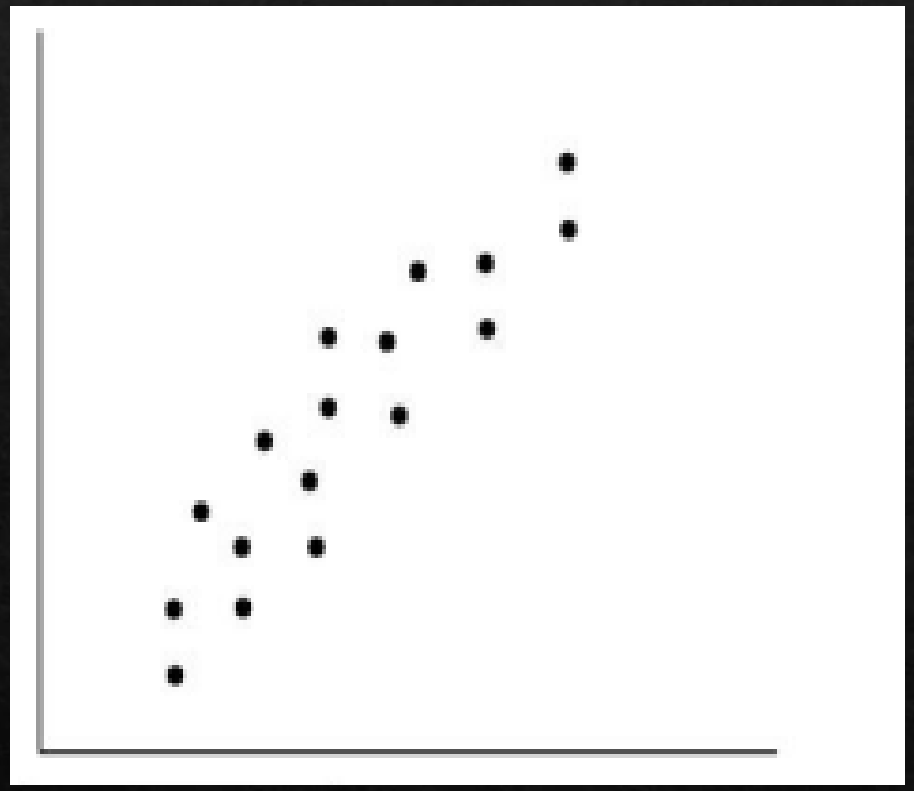
Contents

- Mathematical and Statistical Equation
- Meaning of Intercept and Slope
- Error term
- Measure for Model Fit
- R^2 | MAE | MAPE

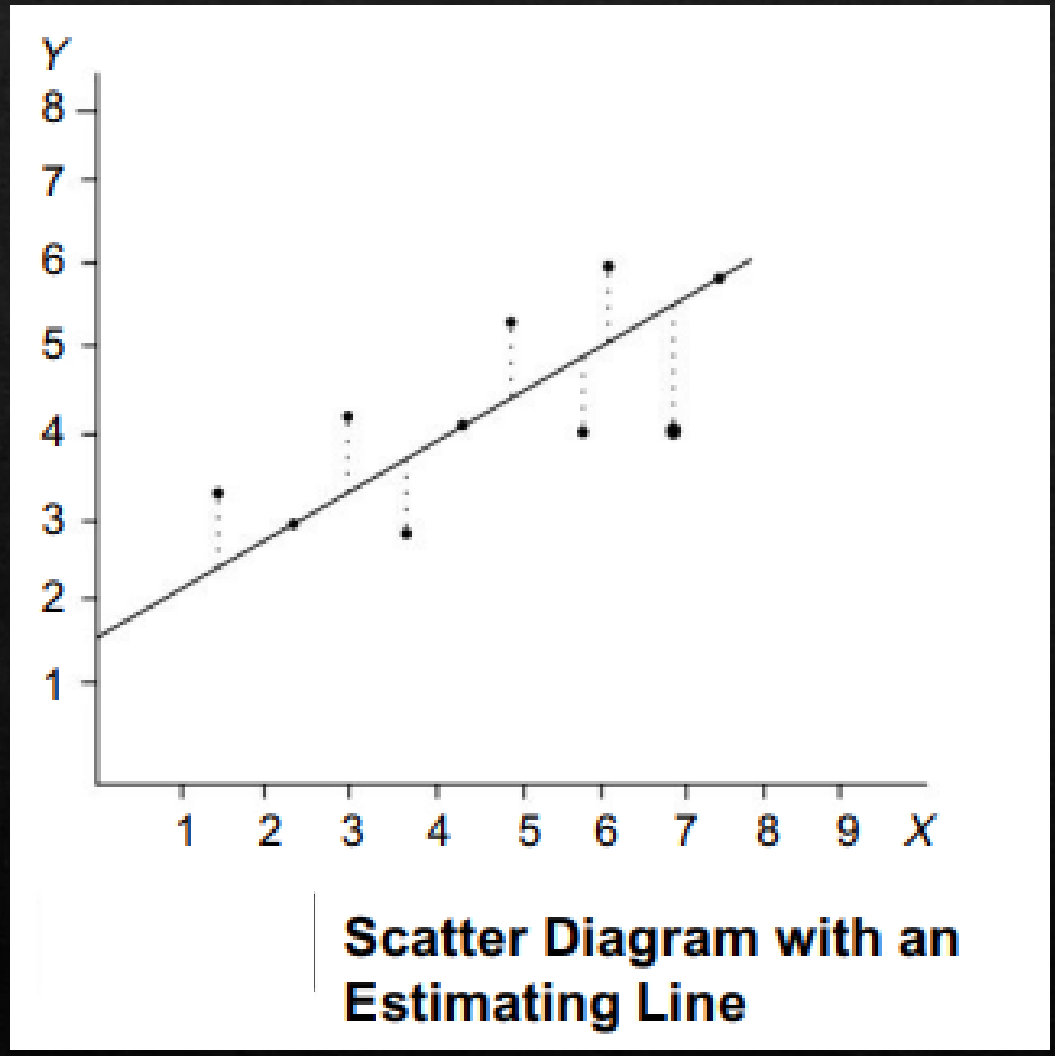
Introduction to Linear Regression and Correlation Analysis

When we have to study the relationship between two variables, X and Y, the very first step we should take is to plot the given data on the graph.
For each pair of X and Y values, there will be a point on the graph. Such a graph or chart is known as a scatter diagram



Scatter Diagram

A scatter diagram can give us a broad idea of the type of relationship (or even absence of any relationship) between the two variables under study.



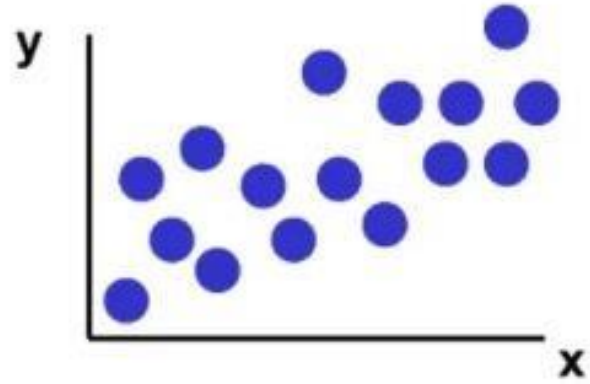
Scatter Diagram with an Estimating Line

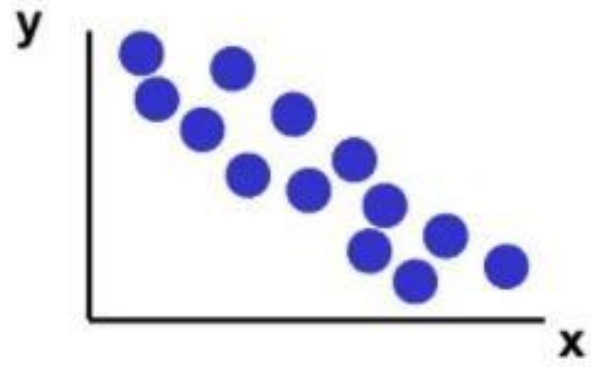
Scatter Plots and Correlation

- A scatter plot (or scatter diagram) is used to show the relationship between two variables
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Only concerned with strength of the relationship
  - No causal effect is implied

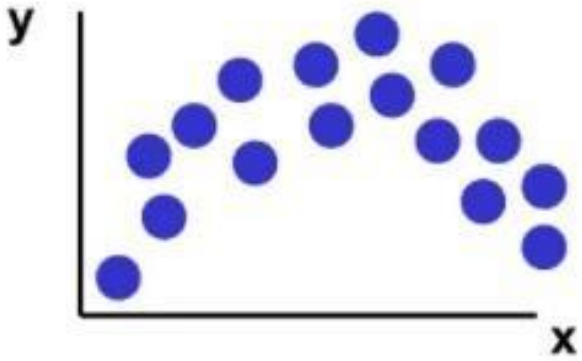
Scatter Plot Examples

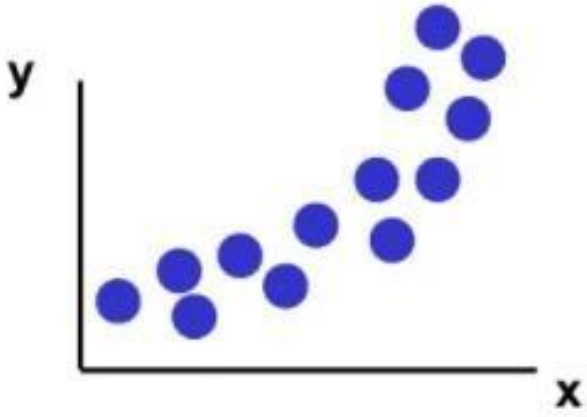
Linear relationships





Curvilinear relationships

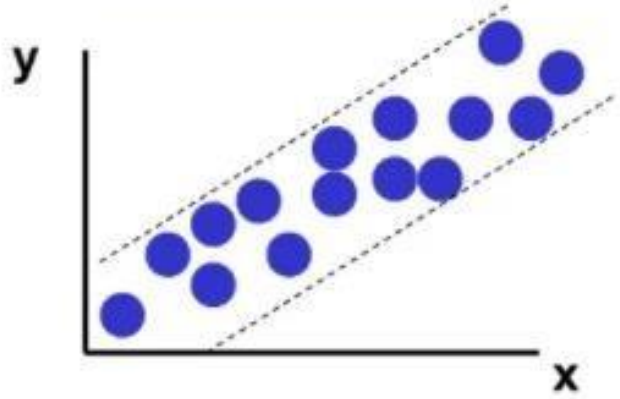


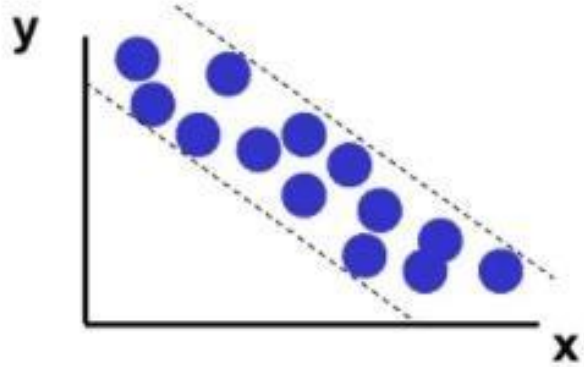


Scatter Plot Examples

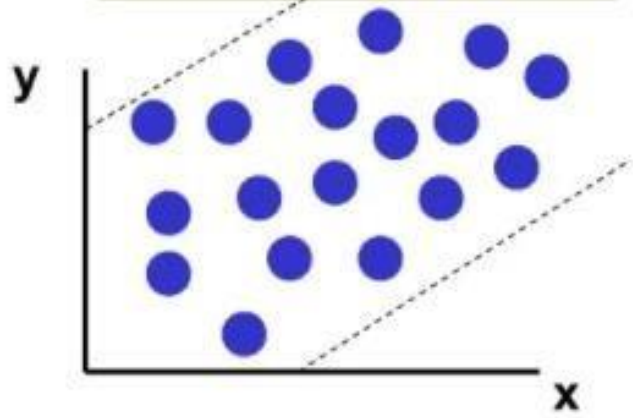
(continued)

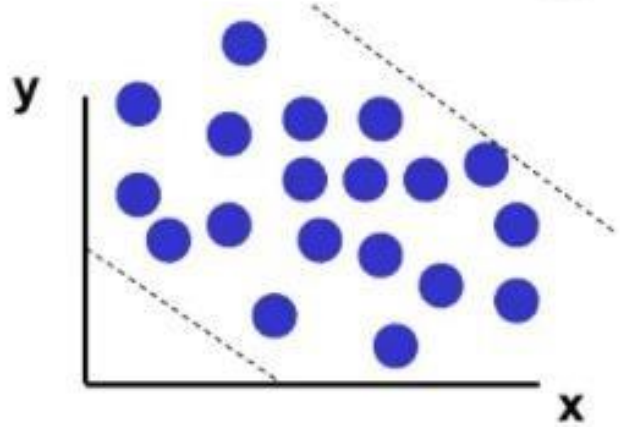
Strong relationships





Weak relationships

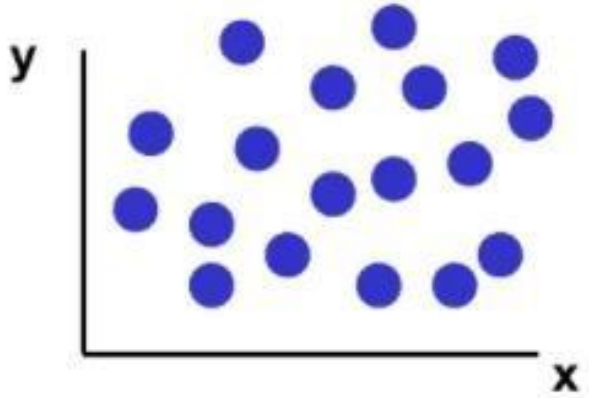


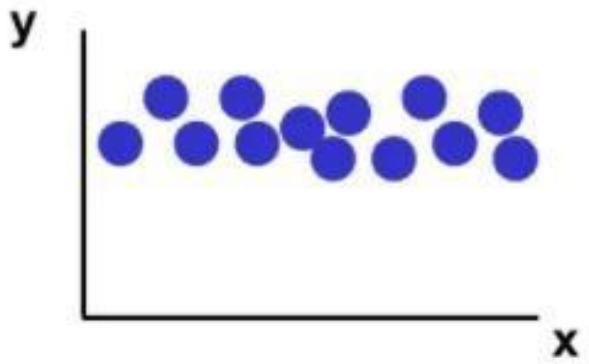


Scatter Plot Examples

(continued)

No relationship





Correlation Coefficient

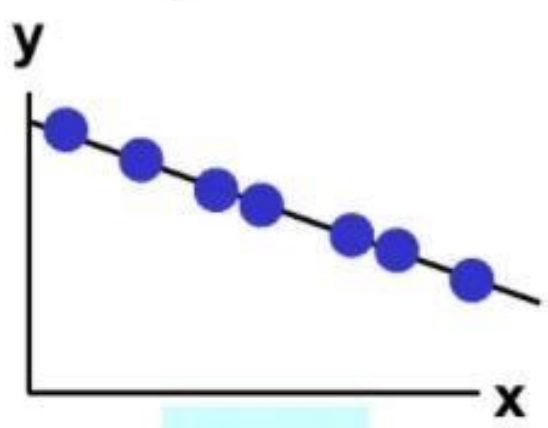
(continued)

- Correlation measures the strength of the linear association between two variables
- The sample correlation coefficient r is a measure of the strength of the linear relationship between two variables, based on sample observations

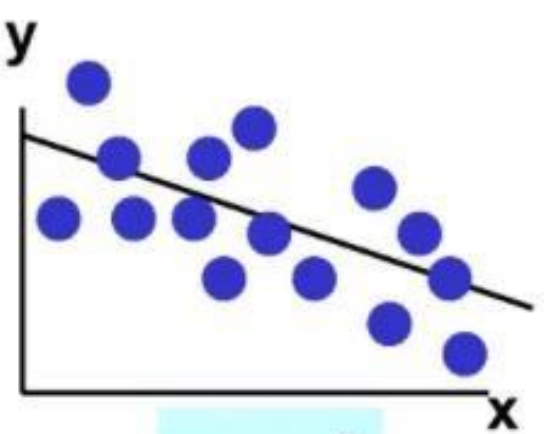
Features of r

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

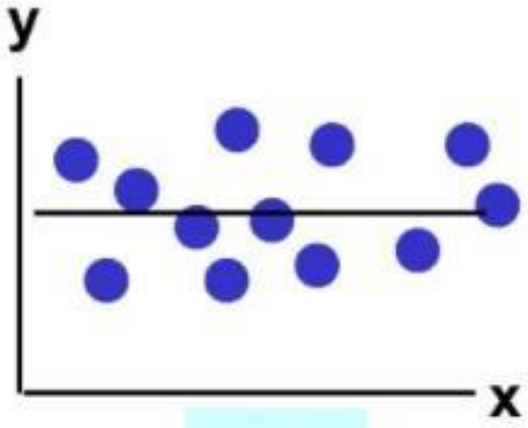
Examples of Approximate r Values



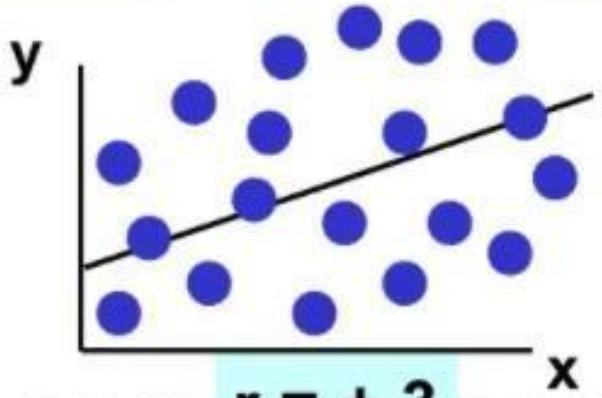
r = -1



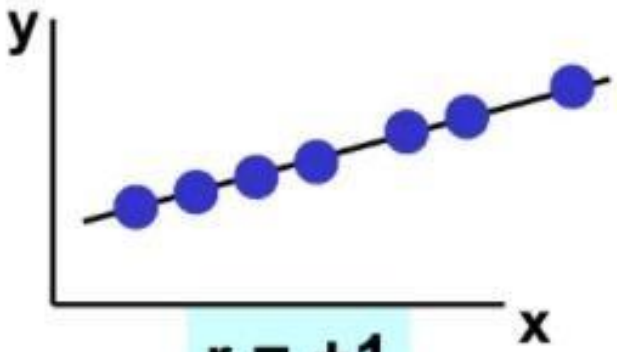
r = -.6



r = 0



r = +.3



r = +1

The following diagrams of the scattered data depict different forms of correlation.

PERFECT POSITIVE CORRELATION

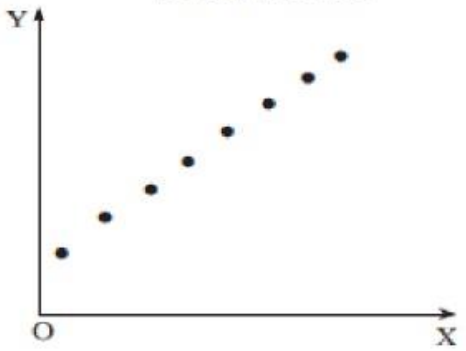


Fig. 8.1.

PERFECT NEGATIVE CORRELATION

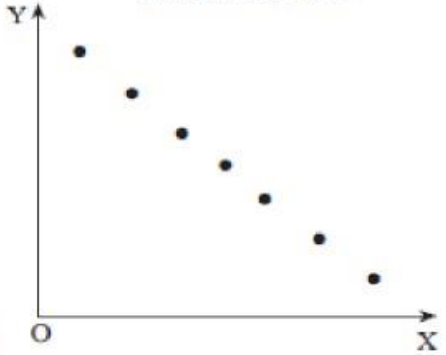


Fig. 8.2.

LOW DEGREE OF POSITIVE CORRELATION

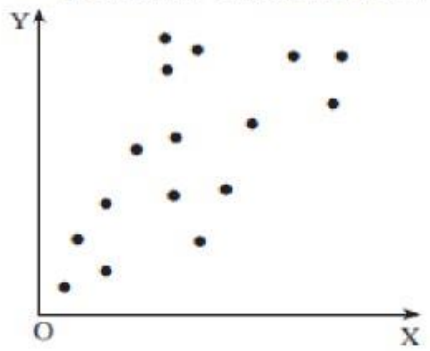


Fig. 8.3.

LOW DEGREE OF NEGATIVE CORRELATION

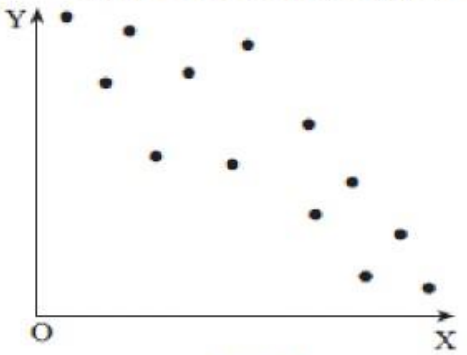


Fig. 8.4.

HIGH DEGREE OF POSITIVE CORRELATION

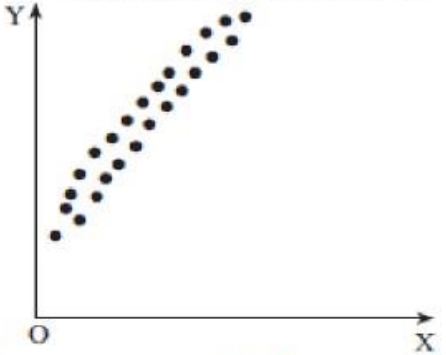


Fig. 8.5.

HIGH DEGREE OF NEGATIVE CORRELATION

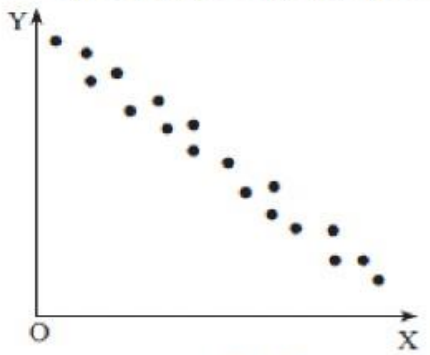


Fig. 8.6.

NO CORRELATION

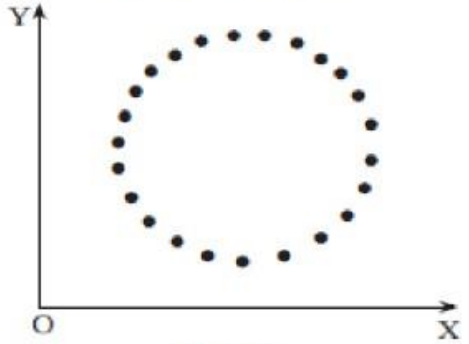


Fig. 8.7.

NO CORRELATION

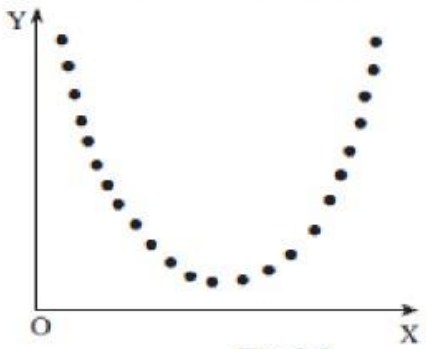
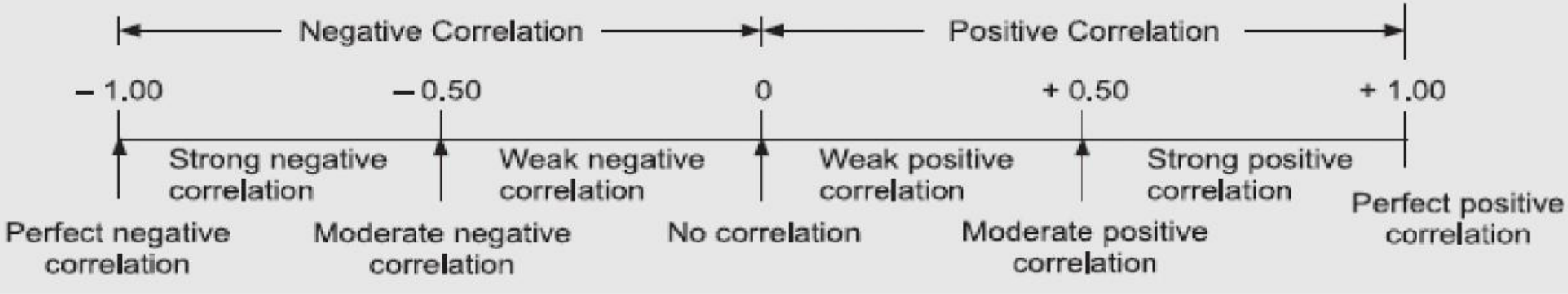


Fig. 8.8.


Negative Correlation
Positive Correlation
– 1.00
– 0.50
0
+ 0.50
+ 1.00
Strong negative correlation
Weak negative correlation
Weak positive correlation
Strong positive correlation
Perfect negative correlation
Moderate negative correlation
No correlation
Moderate positive correlation
Perfect positive correlation

Calculating the Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

or the algebraic equivalent:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where:

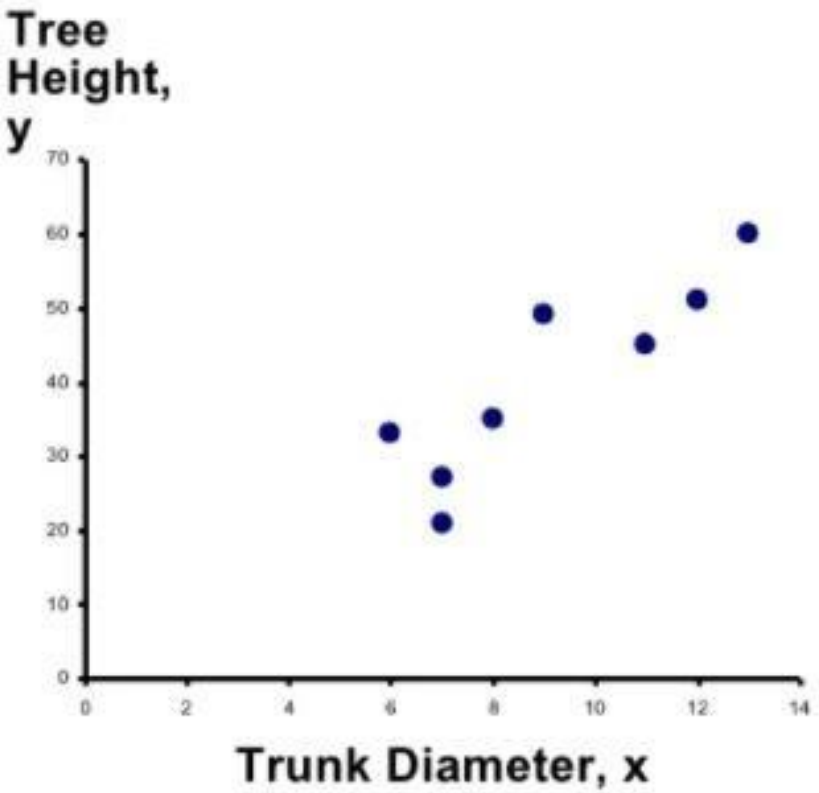
r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

Calculation Example

Tree Height	Trunk Diameter			
y	x	xy	y^2	x^2
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
Σ=321	Σ=73	Σ=3142	Σ=14111	Σ=713

Calculation Example

(continued)



$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$

$$= 0.886$$

r = 0.886 → relatively strong positive linear association between x and y

Cumulative Frequency - Excel

File Home Insert Draw Page Layout Formulas Data Review View Help Tell me what you want to do

Share

Cut Copy Format Painter Paste — Clipboard

Calibri 11 — Font

Wrap Text Merge & Center — Alignment

General — Number

Conditional Formatting Format as Table Cell Styles — Styles

O17

	Tree Height	Trunk Diameter								
	y	x	xy	y2	x2	x-x(bar)	y-y(bar)	x-x(bar) * y-y(bar)	(x-x(bar))^2	(y-y(bar))^2
	dependent	independent								
	35	8	280	1225	64	-1.125	-5.125	5.765625	1.265625	26.265625
	49	9	441	2401	81	-0.125	8.875	-1.109375	0.015625	78.765625
	27	7	189	729	49	-2.125	-13.125	27.890625	4.515625	172.265625
	33	6	198	1089	36	-3.125	-7.125	22.265625	9.765625	50.765625
	60	13	780	3600	169	3.875	19.875	77.015625	15.015625	395.015625
	21	7	147	441	49	-2.125	-19.125	40.640625	4.515625	365.765625
	45	11	495	2025	121	1.875	4.875	9.140625	3.515625	23.765625
	51	12	612	2601	144	2.875	10.875	31.265625	8.265625	118.265625
	321	**73**	**3142**	**14111**	**713**	**63.875**	**280.875**	**17940.89063**	**4080.01563**	**78890.7656**
	40.125	9.125								

Correlation Coefficient | Sheet6 | Sheet10 | Regression | Sheet12 | Sheet17 | Sheet18 | Sheet5 | Formula 1 | _1

Ready Accessibility: Investigate 100%

fx-991ES PLUS C 2nd edition Emulator

Window Tools Help

CASIO

Math

17940.890

√4080.01563x7889

0.999999998

SHIFT ALPHA MODE SETUP ON

CALC SOLVE

hyp sin cos tan

log ln

RCL ENG ( ) S⇔D M+

CONST CONV CLR INS OFF

7 8 9 DEL AC

4 5 6 × ÷

1 2 3 + −

0 • ×10^x Ans =

QA | Road Accident Analys... | Module 3 - PowerPoint | Cumulative Frequency... | fx-991ES PLUS C 2nd ... | 21... | ENG | 10:07 AM

Excel Output

Excel Correlation Output

Tools / data analysis / correlation…

	Tree Height	Trunk Diameter
Tree Height	1	
Trunk Diameter	0.886231	1

Correlation between
Tree Height and Trunk Diameter

Find Pearson's Correlation Coefficient (r)

x	y
17	94
13	73
12	59
15	80
16	93
14	85
16	66
16	79
18	77
19	91

Comment on the relationship between two variables

Calculate Karl Pearson's coefficient of correlation between for the following data

X	1	2	3	4	5	6	7	8	9	10
Y	38	36	34	32	30	28	26	24	22	20

Calculate Karl Pearson's coefficient of correlation between expenditure on advertising and sales from the data given below.

Advertising expenses ('000 Rs.)	:	39	65	62	90	82	75	25	98	36	78
Sales (lakh Rs.)	:	47	53	58	86	62	68	60	91	51	84

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

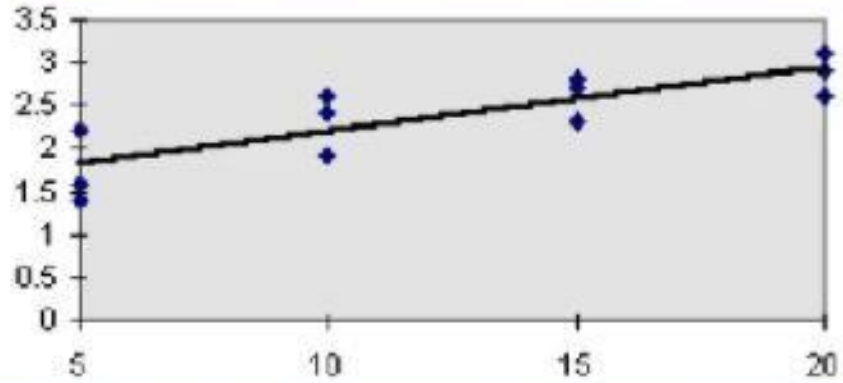
Independent variable: the variable used to explain the dependent variable

Simple Linear Regression Model

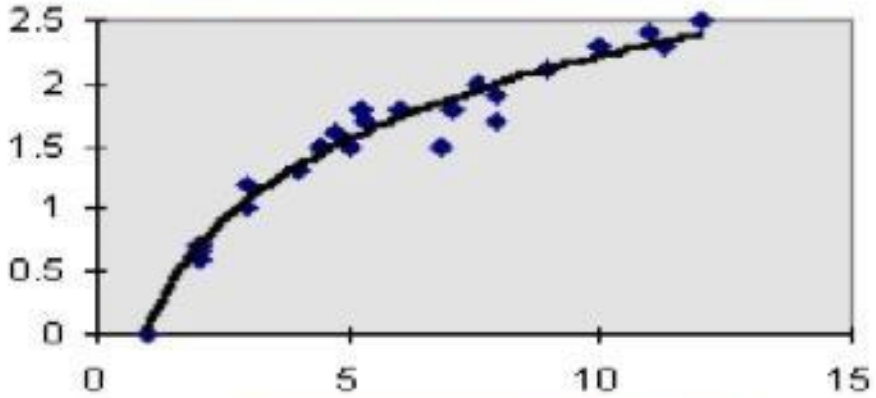
- Only **one** independent variable, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be **caused** by changes in x

Types of Regression Models

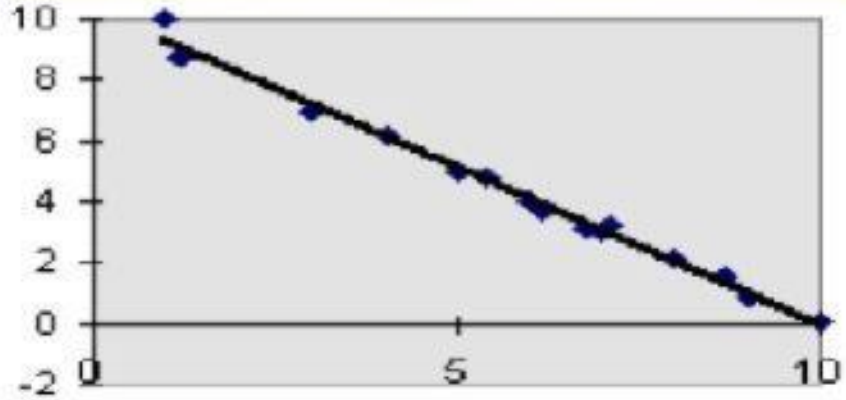
Positive Linear Relationship



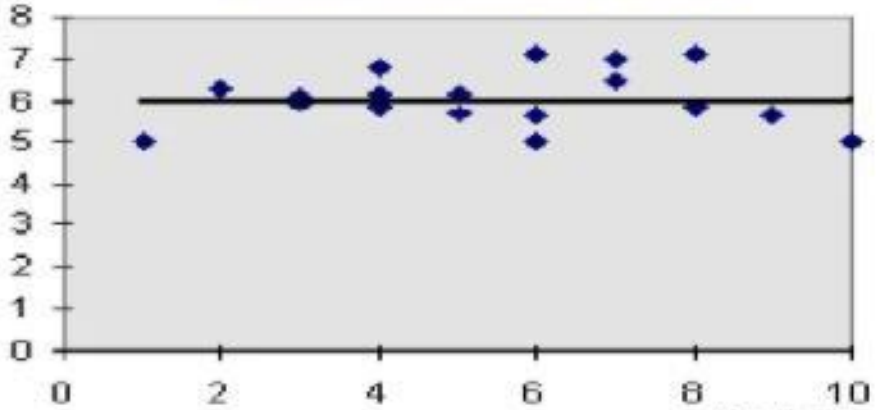
Relationship NOT Linear



Negative Linear Relationship

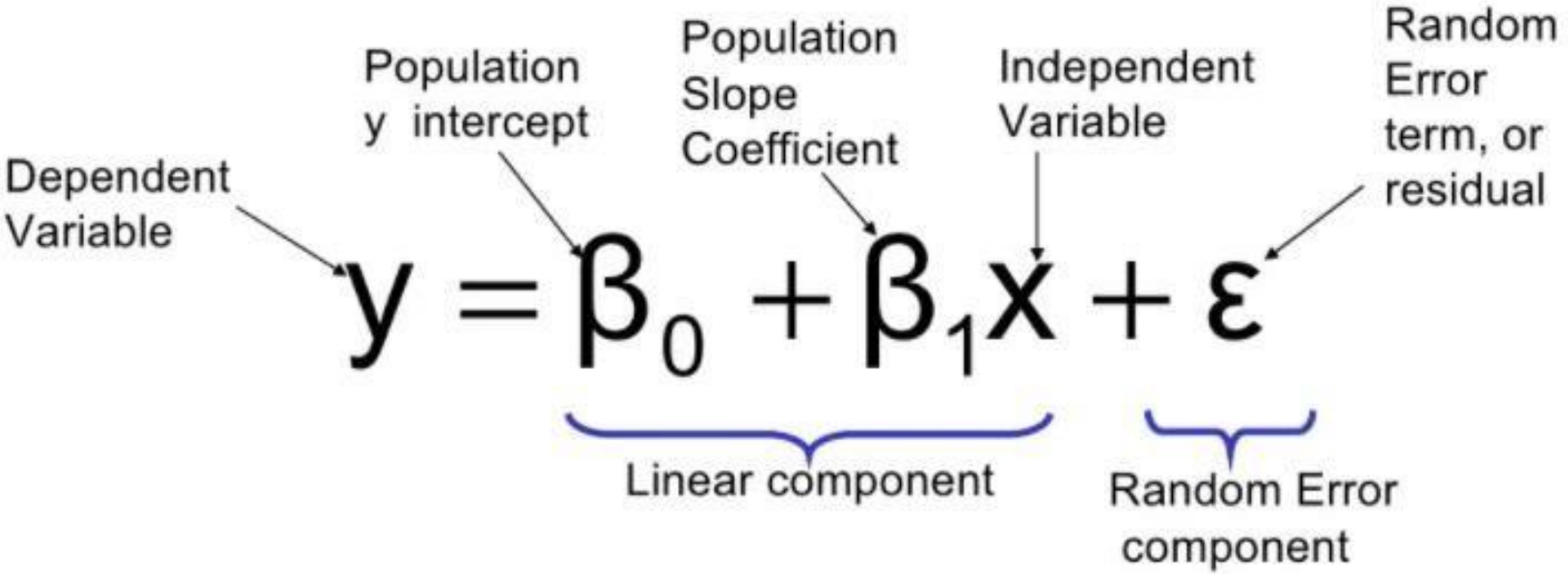


No Relationship



Population Linear Regression

The population regression model:

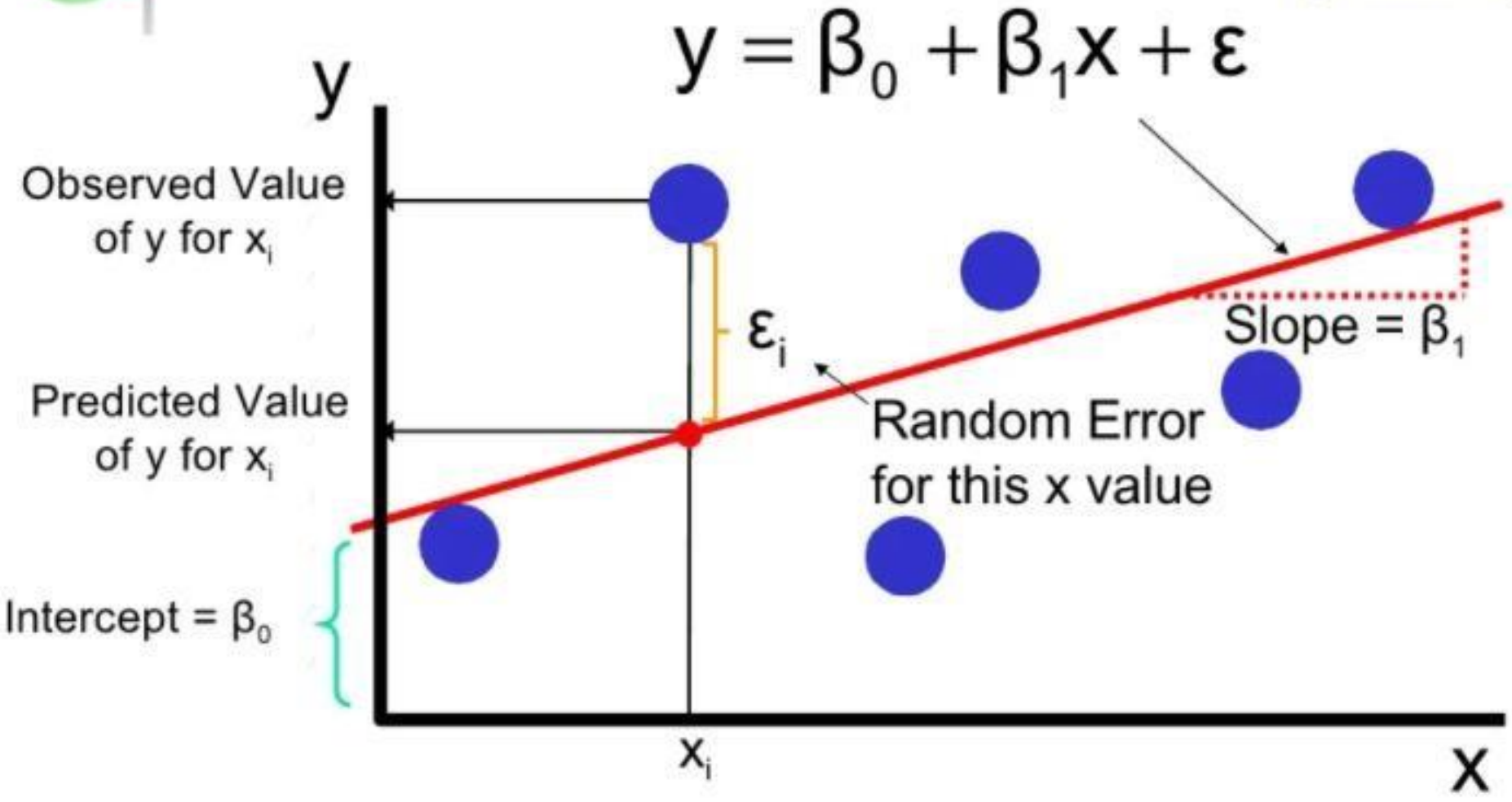


Linear Regression Assumptions

- Error values (ε) are statistically independent
- Error values are normally distributed for any given value of x
- The probability distribution of the errors is normal
- The distributions of possible ε values have equal variances for all values of x
- The underlying relationship between the x variable and the y variable is linear

Population Linear Regression

(continued)



Estimated Regression Model

The sample regression line provides an estimate of the population regression line

Estimated (or predicted) y value

Estimate of the regression intercept

Estimate of the regression slope

Independent variable

$$\hat{y}_i = b_0 + b_1 x$$

The individual random error terms e_i have a mean of zero

Least Squares Criterion

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals

$$\sum e^2 = \sum (y - \hat{y})^2$$

$$= \sum (y - (b_0 + b_1 x))^2$$

The Least Squares Equation

- The formulas for b_1 and b_0 are:

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

algebraic equivalent for b_1:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of y when the value of x is zero

- b_1 is the estimated change in the average value of y as a result of a one-unit change in x

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

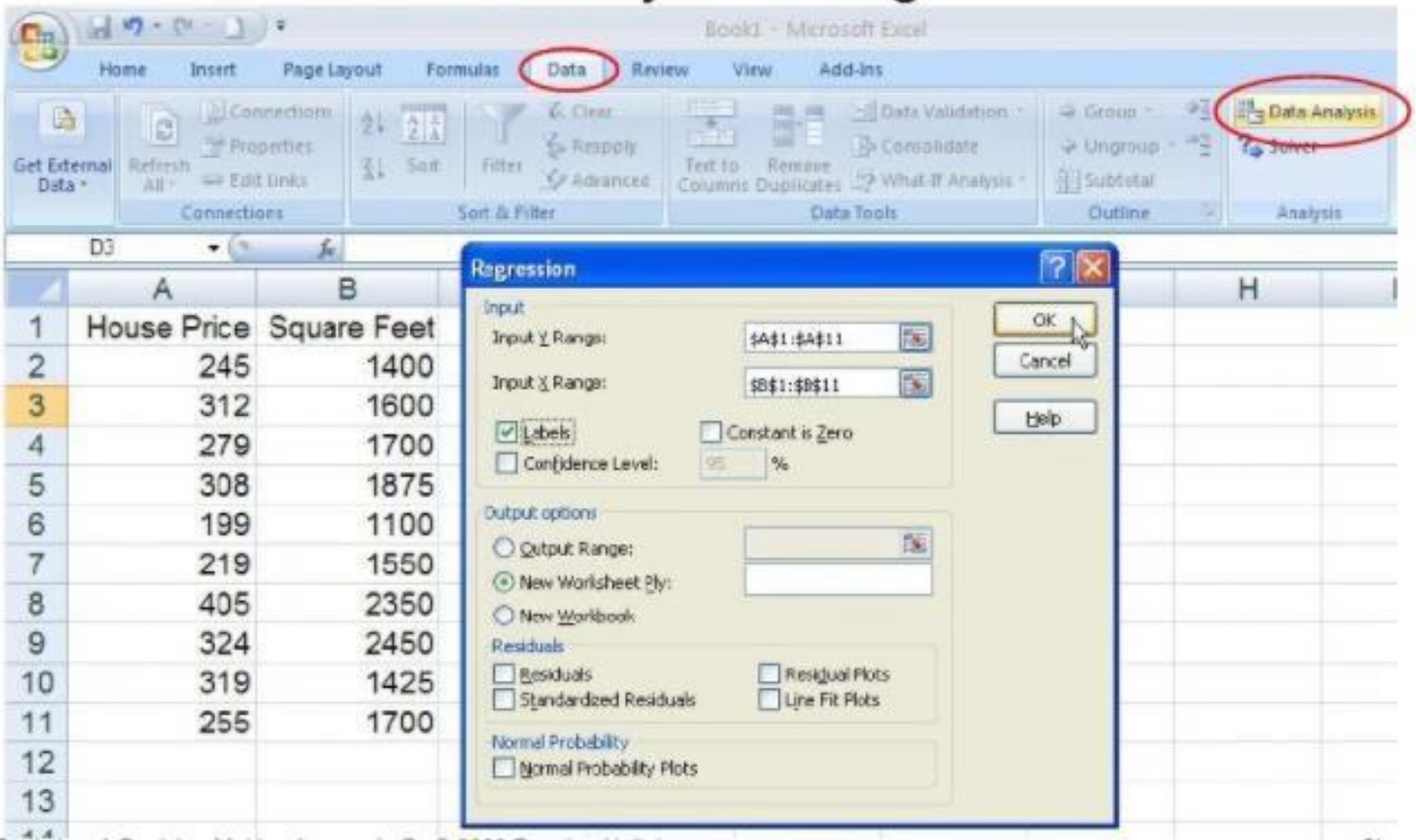
- A random sample of 10 houses is selected
  - Dependent variable (y) = house price in $1000s
  - Independent variable (x) = square feet

Sample Data for House Price Model

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Regression Using Excel

- Data / Data Analysis / Regression



Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

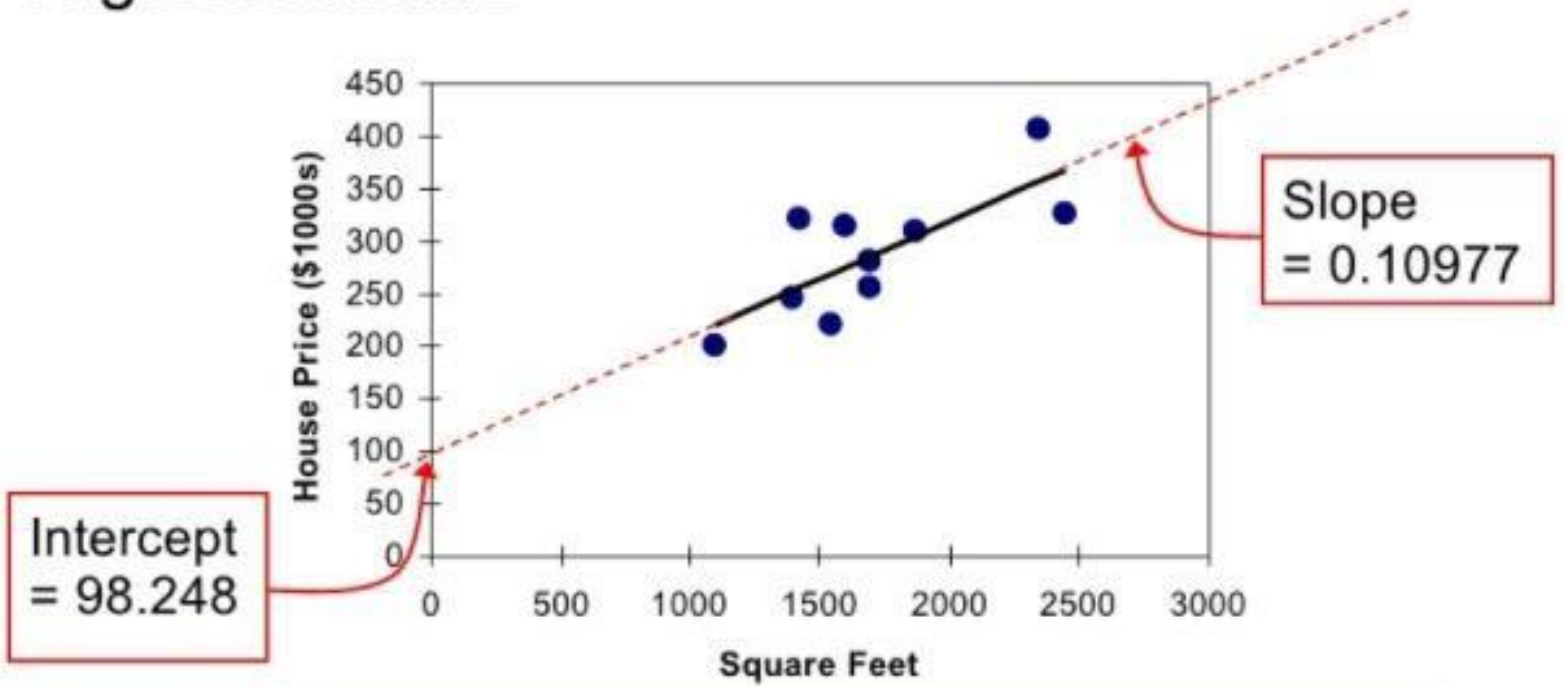
$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Graphical Presentation

- House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if x = 0 is in the range of observed x values)
  - Here, no houses had 0 square feet, so b_0 = 98.24833 just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $b_1 = .10977$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Example: House Prices

(continued)

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\ (\text{sq.ft.})$$

$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Least Squares Regression Properties

- The sum of the residuals from the least squares regression line is 0 ($\sum (y - \hat{y}) = 0$)
- The sum of the squared residuals is a minimum (minimized $\sum (y - \hat{y})^2$)
- The simple regression line always passes through the mean of the y variable and the mean of the x variable
- The least squares coefficients are unbiased estimates of β_0 and β_1

Explained and Unexplained Variation

- Total variation is made up of two parts:

$$SST = SSE + SSR$$

Total sum of Squares	Sum of Squares Error	Sum of Squares Regression

$$SST = \sum (y - \bar{y})^2 \qquad SSE = \sum (y - \hat{y})^2 \qquad SSR = \sum (\hat{y} - \bar{y})^2$$

where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

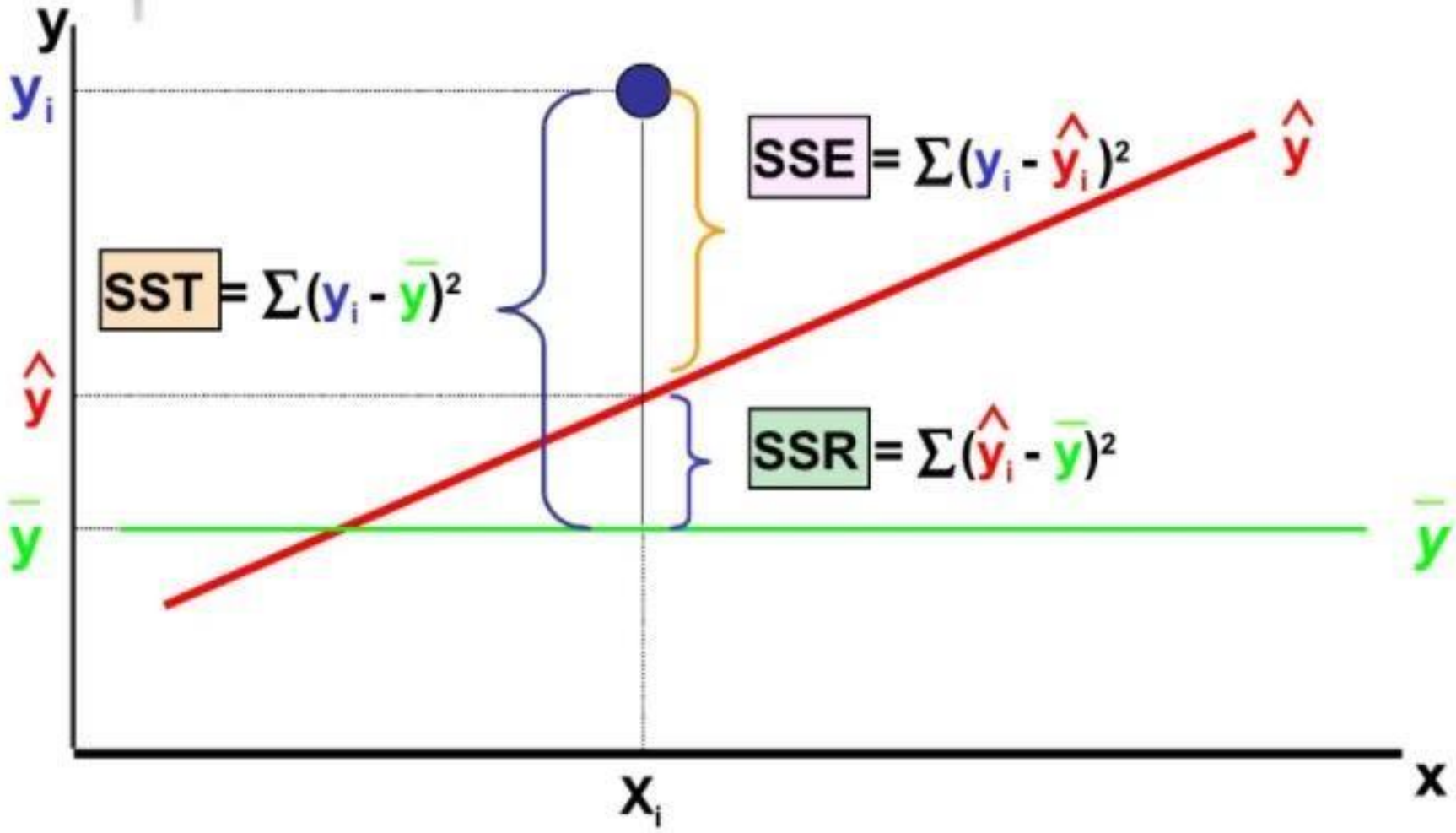
Explained and Unexplained Variation

(continued)

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean y
- SSE = error sum of squares
  - Variation attributable to factors other than the relationship between x and y
- SSR = regression sum of squares
  - Explained variation attributable to the relationship between x and y

Explained and Unexplained Variation

(continued)



Coefficient of Determination, R^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called R-squared and is denoted as R^2

$$R^2 = \frac{SSR}{SST} \quad \text{where} \quad 0 \le R^2 \le 1$$

Coefficient of Determination, R^2

(continued)

Coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

Note: In the single independent variable case, the coefficient of determination is

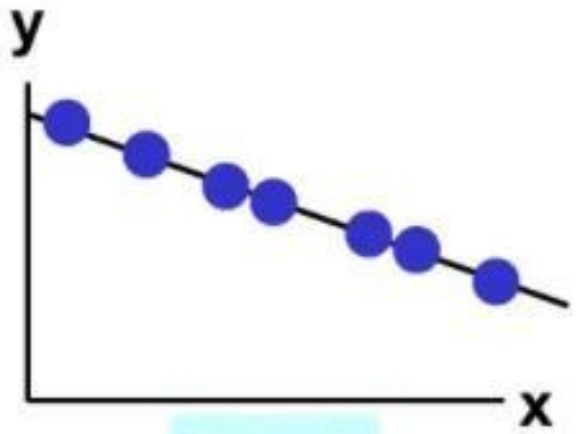
$$R^2 = r^2$$

where:

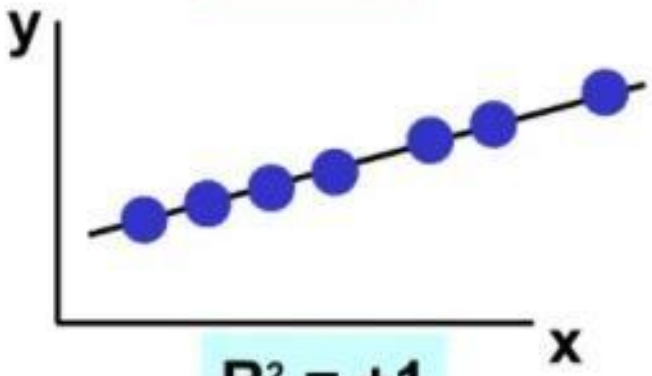
R^2 = Coefficient of determination

r = Simple correlation coefficient

Examples of Approximate R^2 Values



$R^2 = 1$



$R^2 = +1$

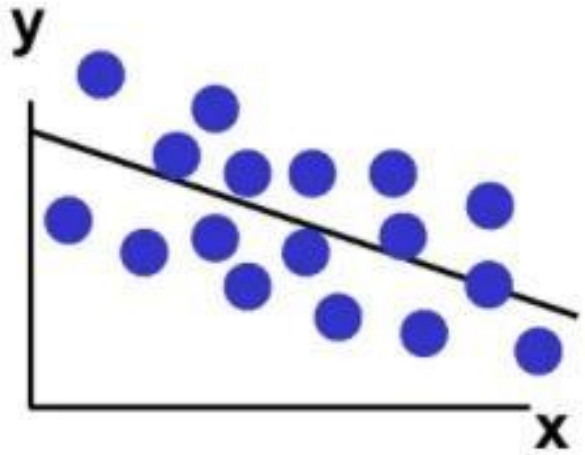
$R^2 = 1$

Perfect linear relationship between x and y:

100% of the variation in y is explained by variation in x

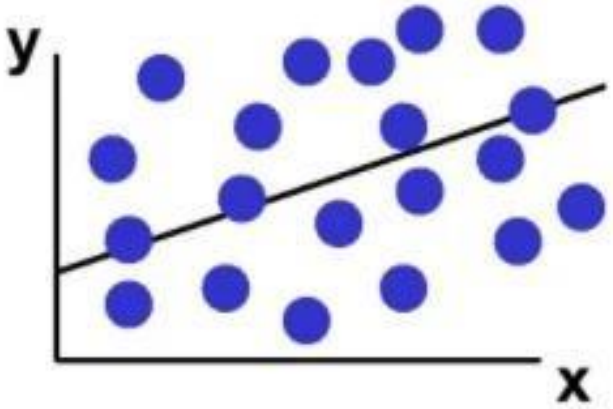
Examples of Approximate R^2 Values

(continued)



$0 < R^2 < 1$

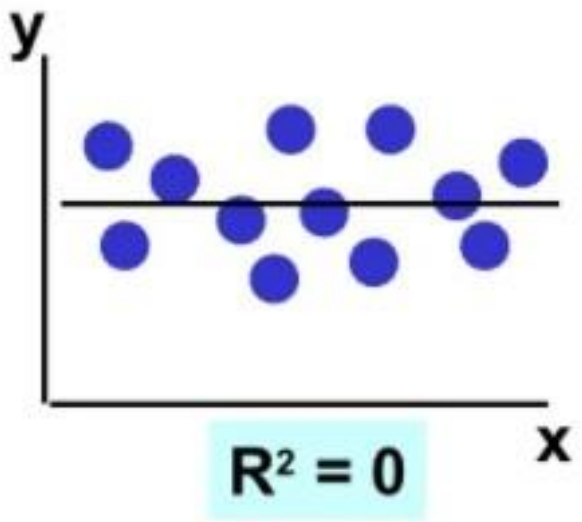
Weaker linear relationship between x and y:



Some but not all of the variation in y is explained by variation in x

Examples of Approximate R^2 Values

(continued)



$R^2 = 0$

No linear relationship between x and y:

The value of Y does not depend on x. (None of the variation in y is explained by variation in x)

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Standard Error of Estimate

- The standard deviation of the variation of observations around the simple regression line is estimated by

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

Where

SSE = Sum of squares error

n = Sample size

The Standard Deviation of the Regression Slope

- The standard error of the regression slope coefficient (b_1) is estimated by

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

where:

s_{b_1} = Estimate of the standard error of the least squares slope

$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ = Sample standard error of the estimate

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$s_\varepsilon = 41.33032$

$s_{b_1} = 0.03297$

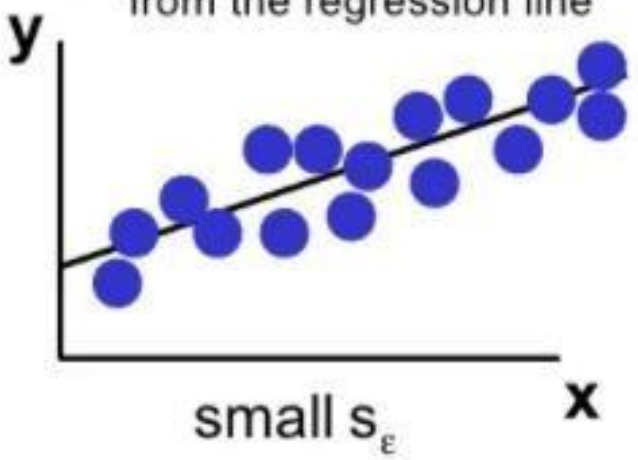
ANOVA

	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

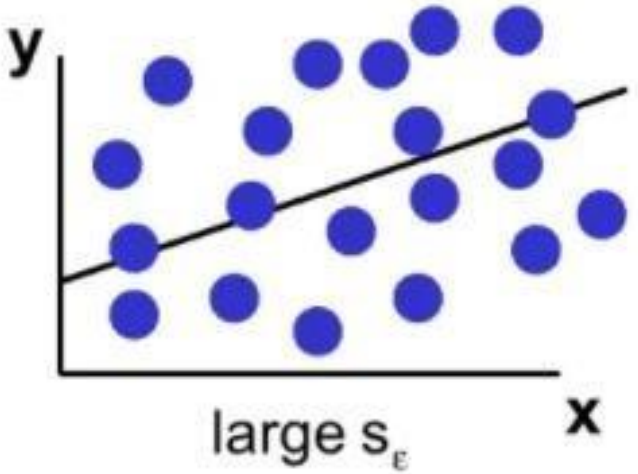
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Comparing Standard Errors

Variation of observed y values from the regression line

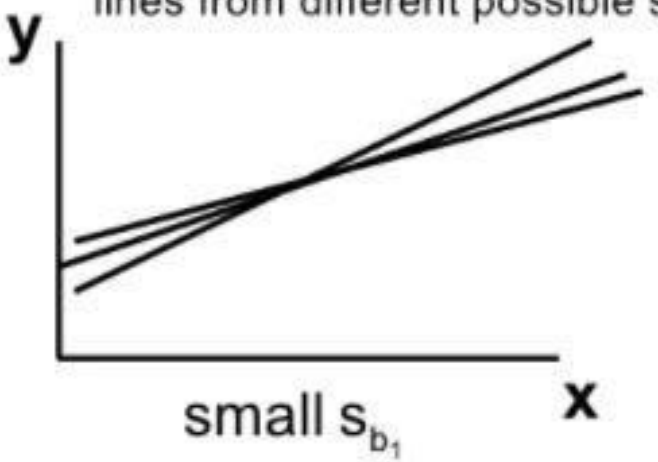


small s_ε

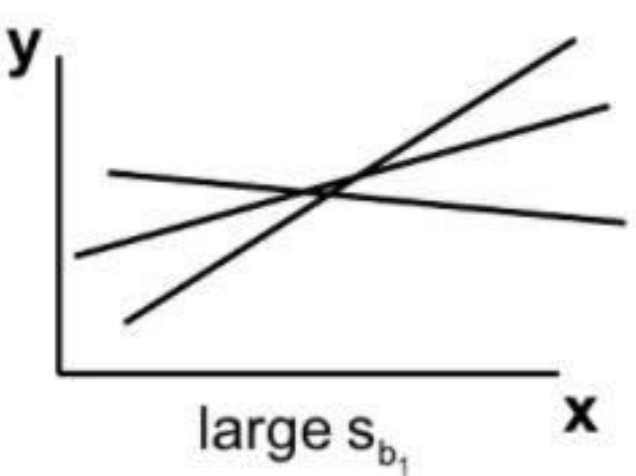


large s_ε

Variation in the slope of regression lines from different possible samples



small s_{b_1}



large s_{b_1}