & \varphi_{r-1} & \varphi_{r-2} & \ldots & \varphi_1 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \\ \vdots \\ X_{t-1,r-1} \\ X_{t-1,r} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} W_t.$$

Define the observation equation as

$$Y_t = \begin{pmatrix} \theta_{r-1} & \theta_{r-2} & \ldots & \theta_1 & 1 \end{pmatrix} X_t + V_t,$$

where V_t is zero for all times t.

ARMA as state space models

We will see how to compute the likelihood in $\mathcal{O}(n)$ operations for linear Gaussian models using the Kalman filter.

In particular this provides an efficient way of evaluating the likelihood of ARMA models. This can be plugged a numerical optimizer to find the MLE.

And thus to obtain parameter estimates, compute forecasts, model selection criteria, etc.

Box–Jenkins methodology

To summarize, the Box–Jenkins toolbox consists of. . .

- preparing and transforming the data so that a linear, stationary process might be an adequate model.
- Identify a model within the ARIMA family, using ACF, PACF, model selection criteria.
- Estimate model parameters and compute predictions along with their errors.
- Inspect the residuals and change the model accordingly, if necessary.

Next time: state space models, which include ARIMA models and many more.