L- 11: Predictive Analytics(Continued) & Forecasting Models

Agenda

- Model validation
- Ridge and lasso models
- Assumptions of Linear regression
- Logistic regression

Classical Linear Regression (OLS)

- Explanatory and Response Variables are Numeric
- Relationship between the mean of the response variable and the level of the explanatory variable assumed to be approximately linear (straight line)
- Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma)$$

- $beta_1 > 0 \Rightarrow$ Positive Association
- $beta_1 < 0 \Rightarrow$ Negative Association
- $beta_1 = 0 \Rightarrow$ No Association

Multiple regression

Numeric Response variable (y)

p Numeric predictor variables

Model:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

- Population Model for mean response:

$$E(Y \mid x_1, \ldots x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Least Squares Fitted (predicted) equation, minimizing *SSE*:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \qquad SSE = \sum \left(Y - \hat{Y} \right)^2$$

Accuracy of a model

By Using the following the strength of the linear model can be tested

1) Coefficient of determination (R^2)

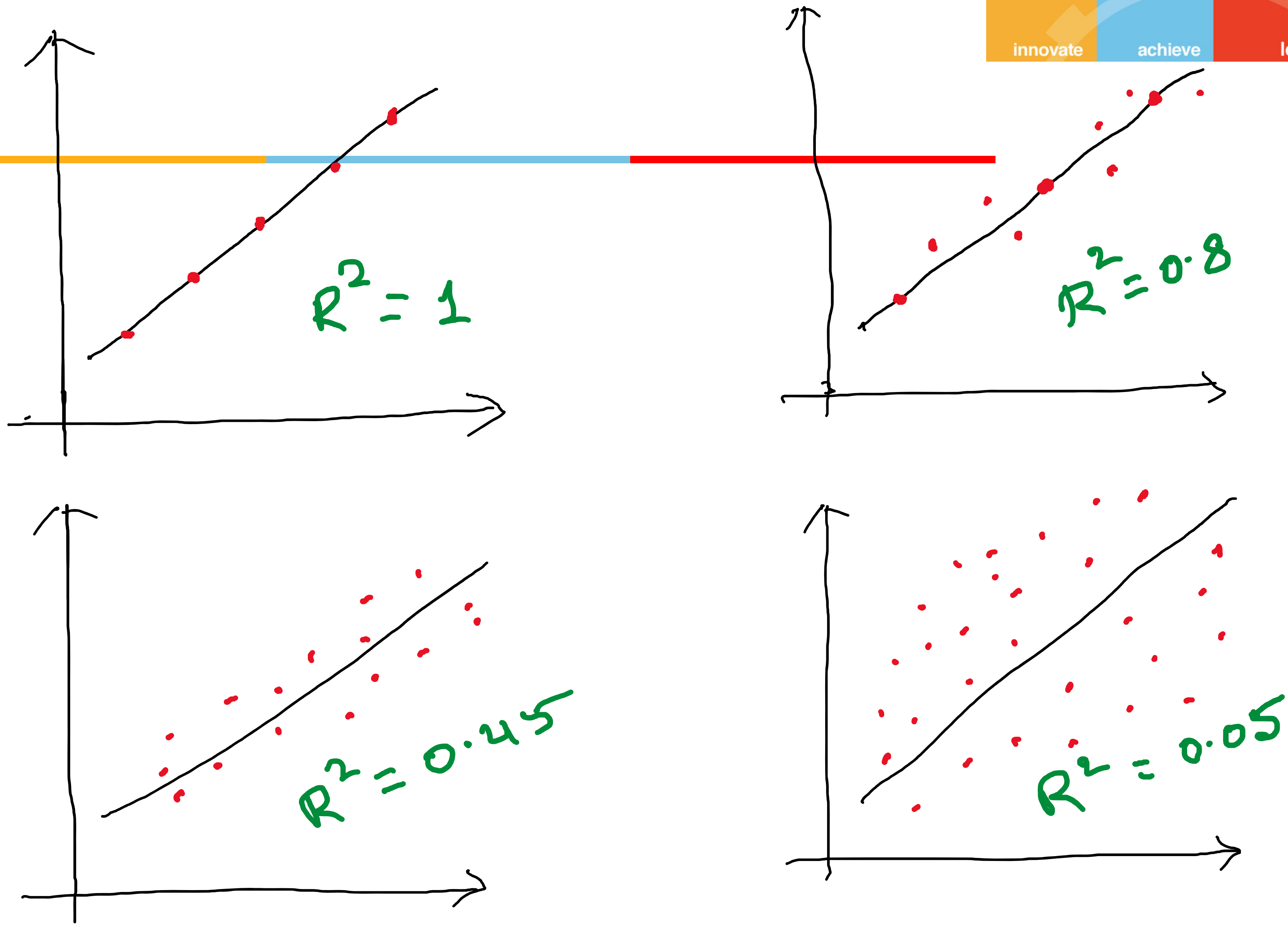
2) Residual Standard error (RSE)

RSS $\longrightarrow$ Residual sum of squares

$$= \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i) \right)^2$$

TSS $\longrightarrow$ $\sum_{i=1}^{n} (y_i - \bar{y})^2$ $\longrightarrow$ mean of respective variables

$$R^2 = 1 - \frac{RSS}{TSS}$$


$R^2 = 1$
$R^2 = 0.8$
$R^2 = 0.45$
$R^2 = 0.05$

R – Squared vs Adjusted R - Squared

- In multiple regression, adjusted R – squared is better metric than R – squared asses the goodness of fit of the model

- R – squared always increases if additional variables are added into model , even if they are not related to the dependent variable

Regularization

- Over fitting can be solved with regularization
- Regularization can be done by putting constraints on the coefficients and variables.
- LASSO: Least Absolute Shrinkage and Selection Operator

  Some coefficients can be dropped(i.e become zero)
- RIDGE: The coefficients will approach zero, but never dropped

Lasso & Ridge

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

- OLS estimation: $$\min SSE = \sum \left(Y - \hat{Y} \right)^2$$
- LASSO estimation: $$\min SSE = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right|$$
- Ridge regression estimation: $$\min SSE = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j^{\,2} \right|$$

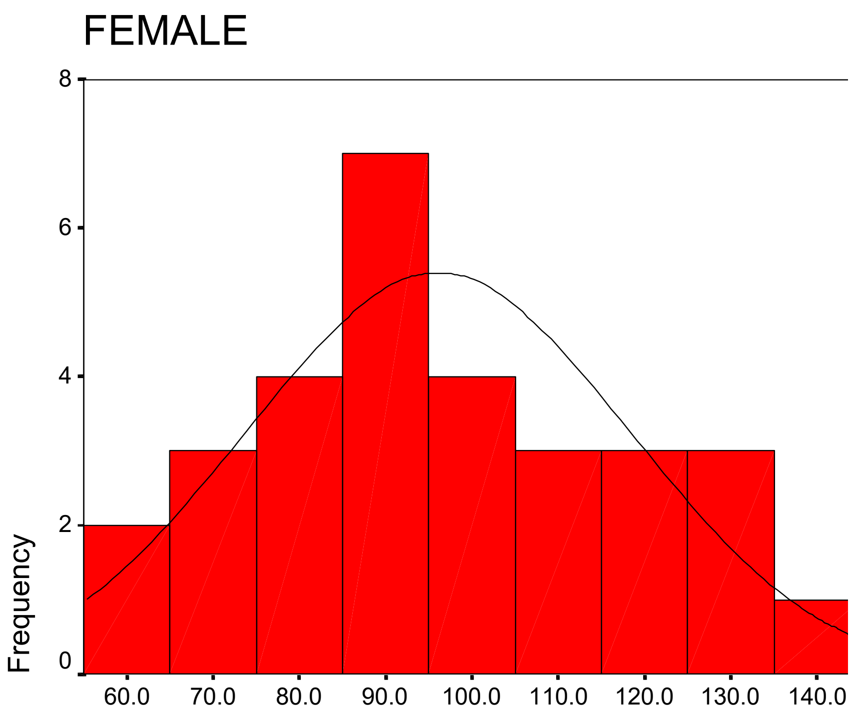
Assumptions in Regression Analysis

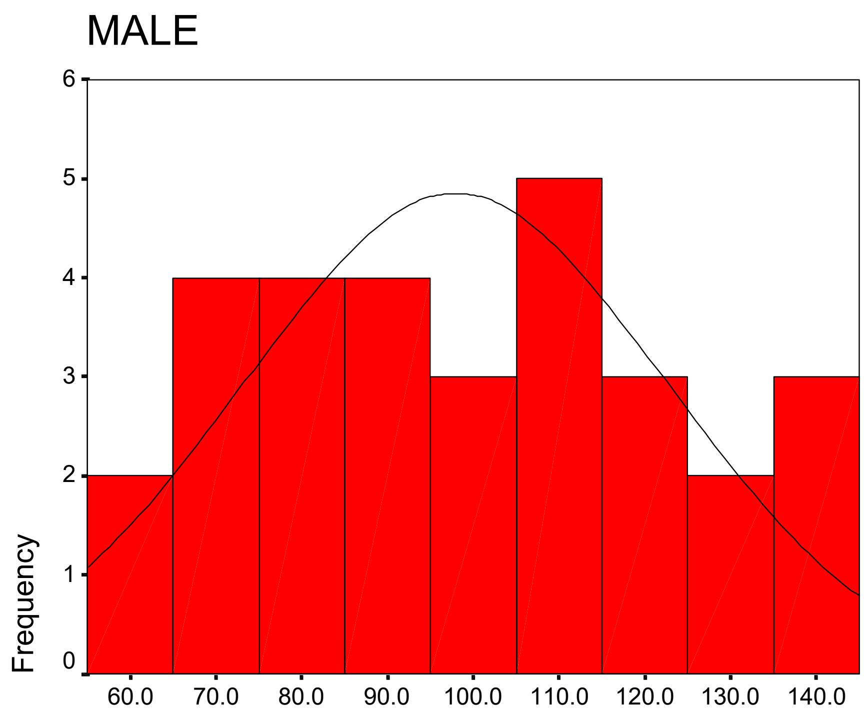
Assumptions

- The distribution of residuals is normal (at each value of the dependent variable).
- The variance of the residuals for every set of values for the independent variable is equal.
  - ✓ violation is called heteroscedasticity.
- The error term is additive
  - ✓ no interactions.
- At every value of the dependent variable the expected (mean) value of the residuals is zero
  - ✓ No non-linear relationships

- The expected correlation between residuals, for any two cases, is 0.
  - The independence assumption (lack of autocorrelation)
- ✓ All independent variables are uncorrelated with the error term.
- ✓ No independent variables are a perfect linear function of other independent variables (no perfect multicollinearity)
- ✓ The mean of the error term is zero.

Assumption 1: The Distribution of Residuals is Normal at Every Value of the Dependent Variable


FEMALE
8
6
4
2
0
Frequency
60.0
70.0
80.0
90.0
100.0
110.0
120.0
130.0
140.0


MALE
6
5
4
3
2
1
0
Frequency
60.0
70.0
80.0
90.0
100.0
110.0
120.0
130.0
140.0

Non-Normality

Skew and Kurtosis

- Skew – much easier to deal with
- Kurtosis – less serious anyway

Transform data

- removes skew
- positive skew – log transform
- negative skew - square

Assumption 2: The variance of the residuals for every set of values for the independent variable is equal.

Heteroscedasticity

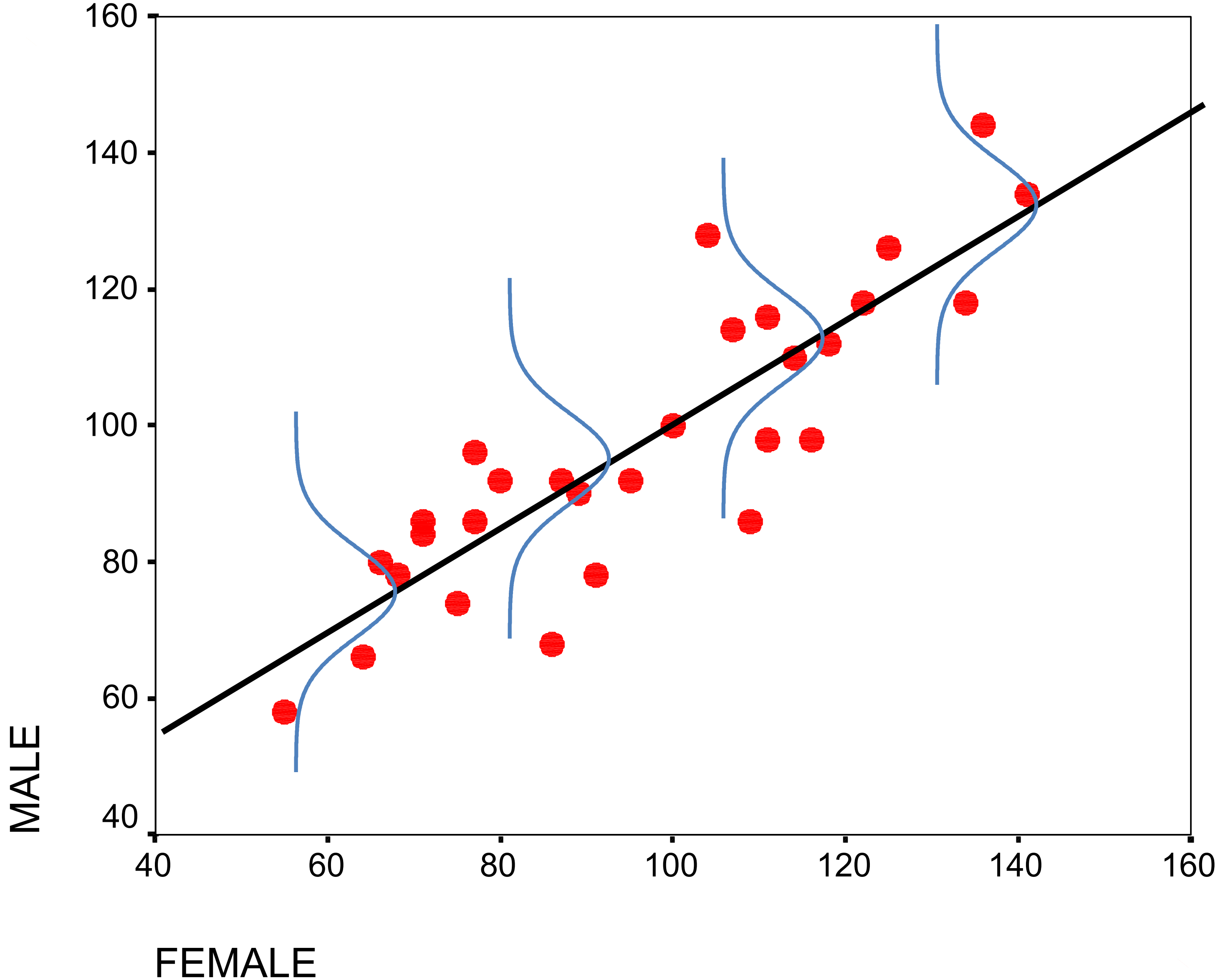
This assumption is a about heteroscedasticity of the residuals

- Hetero=different
- Scedastic = scattered

We don't want heteroscedasticity

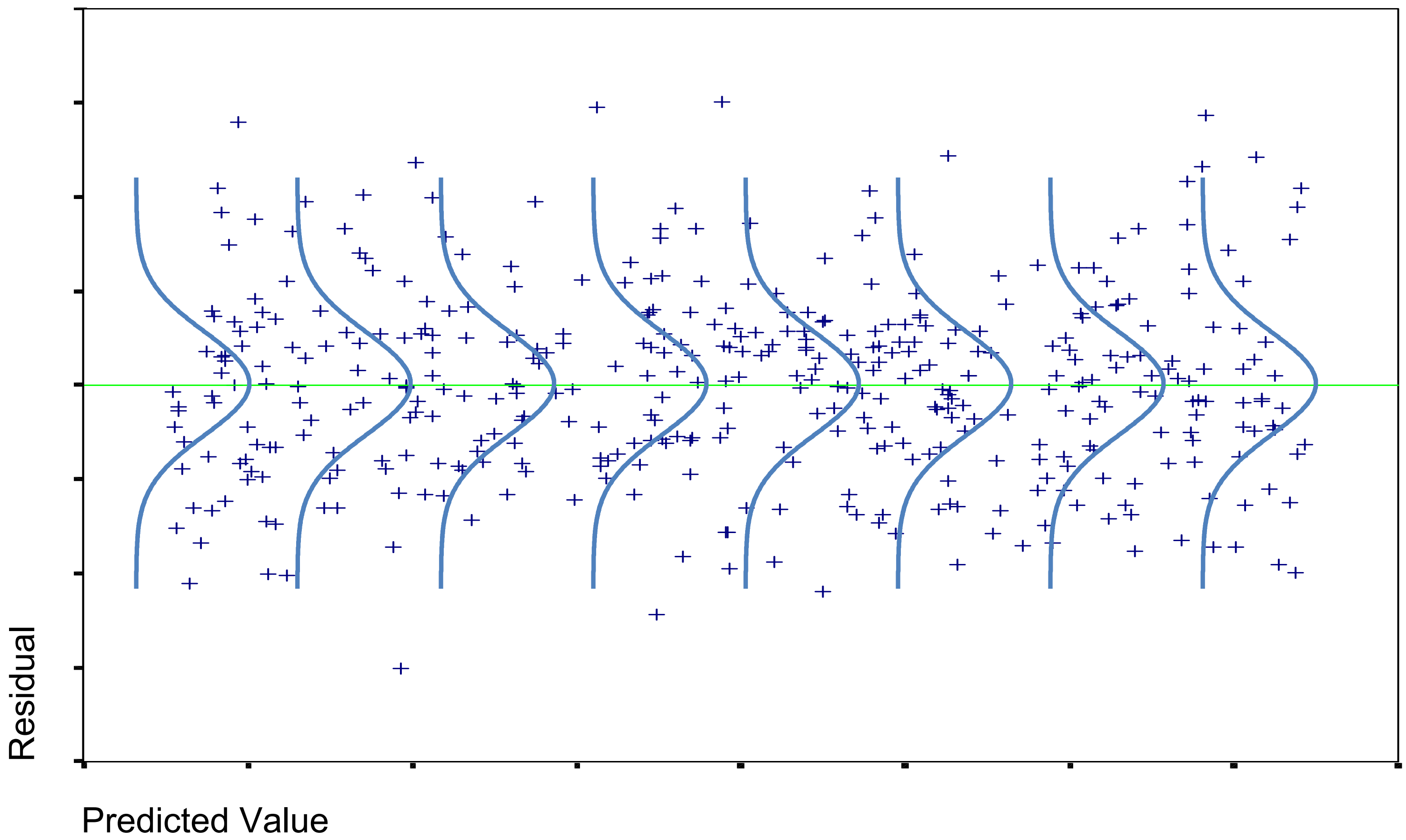
- we want our data to be homoscedastic

Draw a scatterplot to investigate

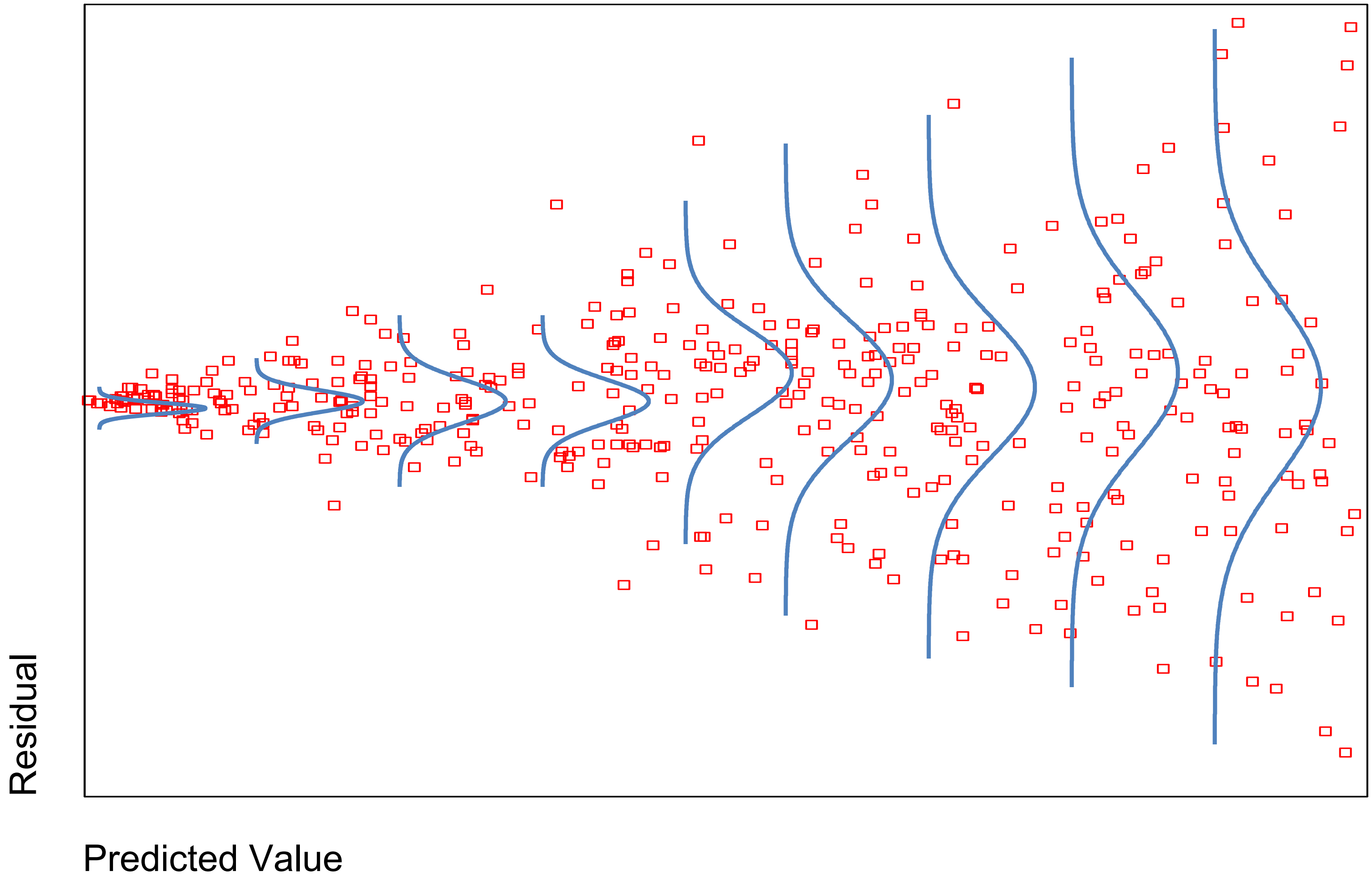


160
140
120
100
80
60
40
MALE
40
60
80
100
120
140
160
FEMALE

Good – no heteroscedasticity



Bad – heteroscedasticity

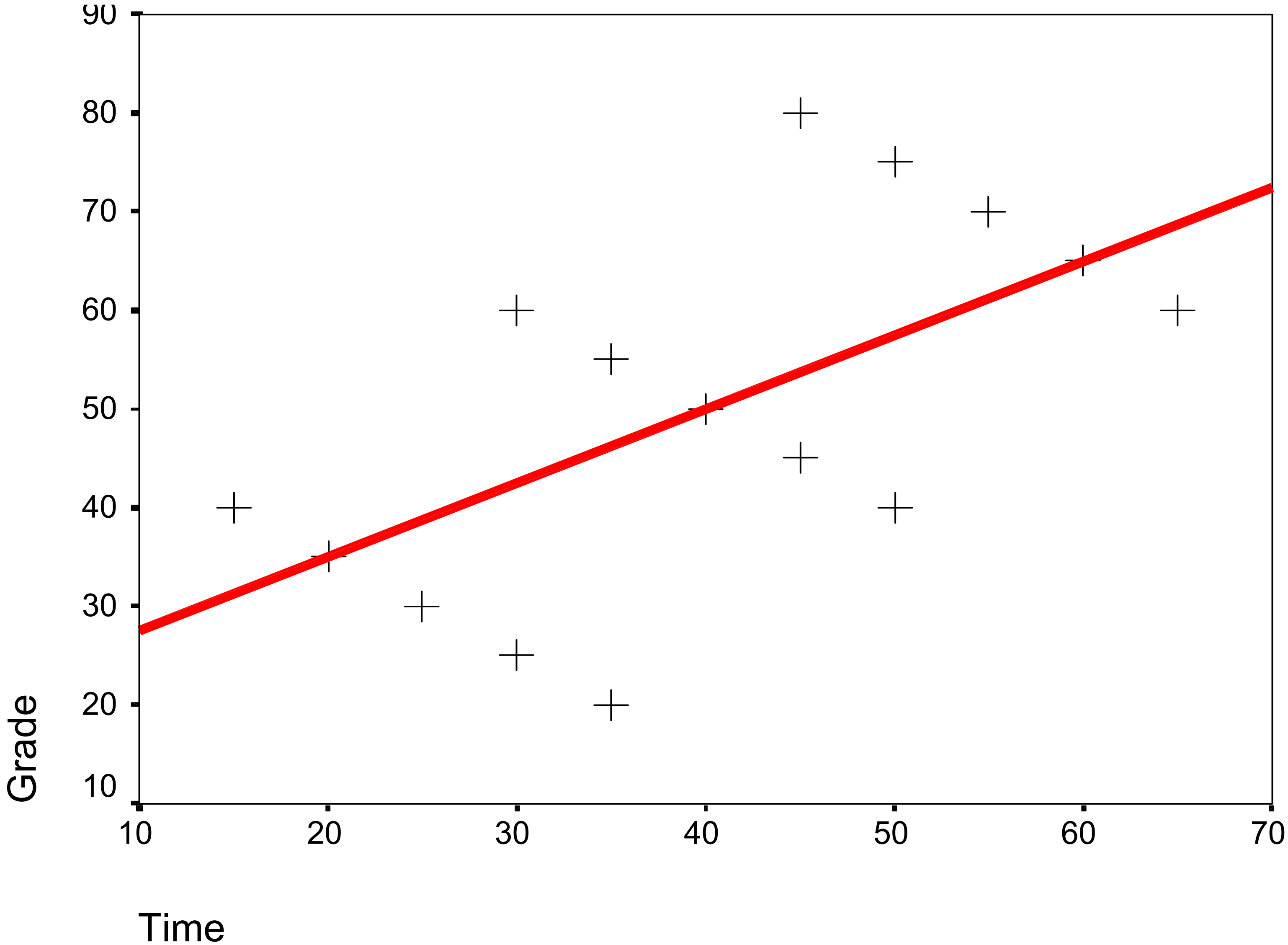


Assumption 3: The Error Term is Additive

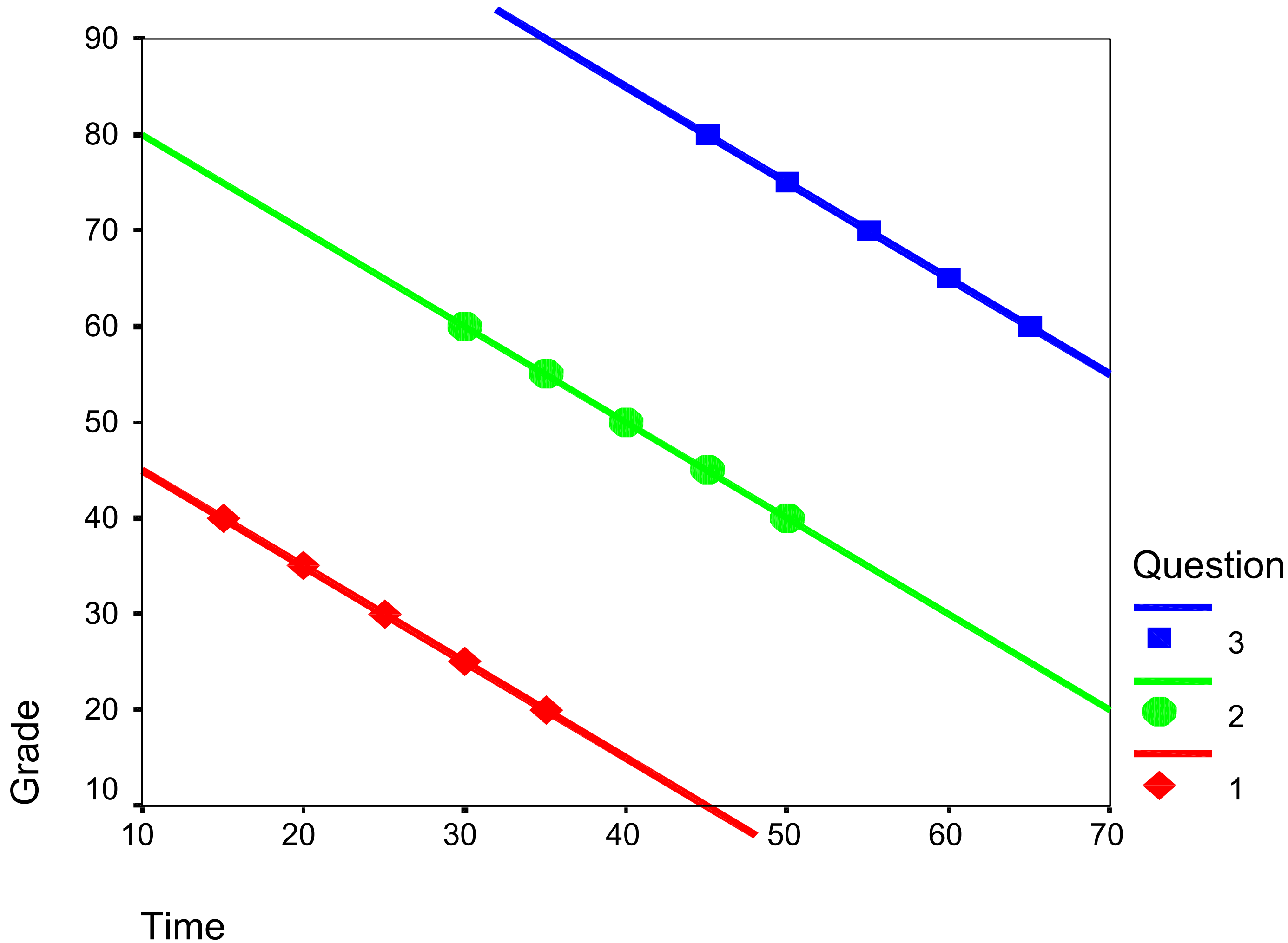
Assumption 4: At every value of the dependent variable the expected (mean) value of the residuals is zero

Assumption 5: The expected correlation between residuals, for any two cases, is 0.

•Result, with line of best fit



Now somewhat different



Assumption 6: All independent variables are uncorrelated with the error term.

Assumption 7: No independent variables are a perfect linear function of other independent variables

Assumption 8: The mean of the error term is zero.

Multicollinearity

Correlation Matrix

	x_1	x_2	x_3	x_4
x_1	1	-0.80	0.98	0.061
x_2	-0.80	1	-0.184	0.103
x_3	0.98	-0.184	1	0.119
x_4	0.061	0.103	0.119	1

VIF(Variance Inflation Factor)

VIF(Variance Inflation Factor)

The better way to assess multi collinearity is to compute the VIF

$$\boxed{VIF = \frac{1}{1 - R^2}}$$

If VIF = 1 then Variables are not correlated

1< VIF < 5 then the variables are moderately correlated

VIF > 5 then highly correlated and need to be eliminated from the model

Logistic Regression

Why use logistic regression?

- There are many important research topics for which the dependent variable is "limited.“

- For example: voting, morbidity or mortality, and participation data is not continuous or distributed normally.

- Logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not vote) or 1(did vote)

Logistic Regression

Logistic regression is a supervised classification model.

This allows us to make predictions from labelled data ,if the target variable is categorical.

Binary classification

Examples

1. A customer will default on a loan or not
2. A particular machine will break down in the next month or not
3. Predicting whether an incoming email is spam or not

Categorical Response Variables

Examples:

Whether or not a person smokes **Binary Response**

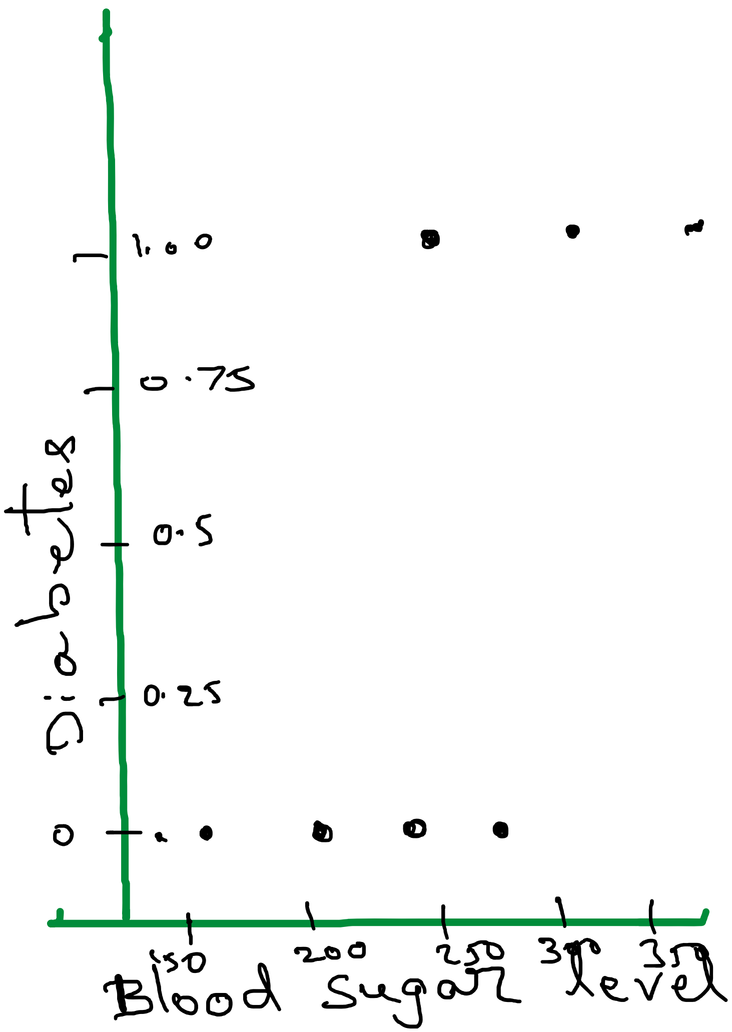
$$Y = \begin{cases} \text{Non} - \text{smoker} \\ \text{Smoker} \end{cases}$$

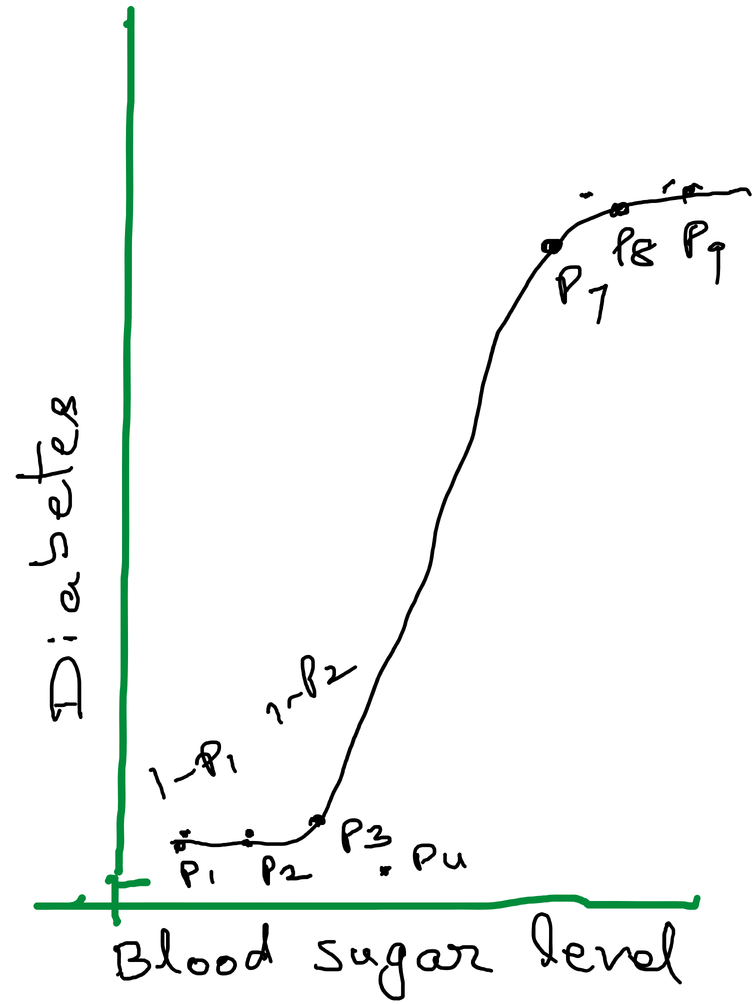
Success of a medical treatment

$$Y = \begin{cases} \text{Survives} \\ \text{Dies} \end{cases}$$

Opinion poll responses **Ordinal Response**

$$Y = \begin{cases} \text{Agree} \\ \text{Neutral} \\ \text{Disagree} \end{cases}$$


1.00
0.75
0.5
0.25
0
Diabetes
150
200
250
300
350
Blood sugar level


Diabetes
1-P1
1-P2
P1
P2
P3
P4
P7
P8
P9
Blood sugar level

$$P(\text{diabetes}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

$$\text{Likelihood} = (1-P_1)(1-P_2)(1-P_3)(1-P_4)$$
$$P_5 (1-P_6) P_7 P_8 P_9 P_{10}$$

ie $\left[(1-P_i)(1-P_j) \ldots \text{for all non diabetics} \right]$.

$* \left[P_i \cdot P_i \ldots \text{for all diabetes} \right]$

Y = Binary response

X = Quantitative predictor

p = proportion of 1's (yes, success) at any X

Equivalent forms of the logistic regression model:

Logit form

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Probability form

$$p = \frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1)}}$$

Binary Logistic Regression via R

```
> logitmodel=glm(Gender~Hgt,family=binomial, data=Pulse)
> summary(logitmodel)
```

```
Call:
glm(formula = Gender ~ Hgt, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.77443  -0.34870  -0.05375   0.32973   2.37928

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14***
---
```

```
Call:
glm(formula = Gender ~ Hgt, family = binomial, data = Pulse)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  64.1416     8.3694   7.664 1.81e-14 ***
Hgt          -0.9424     0.1227  -7.680 1.60e-14***
---
```

$$p = \frac{e^{64.14-0.9424Ht}}{1+e^{64.14-.9424Ht}}$$

proportion of females at that Hgt

Example: TMS for Migraines

Transcranial Magnetic Stimulation vs. Placebo

Pain Free?	TMS	Placebo
YES	39	22
NO	61	78
Total	100	100

$$P_{TMS} = 0.39 \qquad odds_{TMS} = \frac{39/100}{61/100} = \frac{39}{61} = 0.639 \quad P = \frac{0.639}{1+0.639} = 0.39$$

$$P_{Placebo} = 0.22 \qquad odds_{Placebo} = \frac{22}{78} = 0.282$$

$$Odds\ ratio = \frac{0.639}{0.282} = 2.27$$

Odds are 2.27 times higher of getting relief using TMS than placebo

Logistic Regression for TMS data

```
> lmod=glm(cbind(Yes,No)~Group,family=binomial,data=TMS)
> summary(lmod)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2657     0.2414  -5.243 1.58e-07 ***
GroupTMS      0.8184     0.3167   2.584  0.00977 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.8854  on 1  degrees of freedom
Residual deviance: 0.0000  on 0  degrees of freedom
AIC: 13.701
```

Note: $e^{0.8184} = 2.27$ = odds ratio

Binary Logistic Regression Model

Y = Binary	$X_1, X_2, \ldots, X_k$ = Multiple

π = proportion of 1's at any $x_1, x_2, \ldots, x_k$

Equivalent forms of the logistic regression model:

Logit form $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

Probability form

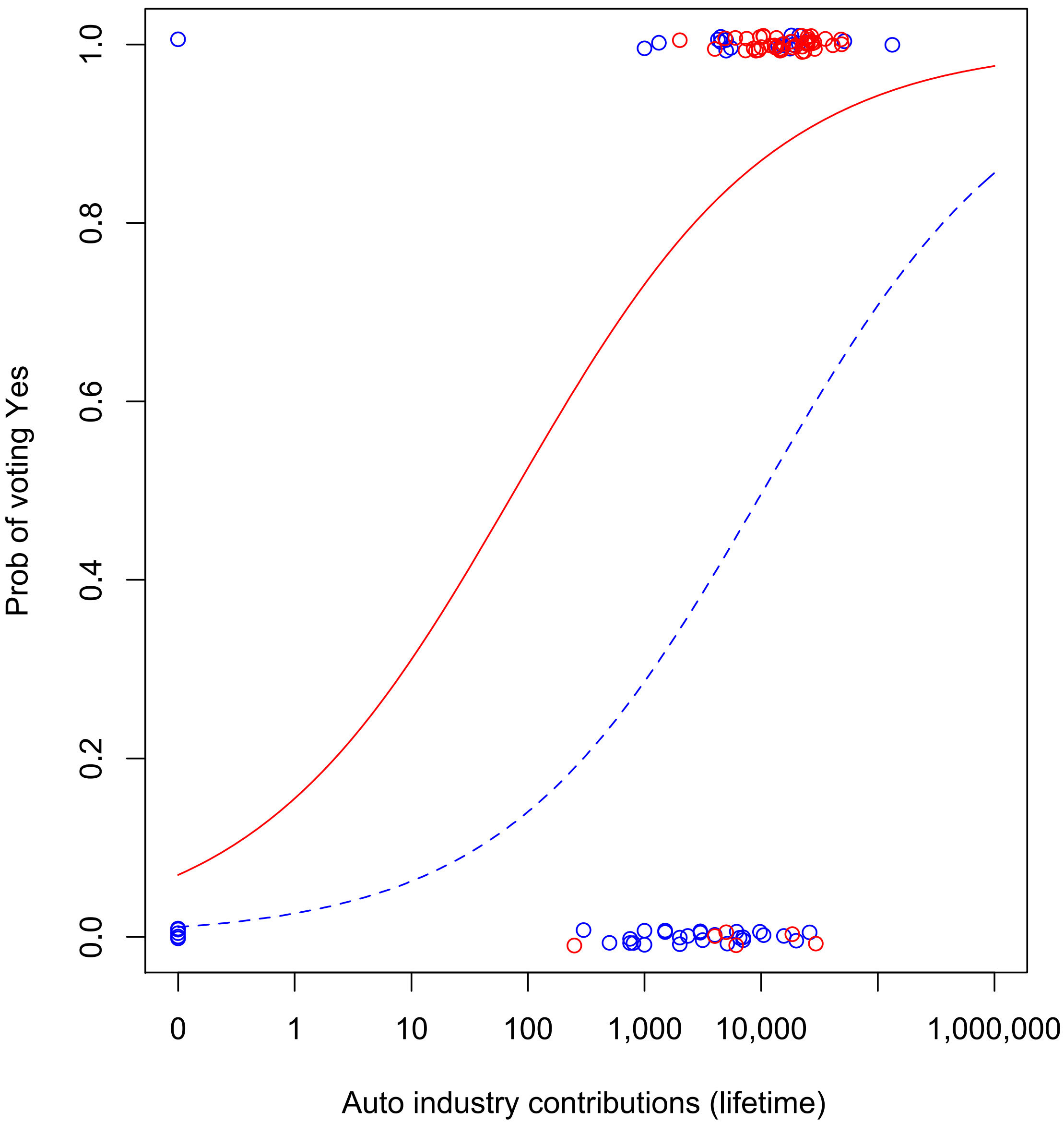
$$p = \frac{e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}{1 + e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k}}$$

$$= \frac{1}{1 + e^{-(\beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

Interactions in logistic regression

Consider Survival in an ICU as a function of SysBP -- BP for short – and Sex

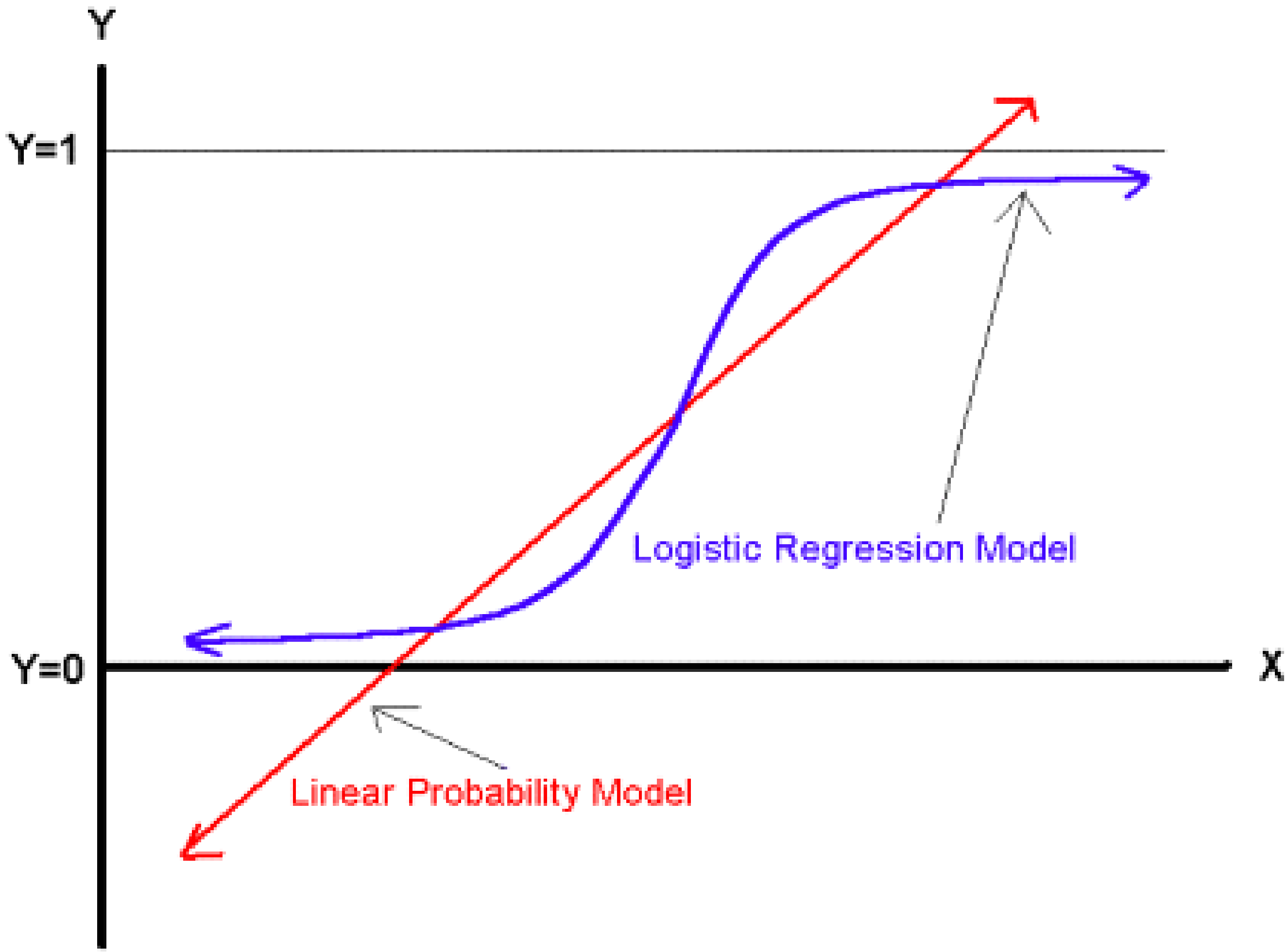
```
> intermodel=glm(Survive~BP*Sex, family=binomial, data=ICU)
> summary(intermodel)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.439304   1.021042  -1.410  0.15865
BP           0.022994   0.008325   2.762  0.00575 **
Sex          1.455166   1.525558   0.954  0.34016
BP:Sex      -0.013020   0.011965  -1.088  0.27653

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 189.99  on 196  degrees of freedom
```



Rep = red,
Dem = blue

Lines are very close to parallel; not a significant interaction

Comparing the LP and Logit Models



Forecasting models

- Principles of forecasting
- Time series analysis
- Smoothing and decomposition methods
- ARIMA
- GARCH
- Holt – winter model
- Casual methods
- Moving averages
- Exponential smoothing

Forecasting

Predict the next number in the pattern:

a) 3.7, 3.7, 3.7, 3.7, 3.7, ?

b) 2.5, 4.5, 6.5, 8.5, 10.5, ?

c) 5.0, 7.5, 6.0, 4.5, 7.0, 9.5, 8.0, 6.5, ?

Forecasting

Predict the next number in the pattern:

a) 3.7, 3.7, 3.7, 3.7, 3.7, 3.7

b) 2.5, 4.5, 6.5, 8.5, 10.5, 12.5

c) 5.0, 7.5, 6.0, 4.5, 7.0, 9.5, 8.0, 6.5, 9.0

What Is Forecasting?

Process of predicting a future event
Underlying basis of all business decisions

- Production
- Inventory
- Personnel
- Facilities

Why do we need to forecast?

Importance of Forecasting

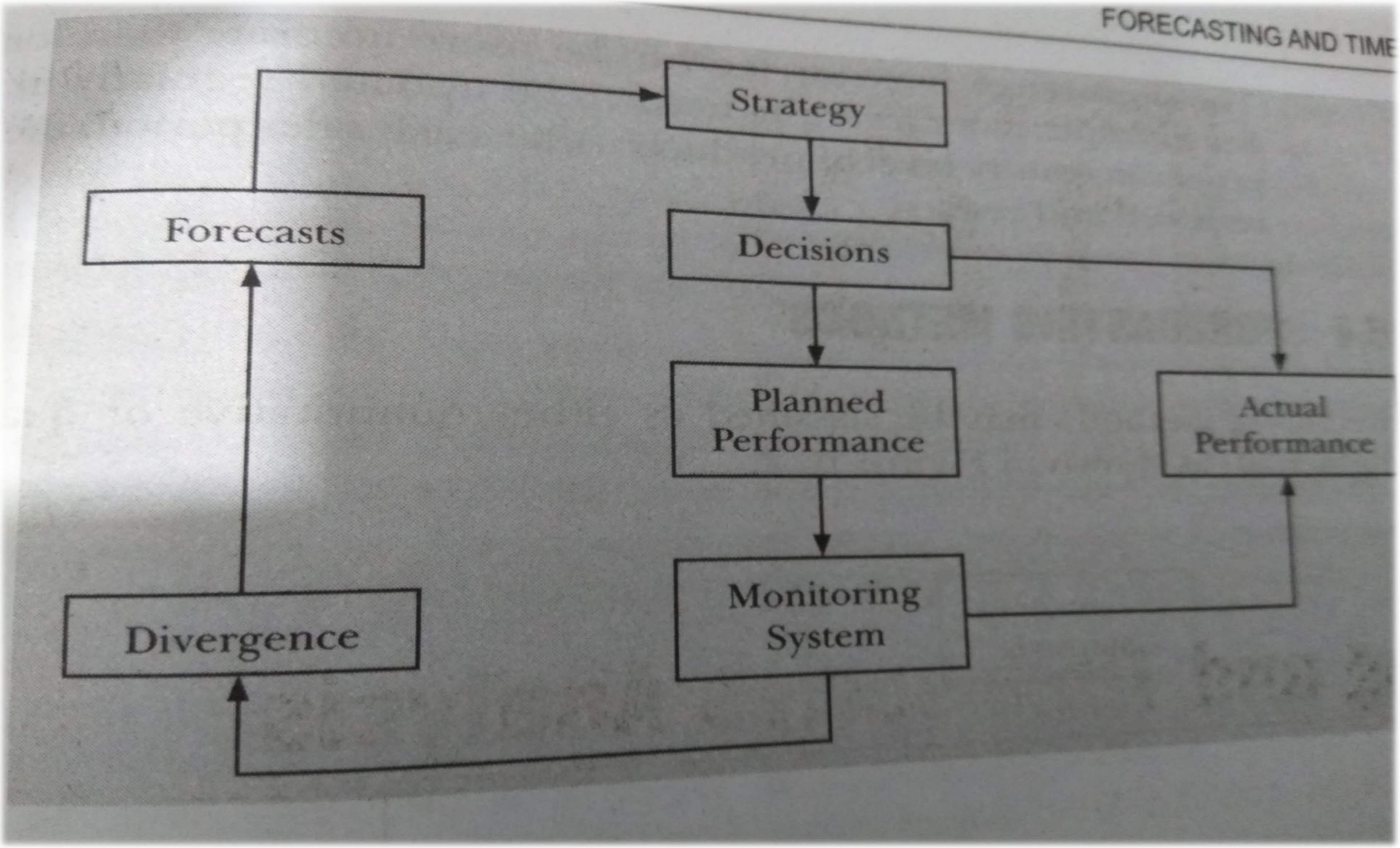
Departments throughout the organization depend on forecasts to formulate and execute their plans.

Finance needs forecasts to project cash flows and capital requirements.

Human resources need forecasts to anticipate hiring needs.

Production needs forecasts to plan production levels, workforce, material requirements, inventories, etc.

- ✓ Demand is not the only variable of interest to forecasters.
- ✓ Manufacturers also forecast worker absenteeism, machine availability, material costs, transportation and production lead times, etc.
- ✓ Besides demand, service providers are also interested in forecasts of population, of other demographic variables, of weather, etc.


FORECASTING AND TIME
Strategy
Forecasts
Decisions
Planned Performance
Actual Performance
Monitoring System
Divergence

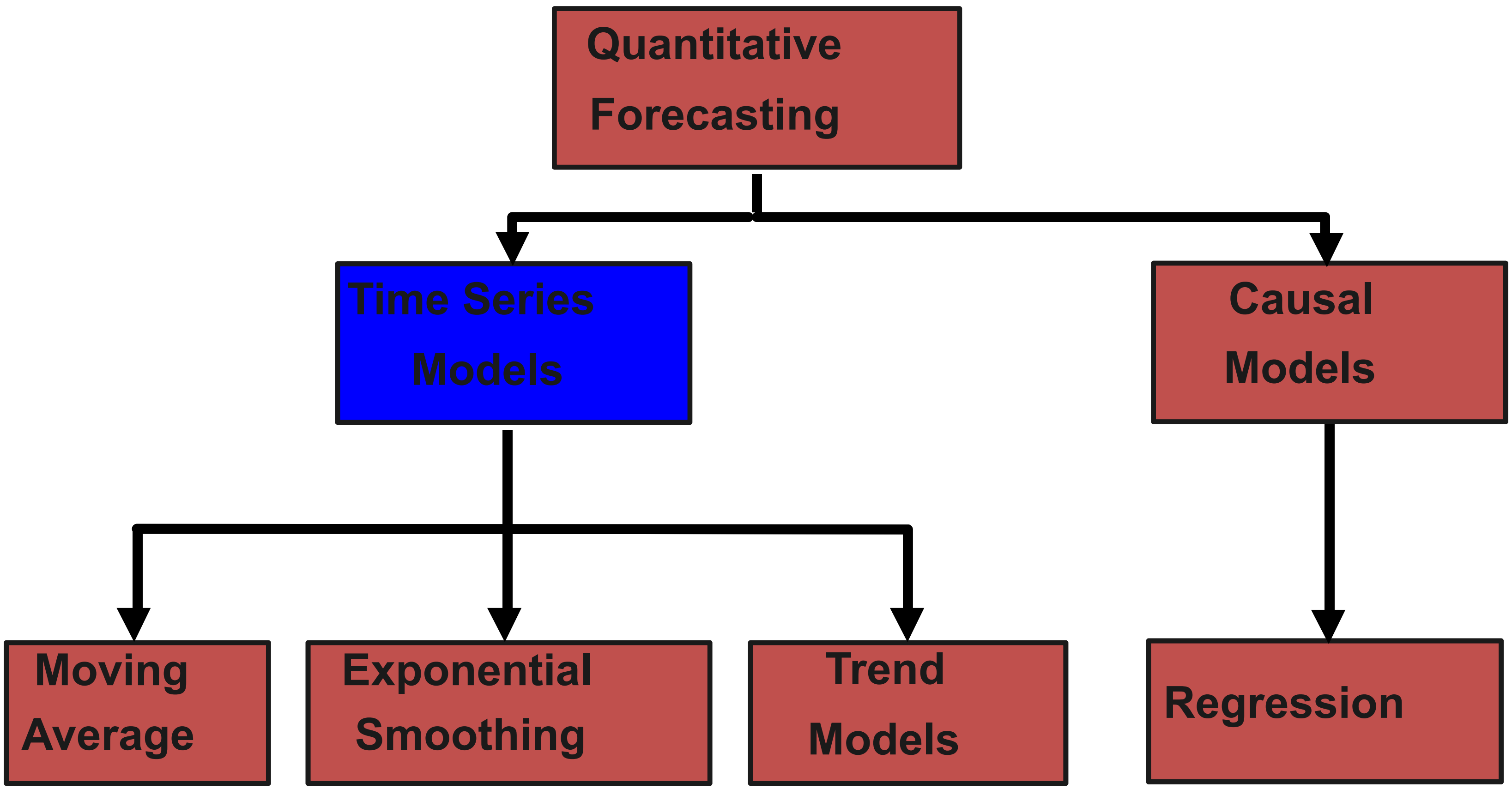
Types of forecasts

- Demand Forecasts
- Environmental Forecasts
- Technological Forecasts

Timing of Forecasts

- ✓ Short-range Forecast
- ✓ Medium – range Forecast
- ✓ Long – range Forecast

Quantitative Forecasting Methods



What is a Time Series?

Set of evenly spaced numerical data

- Obtained by observing response variable at regular time periods

Forecast based only on past values

- Assumes that factors influencing past, present, & future will continue

Example

Year:	1995	1996	1997	1998	1999
Sales:	78.7	63.5	89.7	93.2	92.1

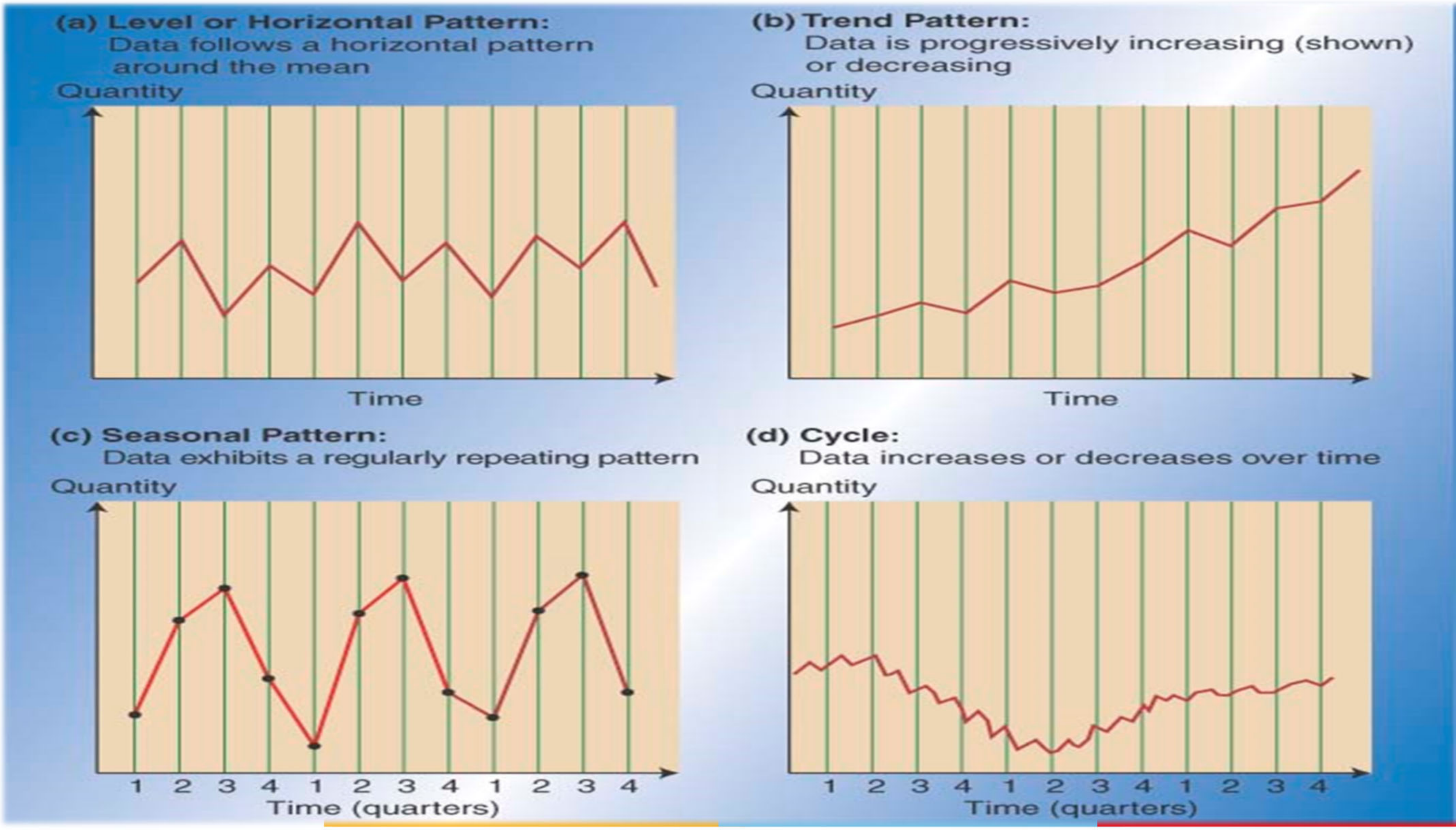
Time Series Models

- Forecaster looks for data patterns as

  Data = historic pattern + random variation

- Historic pattern to be forecasted:
  - Level (long-term average) – data fluctuates around a constant mean
  - Trend – data exhibits an increasing or decreasing pattern
  - Seasonality – any pattern that regularly repeats itself and is of a constant length
  - Cycle – patterns created by economic fluctuations
- Random Variation cannot be predicted

Time Series Patterns



Time Series Components

A time series can be described by models based on the following components

T_t	**Trend Component**
S_t	**Seasonal Component**
C_t	**Cyclical Component**
I_t	**Irregular Component**

Using these components we can define a time series as the sum of its components or an **additive model**

$$X_t = T_t + S_t + C_t + I_t$$

Alternatively, in other circumstances we might define a time series as the product of its components or a **multiplicative model** – often represented as a logarithmic model

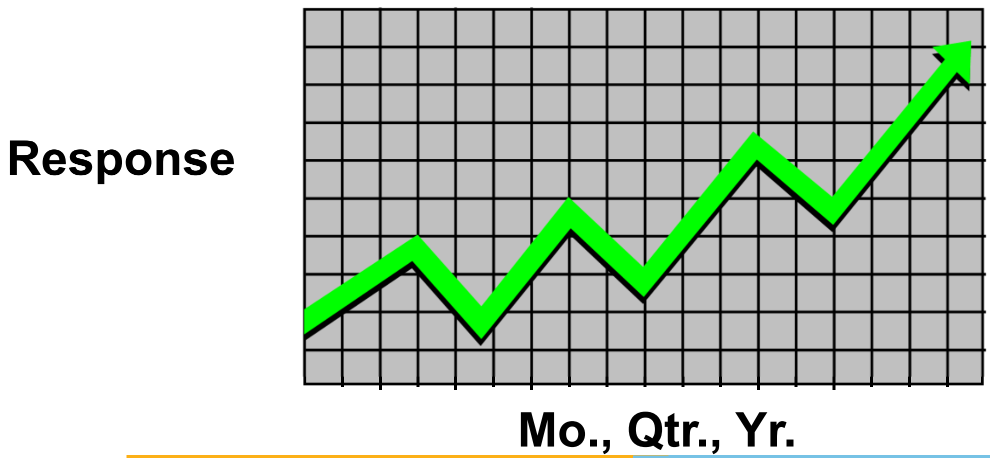
$$X_t = T_t S_t C_t I_t$$

Trend Component

Persistent, overall upward or downward pattern

Due to population, technology etc.

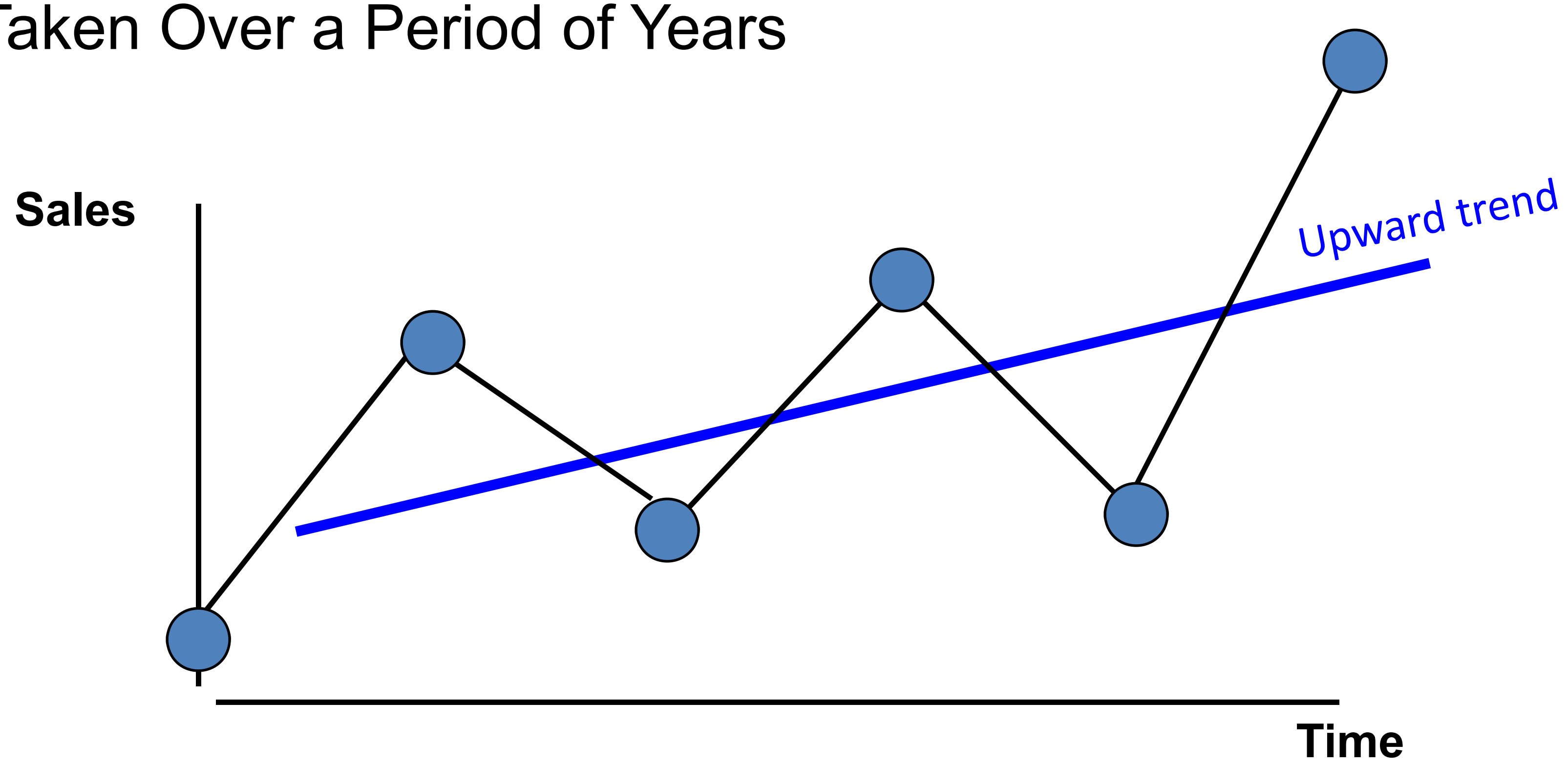
Several years duration



Trend Component

Overall Upward or Downward Movement

Data Taken Over a Period of Years

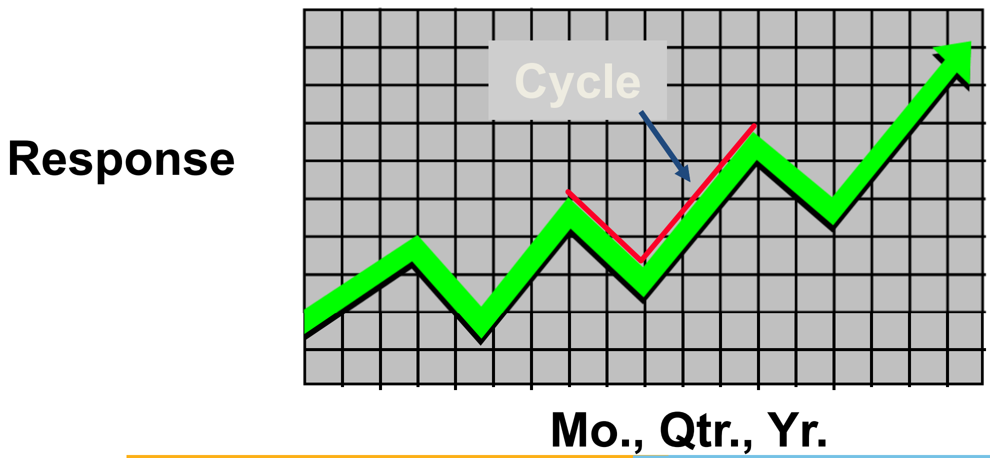


Cyclical Component

Repeating up & down movements

Due to interactions of factors influencing economy

Usually 2-10 years duration

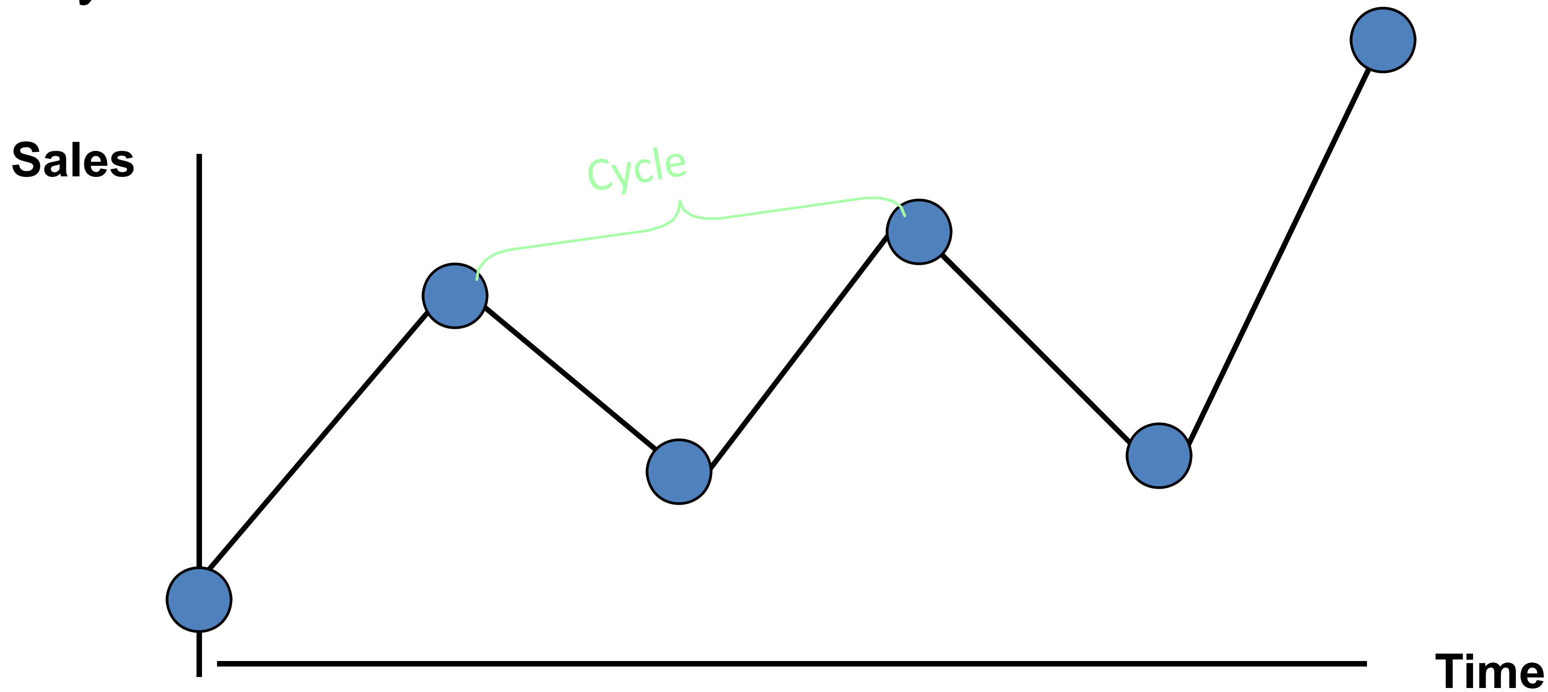


Cyclical Component

Upward or Downward Swings

May Vary in Length

Usually Lasts 2 - 10 Years

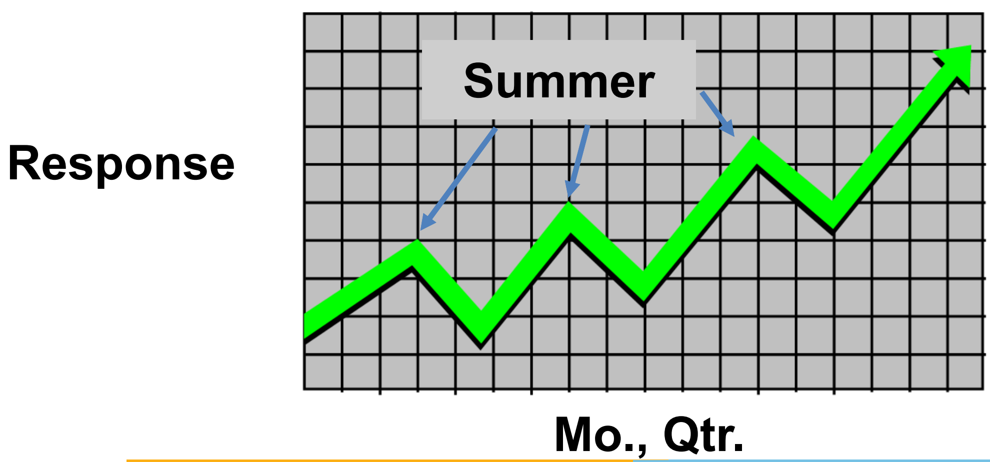


Seasonal Component

Regular pattern of up & down fluctuations

Due to weather, customs etc.

Occurs within one year

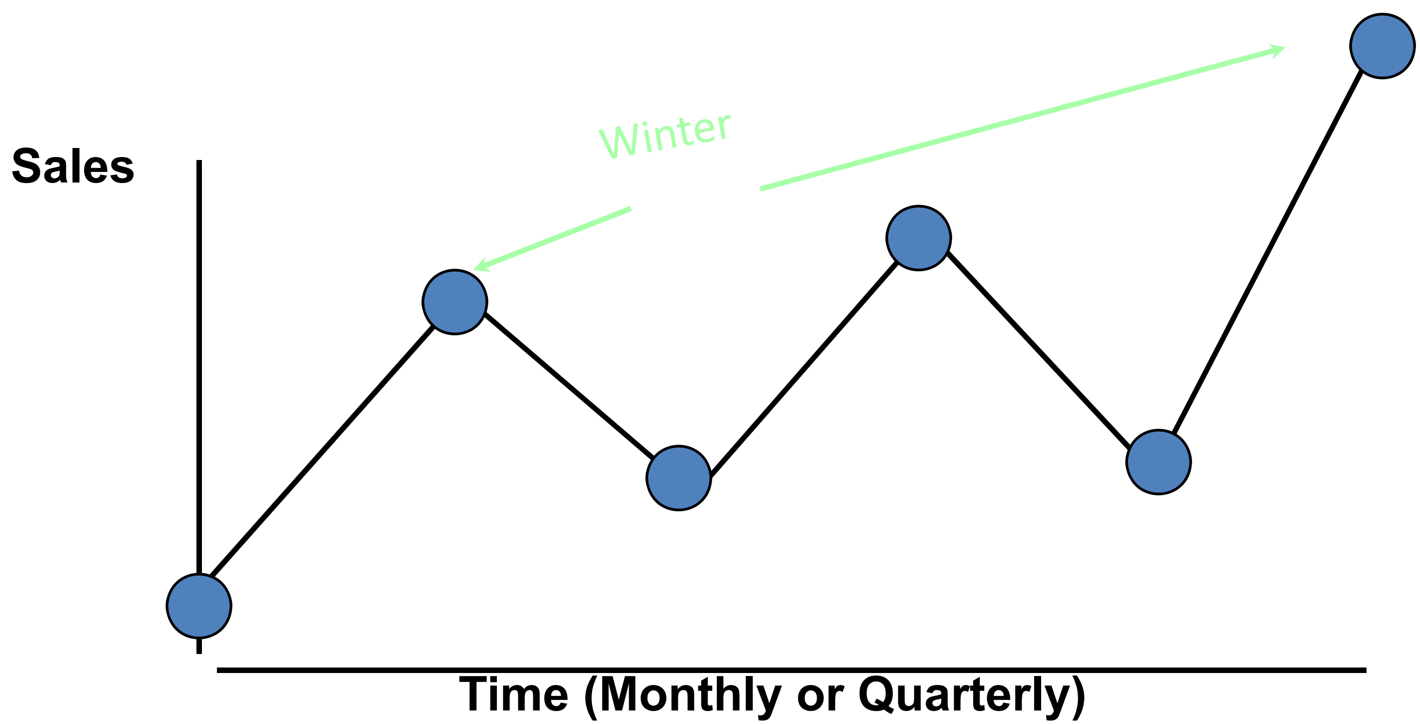


Seasonal Component

Upward or Downward Swings

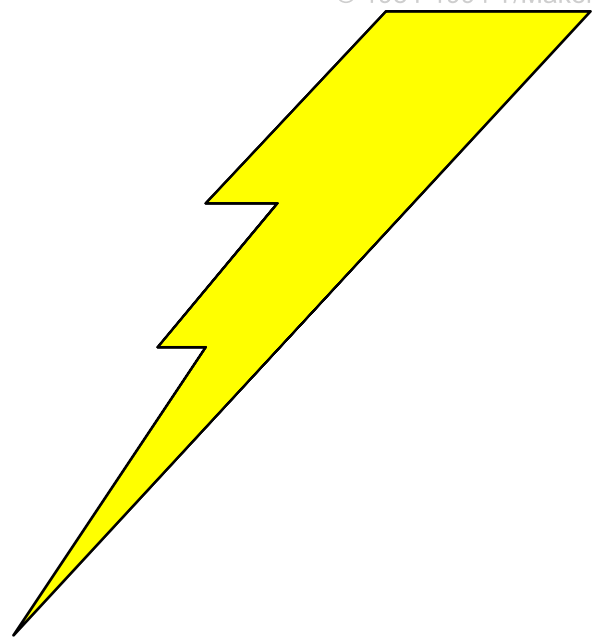
Regular Patterns

Observed Within One Year



Irregular Component

- Erratic, unsystematic, ‘residual’ fluctuations
- Due to random variation or unforeseen events
  - Union strike
  - War
- Short duration & nonrepeating



Moving Average Models

Simple Moving Average Forecast

$$F_t = E(Y_t) = \frac{\sum_{i=t-k}^{t-1} Y_i}{k}$$

Weighted Moving Average Forecast

$$F_t = E(Y_t) = \frac{\sum_{i=t-k}^{t-1} w_i Y_i}{k}$$

Selecting the Right Forecasting Model

1. The amount & type of available data
   - Some methods require more data than others
2. Degree of accuracy required
   - Increasing accuracy means more data
3. Length of forecast horizon
   - Different models for 3 month vs. 10 years
4. Presence of data patterns
   - Lagging will occur when a forecasting model meant for a level pattern is applied with a trend

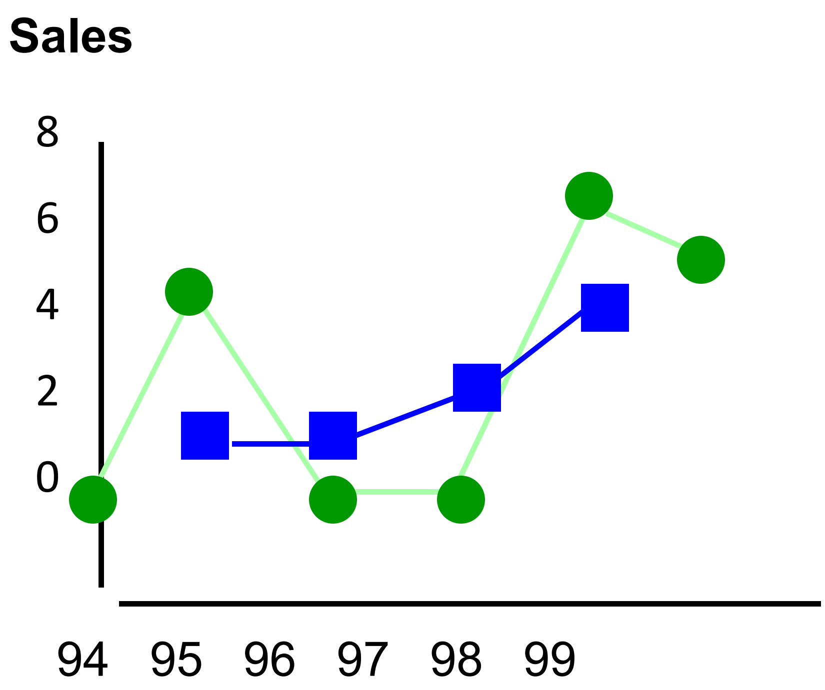
Moving Average

[Solution]

Year	Sales	MA(3) in 1,000
1995	20,000	NA
1996	24,000	(20+24+22)/3 = 22
1997	22,000	(24+22+26)/3 = 24
1998	26,000	(22+26+25)/3 = 24
1999	25,000	NA

Moving Average

Year	Response ●	Moving Ave ■
1994	2	NA
1995	5	3
1996	2	3
1997	2	3.67
1998	7	5
1999	6	NA



Thanks