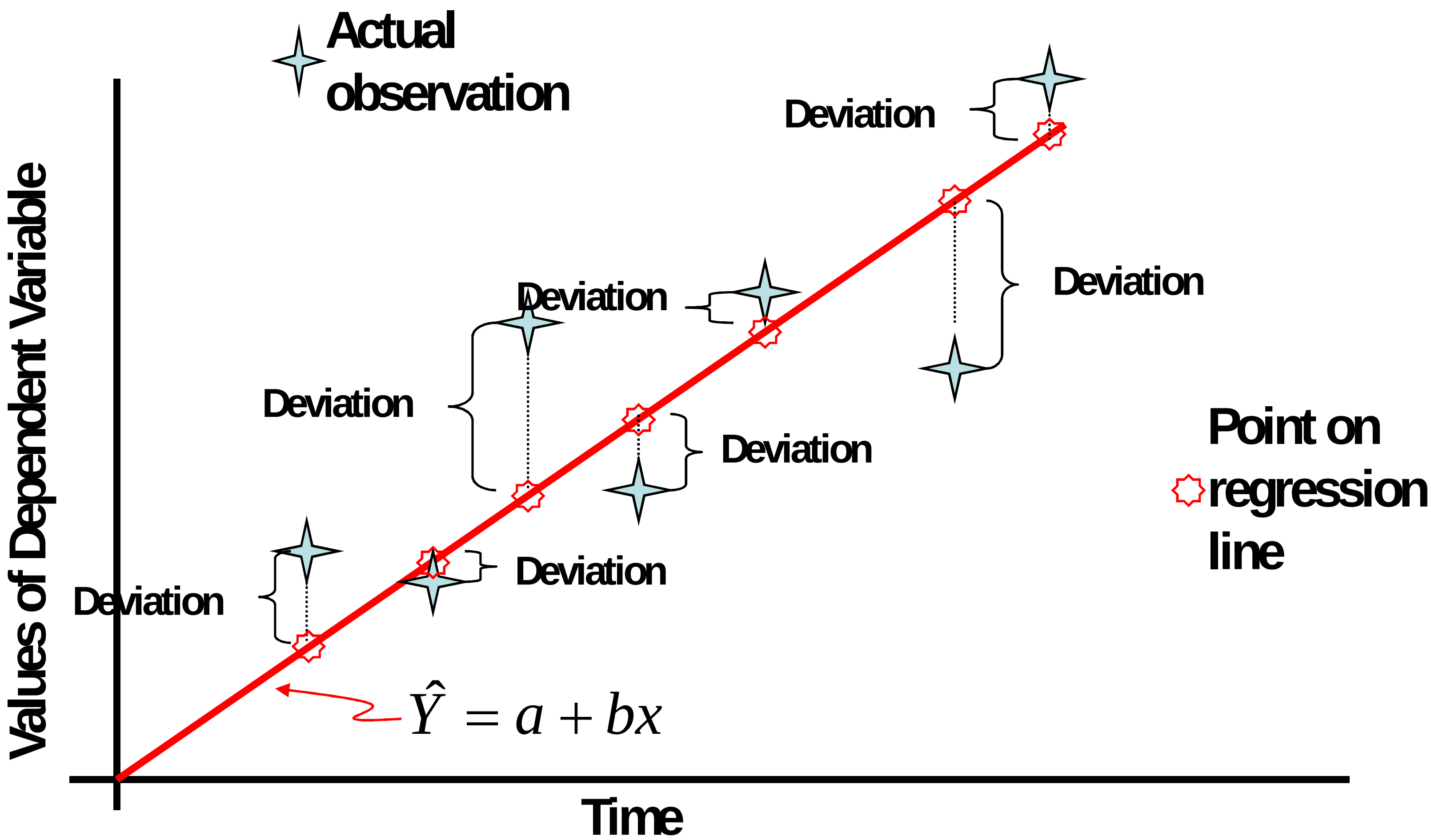
NOTE

- Follow the book formulas
- Formatting error can lead to wrong formula

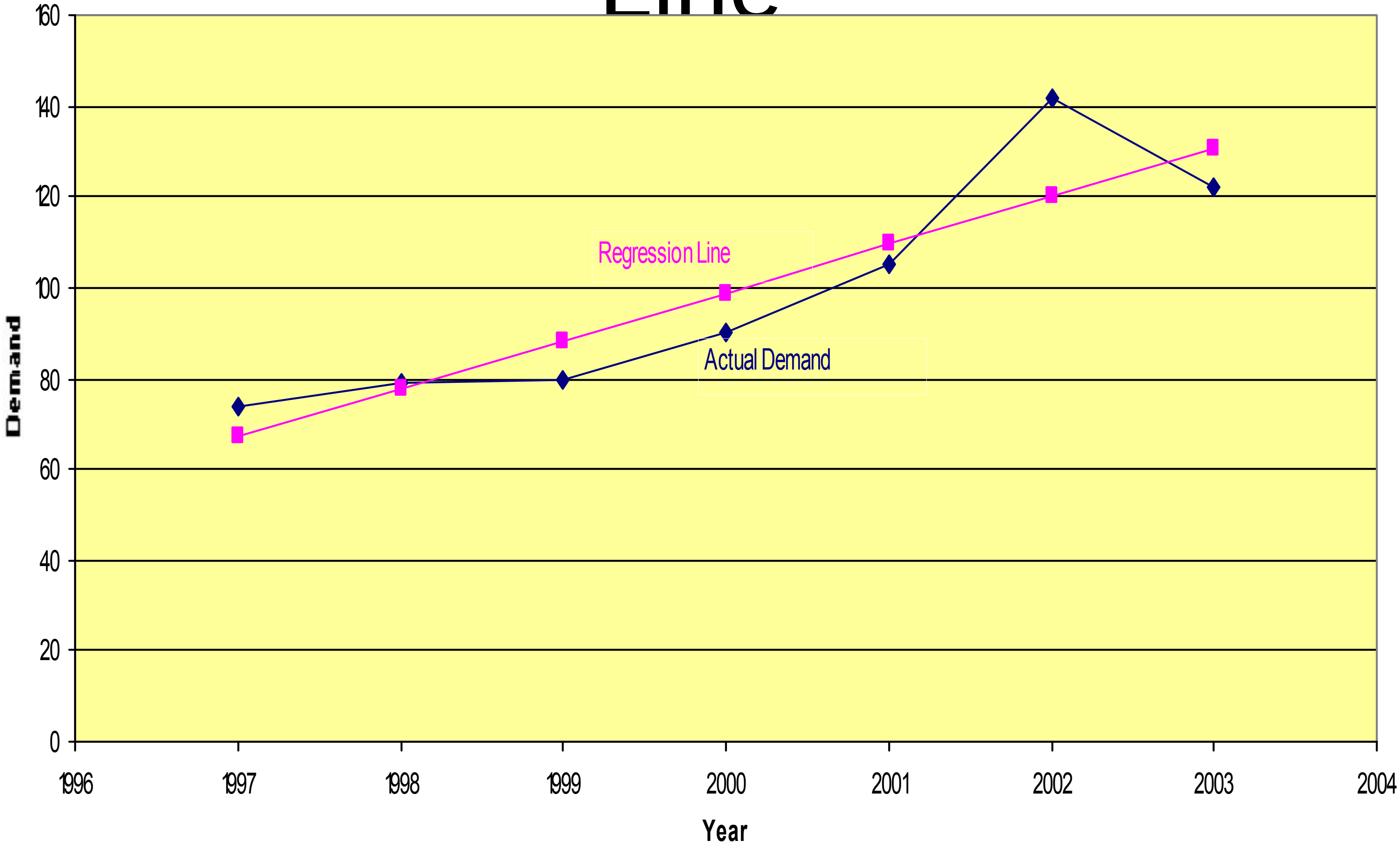
In the slides.

- Strictly follow the book

Least Squares



Actual and the Least Squares Line



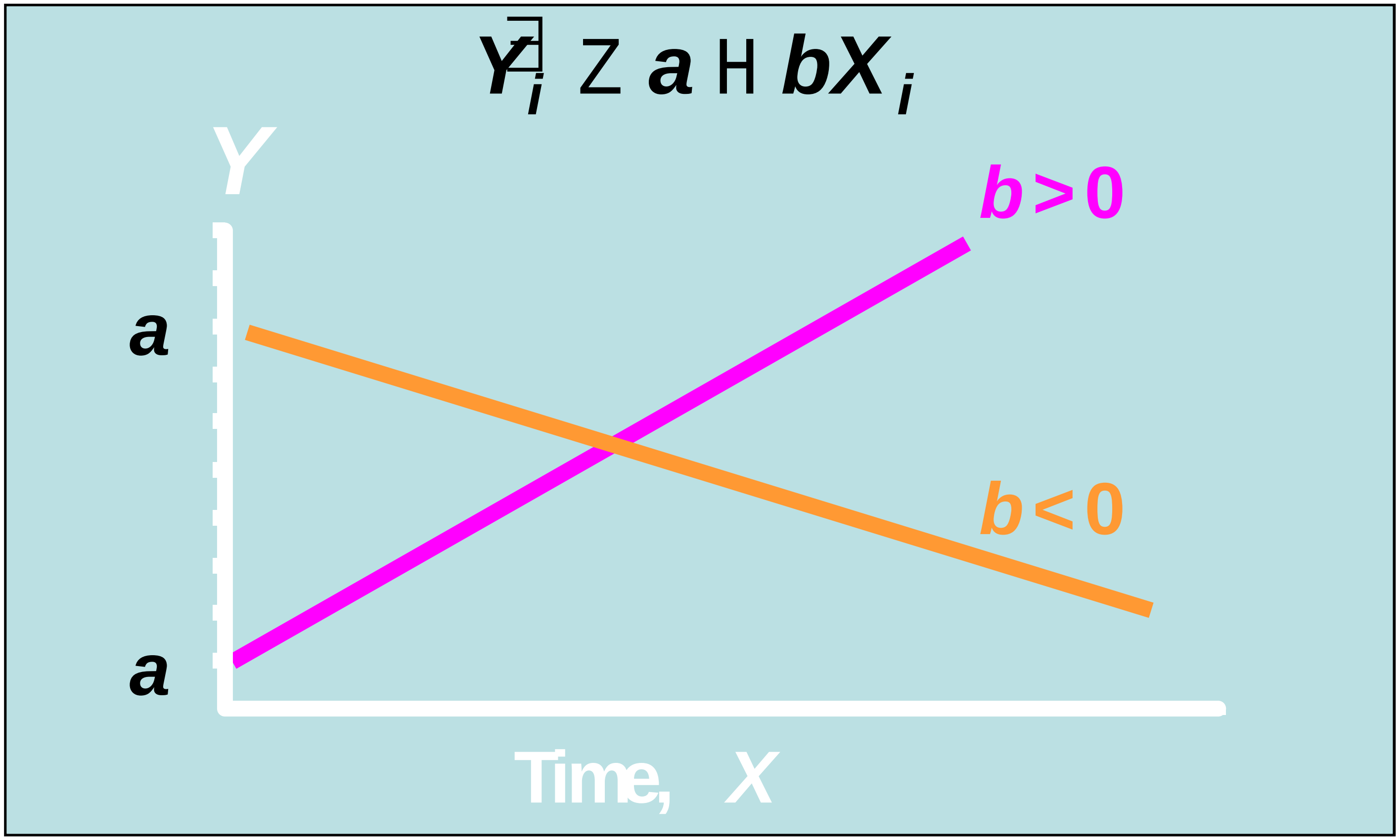
Linear Trend Projection

- Used for forecasting linear trend line
- Assumes relationship between response variable, *Y*, and time, *X*, is a linear function

$$Y_i = a + bX_i$$

- Estimated by least squares method
  - Minimizes sum of squared errors

Linear Trend Projection Model



Least Squares Equations

Equation:

$$\hat{Y}_i = a + bx_i$$

Slope:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Y-Intercept:

$$a = \bar{y} - b\bar{x}$$

Using a Trend Line

Year	Demand
1997	74
1998	79
1999	80
2000	90
2001	105
2002	142
2003	122

The demand for electrical power at N.Y.Edison over the years 1997 – 2003 is given at the left. Find the overall trend.

Linear Regression Model

- Shows linear relationship between dependent & explanatory variables
  - Example: Sales & advertising (*not* time)

Y-intercept

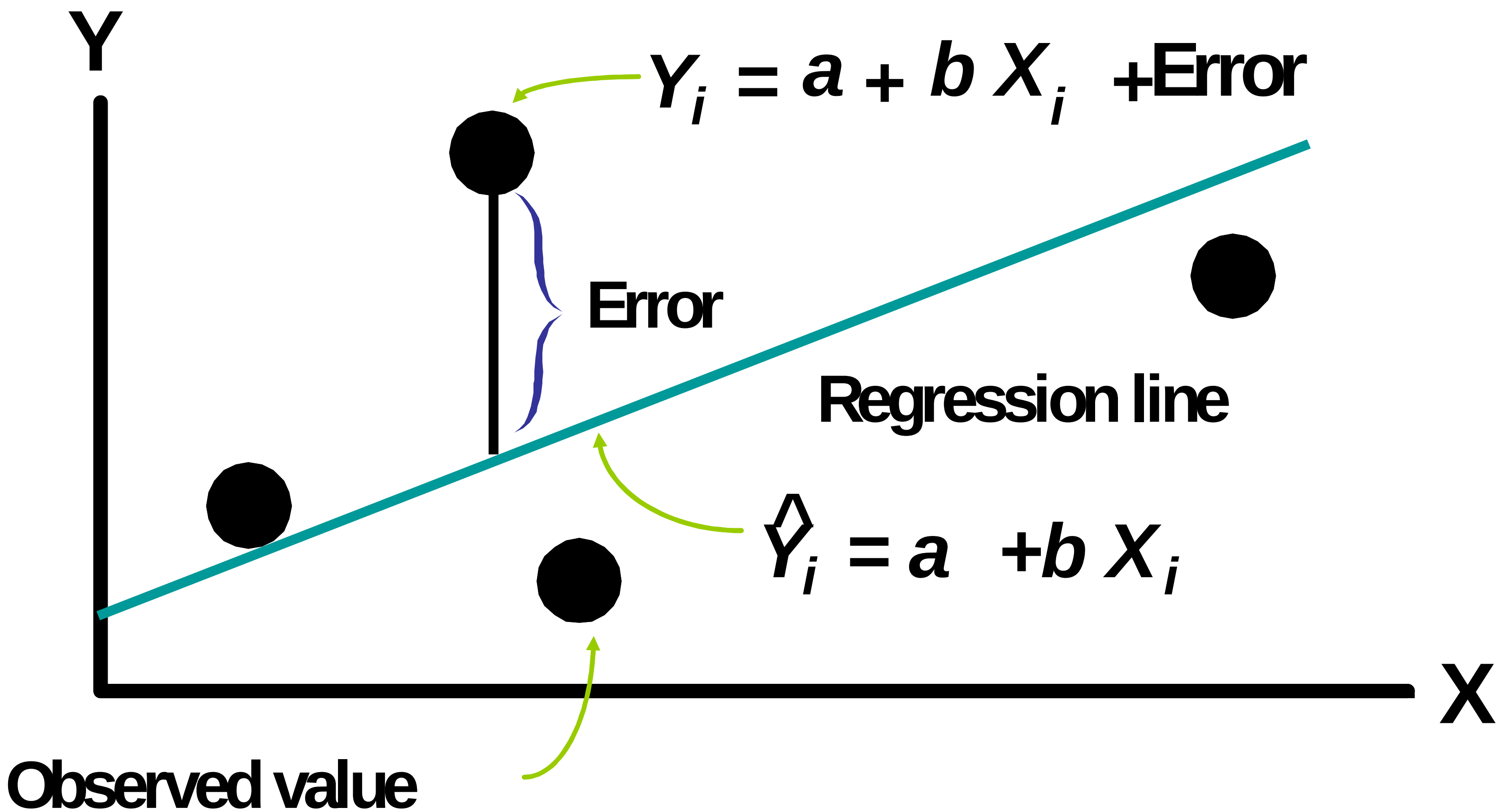
Slope

$$\hat{Y}_i = a + bX_i$$

Dependent (response) variable

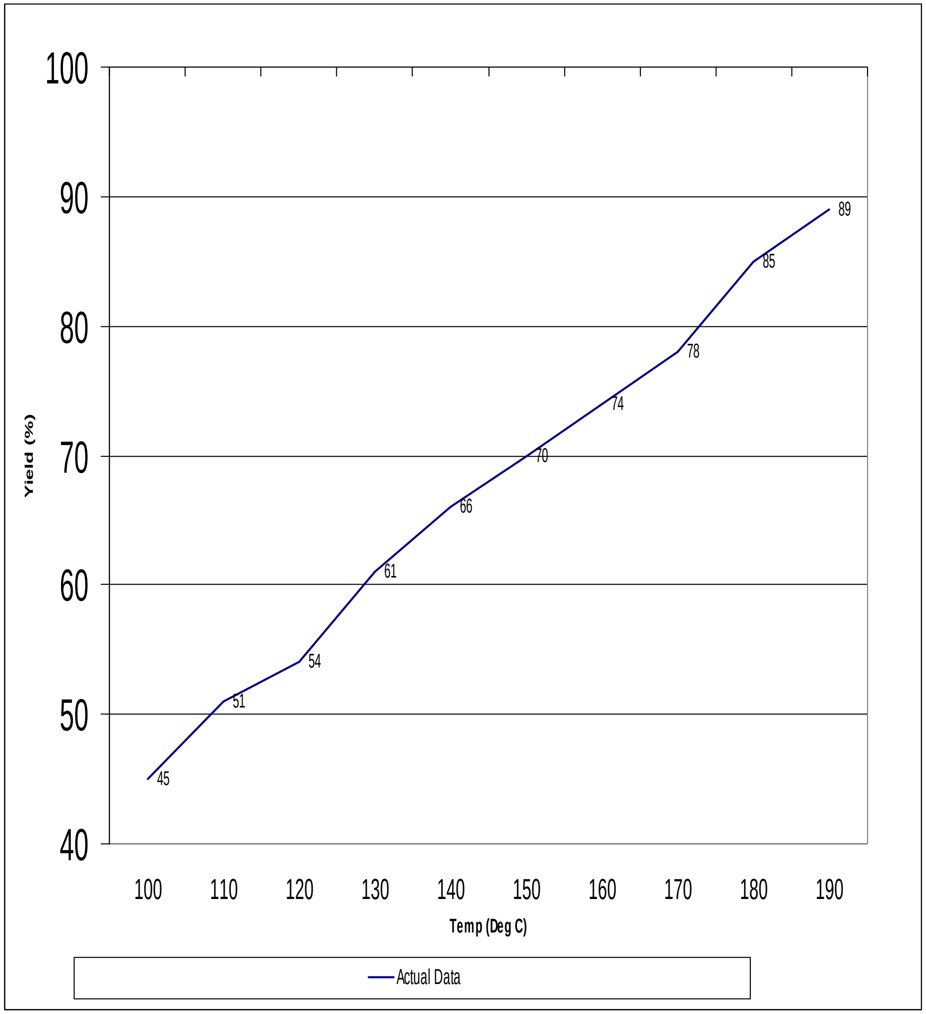
Independent (explanatory) variable

Linear Regression Model

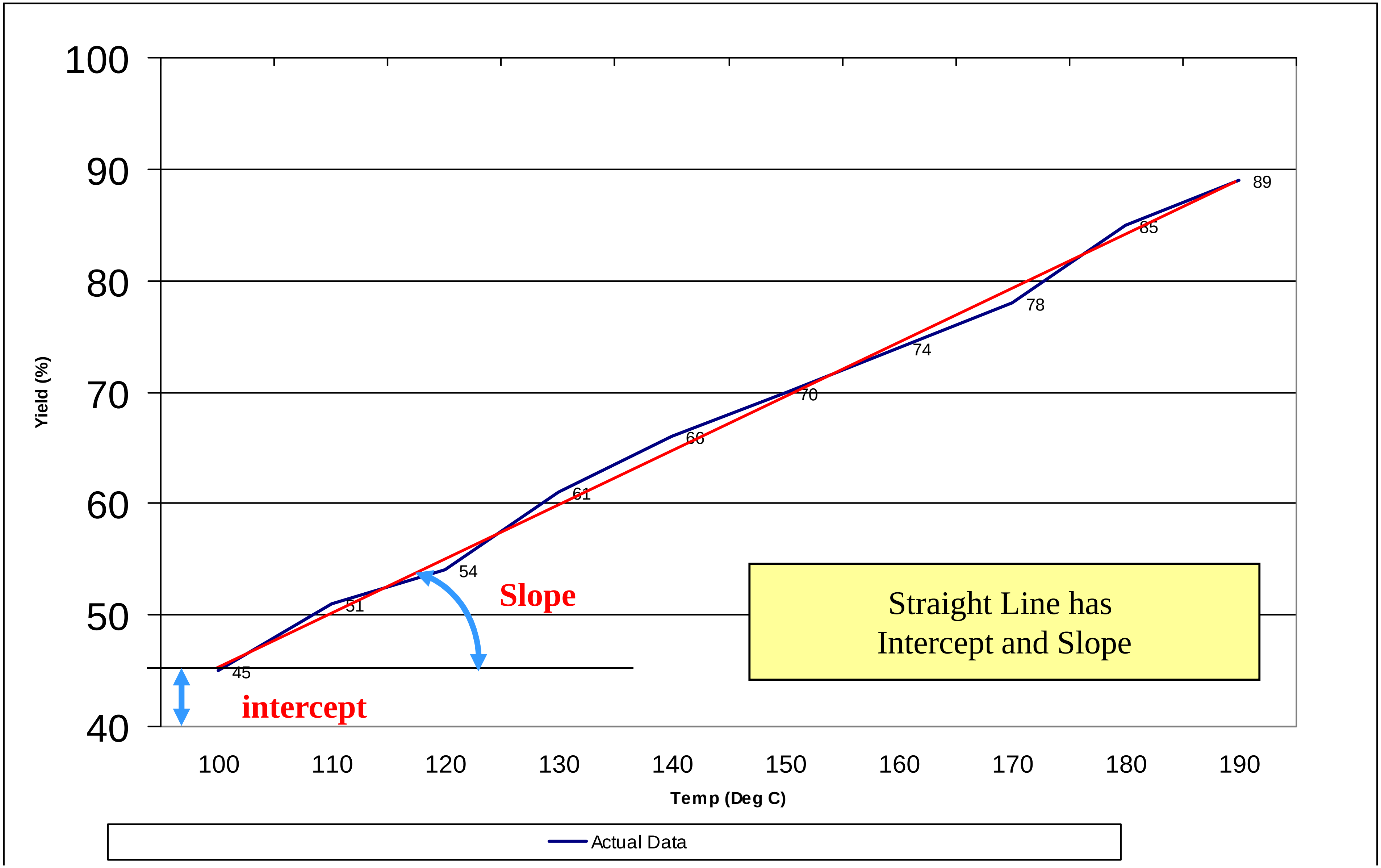


THEORY OF SIMPLE REGRESSION ANALYSIS

Data No	Temperature Deg C (X)	Yield % (Y)
1	100	45
2	110	51
3	120	54
4	130	61
5	140	66
6	150	70
7	160	74
8	170	78
9	180	85
10	190	89



Fit a straight line over the plot



Linear Regression Equations

Equation: $\hat{Y}_i = a + bx_i$

Slope:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Y-Intercept: $a = \bar{y} - b\bar{x}$

Interpretation of Coefficients

- Slope (*b*)
  - Estimated *Y* changes by *b* for each 1 unit increase in *X*
    - If $b = 2$, then sales (*Y*) is expected to increase by 2 for each 1 unit increase in advertising (*X*)
- Y-intercept (*a*)
  - Average value of *Y* when $X = 0$
    - If $a = 4$, then average sales (*Y*) is expected to be 4 when advertising (*X*) is 0

Random Error Variation

- Variation of actual *Y* from predicted *Y*
- Measured by standard error of estimate
  - Sample standard deviation of errors
  - Denoted $S_{Y,X}$
- Affects several factors
  - Prediction accuracy

Least Squares Assumptions

- Relationship is assumed to be linear. Plot the data first - if curve appears to be present, use curvilinear analysis.
- Relationship is assumed to hold only within or slightly outside data range. Do not attempt to predict time periods far beyond the range of the data base.
- Deviations around least squares line are assumed to be random.

Using Regression Analysis to Forecast

Y Nodel's Sales ($100,000's)	X Local Payroll ($100,000,000)
2.0	1
3.0	3
2.5	4
2.0	2
2.0	1
3.5	7

Using Regression Analysis to Forecast

Sales, Y	Payroll, X	X^2	XY
2.0	1	1	2.0
3.0	3	9	9.0
2.5	4	16	10.0
2.0	2	4	4.0
2.0	1	1	2.0
3.5	7	49	24.5
$\Sigma Y = 15$	$\Sigma X = 18$	$\Sigma X^2 = 80$	$\Sigma XY = 51.5$

Using Regression Analysis to Forecast

Calculating the required parameters:

$$\overline{X} = \frac{\Sigma X}{6} = \frac{18}{3} = 3$$

$$\overline{Y} = \frac{\Sigma Y}{6} = \frac{15}{6} = 2.5$$

$$b = \frac{\Sigma XY - n\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2}$$

$$= \frac{51.5 - 6 * 3 * 2.5}{80 - 6 * 3^2} = 0.25$$

$$a = \overline{Y} - b\overline{X} = 2.5 - 0.25 * 3 = 1.75$$

Standard Error of the Estimate

$$S_{Y,X} = \sqrt{\frac{\sum (Y - Y_c)^2}{n - 2}}$$

where

$Y - Y$ − value of each data point

Y_c = value of the dependent variable computed from the regression equation

n = number of data points

or:

$$S_{Y,X} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$

Nodel's Calculations

Y	X	X^2	XY	Y^2
2.0	1	1	2.0	4.0
3.0	3	9	9.0	9.0
2.5	4	16	10.0	6.25
2.0	2	4	4.0	4.0
2.0	1	1	2.0	4.0
3.5	7	49	24.5	12.25
• $Y = 15.0$	• $X = 18$	• $X^2 = 80$	• $XY = 51.5$	• $Y^2 = 39.5$

Standard Error of Estimate

$$S_{Y,X} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$

$$S_{Y,X} = \sqrt{\frac{395 - (1.75)(150) - (0.25)(515)}{6-2}}$$

$$= \sqrt{0.09375} = 0.306$$

Correlation

- Answers: ‘*how strong* is the linear relationship between the variables?’
- Coefficient of correlation Sample correlation coefficient denoted *r*
  - Values range from -1 to +1
  - Measures degree of association
- Used mainly for understanding

Correlation Coefficient

$$r = \frac{n \Sigma X - \Sigma X \Sigma Y}{\sqrt{\left[n \Sigma X^2 - (\Sigma X)^2 \right] \left[n \Sigma Y^2 - (\Sigma Y)^2 \right]}}$$

Nodel's Calculations - continued

$$r = Correlation.Coefficient(see.book)$$

$$= \frac{6 * 51.5 - 18 * 15.0}{\sqrt{(6 * 80 - 18^2)(6 * 39.5 - 15.0^2)}}$$

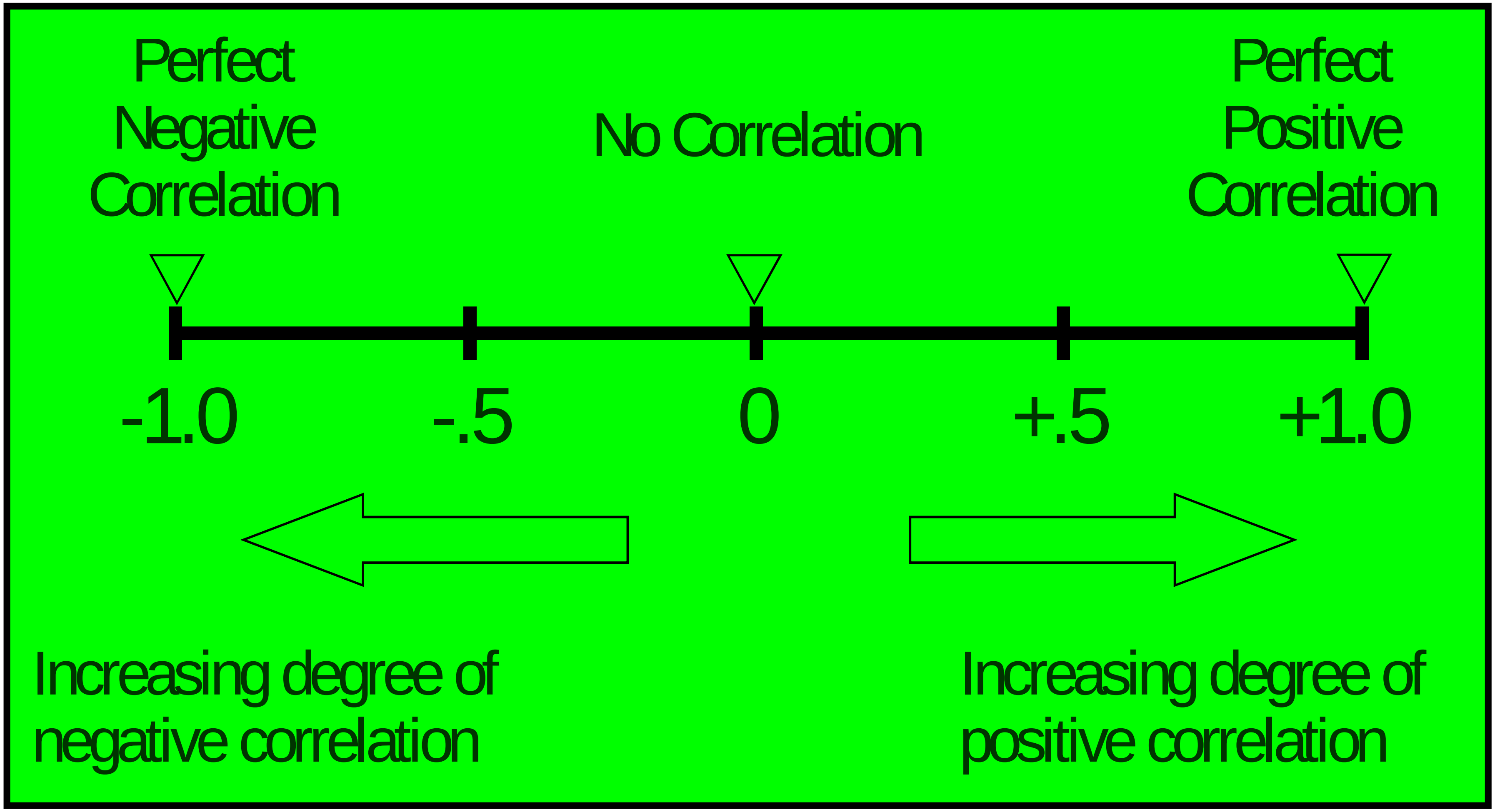
$$= \frac{309 - 270}{\sqrt{156 * 12}}$$

$$= \frac{39}{\sqrt{1872}}$$

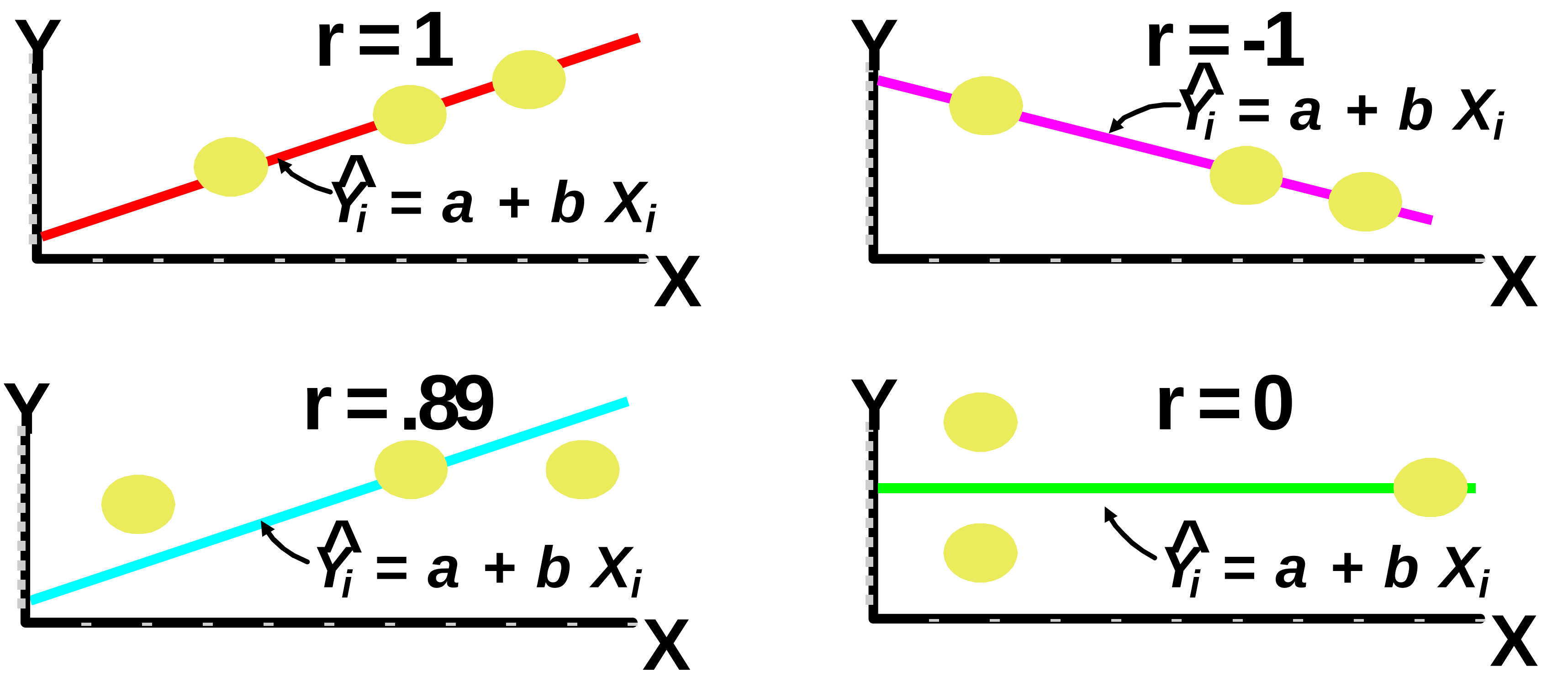
$$= \frac{39}{43.3}$$

$$= 0.901$$

Coefficient of Correlation Values



Coefficient of Correlation and Regression Model



r^2 = square of correlation coefficient (r), is the percent of the variation in *y* that is explained by the regression equation