Data Mining and Business Intelligence

Lecture 9: ARIMA Models

Jing Peng
University of Connecticut

3/26/20

WebEx Session

- Client: https://www.webex.com/downloads.html

- Link for Meeting Room:
  - https://uconn-cmr.webex.com/uconn-cmr/j.php?MTID=m14752071b12dc799faa9b174360ba233
  - Join by phone: Dial +1-415-655-0002 and enter access code: 613 027 204

- Best practice: to ensure you can hear me, everyone will be mute on entrance

- The session will be automatically recorded

Tutorial on How to Use WebEx

- Enter your name while join the meeting
- Using Chat to ask questions
- Polling
- Raising your hand
- Switch to (or back from) Full screen (Alt+Enter)
- If you have difficulty connecting to WebEx during the live session, email Hao

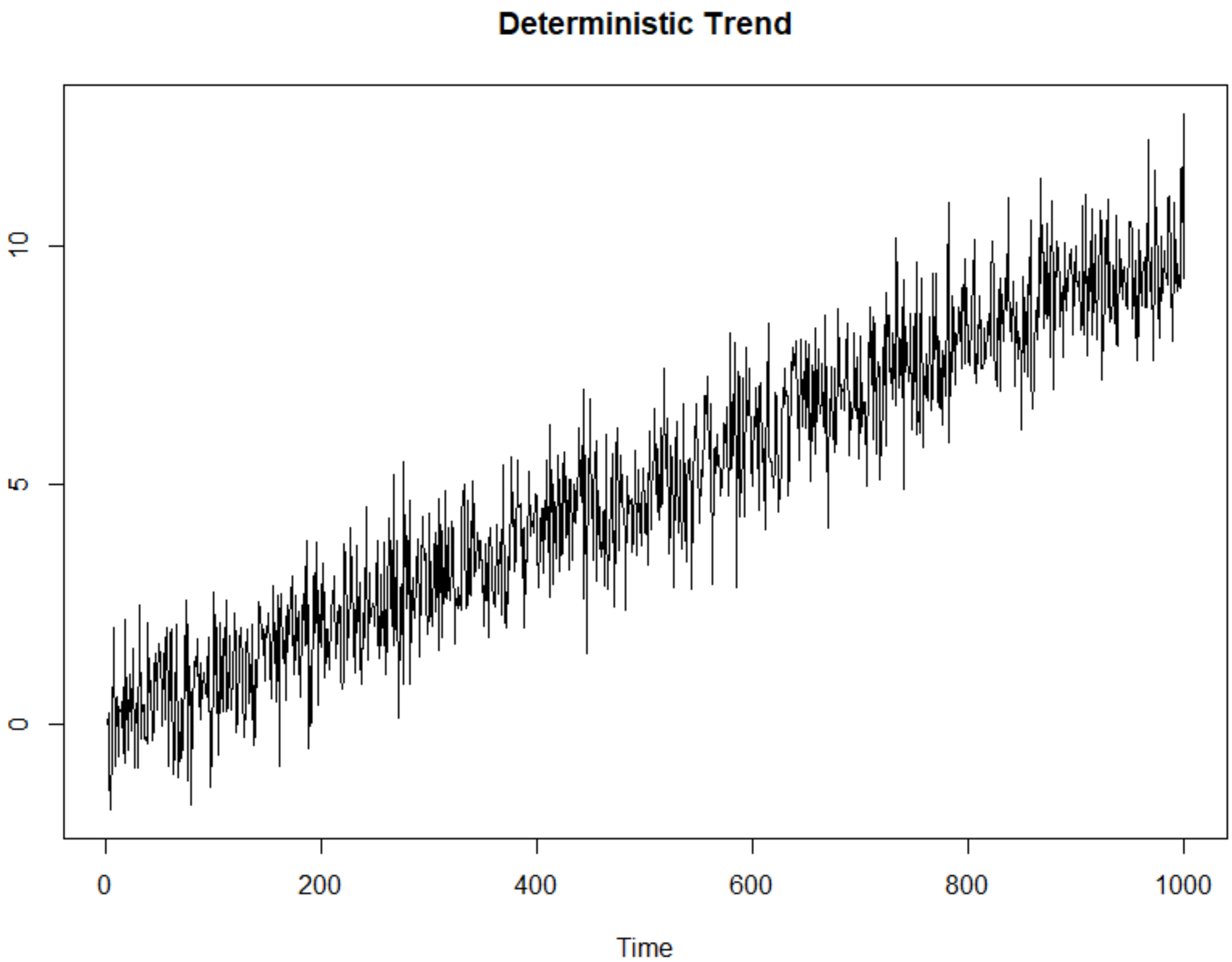
Removing Trend and Seasonality

Two Types of Trend (Seasonality)

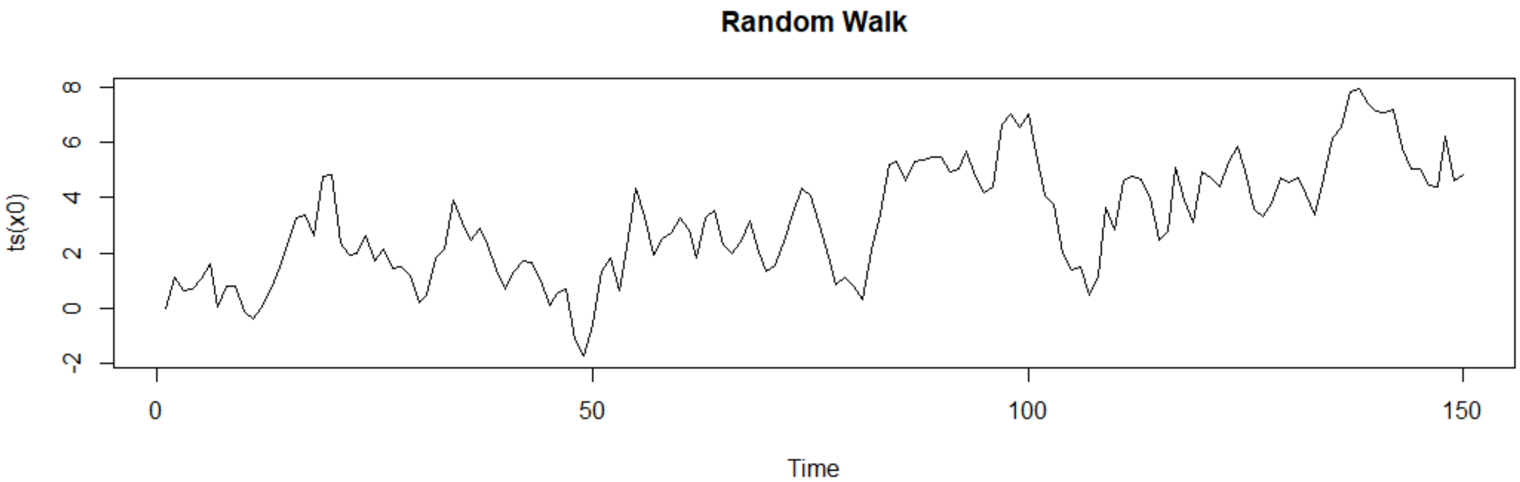
- **Deterministic**: a mathematical function of **time**
  - Linear, quadratic, logarithmic, exponential (e.g., $Y_t = \alpha t + \varepsilon_t$)
  - How to model: mathematical functions

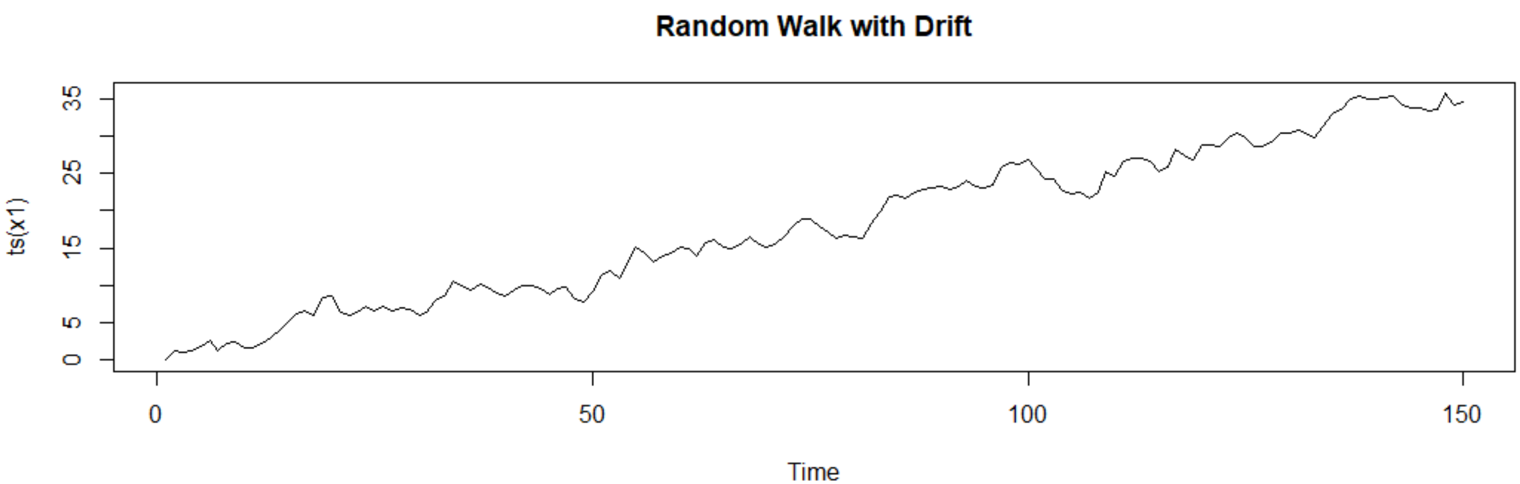
- **Stochastic**: future values depend on **past values** plus error
  - e.g., Random walk with drift ($Y_t = \theta + Y_{t-1} + \varepsilon_t$)
  - How to model: first (seasonal) difference

Deterministic Trend



Stochastic Trend





deterministic vs stochastic.R

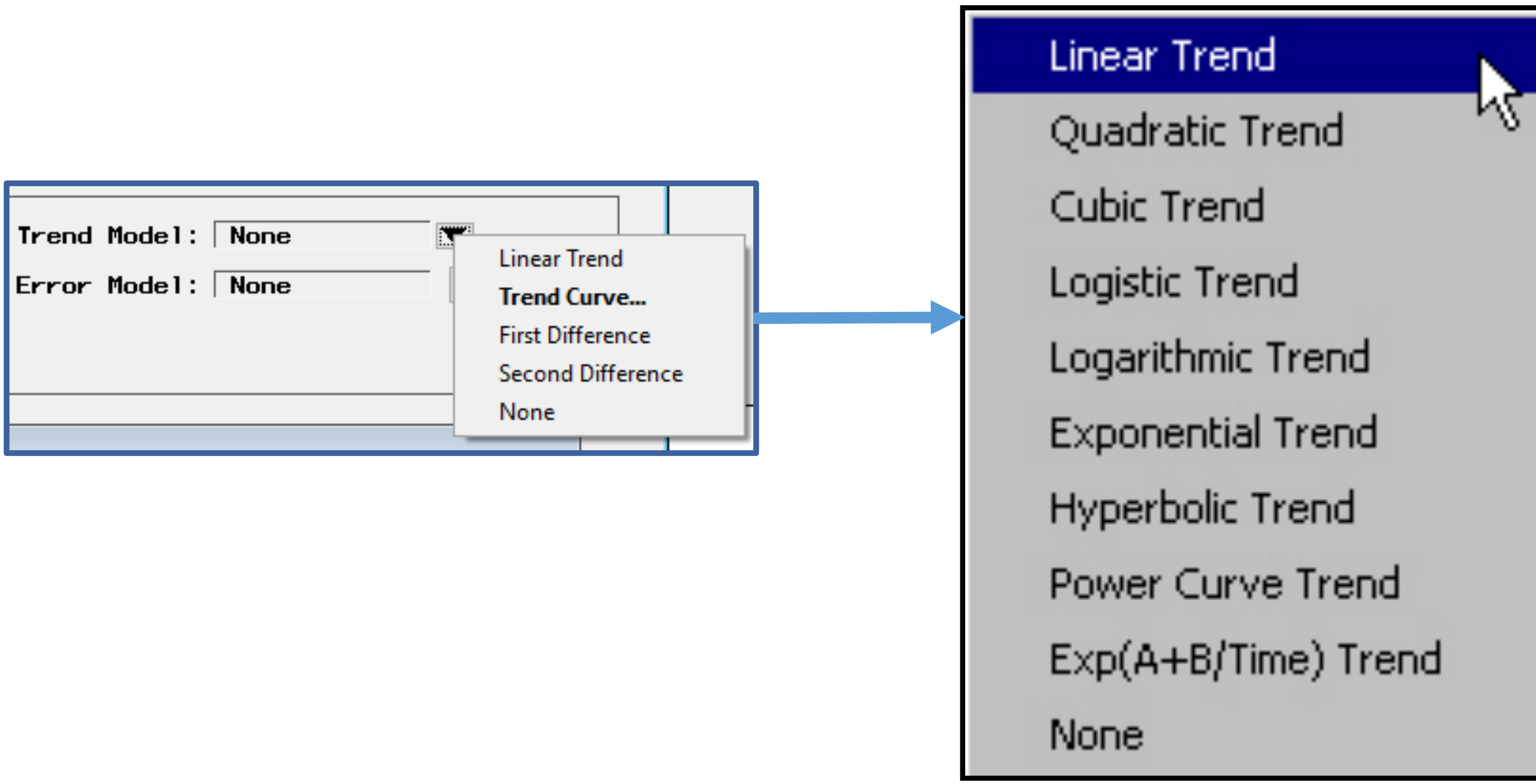
Deterministic vs. Stochastic Trends

	Deterministic	Stochastic
Exemplary Model	$y_t = \alpha t + \varepsilon_t$	$y_t = \alpha + y_{t-1} + \varepsilon_t$
Mean	$E[y_t] = \alpha t$	$E[y_t] = \alpha t$
Variance	$Var(y_t) = \sigma^2$	$Var(y_t) = t\sigma^2$

Read/Watch more:
https://www.youtube.com/watch?v=yCM6N8sRtPY
https://www.youtube.com/watch?v=ouahL4HbwBE
https://stats.stackexchange.com/questions/159650/why-does-the-variance-of-the-random-walk-increase
https://www.quora.com/Is-a-random-walk-the-same-thing-as-a-non-stationary-time-series

How to Model Deterministic Trend?

- Add a deterministic trend component as a regressor into the model



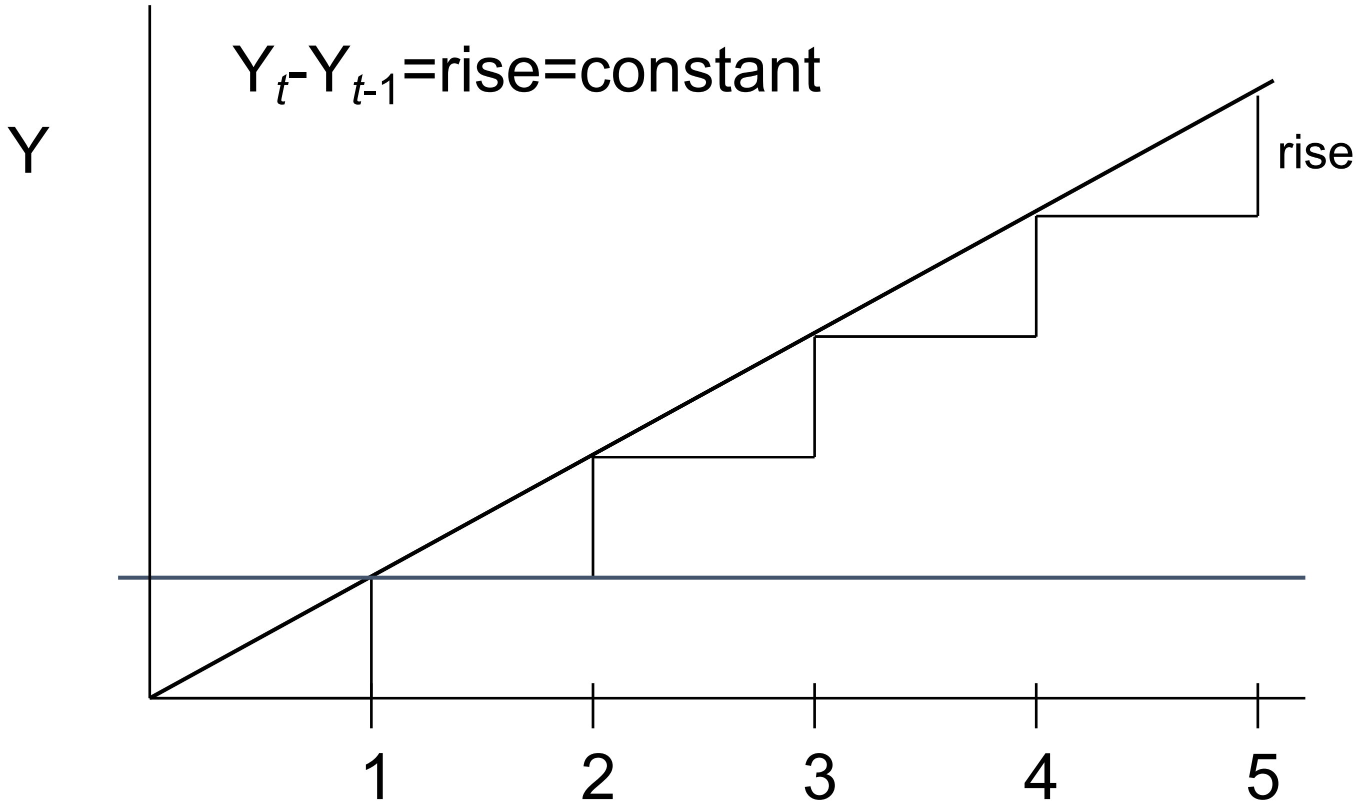
Refer to textbook Forecasting chapter 2.1 for formulas

How to Remove Stochastic Trend?

- First difference:
  - Random walk ($Y_t - Y_{t-1} = \varepsilon_t$)
  - Random walk with drift ($Y_t - Y_{t-1} = \theta + \varepsilon_t$)

- First difference can also remove LINEAR deterministic trend
  - $y_t = \alpha t + \varepsilon_t \Rightarrow y_t - y_{t-1} = \alpha + \varepsilon_t - \varepsilon_{t-1}$
  - The difference of two independent normal variables is still normal

Frist Difference on Straight Line

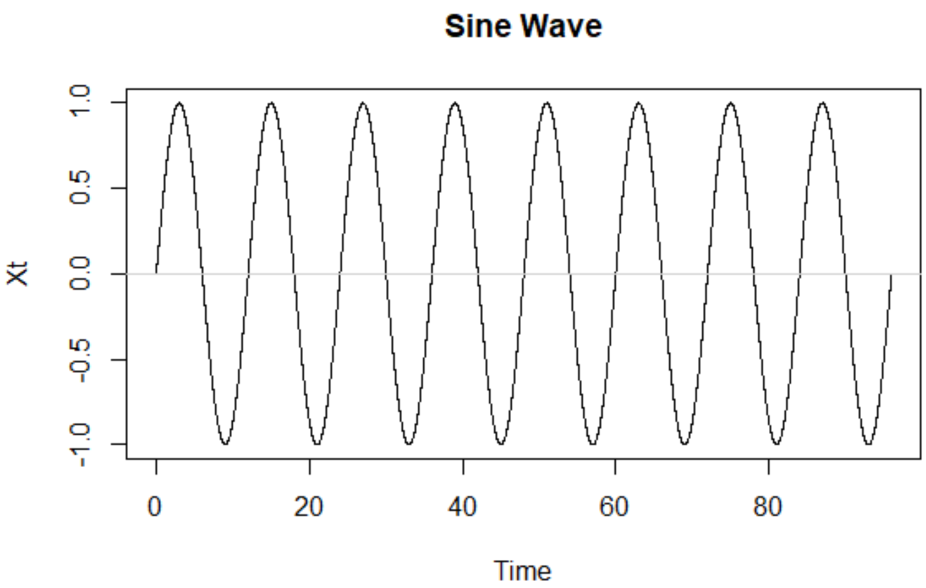


Removing Deterministic Seasonality

- Seasonal Dummy Variables

$$Y_t = \beta_{JAN} I_{JAN} + \beta_{FEB} I_{FEB} + \cdots + \beta_{DEC} I_{DEC}$$
$$\beta_M = \text{effect of month } M$$

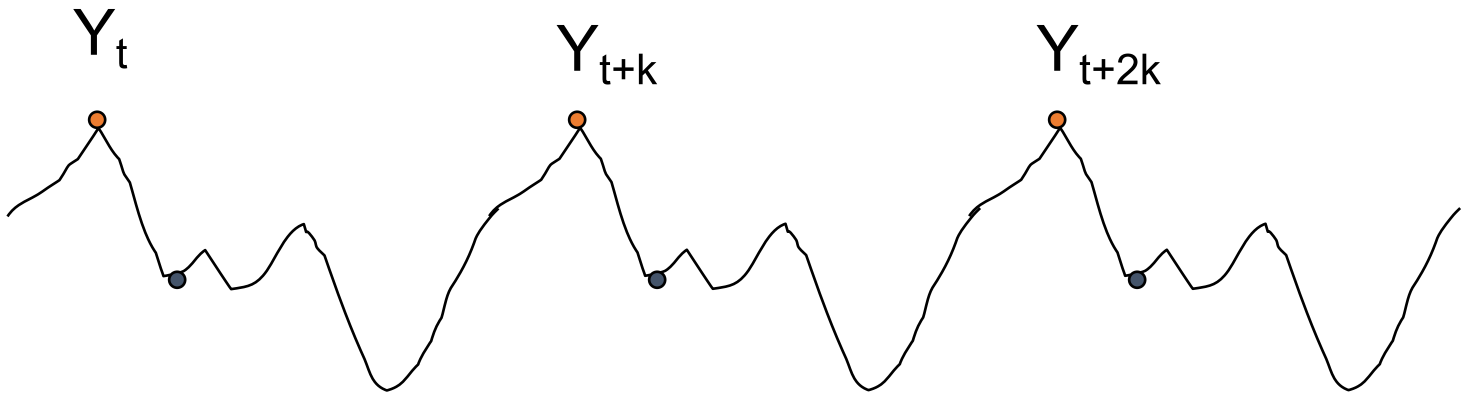
- Trigonometric Functions (not directly supported by TSFS)



$$X_t = \sin\left(\frac{2\pi t}{S}\right)$$

Seasonal Difference

- Can account for both stochastic and deterministic seasonality



$\Delta_k = 0$

Takeaway

- Stochastic trend increases the variance, whereas deterministic trend changes the mean instead of the variance

- Deterministic trend component cannot address stochastic trend

- First difference cannot address nonlinear deterministic trend

- A time series may exhibit both deterministic and stochastic trend

$$Y_t = \theta + \alpha t + Y_{t-1} + \varepsilon_t$$

Diagnosing Trend and Seasonality through

✓ Time series plot

✓ Autocorrelation functions

✓ Unit/Seasonal root test

- Demos on two datasets

Takeaway: Diagnosing Trend

- A time series having a trend component usually exhibits the following
  - a time series plot that is trending up, down, or in a deterministic fashion
  - a highly significant ACF, PACF, and IACF at lag 1
  - an ACF with many significant lags decaying slowly from lag 1
  - an ACF with few significant values after first differencing is applied
  - unit root tests that are not significant but become significant when a first difference is applied

Takeaway: Diagnosing Seasonality

- A time series with a seasonal component having a period S usually exhibits the following:
  - a time series plot that has repetitive behavior every S time units
  - significant ACF, PACF, and IACF values at lag S
  - an ACF with significant values at lags that are multiples of S
  - seasonal root tests that are not significant but become significant when a difference of order S is applied

ARIMA Models

Autoregressive (AR) Model

- AR (1): autoregressive model of order 1

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \varepsilon_t$$

- AR(p): autoregressive model of order p

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

Moving Average (MA) Model

- MA (1): moving average model of order 1

$$Y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

- MA(q): moving average model of order q

$$Y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Autoregressive Integrated Moving Average (ARIMA)

- ARMA

$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

- AR**I**MA: replace Y_t above with $\Delta_d(Y_t)$

- ARIMA(p, d, q)
  - p: order of autoregressive (AR)
  - d: order of differencing (I)
  - q: order of moving average (MA)

ARIMA Encapsulates Some ESMs as Special Cases

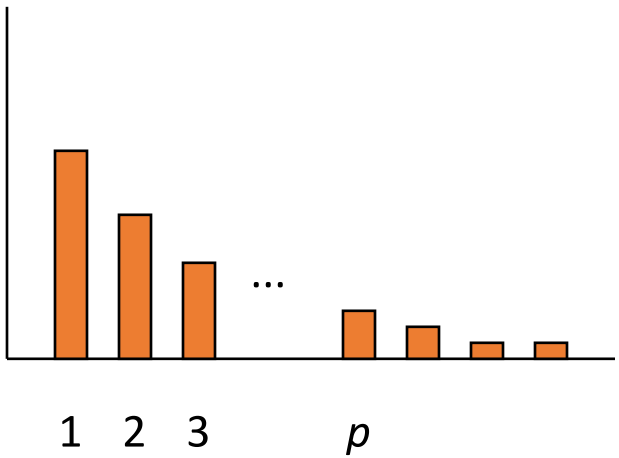
- **ARIMA(0,1,0) = random walk**
- **ARIMA(0,1,1) without constant = simple exponential smoothing**
- **ARIMA (0,2,2) without constant = linear (Holt) exponential smoothing**

- https://people.duke.edu/~rnau/411arim.htm

Identifying AR and MA Models

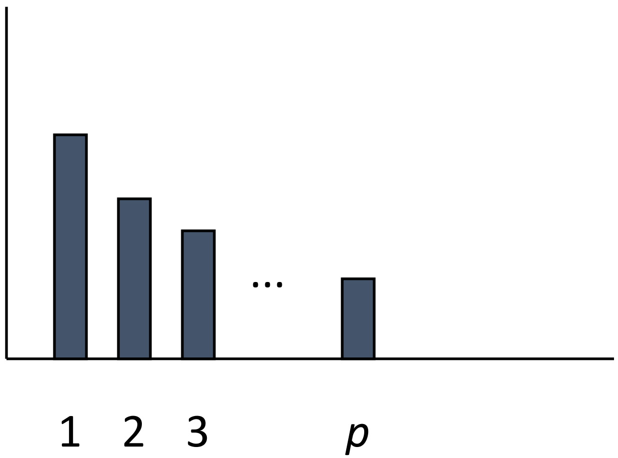
Portfolio of Shapes — AR(p) Model

ACF



Exponential Decay

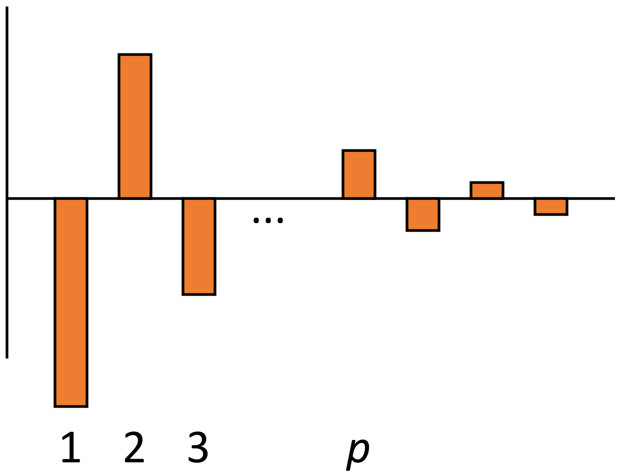
PACF/IACF



Drops to 0 after lag p

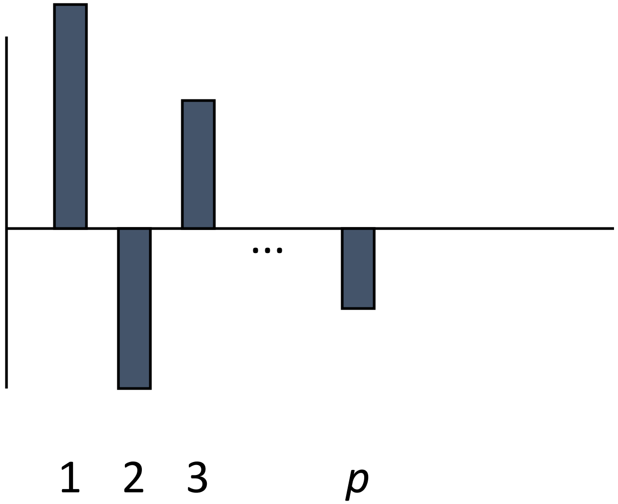
Portfolio of Shapes — AR(p) Model

ACF



Exponential Decay

PACF/IACF



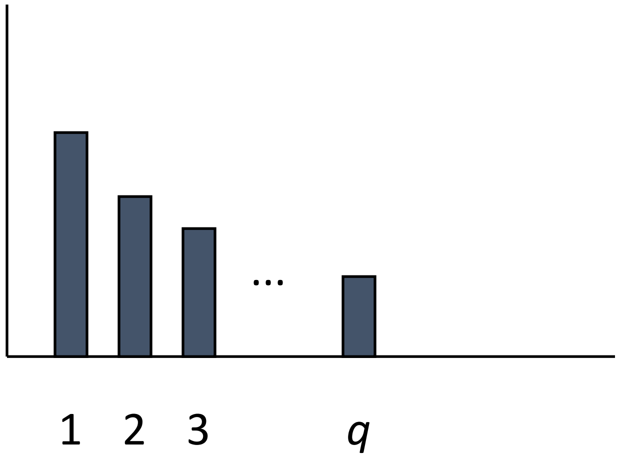
Drops to 0 after lag p

Identifying an AR(*p*) Model

- The PACF and IACF are important in identifying the order of an AR model, namely value of p
- The highest lag of the PACF values or IACF values that are significantly different from zero suggest the appropriate value of *p*
- If PACF and IACF suggest different *p*, consider the larger one

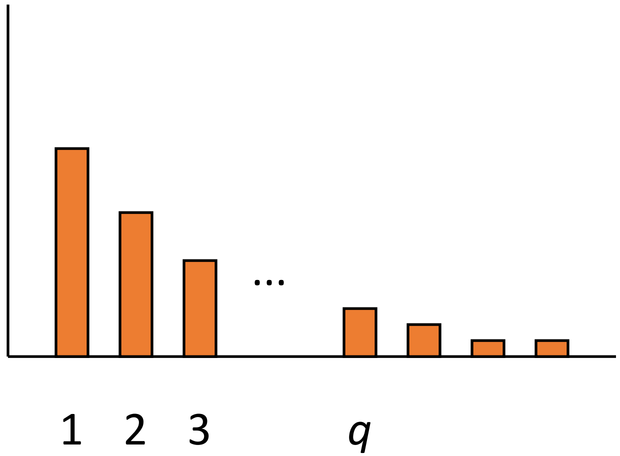
Portfolio of Shapes — MA(q) Model

ACF



Drops to 0 after lag q

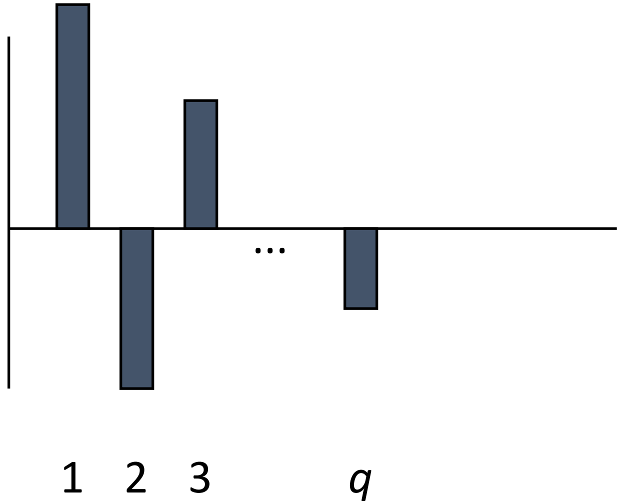
PACF/IACF



Exponential Decay

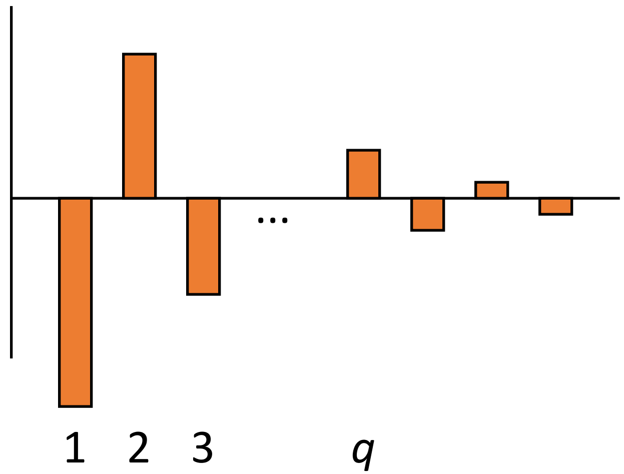
Portfolio of Shapes — MA(q) Model

ACF



Drops to 0 after lag q

PACF/IACF



Exponential Decay

Identifying an MA(q) Model

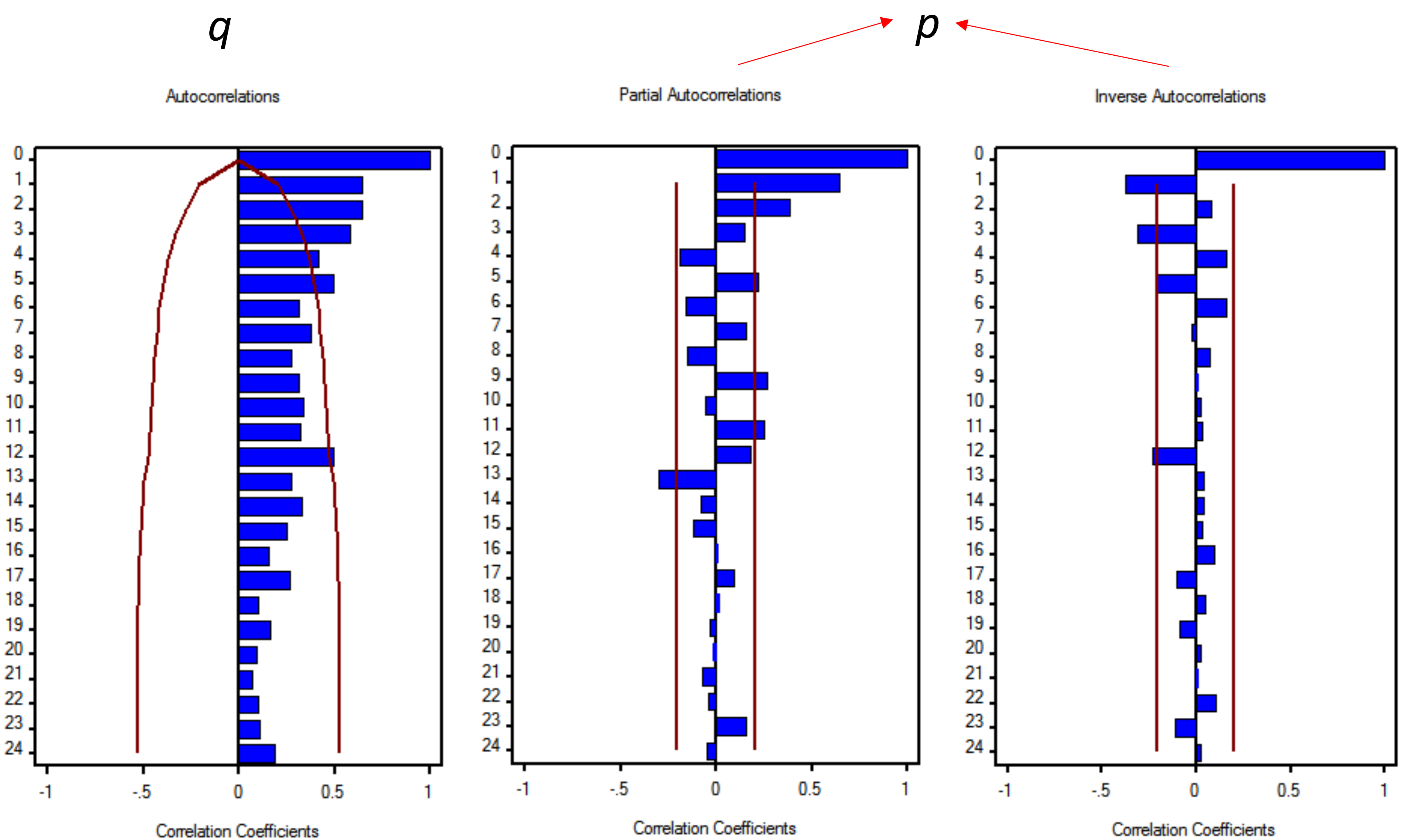
- The ACF is important in identifying the order of an MA model, namely the value of q
- The highest lag of the ACF values that is significantly different from zero suggests the appropriate value of q

Theoretical Patterns of ACF and PACF

Type of Model	Typical Pattern of ACF	Typical Pattern of PACF
AR (*p*)	Decays exponentially or with damped sine wave pattern or both	**Cut-off after lags *p***
MA (*q*)	**Cut-off after lags *q***	Declines exponentially
ARMA (*p,q*)	Exponential decay	Exponential decay

Source: http://slideplayer.com/slide/1507028/

Identifying Orders of ARMA model



Identifying ARIMA Models

Box-Jenkins Modeling Methodology

Identify	Determine ARIMA orders p and q using ACF, PACF, and the IACF. Determine d based on Unit Root Test
Estimate	Fit the ARIMA(p,d,q) model and assess the fit of the model.
Forecast	Produce forecasts using the best ARIMA model that passes assessment.

ARIMA(p,d,q) Model Selection

1. Assumes series is stationary. If not, apply first difference first

2. Find q such that ACF(q) falls outside confidence limits and ACF(k) falls inside confidence limits for all $k>q$.

3. Find p such that PACF(p) / IACF(p) falls outside confidence limits and PACF(k) / IACF(k) falls inside confidence limits for all $k>p$.

ARMA(p,d,q) Model Selection

4 Determine all ordered pairs (j,k) such that $0 \leq j \leq p$ and $0 \leq k \leq q$.

5 For each ordered pair (j,k) found in step 4, fit an ARIMA(j,d,k) model.

6 For all of the models fit in step 5, select the model with the smallest values of RMSE on the holdout sample or AIC or SBC on the fit sample.

Forecasting Steps

- Refit the best model on the entire data set
- Verify that the model parameters and forecast have not changed substantially
- Make sure that the residuals of the model look reasonable (not necessarily perfect)

Application

Demo

Identifying ARMA(p, q) models

Chapter 3 p51-58

Datasets: Groceries

Summary of Identification Demonstration

- Toothpaste series

  ACF implies *q*<=1, PACF suggest *p*<=1, IACF suggest *p*<=2

- Peanut butter series

  ACF implies *q*<=1 PACF and IACF *p*<=1, IACF suggest either *p*<=3, p=(1,3) or *q*=2 because of sine-wave decay pattern.

- Jelly series

  ACF implies *q*<=1, PACF and IACF imply *p*<=2

Demo

Estimation of candidate models

Chapter 3 p62-87

Datasets: Groceries

→ Set Ranges (13 weeks holdout)
→ Identify p and q (done)
→ Fit ARIMA models
→ Examine model performance, residuals, parameters and fit stats
→ View forecasts

Demo

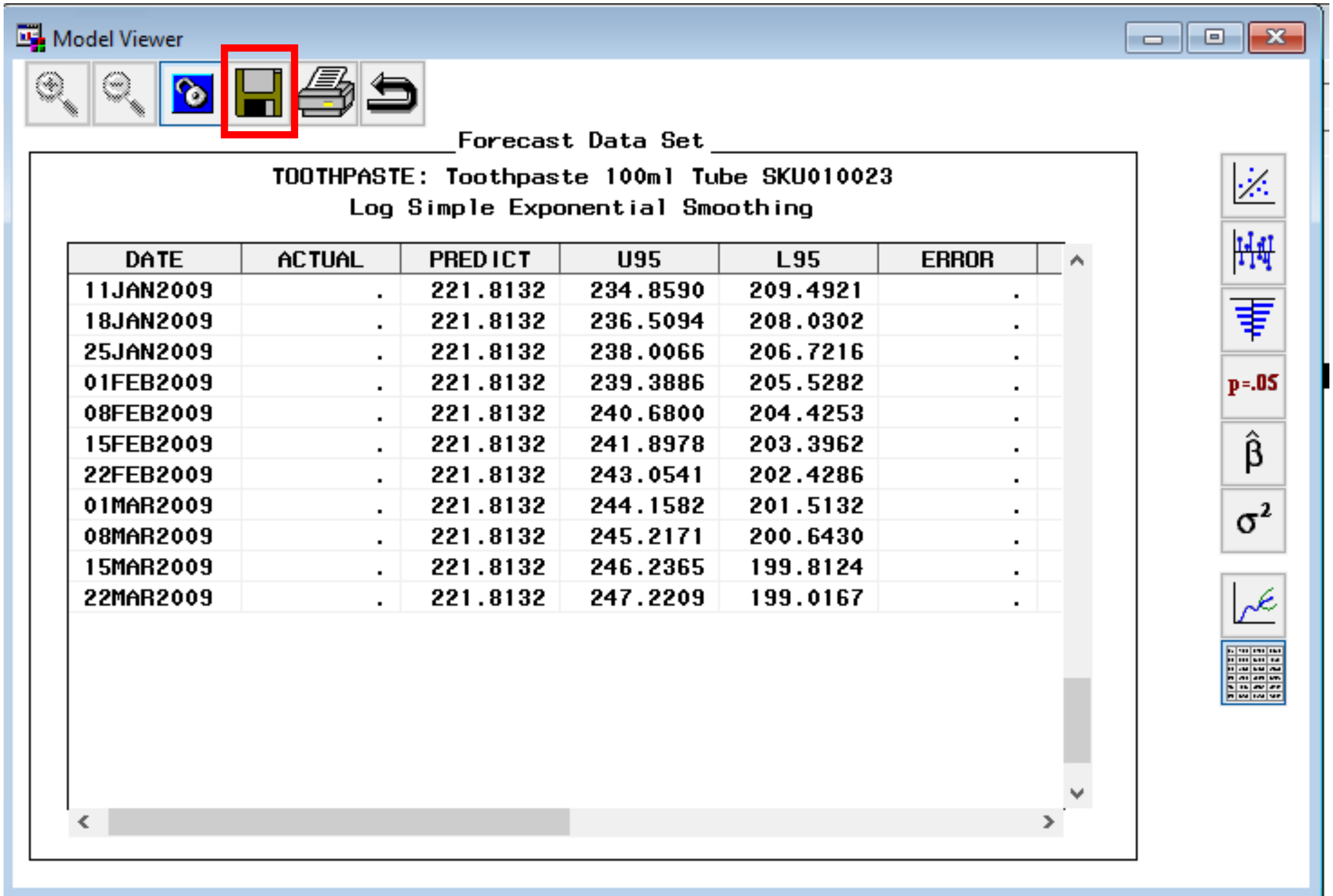
Generate forecast based on the best model

Chapter 3 p90-101

Datasets: Groceries

Reset Ranges → Duplicate best model → Examine changes in model parameters and model residuals → Save Predictions

Saving Predictions



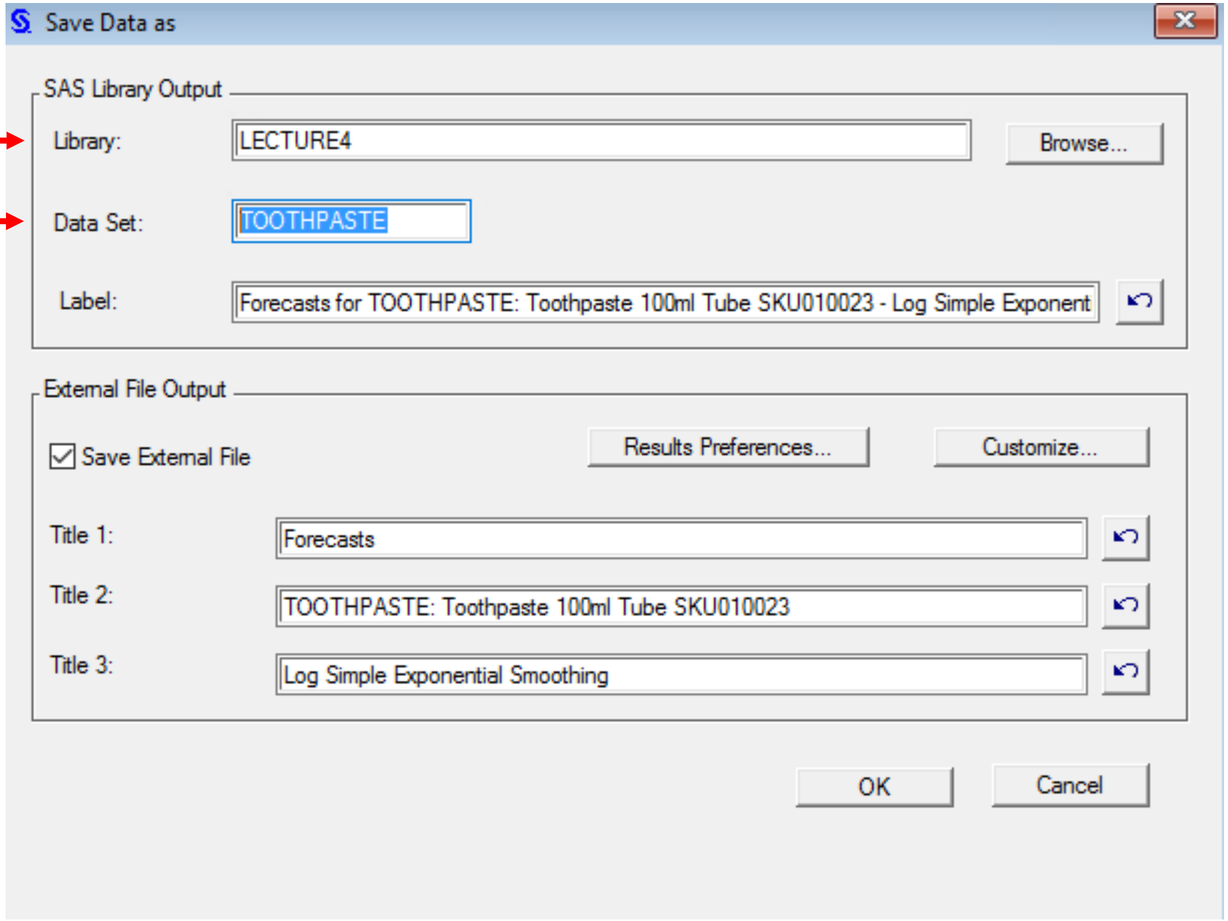
TOOTHPASTE: Toothpaste 100ml Tube SKU010023
Log Simple Exponential Smoothing

DATE	ACTUAL	PREDICT	U95	L95	ERROR
11JAN2009	.	221.8132	234.8590	209.4921	.
18JAN2009	.	221.8132	236.5094	208.0302	.
25JAN2009	.	221.8132	238.0066	206.7216	.
01FEB2009	.	221.8132	239.3886	205.5282	.
08FEB2009	.	221.8132	240.6800	204.4253	.
15FEB2009	.	221.8132	241.8978	203.3962	.
22FEB2009	.	221.8132	243.0541	202.4286	.
01MAR2009	.	221.8132	244.1582	201.5132	.
08MAR2009	.	221.8132	245.2171	200.6430	.
15MAR2009	.	221.8132	246.2365	199.8124	.
22MAR2009	.	221.8132	247.2209	199.0167	.

Save Predictions

Location

Name



Seasonal ARIMA Model

Seasonal ARIMA Model

ARIMA(p,d,q)(P,D,Q)$_S$

- The p, d, and q are the orders of the nonseasonal terms of the model.
- P is the order of the seasonal autoregressive terms.
- Q is the order of the seasonal moving average terms.
- D is the order of the seasonal difference (rarely goes above 1).
- S the length of the seasonal period.

- In practice, you may try (P=0,Q=0),(P=1,Q=0), or (P=0, Q=1), but rarely (P=1,Q=1)

Box-Jenkins Modeling Methodology

Identify	Determine ARIMA orders p and q using ACF, PACF, and the IACF. Determine d/D based on Unit/Seasonal Root Test
Estimate	Fit the ARIMA$(p,d,q)(P,D,Q)_s$ model and assess the fit of the model.
Forecast	Produce forecasts using the best ARIMA model that passes assessment.

Demo

Estimate a seasonal ARIMA model for Airline data

Chapter 4 p52-83

Dataset: DOTAIR9498

Further Readings

- Forecasting Chapters 3 & 4
- Exercise: replicate the in-class demo on your own
- Time series ARIMA models using Stata/R/SAS: video
- Seasonal ARIMA models: https://www.otexts.org/fpp/8/9

Notations for ARIMA models (optional)

- Previous parametrization of ARIMA(p, d, q)

$$\Delta_d(Y_t) = \theta_0 + \phi_1 \Delta_d(Y_{t-1}) + \cdots + \phi_p \Delta_d\left(Y_{t-p}\right) + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

- An alternative and more common parametrization of ARIMA(p, d, q)

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)(\Delta_d(Y_t) - \mu) = \left(1 + \theta_1 B + \cdots + \theta_q B^q\right)\varepsilon_t$$

- Backshift Operator (B)
  - $BY_t = B(Y_t) = Y_{t-1}$
  - $B^k Y_t = Y_{t-k}$
  - $B^k \varepsilon_t = \varepsilon_{t-k}$
  - $\Delta_d(Y_t) = Y_t - Y_{t-d} = \left(1 - B^d\right)Y_t$
  - $B(constant) = constant$

Model Interpretation (optional)

STEELSHP: Steel Shipments Thousands of Net Tons
ARMA(2,2)

Model Parameter	Estimate	Std. Error	T	Prob>¦T¦
Intercept	6450	197.5803	32.6440	<.0001
Moving Average, Lag 1	0.35362	0.4810	0.7352	0.4641
Moving Average, Lag 2	-0.19271	0.2624	-0.7345	0.4646
Autoregressive, Lag 1	0.71490	0.4837	1.4780	0.1429
Autoregressive, Lag 2	0.09856	0.4423	0.2228	0.8242
Model Variance (sigma squared)	205925	.	.	.

$$\left(1-\phi_1 B-\cdots-\phi_p B^p\right)(Y_t-\mu)=\left(1+\theta_1 B+\cdots+\theta_q B^q\right)\varepsilon_t$$

$$(1-0.71B-0.10B^2)(Y_t-6450)=(1+0.35B-0.19B^2)\varepsilon_t$$

$$(Y_t-6450)\ -0.71(Y_{t-1}-6450)-0.10(Y_{t-2}-6450)=\varepsilon_t+0.35\varepsilon_{t-1}-0.19\varepsilon_{t-2}$$

Examples of Seasonal ARIMA Models (optional)

ARIMA(*0,0,0*)(*1,1,1*)$_{12}$

$$(1-\Phi_1 B^{12})(1-B^{12})(Y_t - \mu) = (1-\Theta_1 B^{12})\varepsilon_t$$

ARIMA(*1,1,1*)(*1,1,1*)$_{12}$

$$(1-\phi_1 B)(1-\Phi_1 B^{12})(1-B)(1-B^{12})(Y_t - \mu) = (1-\theta_1 B)(1-\Theta_1 B^{12})\varepsilon_t$$