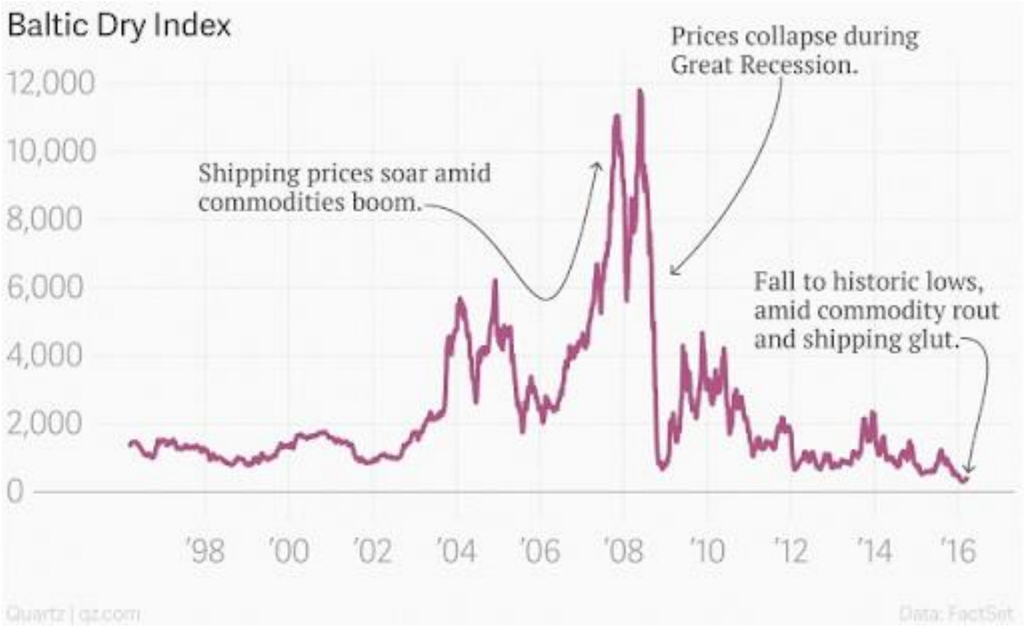
PNU Industrial Data Science

Time series analysis

시계열 분석

Prof. Hyerim Bae

데이터 분석으로 돈벌기


Baltic Dry Index
12,000
10,000
8,000
6,000
4,000
2,000
0
Shipping prices soar amid commodities boom.
Prices collapse during Great Recession.
Fall to historic lows, amid commodity rout and shipping glut.
'98
'00
'02
'04
'06
'08
'10
'12
'14
'16
Quartz | qz.com
Data: FactSet

Contents

산업데이터과학은 산업현장에서 수집된 데이터를 분석하는데 필요한 기초 소양을 강의합니다.

Main ideas

- Forecast future values of a time series
- Distinction between forecasting (main focus) and describing/explaining
- Four components of time series:
  - Level
  - Trend
  - Seasonality
  - noise

Explain vs. Predict

Explanation is the goal of "time series analysis"

- Models are based on causal argument
- Models are not "black-box"

Forecasting (our focus) seeks to predict future values

Control

Time Series Components

Level: 평균

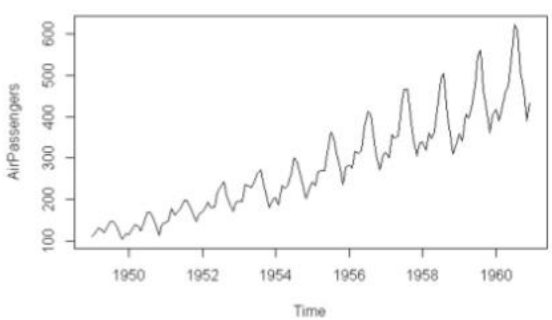
Trend: 추세

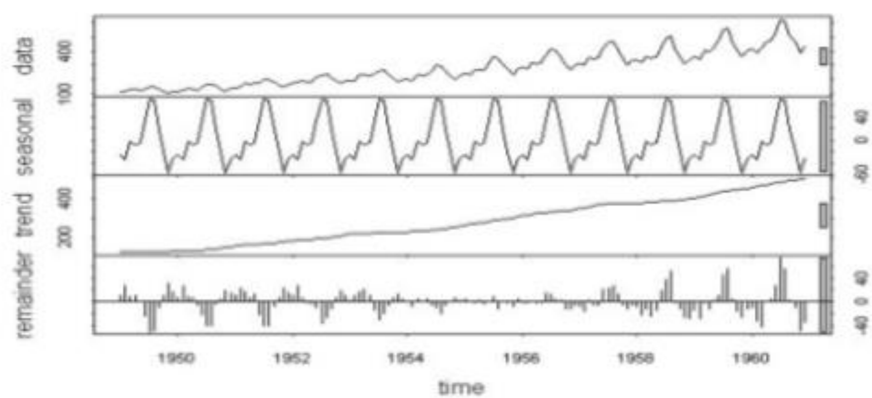
Seasonality: 계절성

Noise: 기타변동

Decomposition of time series data

# of air passengers in 1949 ~ 1951

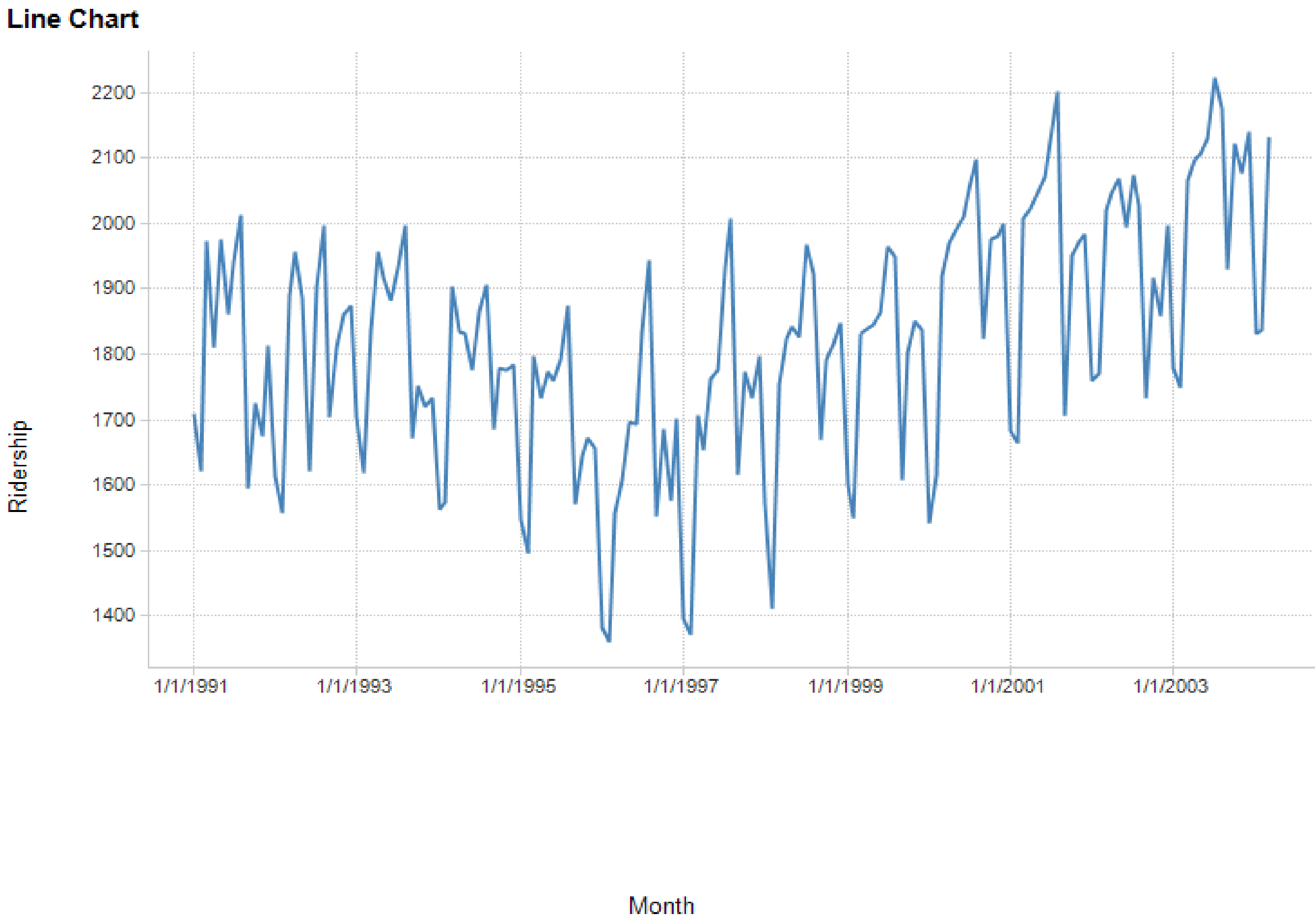




Example: Amtrak Ridership (monthly)

Level - about 1,800,000 passengers per month

Appears to have U-shaped trend



Zoom to 3 years (1997-1999)

Seasonality* appears:

Summer peaks

Noise:

Departure from the general level that is neither trend nor seasonality

*Seasonality is any cyclical pattern. Here it is seasons of the year, but could be any cyclical pattern (daily, weekly, monthly, etc.)

Amtrak Ridership – zoom to 3-years

Partitioning

Divide data into training portion and validation portion

- Test model on the validation portion

Random partitioning would leave holes in the data, which causes problems

- Forecasting methods assume regular sequential data

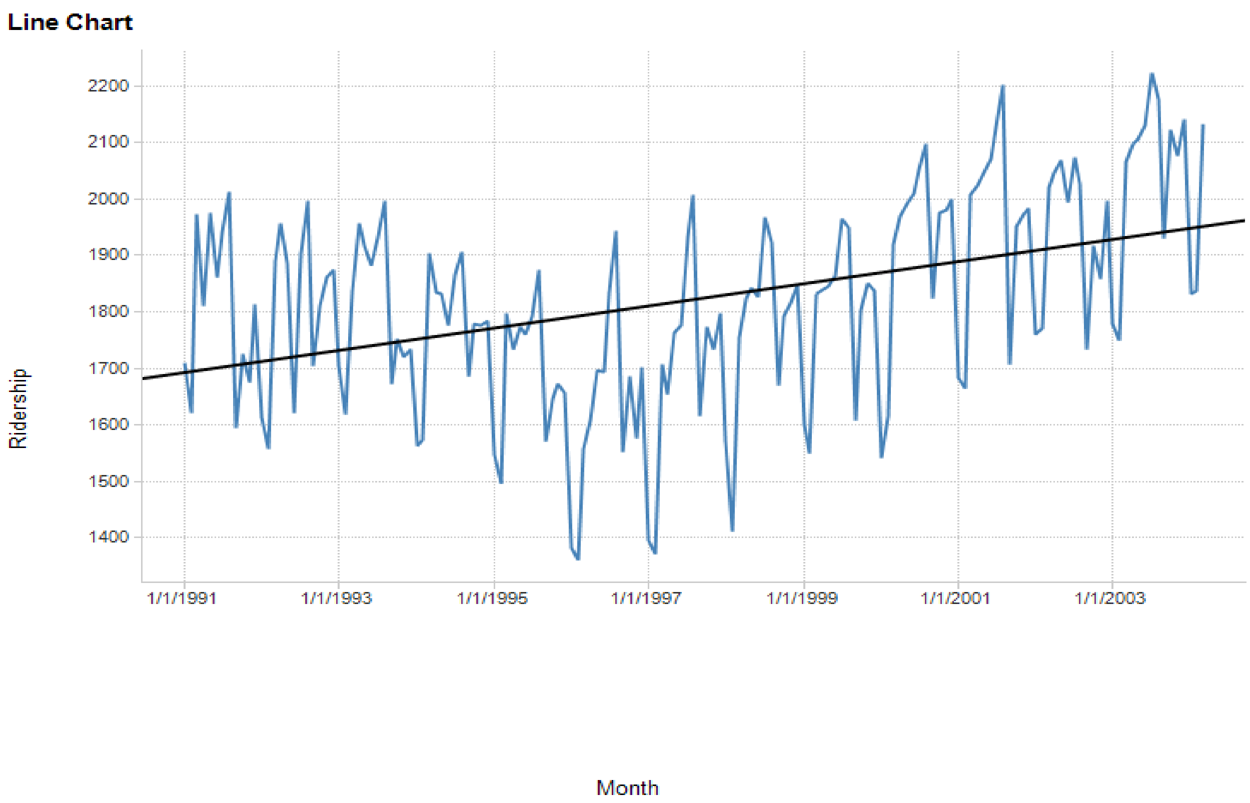
Instead of random selection, divide data into two parts

- Train on early data
- Validate on later data

TS by Regression

- Fit linear trend, time as predictor
- Modify & use also for non-linear trends
  - Exponential
  - Polynomial
- Can also capture seasonality

Linear fit to Amtrak ridership data
(Doesn't fit too well – more later)



The regression model

Ridership Y is a function of time (t) and noise (error = e)

$$Y_i = B_0 + B_1 * t + e$$

Thus we model 3 of the 4 components:

- Level (B_0)
- Trend* (B_1)
- Noise (e)

*Our trend model is linear, which we can see from the graph is not a good fit (more later)

Regression Output

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	1713.028809	27.08552361	0	477456500
t	1.2053107	0.31751993	0.00021544	384546.3125

Training Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
3869551.676	162.2451256	-3.84852E-05

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
529326.616	210.0251207	168.8524156

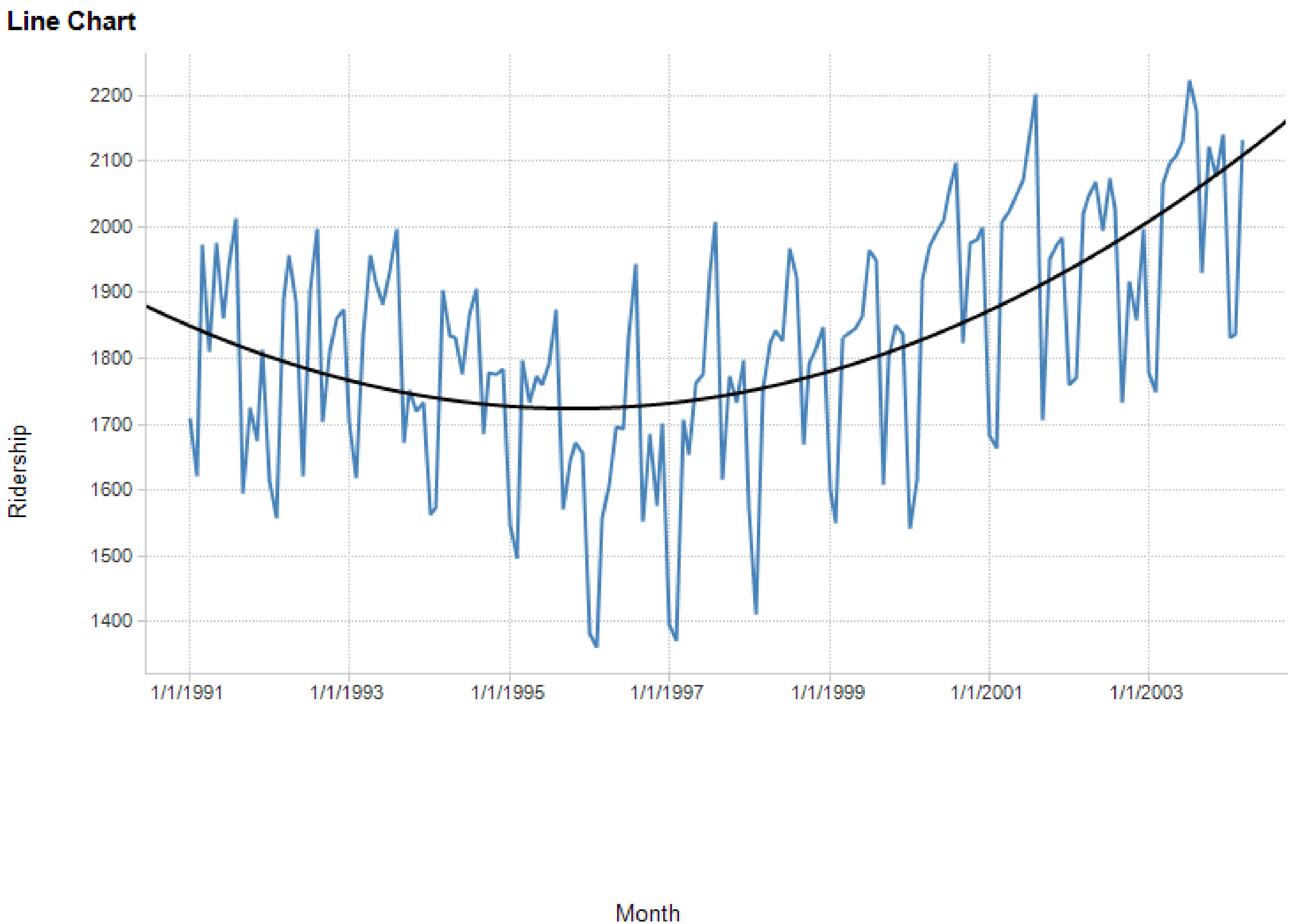
Polynomial Trend

Add additional predictors as appropriate

For example, for quadratic relationship add a t^2 predictor

Fit linear regression using both t and t^2

Quadratic fit to Amtrak data



Quadratic fit to Amtrak Data

Now appears to capture trend

Seasonality remains

Handling Seasonality

- Seasonality is any recurring cyclical pattern of consistently higher or lower values (daily, weekly, monthly, quarterly, etc.)

- Handle in regression by adding categorical variable for season, e.g.,

Month	Ridership	Season
Jan-91	1709	Jan
Feb-91	1621	Feb
Mar-91	1973	March
Apr-91	1812	April

Creating Binary dummies

Logistic regression software usually requires transforming categorical variables into dummies

To avoid multicollinearity problems, use m-1 dummies for m categories

regression output
coefficients for each season

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	1855.235962	33.95079803	0	477456500
season_Aug	139.3903351	48.01367569	0.00431675	483721.3125
season_Dec	-19.82307816	48.01367569	0.68036187	33314.77734
season_Feb	-288.9631348	47.08128357	0	665331.9375
season_Jan	-251.2854462	47.08128357	0.00000034	598841.0625
season_Jul	94.34428406	48.01367569	0.05147372	187691.7656
season_Jun	-10.11090946	48.01367569	0.83352947	11869.09277
season_Mar	11.57308865	47.08128357	0.80620199	48930.94922
season_May	31.24033737	48.01367569	0.51637506	114420.9141
season_Nov	-63.96651077	48.01367569	0.18502063	3121.062012
season_Oct	-54.12883377	48.01367569	0.26158884	14579.31641
season_Sep	-193.6371613	48.01367569	0.00009163	224972.1094

Seasonality types

Additive – described above (model shows amounts by which seasonal values exceed or fall below those in the reference season)

Multiplicative - (model shows percentages by which seasonal values exceed or fall below those in the reference season)

Proceed as above, but use $\log(Y)$ as output

Final model, Amtrak data

Incorporates trend and seasonality

13 predictors

- 11 monthly dummies
- t
- t^2

Regression output - coefficients

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	1932.998779	27.85863113	0	477456500
season_Aug	135.1726227	30.52143288	0.00001955	483721.3125
season_Dec	-29.65872955	30.53801155	0.33320817	33314.77734
season_Feb	-306.3078308	29.94875526	0	665331.9375
season_Jan	-267.444458	29.94642067	0	598841.0625
season_Jul	91.31225586	30.5189991	0.00330446	187691.7656
season_Jun	-12.04474545	30.51724434	0.69370645	11869.09277
season_Mar	-7.04482555	29.95207596	0.81441271	48930.94922
season_May	30.31717491	30.51618195	0.32228076	114420.9141
season_Nov	-72.26641083	30.53282547	0.01938256	3121.062012
season_Oct	-60.98049164	30.52834129	0.04781064	14579.31641
season_Sep	-199.1280975	30.52454758	0	224972.1094
t	-5.246521	0.58674908	0	398979.7188
t^2	0.0437566	0.00384071	0	725213.9375

Model Performance
(superior performance on validation data is unusual)

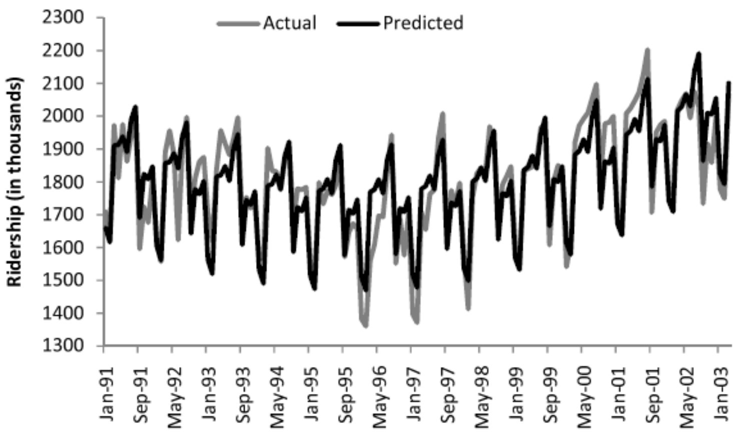
Training Data scoring - Summary Report

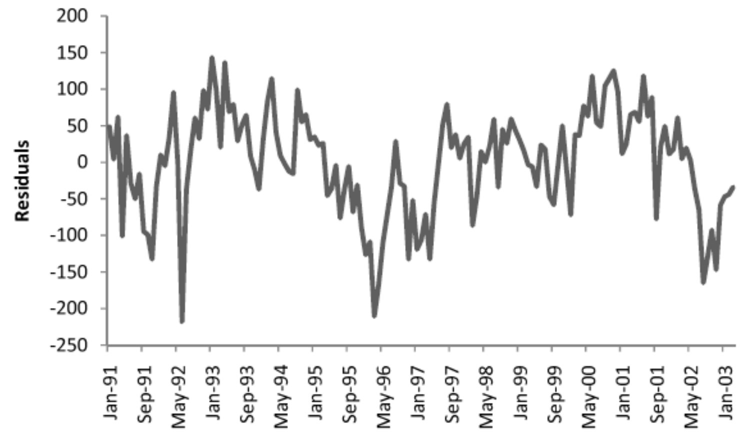
Total sum of squared errors	RMS Error	Average Error
743110.0191	71.0997201	-6.05149E-05

Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
30722.61731	50.59859789	-34.11397564

- Residuals
    - Actual vs. Predicted





Autocorrelation and ARIMA

Autocorrelation

Unlike cross-sectional data, time-series values are typically correlated with nearby values ("autocorrelation")

Ordinary regression does not account for this

Computing autocorrelation

Create "lagged" series

Copy of the original series, offset by one or more time periods

Compute correlation between original series and lagged series (lag-1, lag-2, etc.)

TABLE 16.1 FIRST 24 MONTHS OF AMTRAK RIDERSHIP SERIES

Month	Ridership	Lag-1 Series	Lag-2 Series
Jan-91	1709		
Feb-91	1621	1709	
Mar-91	1973	1621	1709
Apr-91	1812	1973	1621
May-91	1975	1812	1973
Jun-91	1862	1975	1812
Jul-91	1940	1862	1975
Aug-91	2013	1940	1862
Sep-91	1596	2013	1940
Oct-91	1725	1596	2013
Nov-91	1676	1725	1596
Dec-91	1814	1676	1725
Jan-92	1615	1814	1676

Autocorrelation

Positive autocorrelation at lag-1 = stickiness

Strong autocorrelation (positive or negative) at a lag > 1 indicates seasonal (cyclical) pattern

Autocorrelation in residuals indicates the model has not fully captured the seasonality in the data

ACF Values

Lags	ACF
0	1
1	0.64821321
2	0.51890093
3	0.40798336
4	0.31966141
5	0.26237851
6	0.21345751
7	0.22334783
8	0.22640951
9	0.19724335
10	0.14933859
11	0.17307311
12	0.12726976

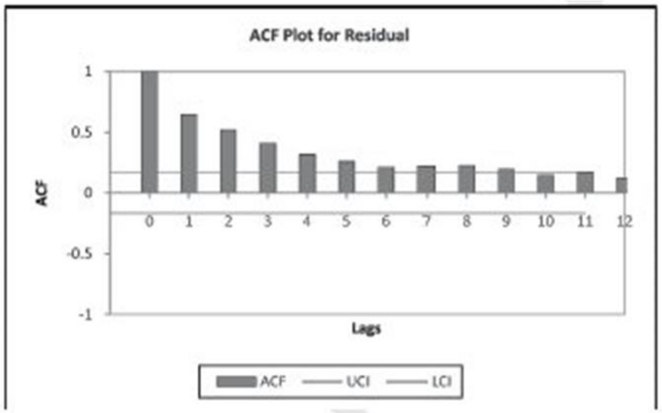
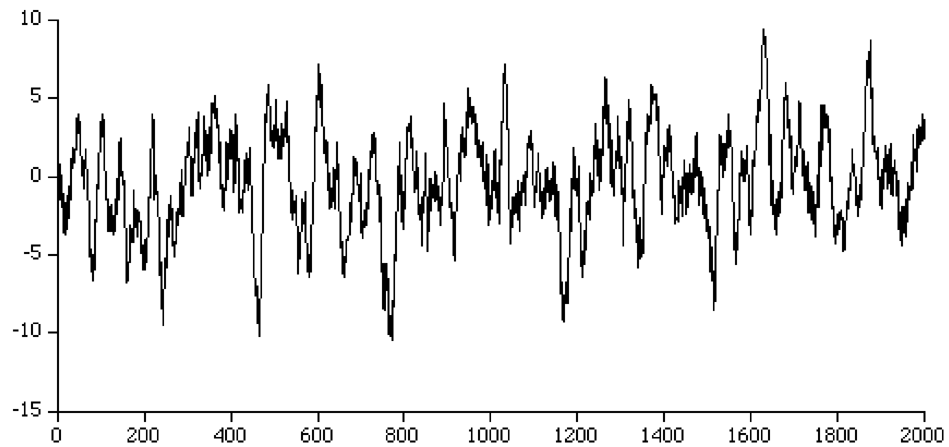


FIGURE 16.11 XLMINER OUTPUT SHOWING AUTOCORRELATION OF RESIDUAL SERIES FROM FIGURE 16.9

ARIMA summary

- AR model

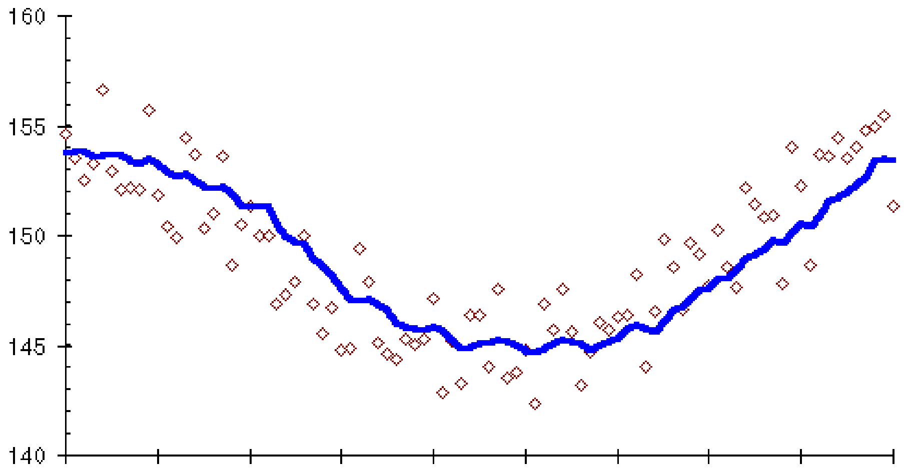
- AR(1)



X(t)={a*X(t-1)+c}+u*e(t)

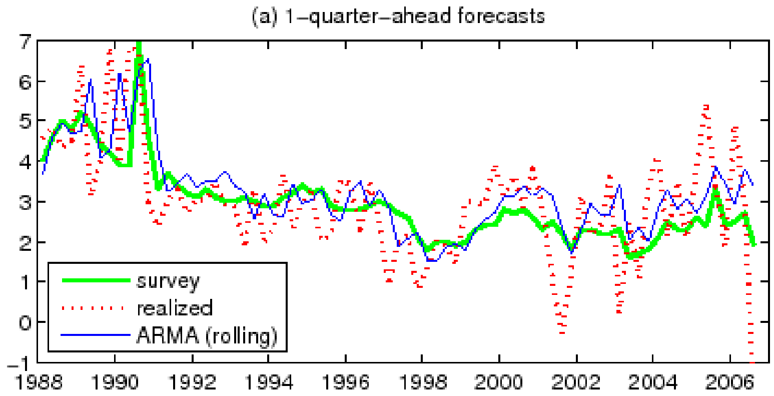
- MA

- MA(1)



X(t)={a*e(t-1)+c}+u*e(t)

- ARMA



X(t)={a*X(t-1)}+{b*e(t-1)}+c+u*e(t)

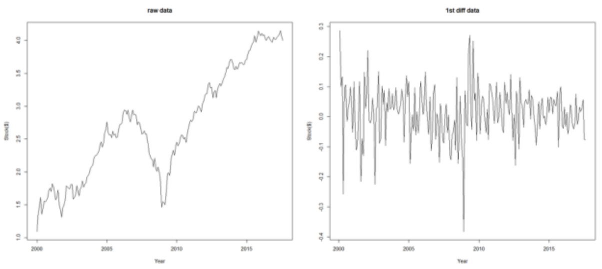
- ARIMA uses 'co-integration' (ARMA is only about correlation)
  - Correlation (Linear)

    if x has a large value, Y tends to have a large value.
  - Co-integration (trend)

    if x increases, Y also increases

→ Consider stationarity

<Figure 2> log(Stock) of Starbucks(2000.01~2017.08)



$$a*\{X(t)-X(t-1)\}=\{b*X(t-1)\}+\{c*e(t-1)\}+d+u*e(t)$$

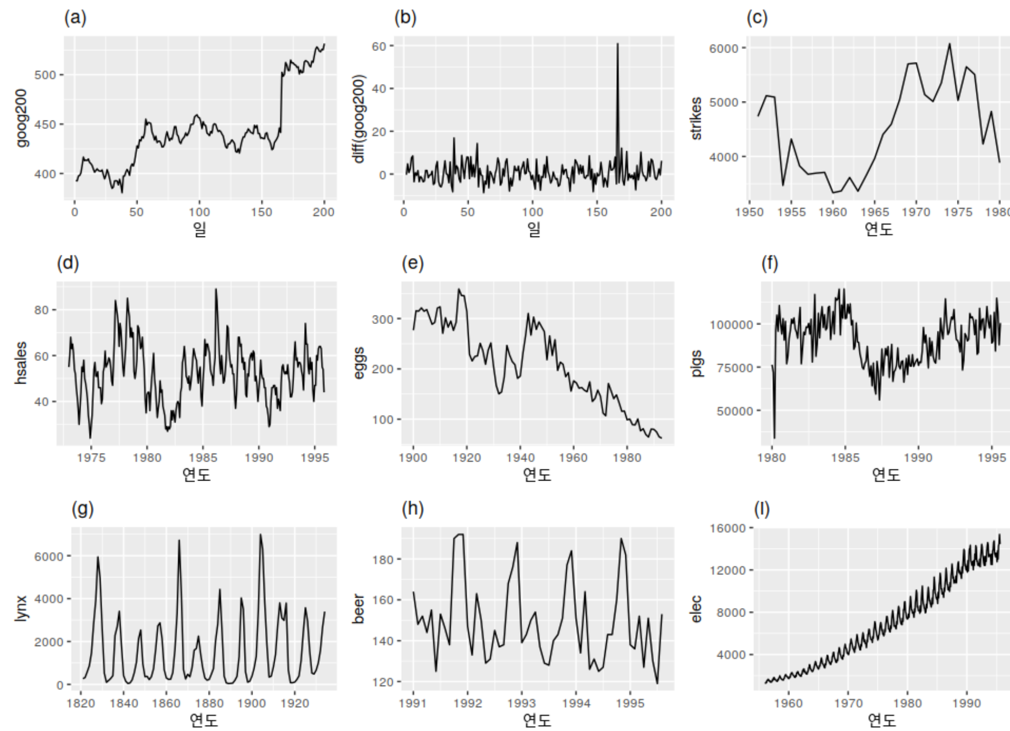
$$X(t)=[X(t-1)+\{b*X(t-1)\}+\{c*e(t-1)\}+d+u*e(t)]/a$$

$$a*[\{X(t)-X(t-1)\}-\{X(t-1)-X(t-2)\}]=\{b*X(t-1)\}+\{c*e(t-1)\}+d+u*e(t)$$

$$X(t)=(2+b/a)*X(t-1)+X(t-2)+(c/a)*e(t-1)+(d/a)+(u/a)*e(t)$$

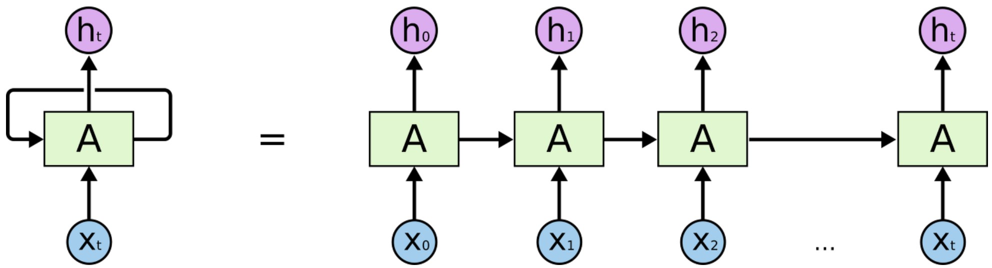
Stationarity in TS

- TS features are independent of time



RNN and LSTM

- RNN



- LSTM

"the clouds are in the *sky*"
"I grew up in France… I speak fluent *French*"