NATIONAL ENGINEERING CENTER
University of the Philippines
Diliman, Quezon City

6.0 Regression and Time Series Analysis

Eugene Rex L. Jalao, Ph.D.

Associate Professor

Department Industrial Engineering and Operations Research

University of the Philippines Diliman

@thephdataminer

Module 1 of the Business Intelligence and Analytics Track of UP NEC and the UP Center of Business Intelligence

Module 1 Outline

1. Intro to Business Intelligence
   - Case Study on Selecting BI Projects
2. Data Warehousing
   - Case Study on Data Extraction and Report Generation
3. Descriptive Analytics
   - Case Study on Data Analysis
4. Classification Analysis
   - Case Study on Classification Analysis
5. **Regression and Time Series Analysis**
   - **Case Study on Regression and Time Series Analysis**
6. Unsupervised Learning and Modern Data Mining
   - Case Study on Text Mining
7. Optimization for BI

Outline for this Session

- Regression and Model Building
- Simple and Multiple Linear Regression
- Model Evaluation
- Indicator Variables
- Alternative Regression Models
- Time Series Analysis
- Components of a Time Series
- Evaluation Methods of Forecast
- Smoothing Methods of Time Series
- Case Study

Regression and Model Building

Definition 6.1: Regression Analysis

- Regression analysis is a statistical technique for investigating and **modeling the relationship between variables.**
- Equation of a straight line (classical)

$$y = mx + b \tag{6.1}$$

- We usually write this as

$$y = \beta_0 + \beta_1 x \tag{6.2}$$

Regression and Model Building

- Input Variables, Regressors, Independent variables or predictor variables (x)
  - Must be **continuous** and have **no missing values**
- Output Variable, Target Variable, Response Variable, or Independent Variable (y)
- Output Variable must be **continuous**

Regression and Model Building

- Not all observations will **fall exactly** on a straight line.

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{6.3}$$

  - where ε represents error
  - it is a variable that accounts for the failure of the model to fit the data exactly.
  - $\varepsilon \sim N(0, \sigma^2)$

Regression and Model Building

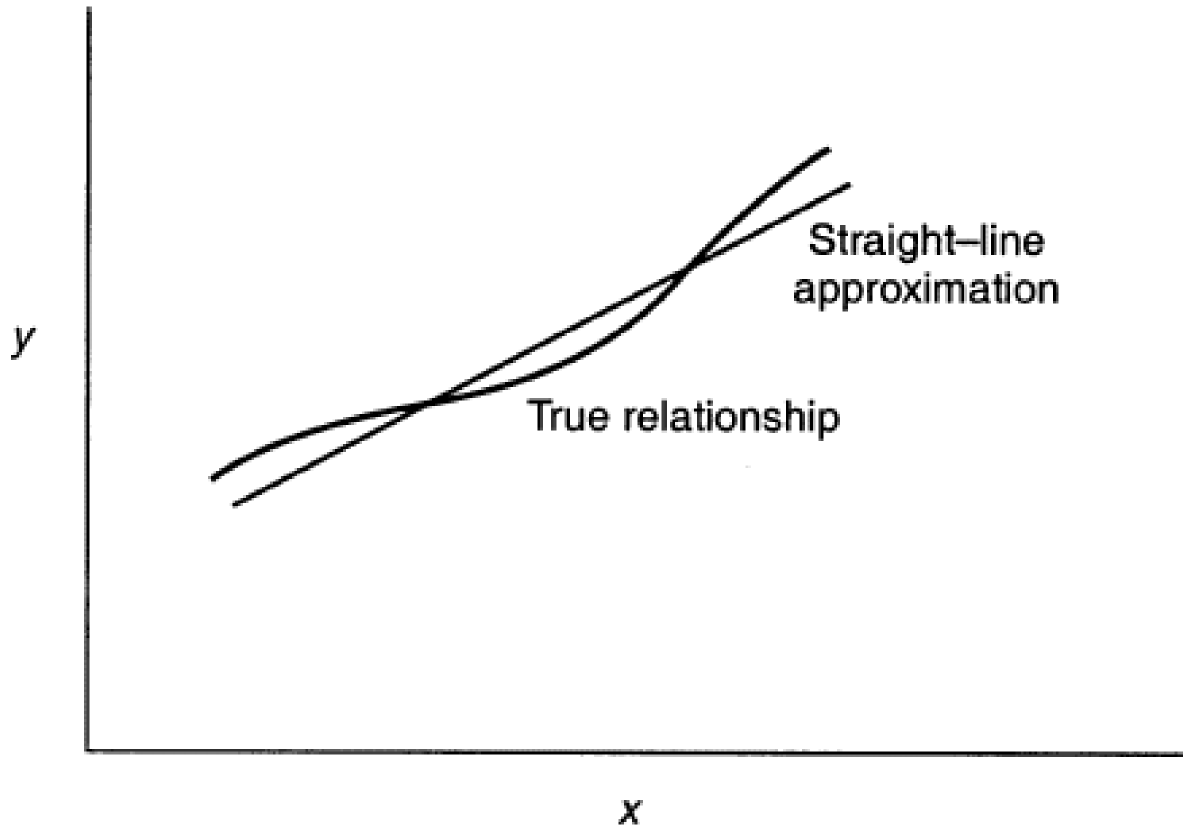


Figure 6.1: Approximation of the True Relationship

Regression and Model Building

- There are many uses of regression, including:
  - Data description
  - Parameter estimation
  - Prediction and estimation
  - Process Control
- Regression analysis is perhaps the **most widely used** statistical technique, and probably the **most widely misused**.

Regression and Model Building

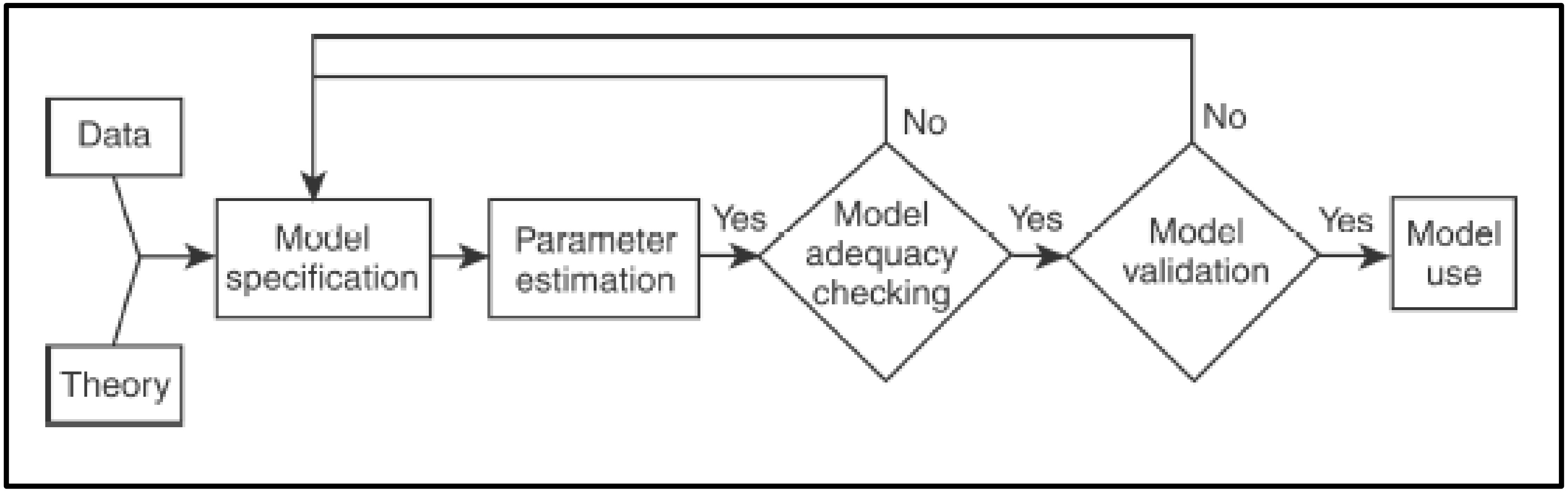


Figure 6.2: Regression Model Building

Outline for this Session

- Regression and Model Building
- **Simple and Multiple Linear Regression**
- Model Evaluation
- Indicator Variables
- Alternative Regression Models
- Time Series Analysis
- Components of a Time Series
- Evaluation Methods of Forecast
- Smoothing Methods of Time Series
- Case Study

Simple and Multiple Linear Regression

- Single predictor, x_1; response, y

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (6.4)$$

- β_0 – **intercept**: then β_0 is the response y, when $x_1 = 0$
- β_1 – **slope**: change in the mean of the distribution of the response produced by a unit change in x
- ε - random **error**: difference between predicted and actual which is distributed $NID(0, \sigma^2)$

Simple and Multiple Linear Regression

Example 6.1: Simple Linear Regression

- In this study, a random sample of **service call records** for a computer repair operation were examined and the length of each call (in minutes) and the number of components repaired were recorded.
- We would like to model the **relationship** between the number of components repaired to the **total time** it took to repair the computer

Minutes	Units
23	1
29	2
49	3
64	4
74	4
87	5
96	6
97	6
109	7
109	7
119	8
149	9
145	9
154	10

Simple and Multiple Linear Regression

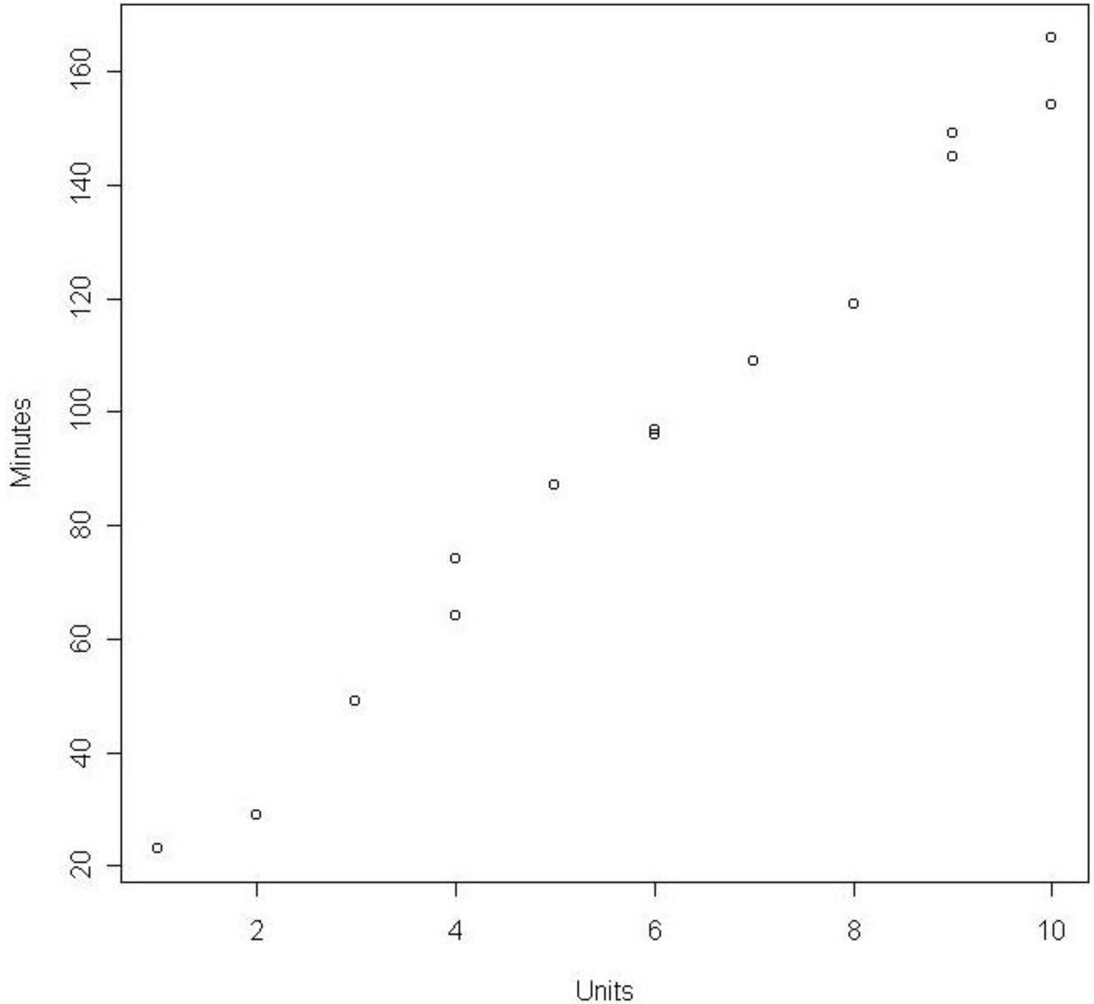
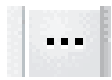


Figure 6.3: Scatter plot of the data

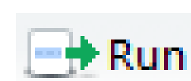
Using R Studio

- Open R Studio from the Programs Menu
- On the file explorer tab click on Files. 

- Click on Explore 

- Go to the Desktop Folder -> BI Training -> 5.0 Regression and Time Series
- Click on More. 

Click on Set as Working Directory.
- Click on File-> New File -> R Script.
- In the new tab script 

, type the following code:
  - `options(scipen=999,digits=2)`
  - `servicecalldata = read.csv("servicecalldata.csv")`
  - `plot(servicecalldata$units,servicecalldata$minutes)`
- Highlight the three lines of code and click on Run 

Simple Regression Using R Studio

- In the new tab script 

, type the following code:
  - `simplelrfit = lm(minutes~units, data=servicecalldata)`
  - `summary(simplelrfit)`
- Highlight the two lines of code and click on Run 


```
Call:
lm(formula = minutes ~ units, data = servicecalldata)

Residuals:
   Min     1Q Median     3Q    Max 
-9.232 -3.341 -0.714  4.777  7.803 

Coefficients:
            Estimate Std. Error t value         Pr(>|t|)    
(Intercept)    4.162      3.355    1.24             0.24    
units         15.509      0.505   30.71 0.00000000000089 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.4 on 12 degrees of freedom
Multiple R-squared:  0.987,	Adjusted R-squared:  0.986 
F-statistic:  943 on 1 and 12 DF,  p-value: 0.000000000000892
```

Simple and Multiple Linear Regression

- Service Call Regression Model:

$$Minutes = 4.162 + 15.509\ Units$$

Simple and Multiple Linear Regression

Definition 6.2: Coefficient of Determination

- R^2 - coefficient of determination

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} \qquad (6.5)$$

- **Proportion of variation** explained by the regressor, x
- $R^2 = \rho_{xy}^2$
- For the service call data

$$R^2 = \frac{SS_R}{SS_T} = 0.987$$

Simple and Multiple Linear Regression

Definition 6.3: Multiple Regression Analysis

- The simple regression model can be extended to have ***k* regressors**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad (6.6)$$

- Number of regressors (k) must be lesser than the number of rows (n)

Simple and Multiple Linear Regression

Example 6.2: Multiple Regression Example

- **Delivery Time Data**
  - A soft drink bottler is analyzing the vending machine service routes in his distribution system.
  - The analyst thinks that delivery time (y) is affected by the number of cases (x_1) and distance walked by the driver (x_2).

Observation Number	Delivery Time (Minutes) y	Number of Cases x_1	Distance (Feet) x_2
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

R Code to Run

```
› deliverytime = read.csv("deliverytime.csv")
› LRFit=lm(deltime ~ ncases + distance, data= deliverytime)
› summary(LRFit)
```

R Output

```
Call:
lm(formula = deltime ~ ncases + distance, data = deliverytime)

Residuals:
   Min     1Q Median     3Q    Max 
-5.788 -0.663  0.436  1.157  7.420 

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)    
(Intercept)  2.34123    1.09673    2.13     0.04417 *  
ncases       1.61591    0.17073    9.46 0.000000033 ***
distance     0.01438    0.00361    3.98     0.00063 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.3 on 22 degrees of freedom
Multiple R-squared:  0.96,	Adjusted R-squared:  0.956 
F-statistic:  261 on 2 and 22 DF,  p-value: 0.0000000000000469
```

Simple and Multiple Linear Regression

- Delivery Time Data

$$deltime = 2.3412 + 1.61591\, ncases + 0.014\, distance$$

Outline for this Session

- Regression and Model Building
- Simple and Multiple Linear Regression
- **Model Evaluation**
- Indicator Variables
- Alternative Regression Models
- Time Series Analysis
- Components of a Time Series
- Evaluation Methods of Forecast
- Smoothing Methods of Time Series
- Case Study

Model Evaluation

- Testing the Global Significance of Regression
  - To know if the x predictor variables **influences** y we consider the F Statistic from the ANOVA table output from R
  - We usually test for:
    - H_0 : There is no relationship between all x and y.
    - H_a : There is some relationship between some x and y.
  - **p-Value Methodology**
    - If $p < \alpha = 0.05$, Reject H_0

Model Evaluation

```
Call:
lm(formula = deltime ~ ncases + distance, data = deliverytime)

Residuals:
   Min     1Q Median     3Q    Max 
-5.788 -0.663  0.436  1.157  7.420 

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)    
(Intercept)  2.34123    1.09673    2.13      0.04417 *  
ncases       1.61591    0.17073    9.46 0.0000000033 ***
distance     0.01438    0.00361    3.98      0.00063 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.3 on 22 degrees of freedom
Multiple R-squared:  0.96,	Adjusted R-squared:  0.956 
F-statistic:  261 on 2 and 22 DF,  p-value: 0.0000000000000469
```

Model Evaluation

- Least-Squares Estimation of the Parameters
  - How well does this equation **fit the data**?
  - Is the model likely to be useful as a **predictor**?

Model Evaluation

- Residuals: $e_i = y_i - \hat{y}_i$

$$e_i = y_i - \hat{y}_i \tag{6.7}$$

- Residuals will be used to determine the **adequacy** of the model

Model Evaluation

- Some issues with R^2

```
Call:
lm(formula = DelTime ~ Ncases + Distance + Gibber, data = DeliveryTime)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6351 -0.7624  0.5539  1.2116  7.3706 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.579657   1.721687   1.498 0.148930    
Ncases       1.610432   0.177172   9.090    1e-08 ***
Distance     0.014470   0.003725   3.885 0.000855 ***
Gibber      -0.449819   2.464269  -0.183 0.856912    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.334 on 21 degrees of freedom
Multiple R-squared:  0.9597, Adjusted R-squared:  0.9539 
F-statistic: 166.5 on 3 and 21 DF,  p-value: 8.52e-15
```

Model Evaluation

Definition 6.4: Adjusted Coefficient of Determination

- Penalizes for **added terms** to the model that are not significant

$$R_{adj}^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2) \qquad (6.8)$$

- For the Delivery Time Data

$$R_{adj}^2 = 95.59\%$$

- With Gibberish

$$R_{adj}^2 = 95.39\%$$

Outline for this Session

- Regression and Model Building
- Simple and Multiple Linear Regression
- Model Evaluation
- **Indicator Variables**
- Alternative Regression Models
- Time Series Analysis
- Components of a Time Series
- Evaluation Methods of Forecast
- Smoothing Methods of Time Series
- Case Study

Indicator Variables

Definition 6.5: Indicator Variables

- **Indicator variables** – a variable that assigns levels to the qualitative variable (also known as dummy variables).
- Example Variable:
  - Red
  - Green
  - Blue
- Qualitative variables do not have a **scale of measurement**.
- We **cannot assign** numerical values as follows
  - Red= 1
  - Green=2
  - Blue=3

Indicator Variables

Example 6.3: Tool Life Data

- Relate the effective life of a cutting tool (y) used on a lathe to the lathe speed in revolutions per minute (x_1) and type of cutting tool used.
- Tool type is **qualitative** and can be represented as:

$$x_2 = \begin{cases} 0 & ToolA \\ 1 & ToolB \end{cases}$$

- If a **first-order model** is appropriate:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Indicator Variables

- If Tool **type A** is used, model becomes:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- If Tool **type B** is used, model becomes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 + \varepsilon$$

  - Then:

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$$

- Changing from A to B induces a change in the **intercept** (slope is unchanged and identical).
- We assume that the **variance is equal** for all levels of the qualitative variable.

Indicator Variables

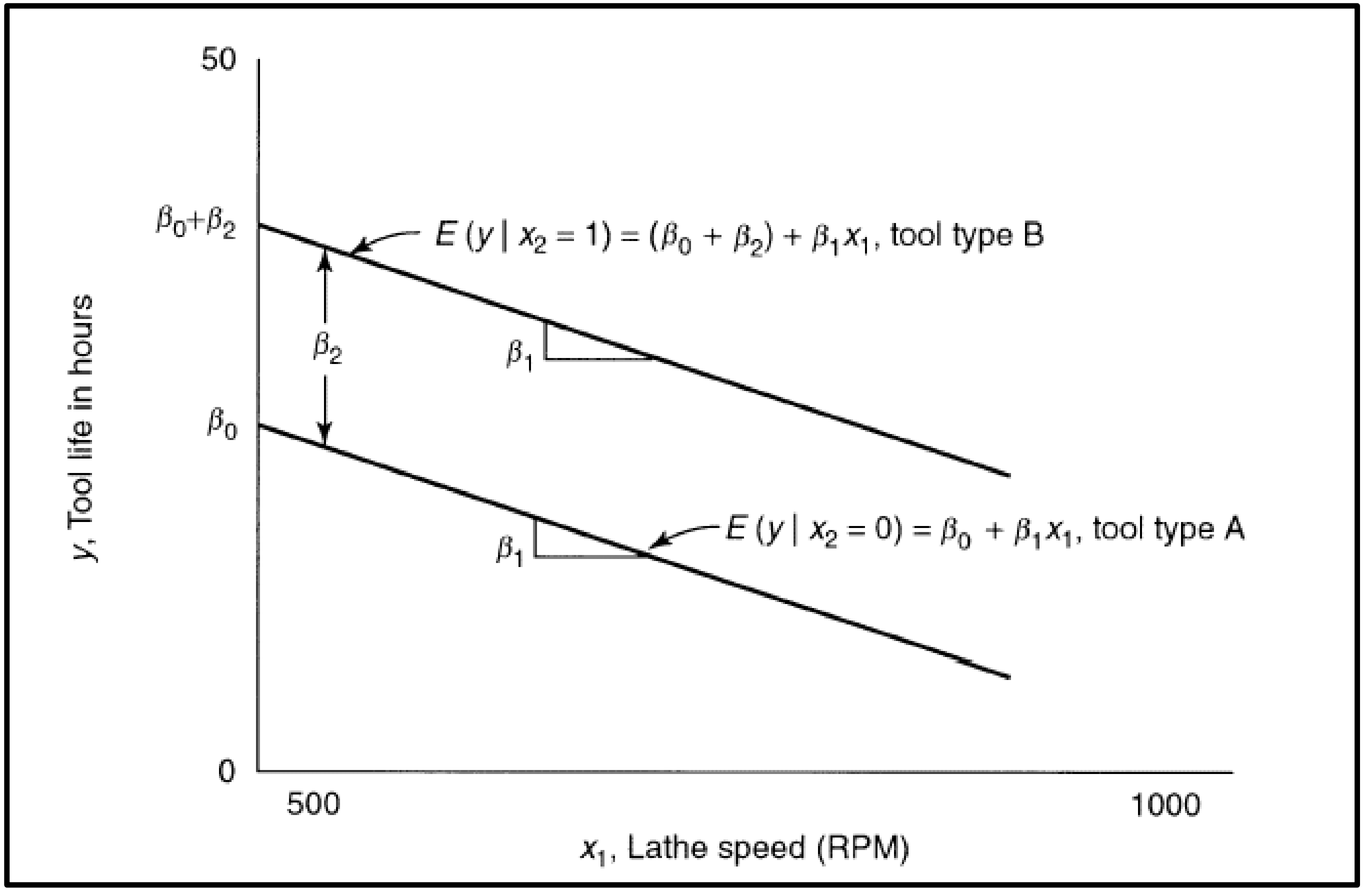


Figure 6.4: Tool Life Data

Indicator Variables

Example 6.3 (Cont.): Tool Life Data

– Twenty observations on tool life and lathe speed are presented and the scatter diagram is shown as follows. Use regression to predict tool life.

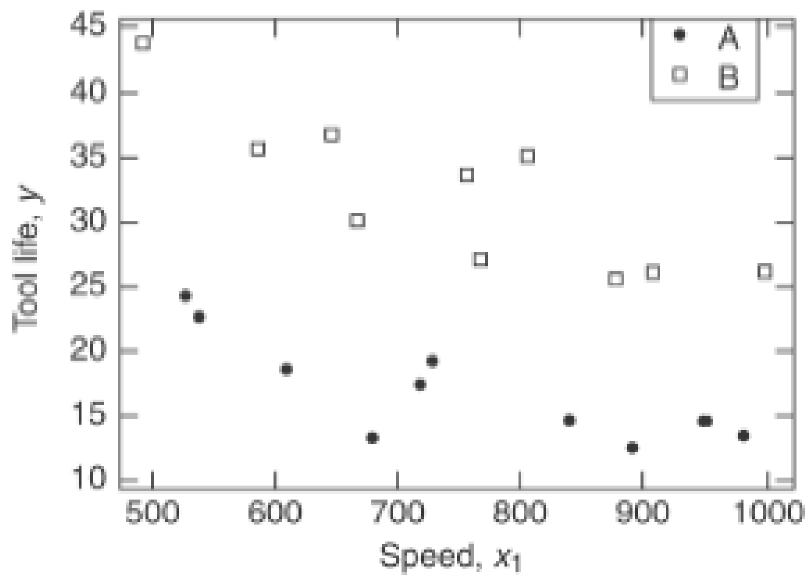


Figure 6.5: Tool Life Data Scatter Plot

Indicator Variables

- toollife = read.csv(“toollife.csv")
- toollifefit=lm(hours~rpm+tooltype,data=toollife)
- summary(toollifefit)

```
Call:
lm(formula = Hours ~ RPM + ToolTypeB, data = ToolLife)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.6255 -1.6308  0.0612  2.2218  5.5044 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.208726   3.738882   9.417 3.71e-08 ***
RPM         -0.024557   0.004865  -5.048 9.92e-05 ***
ToolTypeB   15.235474   1.501220  10.149 1.25e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.352 on 17 degrees of freedom
Multiple R-squared:  0.8787, Adjusted R-squared:  0.8645 
F-statistic:  61.6 on 2 and 17 DF,  p-value: 1.627e-08
```

Indicator Variables

- Tool Type Regression Model

$$Hours = 35.208 - 0.024\,RPM + 15.235\,ToolTypeB$$

Indicator Variables

- For qualitative variables with a levels, we would need $\boldsymbol{a-1}$ ***indicator variables*.**
  - For example, say there were three tool types, A, B, and C. Then two indicator variables (called x2 and x3) will be needed:

x_2	x_3	
0	0	if the observation is from tool type A
1	0	if the observation is from tool type B
0	1	if the observation is from tool type C

the regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

eljalao@up.edu.ph

Outline for this Session

- Regression and Model Building
- Simple and Multiple Linear Regression
- Model Evaluation
- Indicator Variables
- **Alternative Regression Models**
- Time Series Analysis
- Components of a Time Series
- Evaluation Methods of Forecast
- Smoothing Methods of Time Series
- Case Study

Alternative Models of Regression

- Logistic Regression
- Stepwise/Best Subsets Regression

Logistic Regression

Definition 6.6: Logistic Regression

- Logistic regression predicts the **probability** of an outcome that can only have **two values**
- The prediction is based on the use of **one or several predictors** (numerical and categorical).
- Logistic regression produces a **logistic curve**, which is limited to values between 0 and 1.

Logistic Regression

- Logit Function

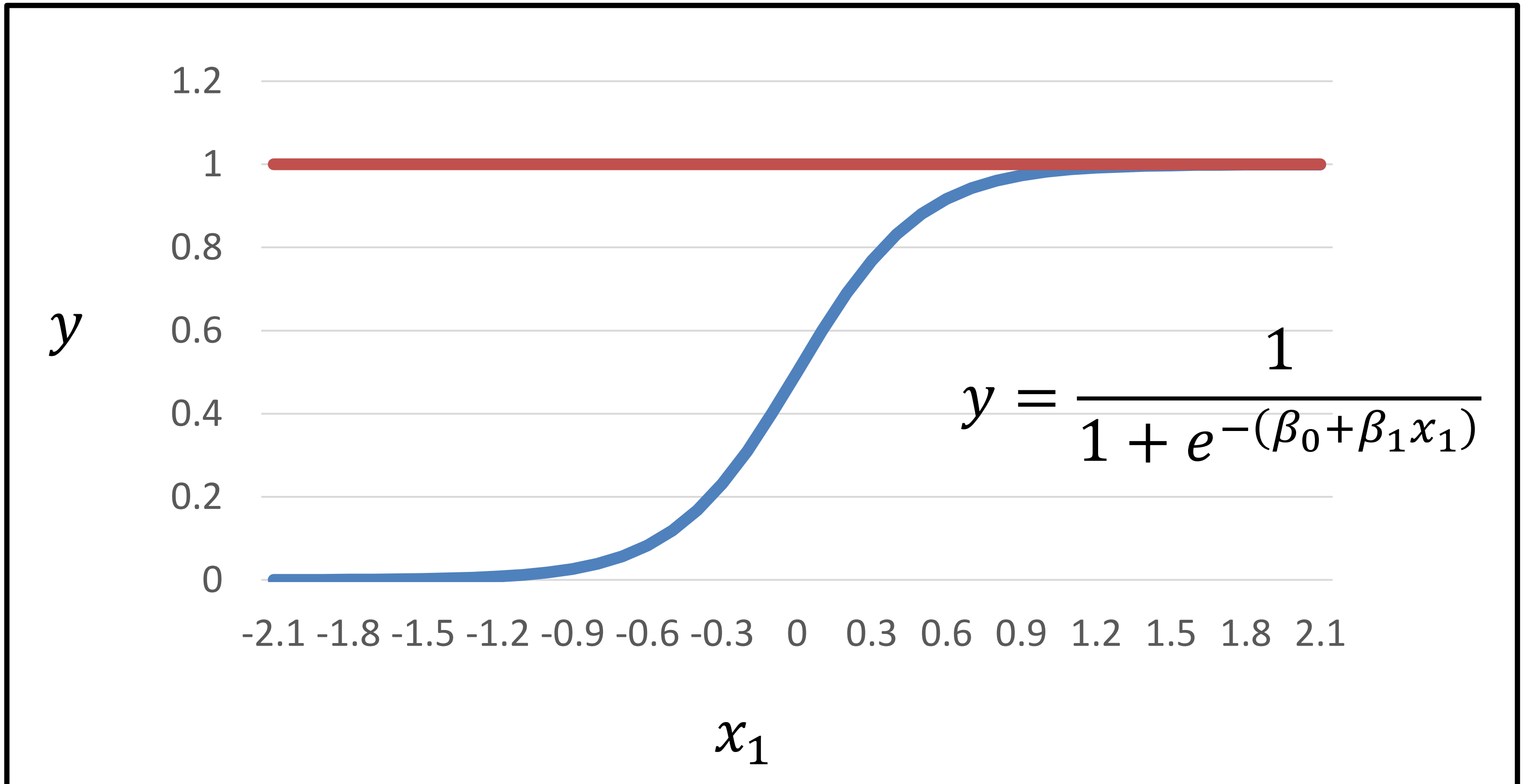


Figure 6.6: Logistic Regression Logit Function

Logistic Regression

Example 6.4 (Cont.): Menarche Data

- Data **contains**:
  - "Age" (average age of age homogeneous groups of girls),
  - "Total" (number of girls in each group),
  - "Menarche" (number of girls in the group who have reached menarche)
- Sources: (Milicer, H. and Szczotka, F., 1966, Age at Menarche in Warsaw girls in 1965, Human Biology, 38, 199-203)

R Code

```
› menarchedata =
  read.csv("menarchedata.csv")
› menarchedata.fit = glm(cbind(menarche,
  total-menarche) ~ age,
  family=binomial(logit), data=menarchedata)
› summary(menarchedata.fit)
› plot(menarche/total ~ age,
  data=menarchedata)
› lines(menarchedata$age,
  menarchedata.fit$fitted, type="l",
  col="red")
```

R Output

```
Call:
glm(formula = cbind(Menarche, Total - Menarche) ~ Age, family = binomial(logit),
    data = menarche)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0363  -0.9953  -0.4900   0.7780   1.3675

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -21.22639    0.77068  -27.54   <2e-16 ***
Age           1.63197    0.05895   27.68   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3693.884  on 24  degrees of freedom
Residual deviance:   26.703  on 23  degrees of freedom
AIC: 114.76

Number of Fisher Scoring iterations: 4
```

Logistic Regression

$$Probability\ of\ Menarchy = \frac{1}{1 + e^{-(-21+1.61\ Age)}}$$

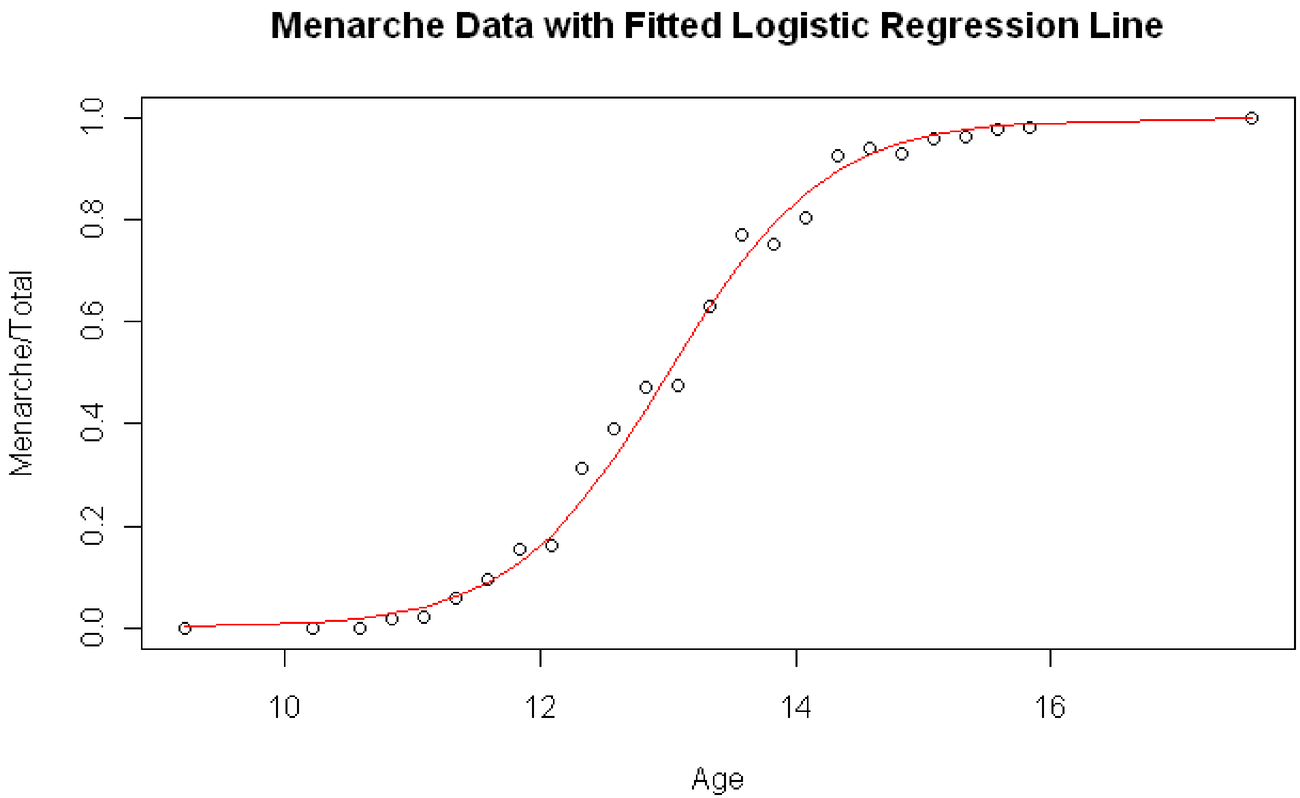


Figure 6.7: Menarche Data

Logistic Regression

- Generated **Model**

$$Probability\ of\ Menarchy = \frac{1}{1 + e^{-(-21+1.61\ Age)}}$$

- The coefficient of "Age" can be **interpreted** as "for every one year increase in age the odds of having reached menarche increase by exp(1.632) = 5.11 times.“
- Prediction for Age = 12

$$Probability\ of\ Menarchy = \frac{1}{1 + e^{-(-21+1.61\ *12)}}$$

$$Probability\ of\ Menarchy = 15.71\%$$

Stepwise Regression

Definition 6.7: Stepwise Regression

- **Stepwise regression**: Enter and remove predictors, in a stepwise manner, until there is no justifiable reason to enter or remove more.
- **Best subsets regression**: Select the subset of predictors that do the best at meeting some well-defined objective criterion.

Stepwise Regression

- Start with no predictors in the "**stepwise model**."
- At each step, enter or remove a predictor based on partial *F*-tests (that is, the *t*-tests).
- Stop when no more predictors can be justifiably entered or removed from the stepwise model.

Stepwise Regression: y versus x1, x2, x3, x4

Alpha-to-Enter: 0.15 **Alpha-to-Remove:** 0.15

Response is **y** on 4 predictors, with N = 13

Step	**1**	**2**	**3**	**4**
Constant	117.57	103.10	71.65	52.58
x4	-0.738	-0.614	-0.237	
T-Value	-4.77	-12.62	-1.37	
P-Value	0.001	0.000	0.205	
x1		1.44	1.45	1.47
T-Value		10.40	12.41	12.10
P-Value		0.000	0.000	0.000
x2			0.416	0.662
T-Value			2.24	14.44
P-Value			0.052	0.000
S	8.96	2.73	2.31	2.41
R-Sq	67.45	97.25	98.23	97.87
R-Sq(adj)	64.50	96.70	97.64	97.44
C-p	138.7	5.5	3.0	2.7

Outline for this Session

- Regression and Model Building
- Simple and Multiple Linear Regression
- Model Evaluation
- Indicator Variables
- Alternative Regression Models
- **Time Series Analysis**
- Components of a Time Series
- Evaluation Methods of Forecast
- Smoothing Methods of Time Series
- Case Study

Time Series Analysis

Definition 6.8: Time Series

- A time series is a collection of observations made **sequentially in time**.

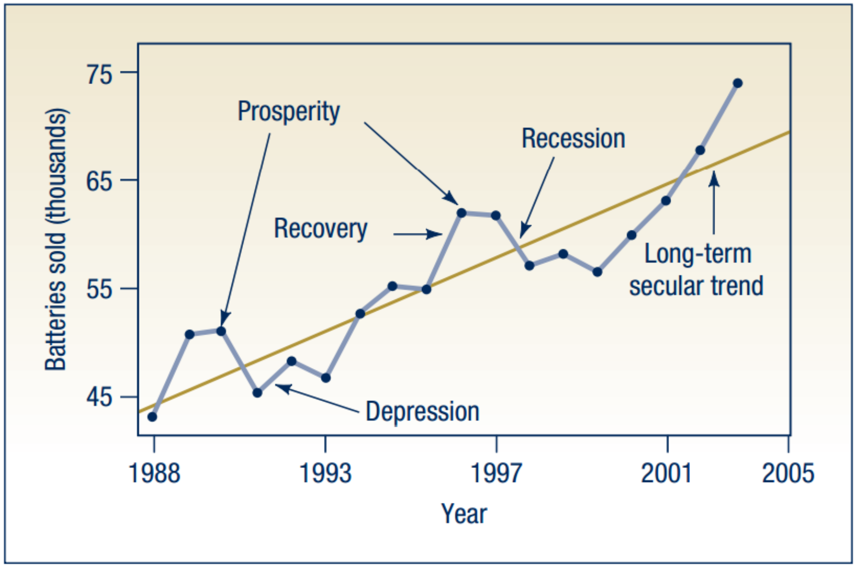


Figure 6.8: Battery Sales by National Battery Sales, Inc., 1988–2005

Time Series Analysis

- A word can be represented by two time series created by moving over and under the word

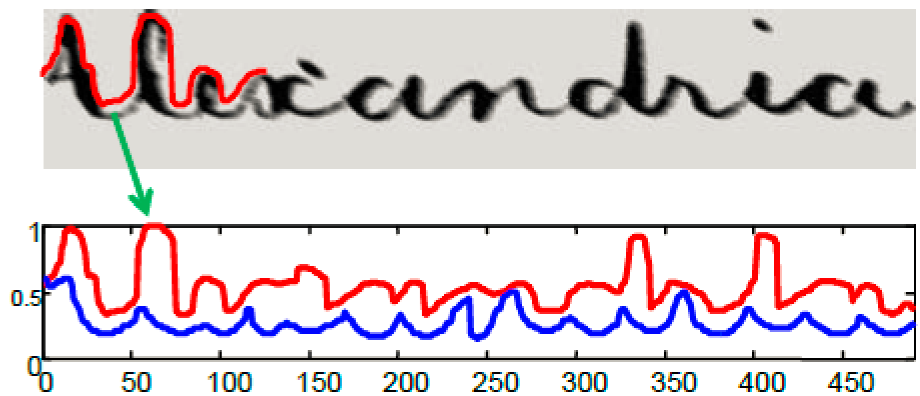


Figure 6.9: Recognizing Words

Time Series Analysis

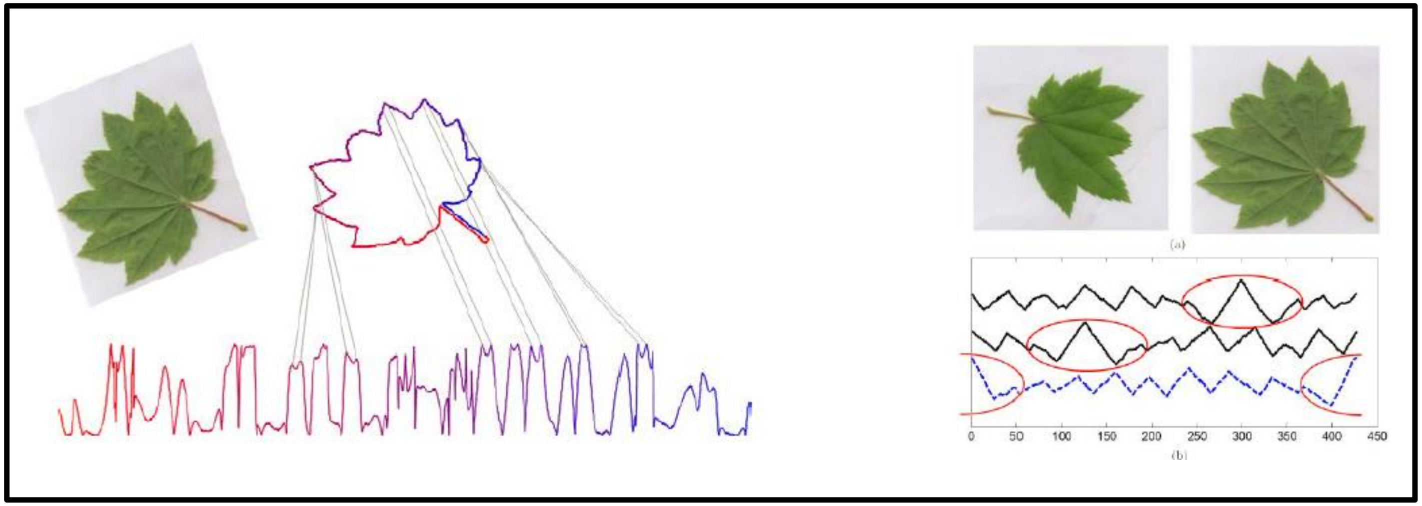


Figure 6.10: Recognizing trees from the leaf images

Time Series Analysis

- Some Time Series Data Mining Tasks
  - Clustering
  - Classification
  - Rule Discovery
  - Anomaly Detection

Time Series Analysis

- Time Series **Clustering**: Identify which time series are similar to each other

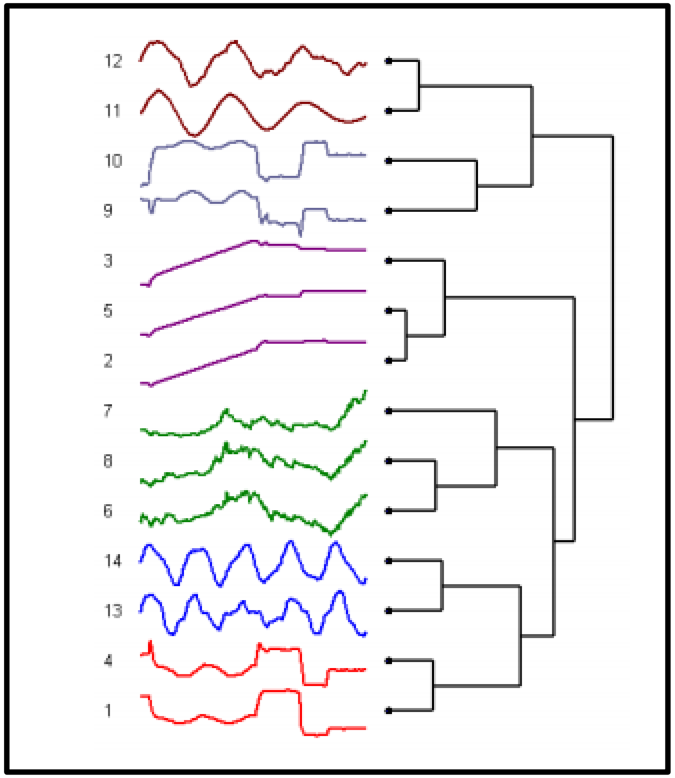


Figure 6.11: Time Series Clustering

Time Series Analysis

- A supervised learning problem aimed at **labeling temporally** structured univariate (or multivariate) sequences of certain (or variable) length.

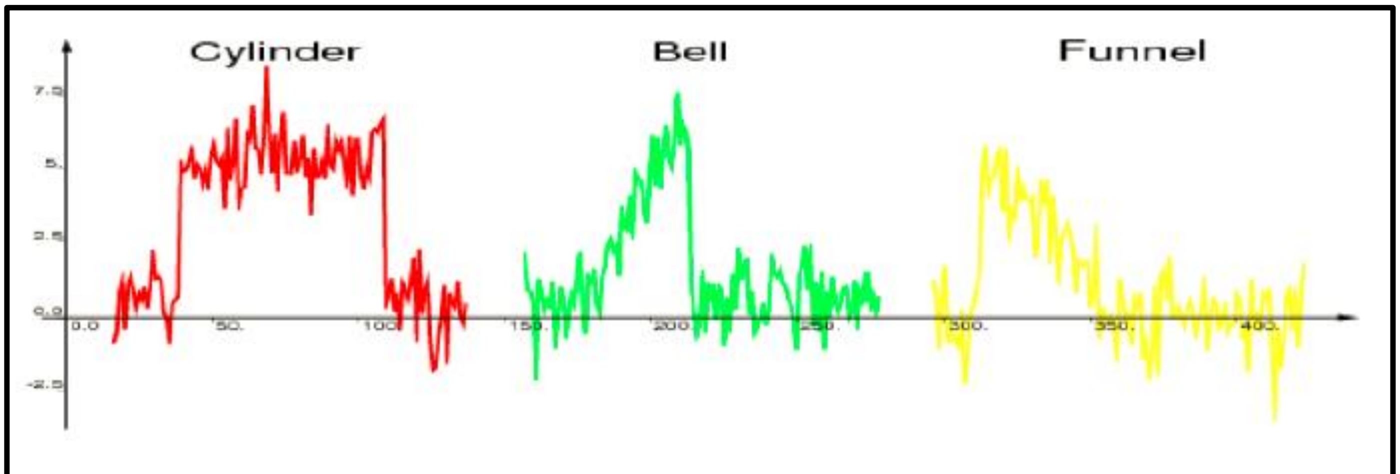


Figure 6.12: Time Series Classification

Time Series Analysis

- Task: Classify grad students based on their faces images transformed into “time series”

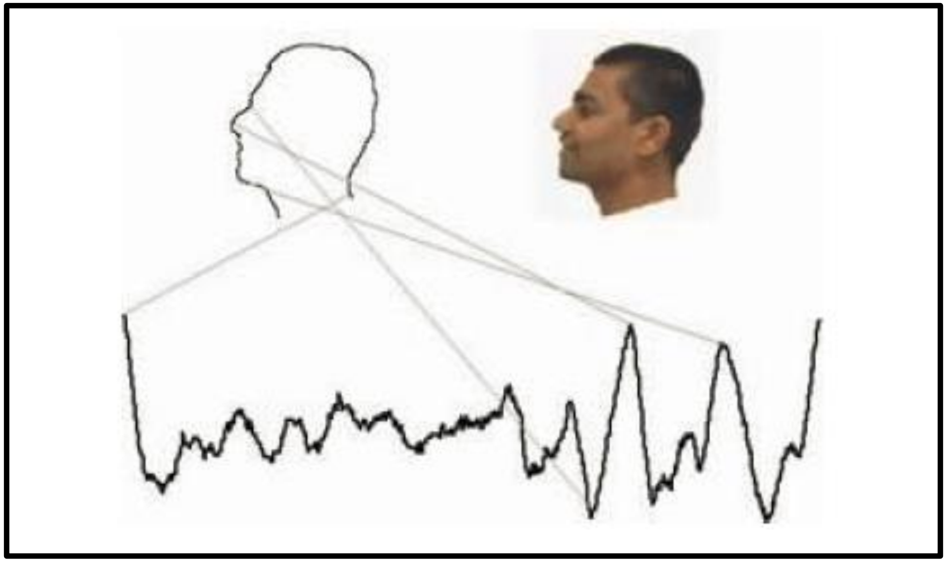


Figure 6.13: Time Series Classification

Time Series Analysis

- Identify sequence of sales:
- - 60% of clients who placed an online order in */company/products/product1.html*, also placed an online order in */company1/products/product4* within 15 days.

Murali K. Kadimi

Time Series Analysis

- Identify anomalous transactions

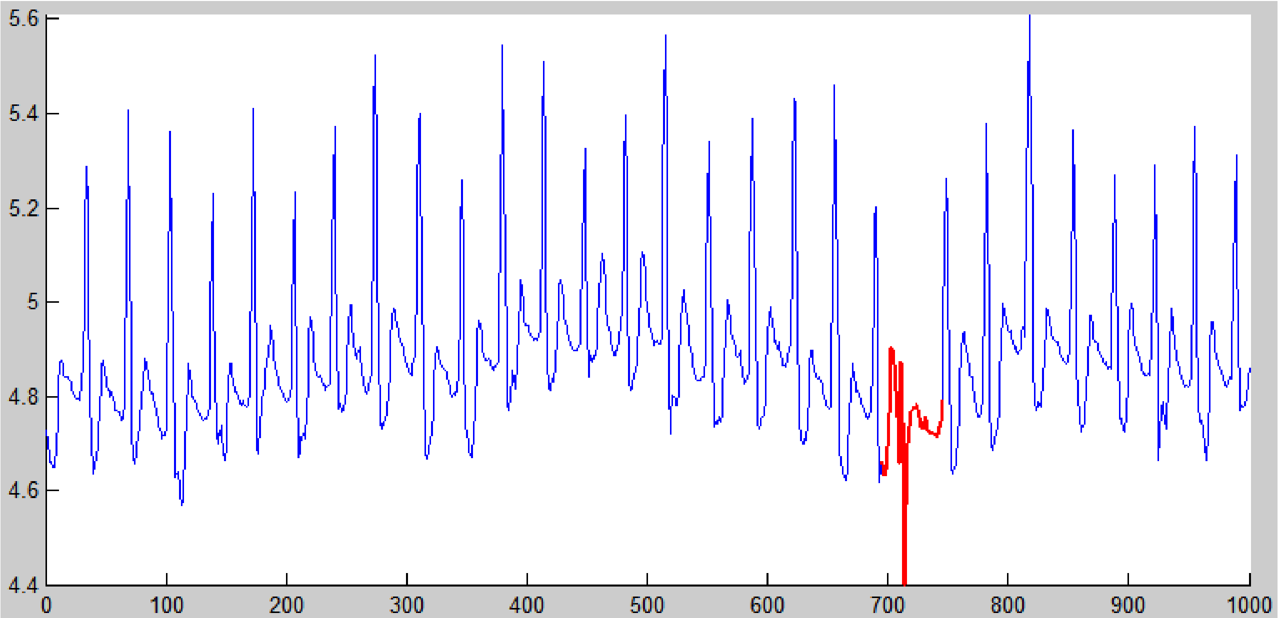


Figure 6.14: Time Series Classification: Anomaly Detection

http://www.anomalydetectionresearch.com/

Time Series Analysis

- Qualitative Modeling
  - Expert Opinion
    - Informed personal insight is always useful.
    - Panel consensus reconciles different views.
    - Delphi method seeks informed consensus.
  - Survey Techniques
    - Random samples give population profile.
    - Stratified samples give detailed profiles of population segments.
- Quantitative Modeling
  - Deterministic modeling
  - Regression modeling
  - Stochastic modeling

Outline for this Session

- Regression and Model Building
- Simple and Multiple Linear Regression
- Model Evaluation
- Indicator Variables
- Alternative Regression Models
- Time Series Analysis
- **Components of a Time Series**
- Evaluation Methods of Forecast
- Smoothing Methods of Time Series
- Case Study

Components of a Time Series

- The pattern or behavior of the data in a time series has **several components.**
- Theoretically, any time series can be decomposed into:
  - Trend
  - Cyclical
  - Seasonal
  - Irregular
- However, this decomposition is often **not straight-forward** because these factors interact.

Components of a Time Series

Definition 6.9: Trend Component

- Accounts for the **gradual shifting** of the time series to relatively higher or lower values over a long period of time.
- Trend is usually the result **of long-term factors** such as changes in the population, demographics, technology, or consumer preferences.

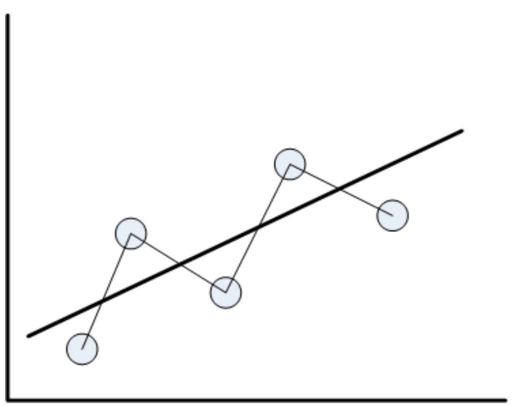


Figure 6.15: Trend Component

Components of a Time Series

Definition 6.10: Seasonal Component

- Accounts for **regular patterns of variability** within certain time periods, such as a year.
- The variability does not always correspond with the seasons of the year (i.e. winter, spring, summer, fall).

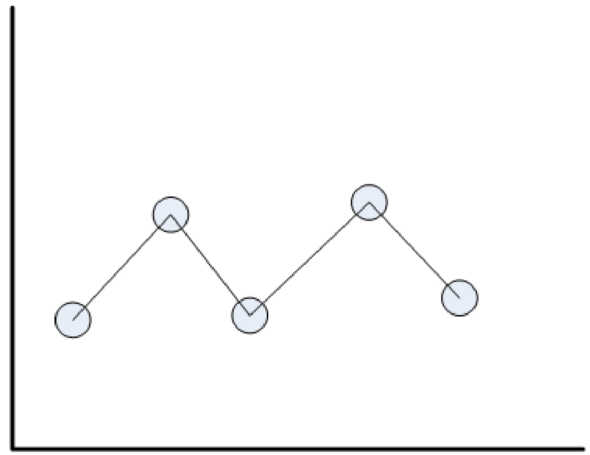


Figure 6.16: Seasonality Component

Components of a Time Series

Definition 6.11: Cyclical Component

- Any regular pattern of sequences of values above and below the trend line lasting more than one year can be attributed to the **cyclical component.**
- Usually, this component is due to multiyear cyclical movements in the economy.

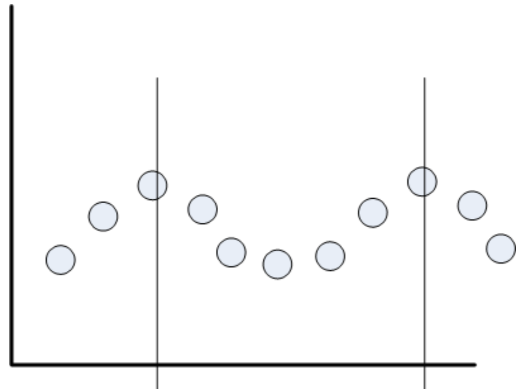


Figure 6.17: Cyclical Component

Components of a Time Series

Definition 6.12: Irregular Component

- Component of a time series **not accounted** for by the other three components
- **Random Error**
- Usually **ignored** in analysis but forms the basis for model evaluation (regression)

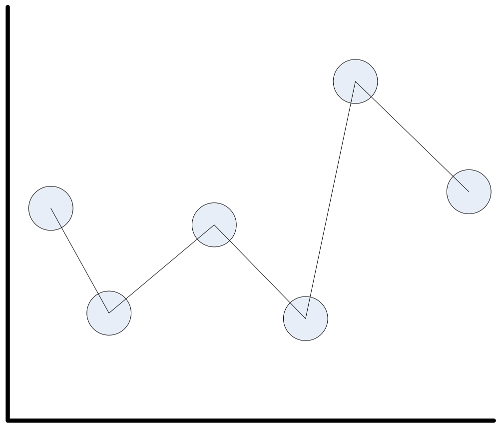


Figure 6.18: Irregular Component

Outline for this Session

- Regression and Model Building
- Simple and Multiple Linear Regression
- Model Evaluation
- Indicator Variables
- Alternative Regression Models
- Time Series Analysis
- Components of a Time Series
- **Evaluation Methods of Forecast**
- Smoothing Methods of Time Series
- Case Study

Evaluation Methods of Forecasts

Definition 6.13: Definition of Errors

- Given the observations y_t of a time series and the corresponding forecasts $\hat{y}_t$ using the k past periods, **the prediction error** at time t

$$e_t = y_t - \hat{y}_t \tag{6.9}$$

- The percentage prediction error at time t is

$$e_t^p = \frac{y_t - \hat{y}_t}{y_t} \times 100\% \tag{6.10}$$

Evaluation Methods of Forecasts

- There are **three measures of accuracy** of the fitted models: MAPE, MAD and MSD for each of the sample forecasting and smoothing methods.
- For all three measures, the smaller the value, the better the fit of the model.
- Use these statistics to compare the **fit of the different methods.**

Evaluation Methods of Forecasts

- Mean Absolute Percentage Error:

$$MAPE = \frac{\sum_{t=1}^{k} \left| e_t^p \right|}{k} \qquad (6.11)$$

- Mean Absolute Deviation:

$$MAD = \frac{\sum_{t=1}^{k} |e_t|}{k} \qquad (6.12)$$

- Mean Squared Deviation:

$$MSD = \frac{\sum_{t=1}^{k} e_t^2}{k} \qquad (6.13)$$

Evaluation Methods of Forecasts

- MAPE
  - Expresses accuracy as a percentage of the error. For example, if the MAPE is 0.05, on average, the forecast is off by 5%.
- MAD
  - Expresses accuracy in the same units as the data, which helps conceptualize the amount of error.
- MSD
  - A commonly-used measure of accuracy of fitted time series values. This is differentiable hence a minimum can be obtained.

Evaluation Methods of Forecasts

Example 6.5 : Forecast Evaluation Example

Actual Sales	Forecasted Sales	
	Model 1	Model 2
56	54	50
43	44	40
22	20	22
24	19	20
55	50	49
MAPE	0.0898	0.0905
MAD	2.6	3.8
MSD	11.8	19.4

eljalao@up.edu.ph

Outline for this Session

- Regression and Model Building
- Simple and Multiple Linear Regression
- Model Evaluation
- Indicator Variables
- Alternative Regression Models
- Time Series Analysis
- Components of a Time Series
- Evaluation Methods of Forecast
- **Smoothing Methods of Time Series**
- Case Study

Smoothing Methods for Time Series

- **Smoothing** a time series: to eliminate some of short-term fluctuations.
- Smoothing also can be done to **remove seasonal fluctuations**, i.e., to deseasonalize a time series.
  - Arithmetic Moving Average
  - Exponential Smoothing Methods
  - Holt-Winters method for Exponential Smoothing
- These models are **deterministic**

Smoothing Methods for Time Series

- Simple Averages - **quick, inexpensive**
- Moving Average method
  – Consists of computing an average of the **most recent n data** values for the series and using this average for forecasting the value of the time series for the next period.

$$\hat{y}_{t+1} = \frac{y_t + y_{t-1} + \cdots + y_{t-n}}{n} \qquad (6.14)$$

Smoothing Methods for Time Series

- **Moving averages** are useful if one can assume item to be forecast will stay steady over time.
- **Series of arithmetic means** – used only for smoothing, provides overall impression of data over time
  - The smaller the number, the more weight given to recent periods. This is desirable when there are sudden shifts in the level of the series.
  - The greater the number, less weight is given to more recent periods and the greater the smoothing effect.

Smoothing Methods for Time Series

Example 6.6 : Births Dataset

- An example is a data set of the number of births per month in New York city, from January 1946 to December 1959

R Code

- `library("TTR")`
- `births = read.csv("births.csv")`
- `birthsts = ts(births[,2], frequency=12, start=c(1946,1))`
- `birthstsSMA2 = SMA(birthsts,n=2)`
- `birthstsSMA10 = SMA(birthsts,n=5)`
- `birthstsSMA20 = SMA(birthsts,n=10)`
- `total = cbind(birthsts,birthstsSMA2,birthstsSMA10,birthstsSMA20)`
- `plot(total, plot.type="single", col = 1:ncol(total), lwd = c(2, 2, 2,2))`
- `legend("bottomright", colnames(total), col=1:ncol(total), lty = c(1, 1, 1,1), cex=.5, y.intersp = 1)`

Smoothing Methods for Time Series

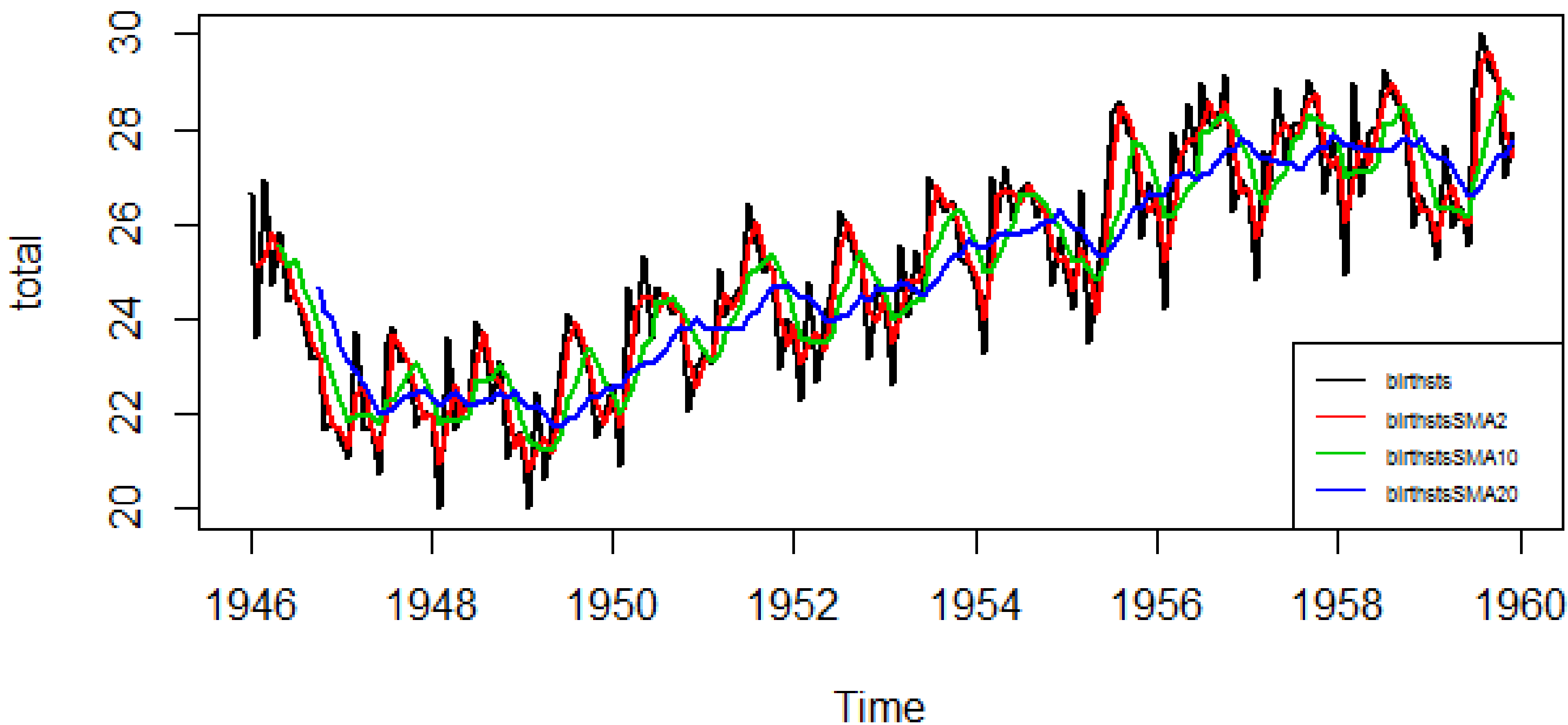


Figure 6.19: Smoothing Methods with Different n

Smoothing Methods for Time Series

- Notes on Moving Averages
  - MA models do not provide information about **forecast confidence**.
  - We can not calculate **standard errors.**
  - We can not explain the stochastic component of the time series. This stochastic component creates the error in our forecast.

Smoothing Methods for Time Series

- Exponential Smoothing Methods
  - **Single Exponential Smoothing (Averaging)**
    - Used for a series without a trend and a seasonal component.
  - **Double Exponential Smoothing**
    - Double Exponential Smoothing is for a series with a trend but without a seasonal component.
  - **Winter's Model.**
    - Winter's model is for a series with a trend and seasonal component.

Smoothing Methods for Time Series

Definition 6.14: Single Exponential Smoothing

- Averaging (smoothing) past values of a **series in a decreasing (exponential) manner.**
- The **observations are weighted** with more weight being given to the more recent observations

$$\hat{y}_{t+1} = \alpha y_t + (1 - \alpha)\hat{y}_t \qquad (6.16)$$

- New forecast = α × (old observation) + (1- α) × old forecast.
- The equation can be **rewritten as:**

$$\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t) \qquad (6.17)$$

Smoothing Methods for Time Series

- We need a **smoothing constant** α, an initial forecast, and an actual value.
- The smoothing constant serves as the **weighting factor**.
  - When a is close to 1, the new forecast will include a substantial adjustment for any error that occurred in the preceding forecast.
  - When a is close to 0, the new forecast is very similar to the old forecast.
- The smoothing constant α is not an arbitrary choice - but generally falls between 0.1 and 0.4.

R Code

- `birthstssesa01 =  HoltWinters(birthsts,alpha=0.1, beta=FALSE, gamma=FALSE)`
- `birthstssesa02 =  HoltWinters(birthsts,alpha=0.2, beta=FALSE, gamma=FALSE)`
- `birthstssesa09 =  HoltWinters(birthsts,alpha=0.9, beta=FALSE, gamma=FALSE)`
- `total = cbind(birthsts,birthstssesa01$fitted[,1],birthstssesa02$fitted[,1],birthstssesa09$fitted[,1])`
- `plot(total, plot.type="single", col = 1:ncol(total), lwd = c(2, 2,2,2))`
- `legend("bottomright", c("Original","0.1","0.2","0.9"), col=1:ncol(total), lty = c(1, 1), cex=.5, y.intersp = 1)`

Choice of Alpha

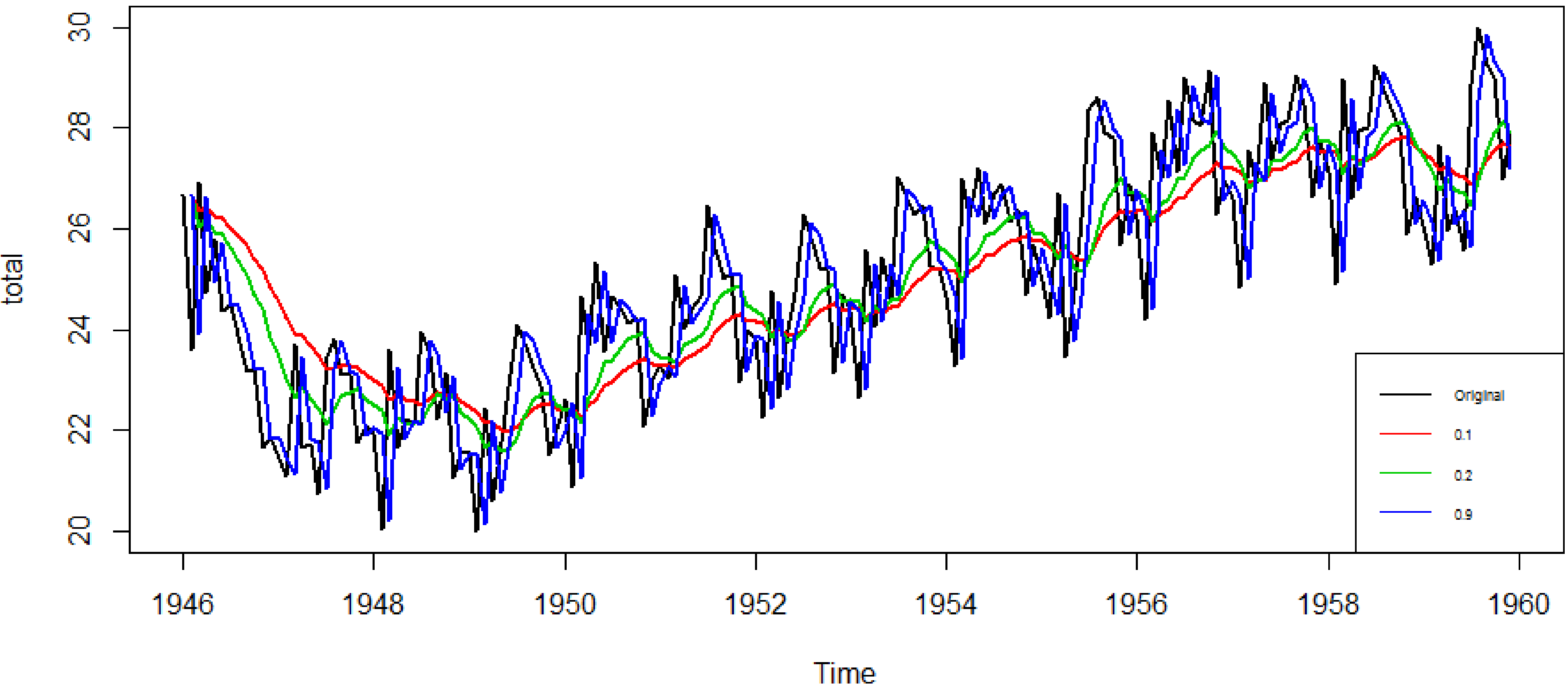


Figure 6.20: Choice of Alpha

Smoothing Methods for Time Series

- Use a **tracking signal** (measure of errors over time) and setting limits.
- For example, if we forecast n periods, count the number of negative and positive errors.
- If the number of positive errors is substantially less or greater than $n/2$, then the process is out of control.

Smoothing Methods for Time Series

Example 6.7 : Use of Tracking Signal

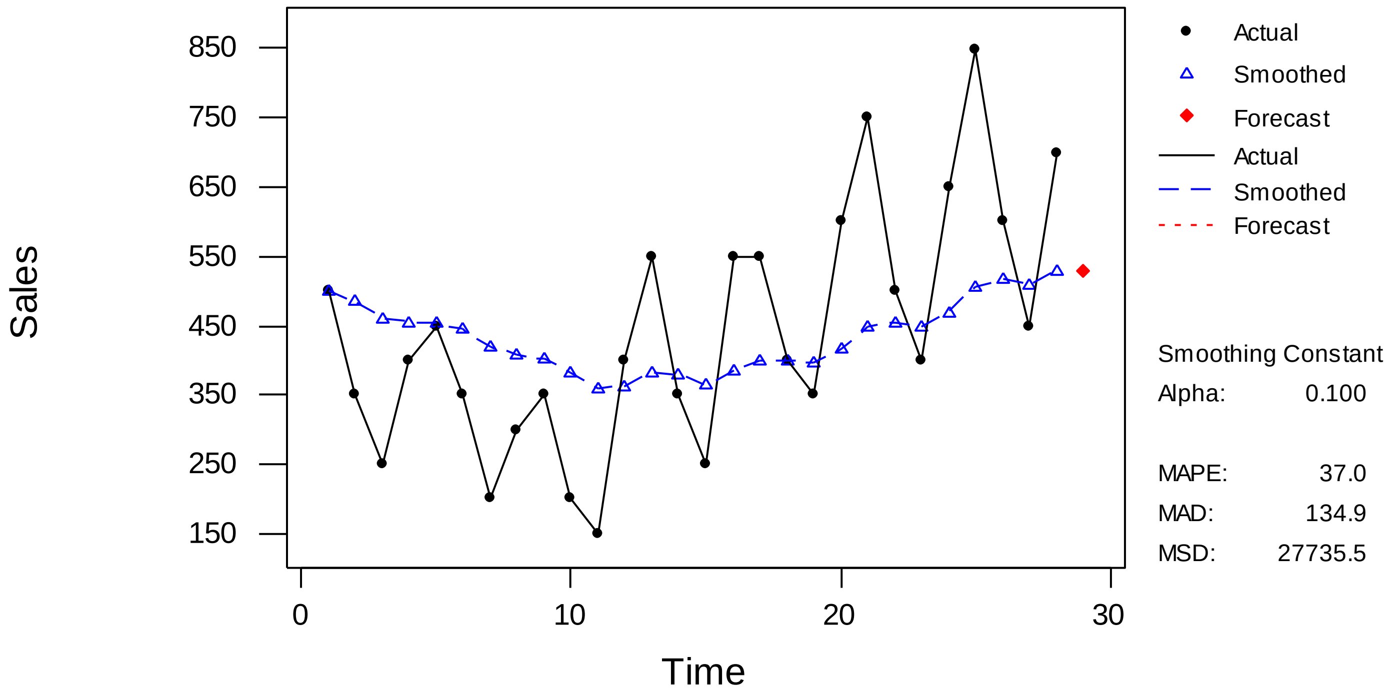


Figure 6.21: Tracking Sales Data

Smoothing Methods for Time Series

- Can also use 95% *prediction interval*

$$\hat{y} \pm z_{\frac{\alpha}{2}} \sqrt{MSD} \tag{6.18}$$

- If the forecast error is **outside of the interval**, use a new optimal α.
- Looking back at the 0.1 single exponential smoothing:

$$\hat{y} \pm 1.96 * \sqrt{27735.5} = \hat{y} \pm 326.4$$

  – Observation #21 is out-of-control. We need to re-evaluate alpha level because this technique is biased.

Outline for this Session

- Regression and Model Building
- Simple and Multiple Linear Regression
- Model Evaluation
- Indicator Variables
- Alternative Regression Models
- Time Series Analysis
- Components of Time Series
- Evaluation Methods of Forecast
- Smoothing Methods of Time Series
- **Case Study**

Case Study 6

- House Data and Airline Data

References

- Notes and Datasets from Montgomery, Peck and Vining, Introduction to Linear Regression Analysis 4th Ed. Wiley
- Notes from G. Runger, ASU IEE 578
- Trevor Hastie, Rob Tibshirani, Friedman: Elements of Statistical Learning (2nd Ed.) 2009
- *Regression Analysis by Example* (4th ed.) by Chatterjee and Hadi (Wiley, New York, 2006).
- http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html
- http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/src/timeseries.html