CSCI446/946 Big Data Analytics

Week 11 Advanced Analytical Theory and Methods: Time Series Analysis

School of Computing and Information Technology
University of Wollongong Australia

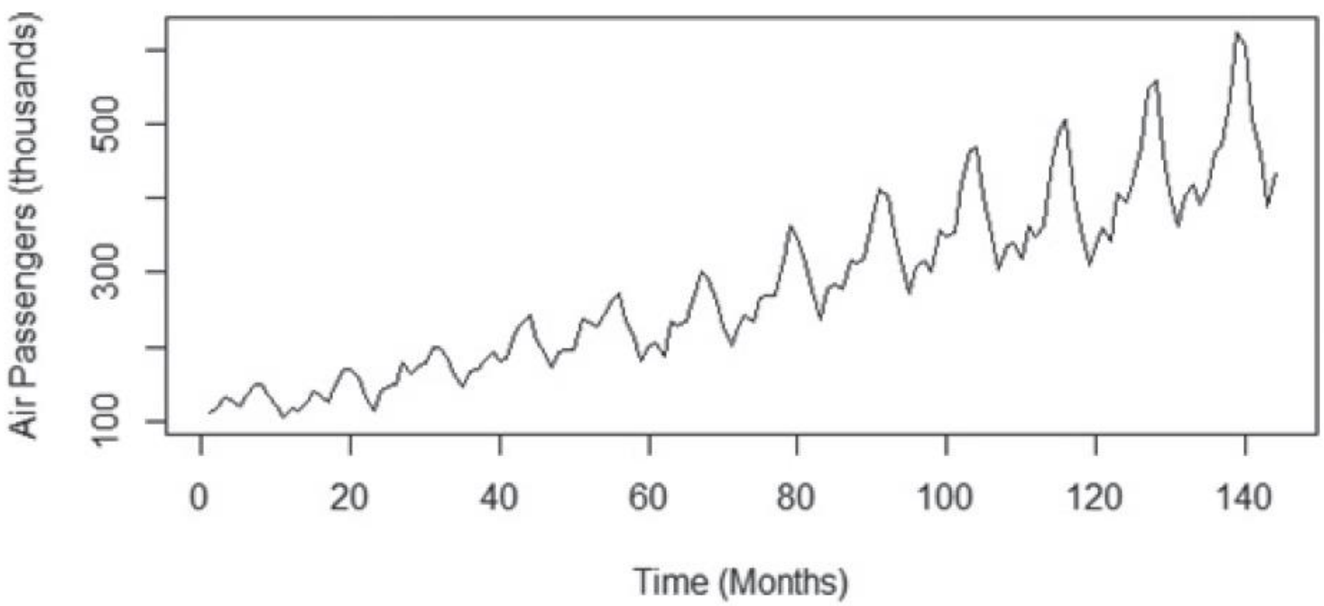
Advanced Analytical Theory and Methods: Time Series Analysis

- Overview of Time Series Analysis
- Box-Jenkins Methodology
- Autoregressive (AR) Models
- Moving Average (MA) Models
- Building and evaluating ARIMA models
- Additional Models

All the figures, tables and codes are from the book "Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data" unless indicated otherwise.

Advanced Analytical Theory and Methods: Time Series Analysis

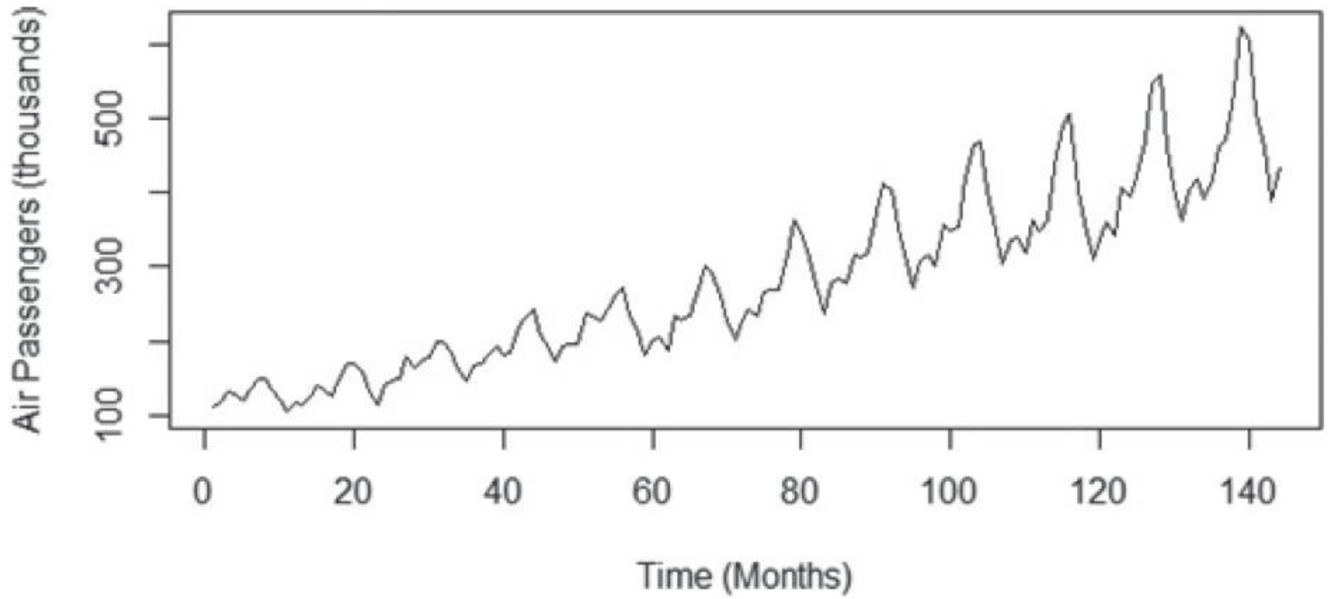
- Time series
  - An ordered sequence of equally spaced values over time



Advanced Analytical Theory and Methods: Time Series Analysis

- Applications in finance, economics, biology, engineering, retail, manufacturing, etc.
- Understand the underlying process that generates the observed data
- **Forecast** and monitor the future data
- Three examples
  - Retail sales, Spare parts planning
  - Stock trading

Overview of Time Series Analysis



- Identify and model the structure of time series
- Forecast future values in the time series

Overview of Time Series Analysis

- Box-Jenkins methodology
  1. Condition data and select a model
     - Identify and account for any trends or seasonality in the time series
     - Examine the remaining time series and determine a suitable model
  2. Estimate the model parameters
  3. Assess the model and return to Step 1, if needed

Overview of Time Series Analysis

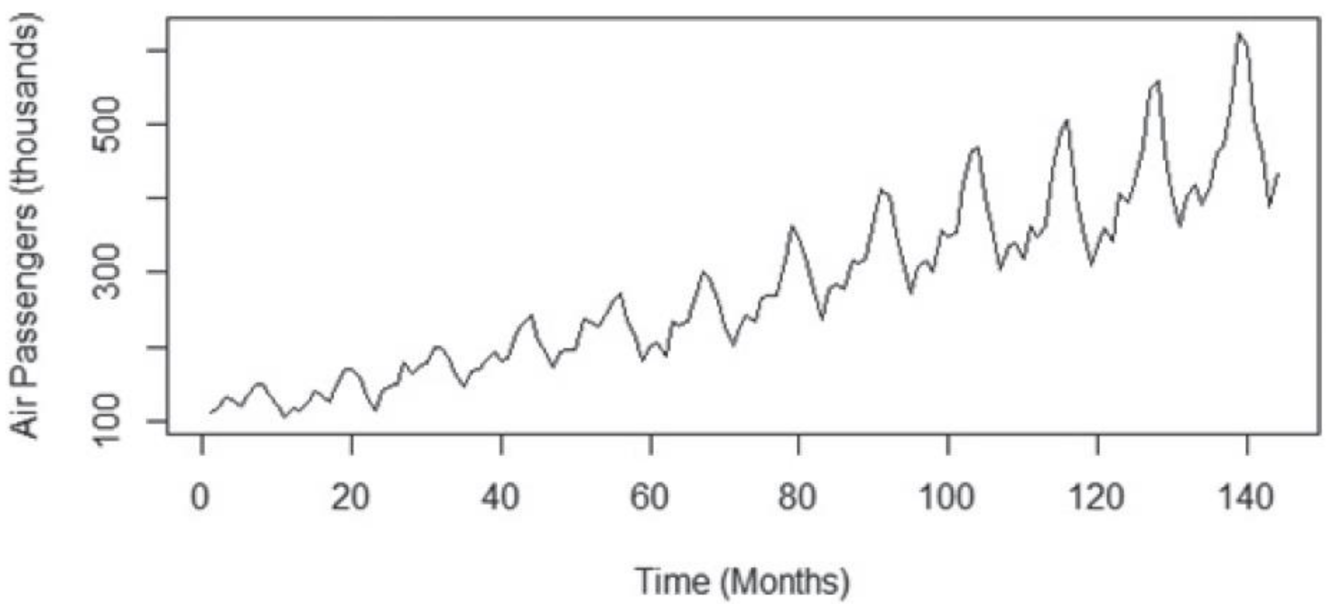
- A time series can consists of
  - Trend, Seasonality, Cyclic, and Random
- Trend
  - The long-term movement in a time series
  - The values increase or decrease over time
- Seasonality
  - The fixed, periodic fluctuation over time
  - Often related to the calendar

Overview of Time Series Analysis

- A time series can consists of
  - Trend, Seasonality, Cyclic, and Random
- Cyclic
  - Periodic, but not fixed fluctuation over time
  - Say, the boom-bust cycles of the economy
- Random
  - What remains after the above three components
  - Noise + underlying structure to be modelled

Overview of Time Series Analysis

- A time series can consists of
  - Trend, Seasonality, Cyclic, and Random

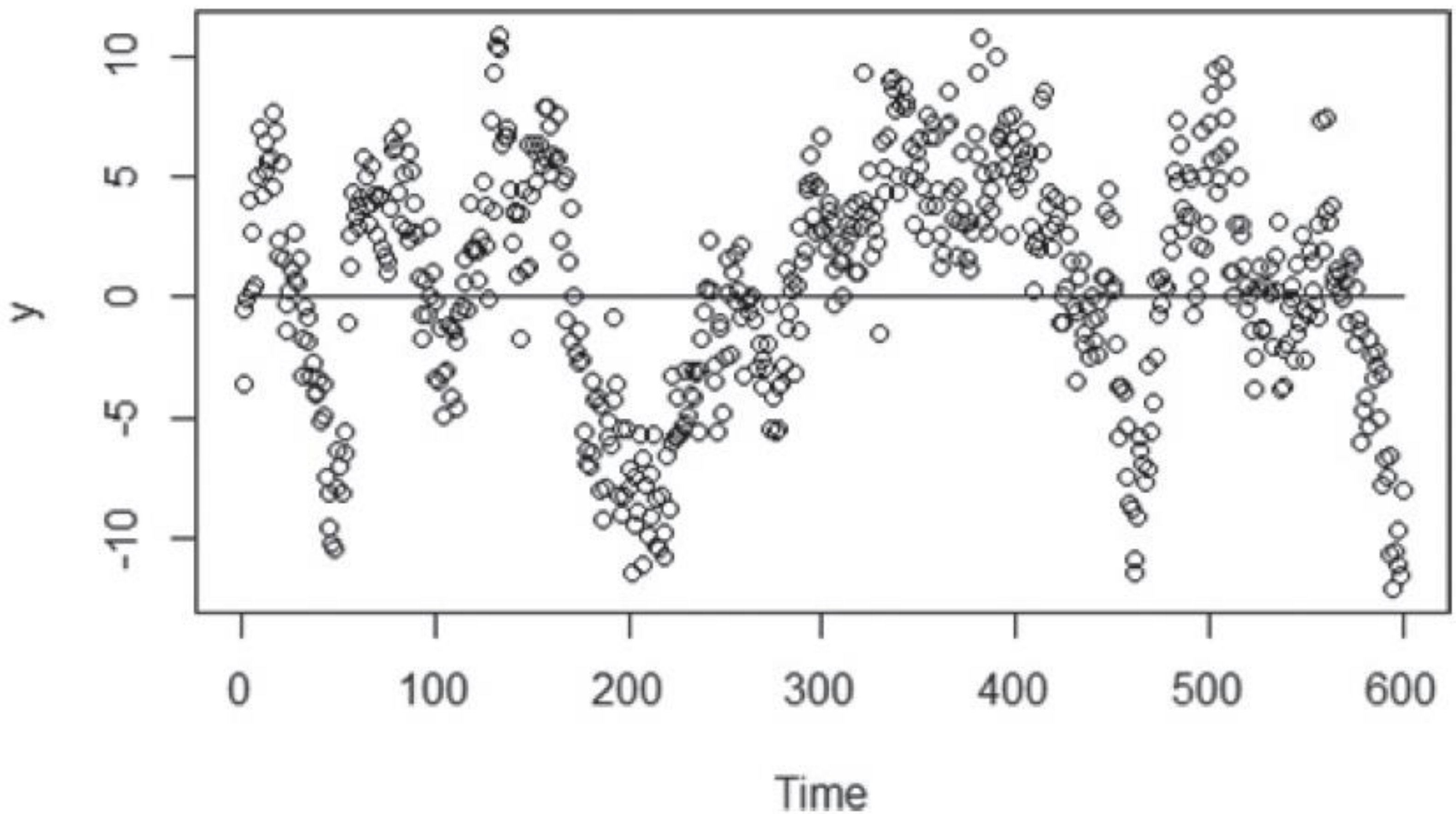


A plot of the monthly number of international airline passengers over a 12-year period

ARIMA Model

- ARIMA
  - AutoRegressive Integrated Moving Average
  - Shall be applied to stationary time series
- Stationary time series
  - The mean of y_t is a constant over time
  - The variance of y_t is finite (i.e., $cov(y_t, y_t)$)
  - The covariance of y_t and y_{t+h} (i.e., $cov(y_t, y_{t+h})$) depend only on h

ARIMA Model

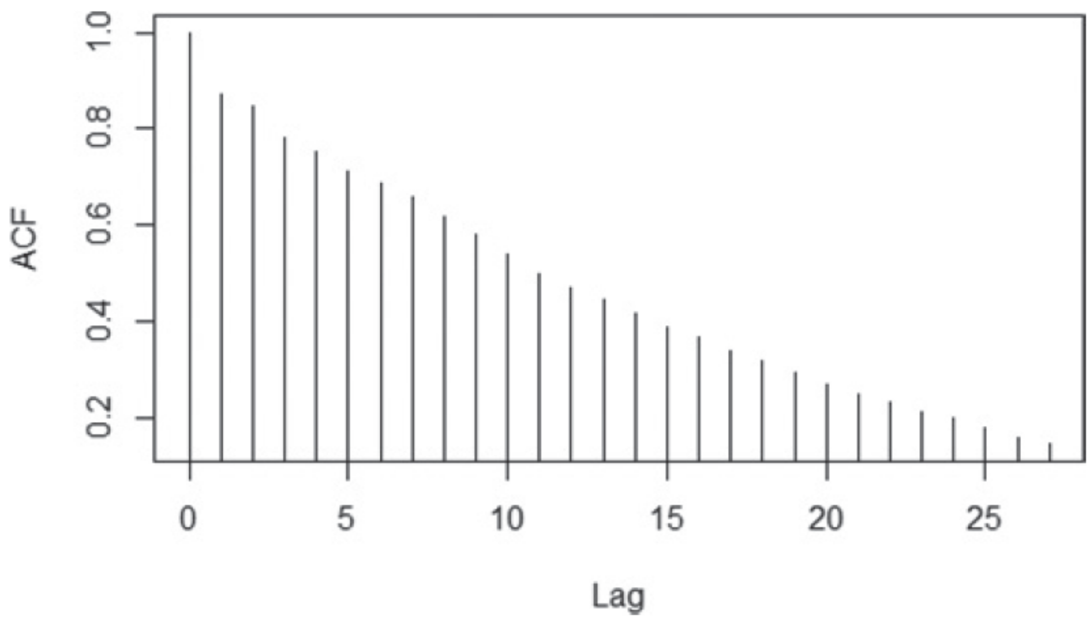


Flat looking (No trend); Constant variance (Similar scattering); Constant covariance over time

ARIMA Model

- Autocorrelation function (ACF) ([-1, +1])

$$ACF(h) = \frac{cov(y_t, y_{t+h})}{\sqrt{cov(y_t, y_t)\, cov(y_{t+h}, y_{t+h})}} = \frac{cov(h)}{cov(0)}$$



https://www.itl.nist.gov/div898/handbook/eda/section3/eda35c.htm

ARIMA Model

- Autoregressive (AR) Models
  - For a stationary time series, AR(p) is expressed as

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

where δ is a constant for a nonzero-centered time series:

ϕ_j is a constant for j = 1, 2, . . ., p

y_{t-j} is the value of the time series at time $t - j$

$\phi_p \neq 0$

$\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ for all t

A point y_t is a linear combination of the prior p values

ARIMA Model

- How to identify the order p in AR(p)?
- Partial autocorrelation function (PACF)
  - "The PACF of an AR(p) process is zero at lag p + 1 and greater"
  - So, let check the PACF of our data and find the lag after which the PACF becomes zero
- Why partial autocorrelation?
  - Even the simple AR(1) model leads to considerable autocorrelation

ARIMA Model

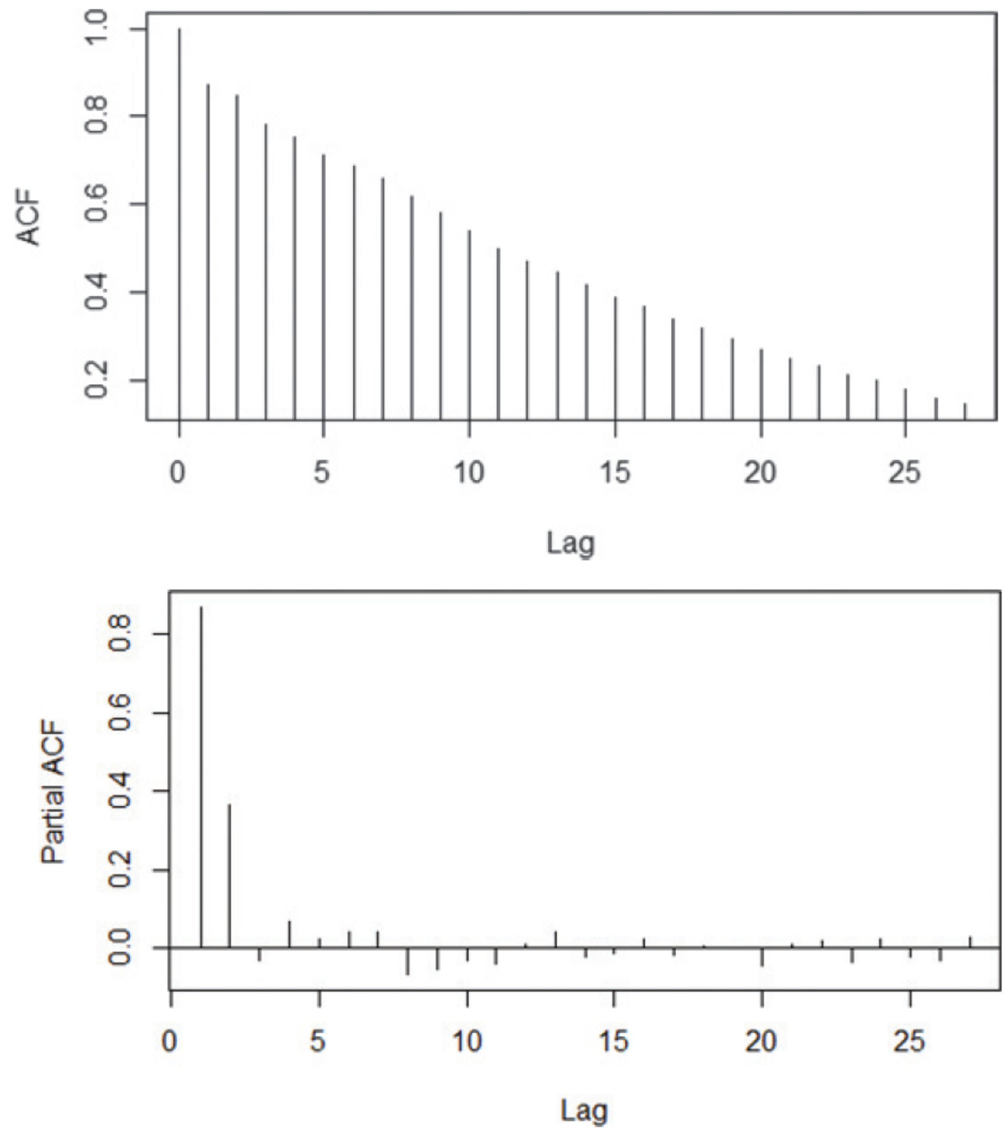
- Partial autocorrelation function (PACF)

$$
\begin{aligned}
PACF(h) &= corr(y_t - y_t^*, y_{t+h} - y_{t+h}^*) \text{ for } h \geq 2 \\
&= corr(y_t, y_{t+1}) \qquad\qquad \text{for } h = 1
\end{aligned}
$$

$$
\begin{aligned}
\text{where } y_t^* &= \beta_1 y_{t+1} + \beta_2 y_{t+2} \cdots + \beta_{h-1} y_{t+h-1}, \\
y_{t+h}^* &= \beta_1 y_{t+h-1} + \beta_2 y_{t+h-2} \cdots + \beta_{h-1} y_{t+1}, \text{ and}
\end{aligned}
$$

the $h-1$ values of the βs are based on linear regression

ARIMA Model


ACF
1.0
0.8
0.6
0.4
0.2
0
5
10
15
20
25
Lag
Partial ACF
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
Lag

ARIMA Model

- Moving Average (MA) models
  - For a time series, y_t, centred at zero, a MA(q) is expressed as

$$y_t = \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \ldots + \Theta_q \varepsilon_{t-q}$$

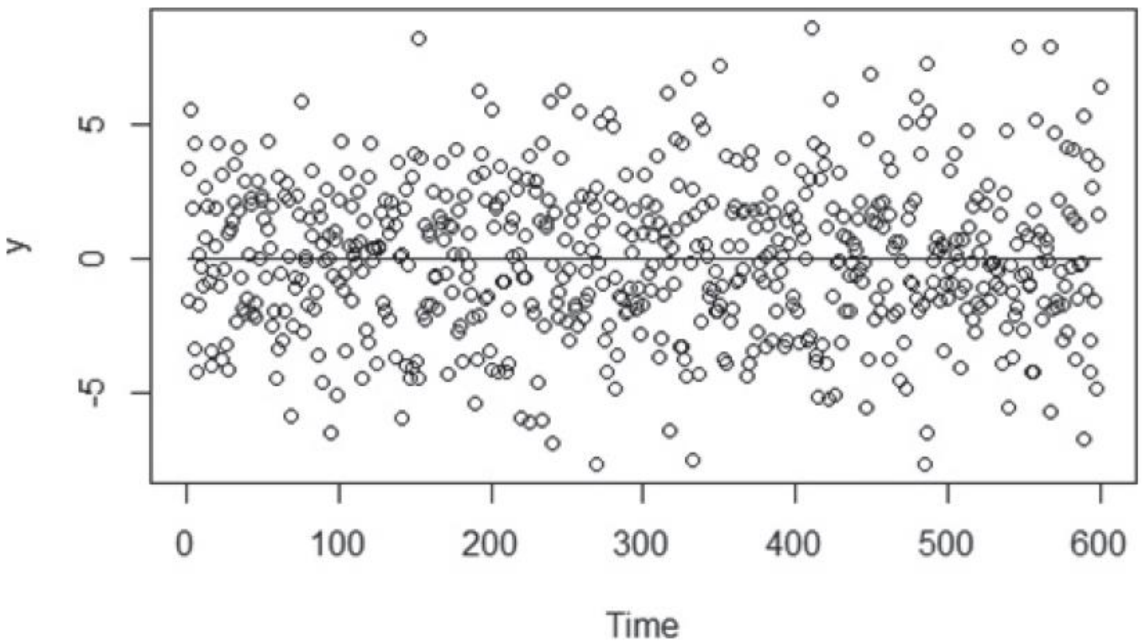
where θ_k is a constant for k = 1, 2, ..., q
$\theta_q \neq 0$
$\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ for all t

ARIMA Model

- Moving Average (MA) models
  - A linear regression of y_t against current and previous (observed) white noise error terms or random shocks
- Two differences from AR(p) models
  - ε_{t-1} are propagated to y_t directly
  - ε_t only affects the current and future p values

ARIMA Model

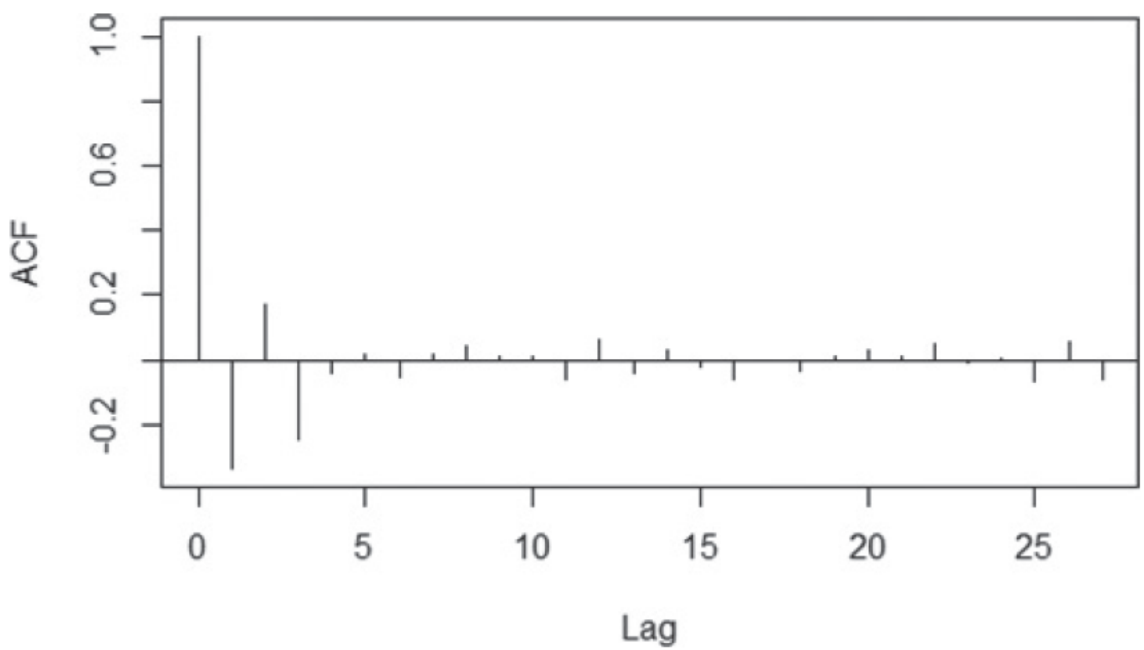
- Finite MA models are always stationary



$$y_t = \varepsilon_t - 0.4\,\varepsilon_{t-1} + 1.1\varepsilon_{t-2} - 2.5\,\varepsilon_{t-3} \quad \text{where } \varepsilon_t \sim N(0,1)$$

ARIMA Model

- The ACF of this MA(3) time series



$$y_t = \varepsilon_t - 0.4\,\varepsilon_{t-1} + 1.1\varepsilon_{t-2} - 2.5\,\varepsilon_{t-3} \quad \text{where } \varepsilon_t \sim N(0,1)$$

* ACF can help to identify q in MA(q)

ARIMA Model

- Why is this case?

$$
\begin{aligned}
y_t &= \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} \\
y_{t-1} &= \varepsilon_{t-1} + \theta_1 \varepsilon_{t-2} + \theta_2 \varepsilon_{t-3} + \theta_3 \varepsilon_{t-4} \\
y_{t-2} &= \varepsilon_{t-2} + \theta_1 \varepsilon_{t-3} + \theta_2 \varepsilon_{t-4} + \theta_3 \varepsilon_t \\
y_{t-3} &= \varepsilon_{t-3} + \theta_1 \varepsilon_{t-4} + \theta_2 \varepsilon_{t-5} + \theta_3 \varepsilon \\
y_{t-4} &= \varepsilon_{t-4} + \theta_1 \varepsilon_{t-5} + \theta_2 \varepsilon_{t-6} + \theta_3
\end{aligned}
$$

Because y_t and y_{t-4} have no overlapping at all

$$y_t = \varepsilon_t - 0.4\,\varepsilon_{t-1} + 1.1\varepsilon_{t-2} - 2.5\,\varepsilon_{t-3} \quad \text{where } \varepsilon_t \sim N(0, 1)$$

ARIMA Model

- AR(p) and MA(q) are often combined into one model for time series, resulting in ARMA(p,q).

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p}$$
$$+ \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$$

where δ is a constant for a nonzero-centered time series

ϕ_j is a constant for j = 1, 2, ..., p

$\phi_p \neq 0$

θ_k is a constant for k = 1, 2, ..., q

$\theta_q \neq 0$

$\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ for all t

ARIMA Model

- Transformations to stationary time series
  - Train a linear or higher-order regression model to remove trends
  - Or, compute the difference between successive y-value, which is known as differencing

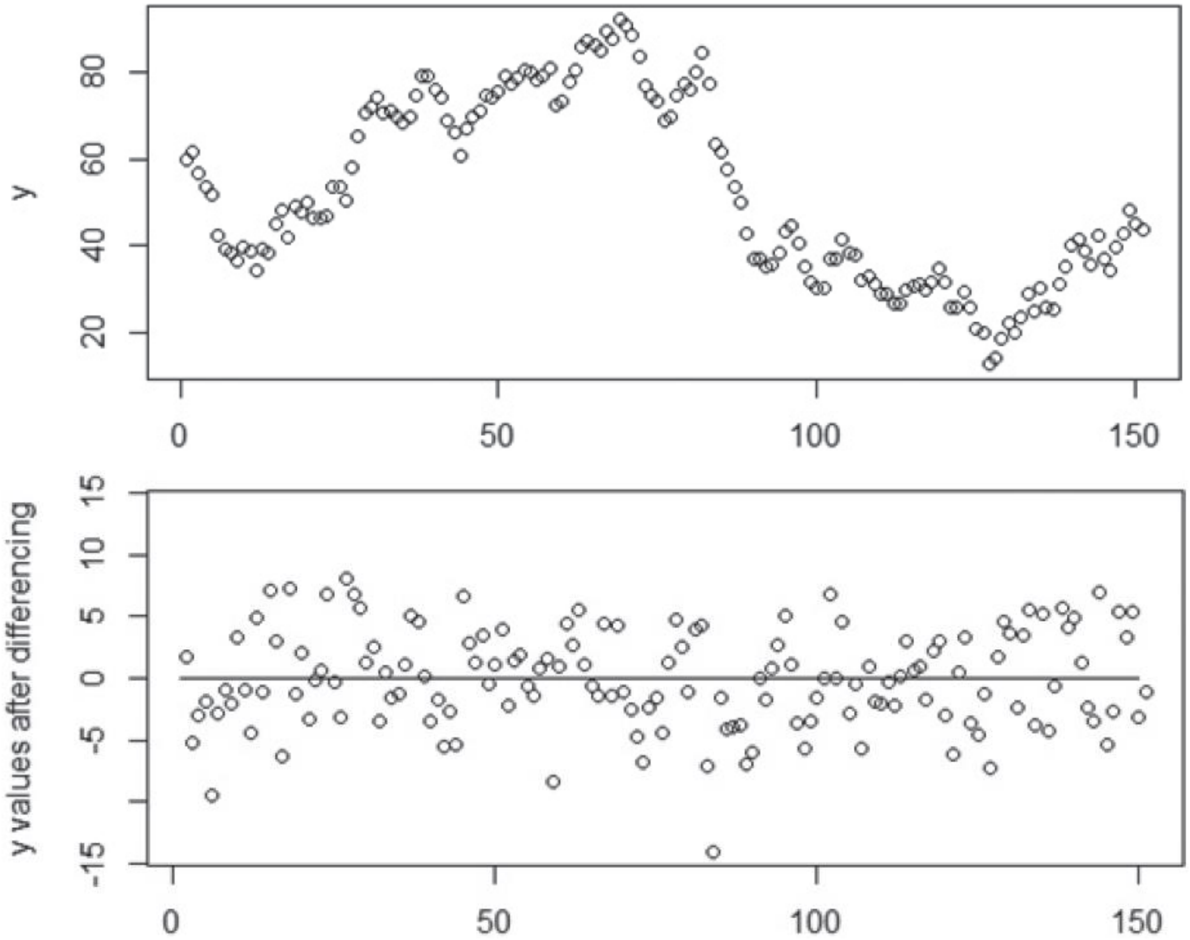
$$d_t = y_t - y_{t-1} \quad \text{for } t = 2, 3, \ldots, n$$

  - If still not stationary, apply differencing more times

$$\begin{aligned} d_{t-1} - d_{t-2} &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned}$$

ARIMA Model

- Transformations to stationary time series

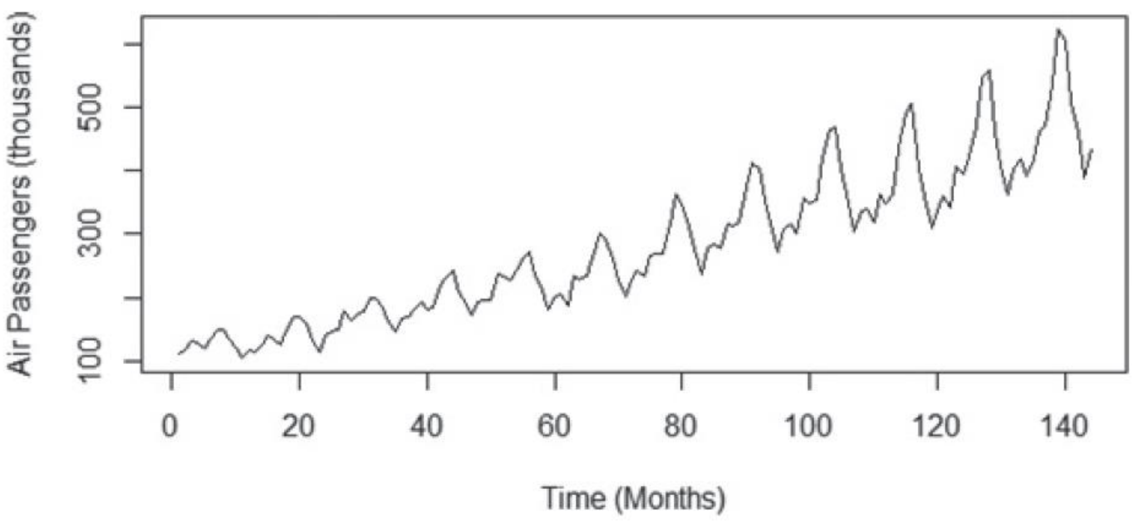


ARIMA Model

- ARIMA model
  - Autoregressive Integrated Moving Average
  - Differencing is included in ARMA model
- ARIMA(p,d,q)
  - ARMA(p,q) mode is applied after applying differencing d times

ARIMA Model

- Seasonal ARIMA model
  - ARIMA(p,d,q) x $(P,D,Q)_s$
- It is used to account for seasonal patterns
- An alternative to regression analysis



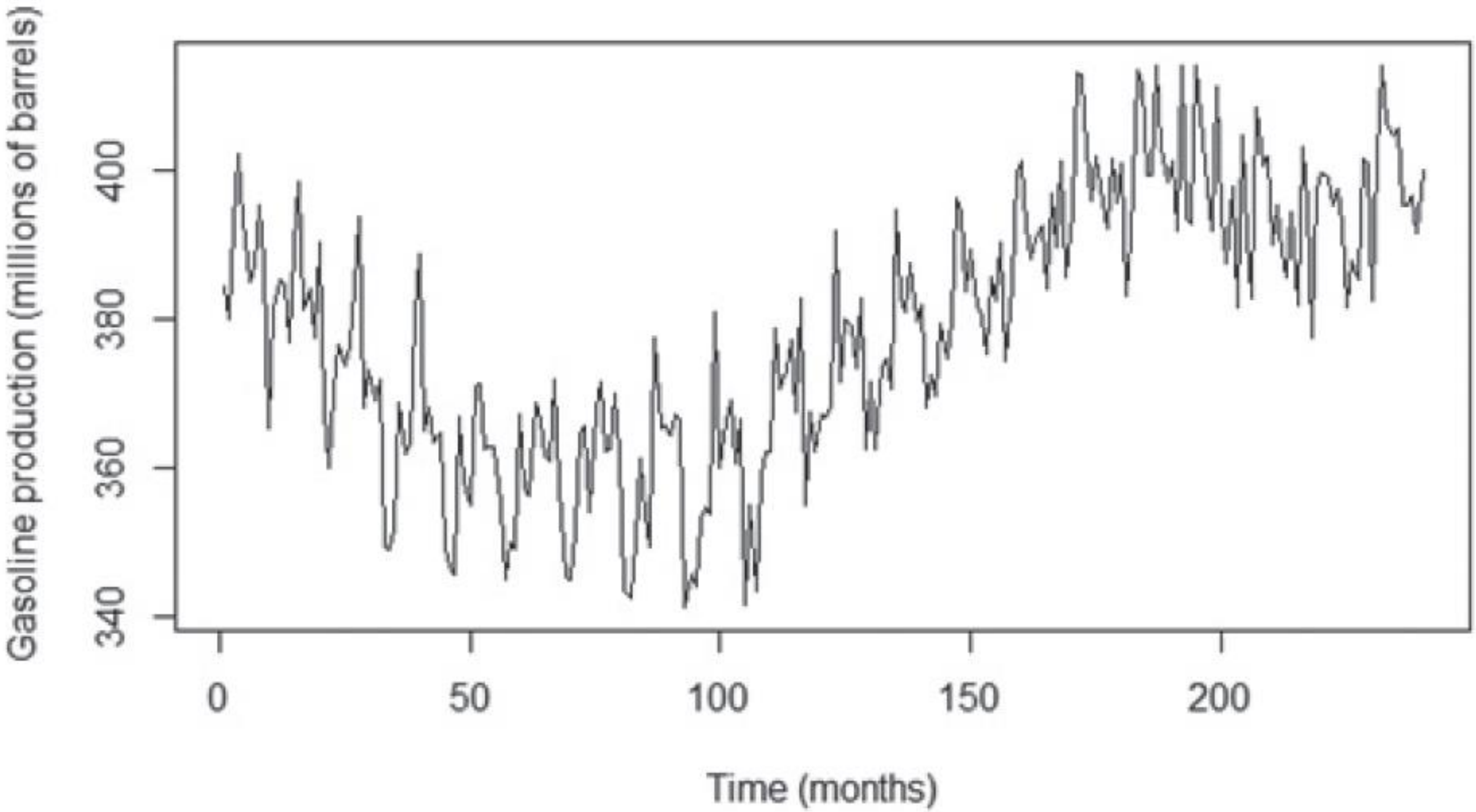
Building and evaluating ARIMA models

- An example
  - Monthly gasoline production (millions of barrels)
- Need to forecast short-term production

Building and evaluating ARIMA models

```
library(forecast)

# read in gasoline production time series
# monthly gas production expressed in millions of barrels
gas_prod_input <- as.data.frame( read.csv("c:/data/gas_prod.csv") )

# create a time series object
gas_prod <- ts(gas_prod_input[,2])

#examine the time series
plot(gas_prod, xlab = "Time (months)",
     ylab = "Gasoline production (millions of barrels)")
```

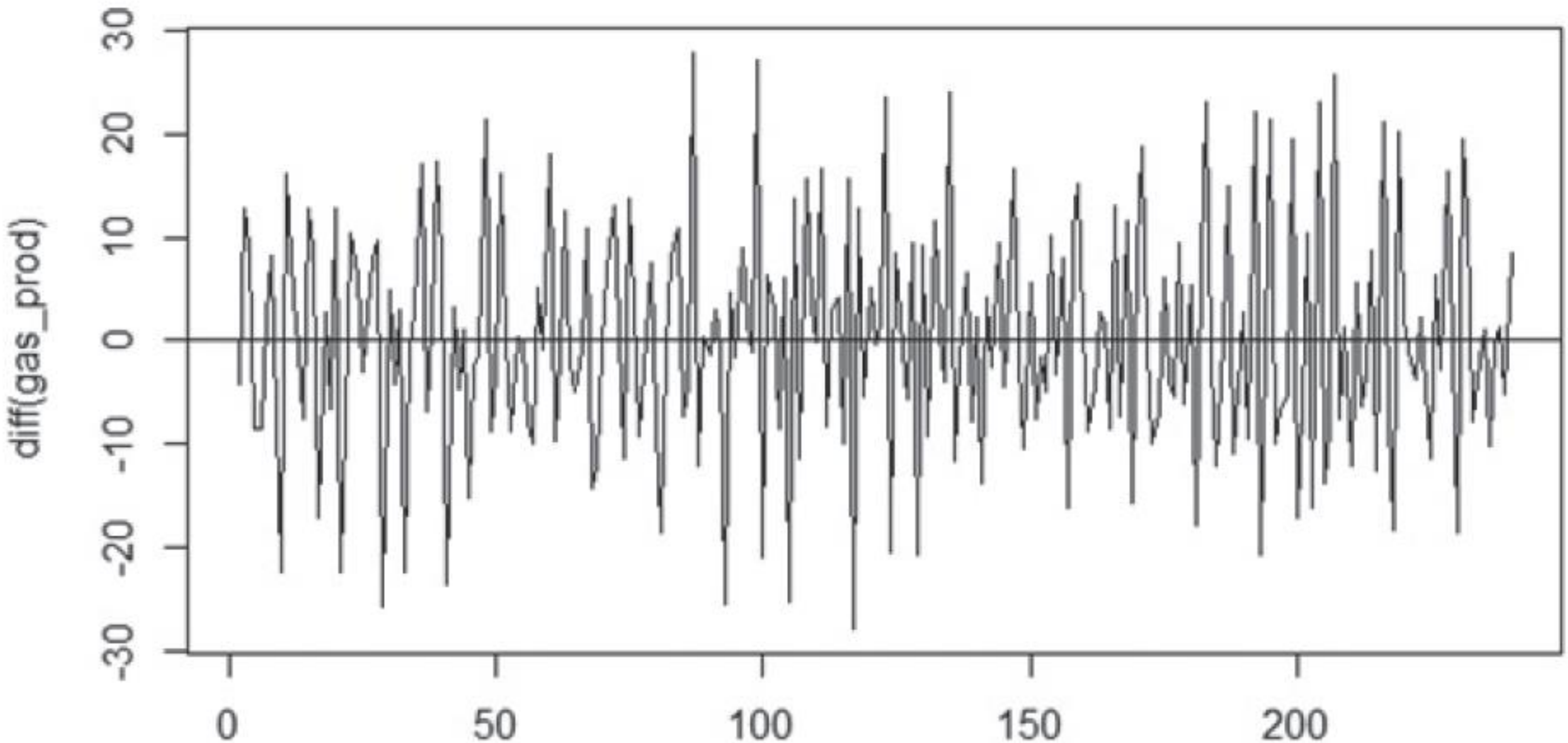
Building and evaluating ARIMA models



What is your observation on this time series?

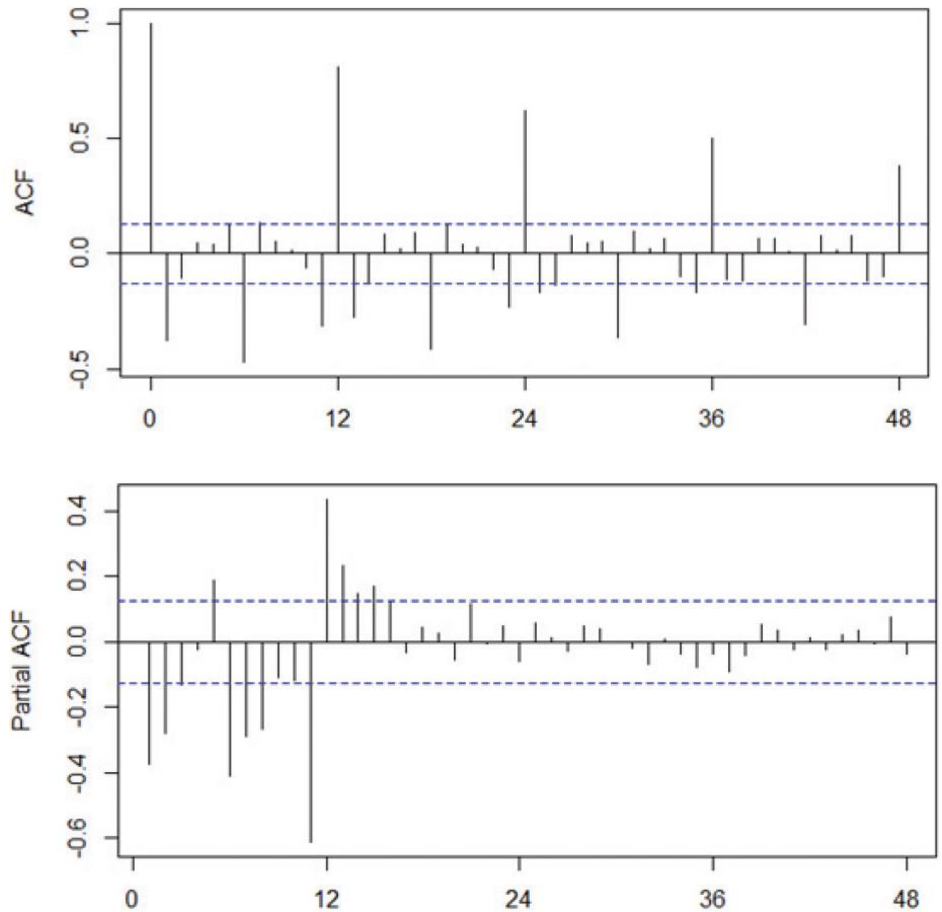
Building and evaluating ARIMA models

```
plot(diff(gas_prod))
abline(a=0, b=0)
```



Building and evaluating ARIMA models

```
# examine ACF and PACF of differenced series
acf(diff(gas_prod), xaxp = c(0, 48, 4), lag.max=48, main="")
pacf(diff(gas_prod), xaxp = c(0, 48, 4), lag.max=48, main="")
```



Building and evaluating ARIMA models

- Differencing + a seasonal AR(1) model
  - ARIMA(0,1,0) x $(1,0,0)_{12}$

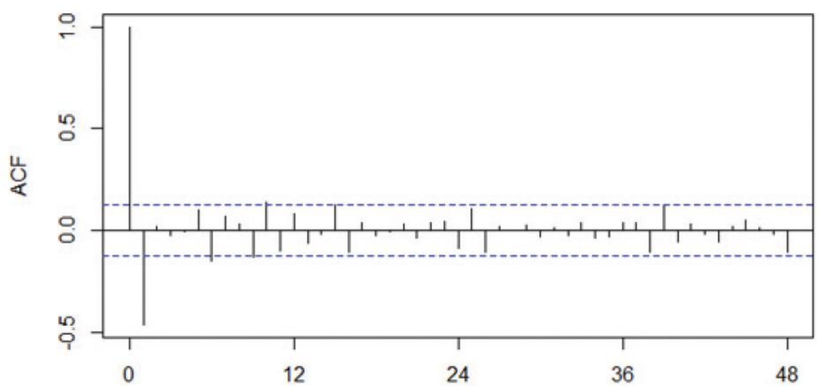
```
arima_1 <- arima (gas_prod,
                  order=c(0,1,0),
                  seasonal = list(order=c(1,0,0),period=12))
arima_1
```

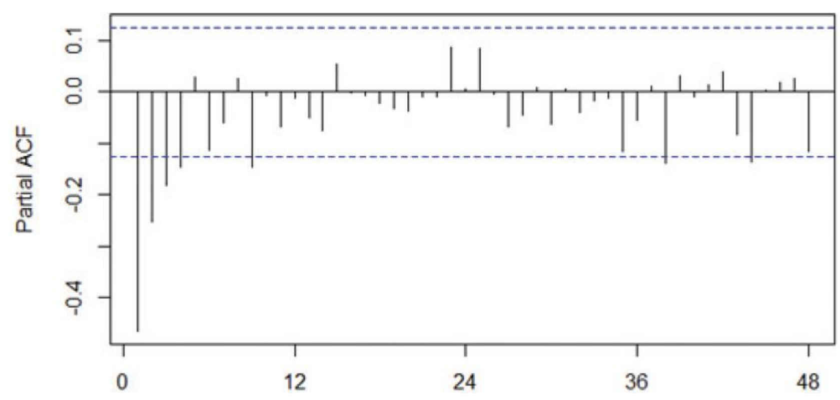
```
Series: gas_prod
ARIMA(0,1,0)(1,0,0)[12]

Coefficients:
        sar1
      0.8335

s.e.   0.0324

sigma^2 estimated as 37.29:  log likelihood=-778.69
AIC=1561.38   AICc=1561.43   BIC=1568.33
```

Building and evaluating ARIMA models

- Examine the residuals after fitting ARIMA(0,1,0) x (1,0,0)$_{12}$

```
# examine ACF and PACF of the (0,1,0)x(1,0,0)12 residuals
acf(arima_1$residuals, xaxp = c(0, 48, 4), lag.max=48, main="")
pacf(arima_1$residuals, xaxp = c(0, 48, 4), lag.max=48, main="")
```





Building and evaluating ARIMA models

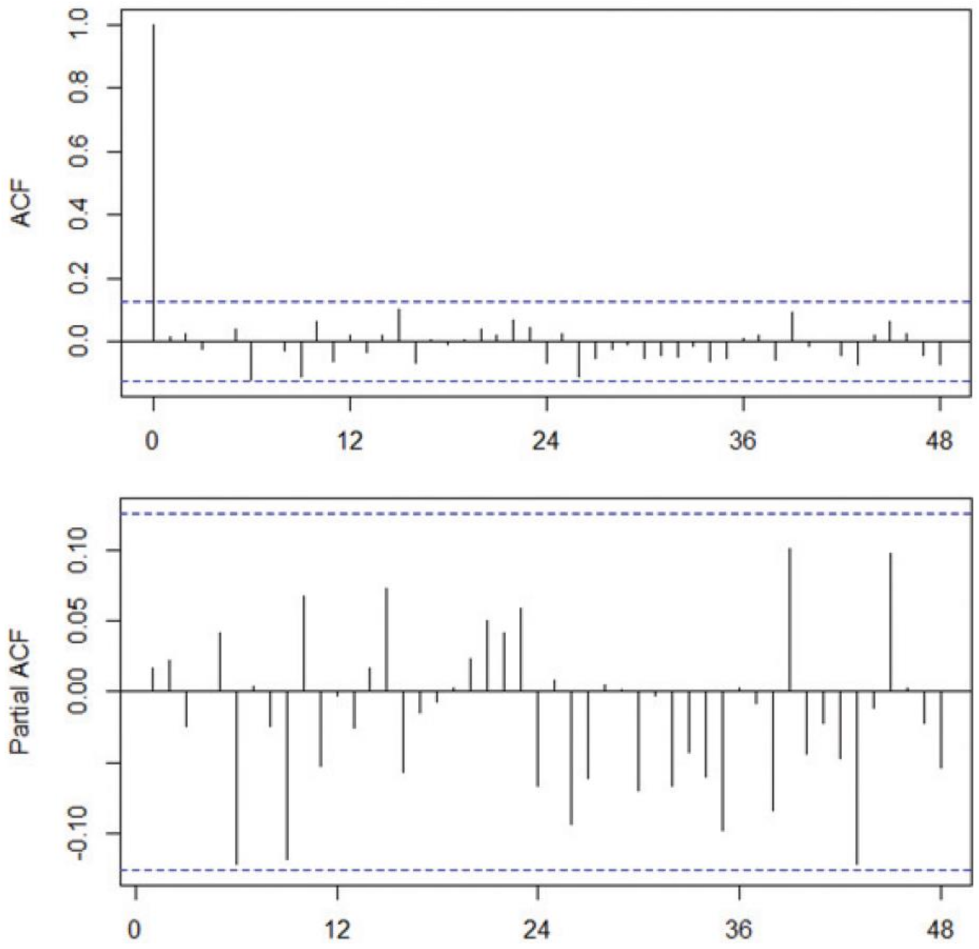
- Fit with a new model ARIMA(0,1,1) x $(1,0,0)_{12}$

```
arima_2 <- arima (gas_prod,
                  order=c(0,1,1),
                  seasonal = list(order=c(1,0,0),period=12))
arima_2
```

```
Series: gas_prod
ARIMA(0,1,1)(1,0,0)[12]

Coefficients:
          ma1    sar1
      -0.7065  0.8566
s.e.   0.0526  0.0298

sigma^2 estimated as 25.24:  log likelihood=-733.22
AIC=1472.43   AICc=1472.53   BIC=1482.86
```

Building and evaluating ARIMA models

- Fit with a new model ARIMA(0,1,1) x $(1,0,0)_{12}$



Building and evaluating ARIMA models

- Comparing fitted time series models
  - Model coefficients are estimated via Maximum Likelihood Estimation (MLE)
  - R provides several measures based on MLE value
    - AIC (Akaike Information Criterion)
    - AICc (AIC corrected)
    - BIC (Bayesian Information Criterion)
  - The preferred model has the smallest AIC, AICc, or BIC value

Building and evaluating ARIMA models

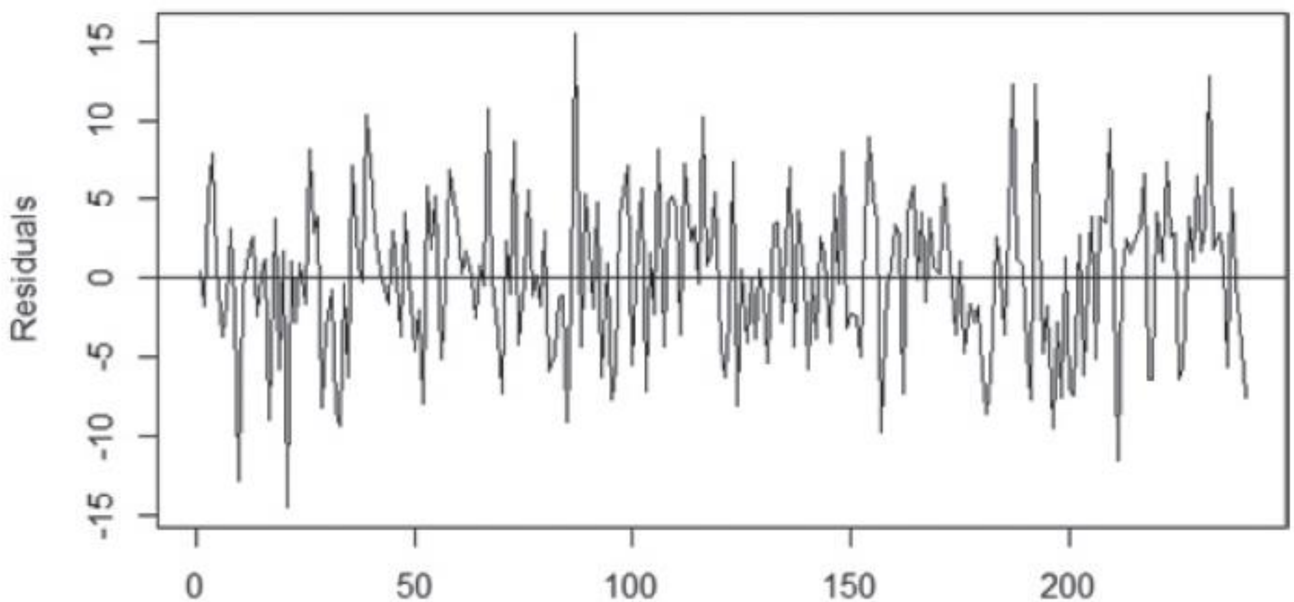
- Comparing fitted time series models

ARIMA Model $(p,d,q) \times (P,Q,D)_S$	AIC	AICc	BIC
$(0,1,0) \times (1,0,0)_{12}$	1561.38	1561.43	1568.33
$(0,1,1) \times (1,0,0)_{12}$	1472.43	1472.53	1482.86
$(0,1,2) \times (1,0,0)_{12}$	1474.25	1474.42	1488.16
$(1,1,0) \times (1,0,0)_{12}$	1504.29	1504.39	1514.72
$(1,1,1) \times (1,0,0)_{12}$	1474.22	1474.39	1488.12

Building and evaluating ARIMA models

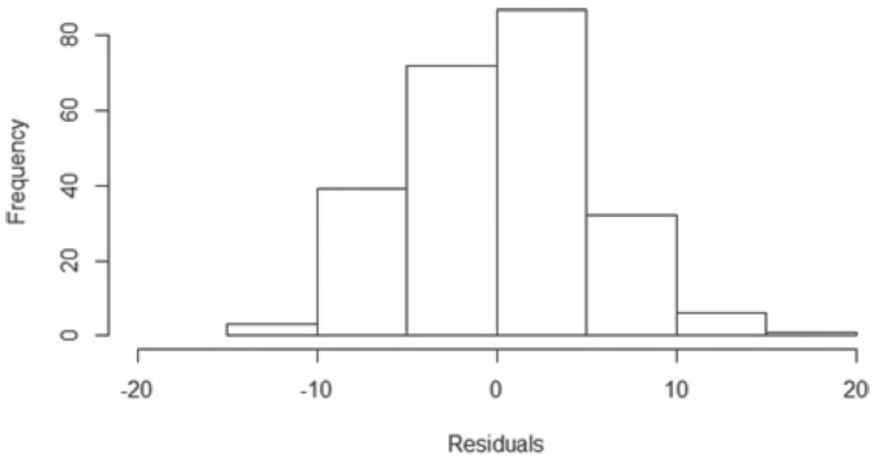
- Normality and constant variance

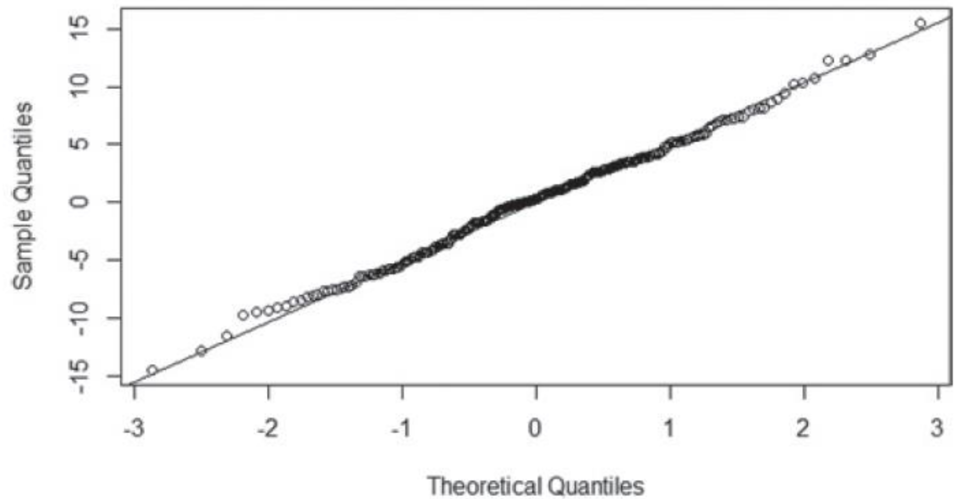
```
plot(arima_2$residuals, ylab = "Residuals")
abline(a=0, b=0)

hist(arima_2$residuals, xlab="Residuals", xlim=c(-20,20))

qqnorm(arima_2$residuals, main="")
qqline(arima_2$residuals)
```



Building and evaluating ARIMA models

- Normality and constant variance



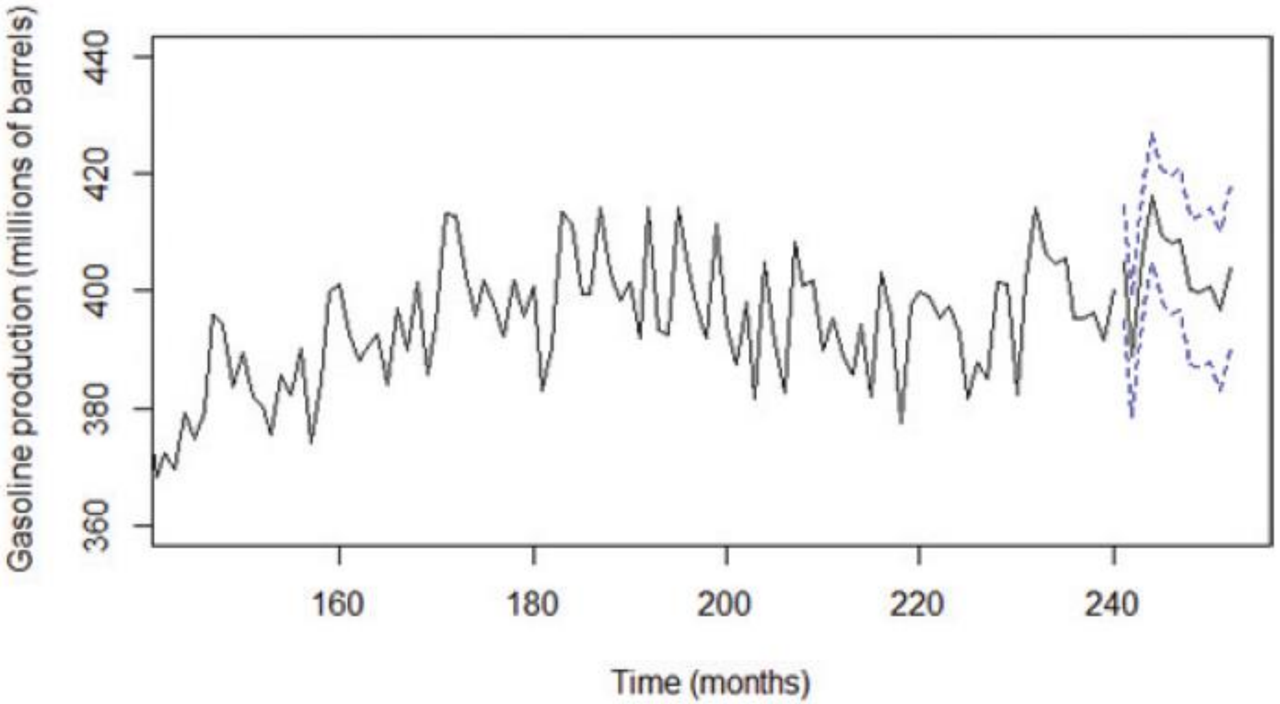


- If not satisfied, need to transform the time series before model fitting
  – Say, apply a logarithm function

Building and evaluating ARIMA models

- Forecast
  - The next 12 months of gasoline production

```
#predict the next 12 months
arima_2.predict <- predict(arima_2,n.ahead=12)
```



Reasons to Choose and Cautions

- Time series analysis is based on historical data for the variable of interest
- Forecasting process is simplified with ARIMA
  - What if regression analysis is used instead?
- Disadvantage
  - Not know underlying variables affect the outcome
- Cautions
  - Impact of severe shocks to the system
  - Shall only be used for short-term forecast

Additional Methods

- Autoregressive Moving Average with Exogenous inputs (ARMAX)
- Spectral analysis
- Generalised Autoregressive Conditionally Heteroscedastic (GARCH)
- Kalman filtering
- Multivariate time series analysis (VARIMA)

Summary

- Box-Jenkins methodology
  - Condition, identify, model
  - Stationary time series, ACF, PACF plot
- ARIMA model
  - AR, MA, Differencing
  - ARIMA(p,d,q) x $(P,D,Q)_s$
- Build ARIMA model in R
- Assess and refine the model

Reference books

1. The Analysis of Time Series: An Introduction, Sixth Edition, Chris Chatfield
2. Time Series Analysis and Its Applications With R Examples ,Shumway, Robert H., Stoffer, David S.

Q&A

Images Courtesy of Google Image