University of Basra

College of Engineering

Chemical Engineering Department

Engineering Statistics

The second stage

Dr. Mohammad N. Fares

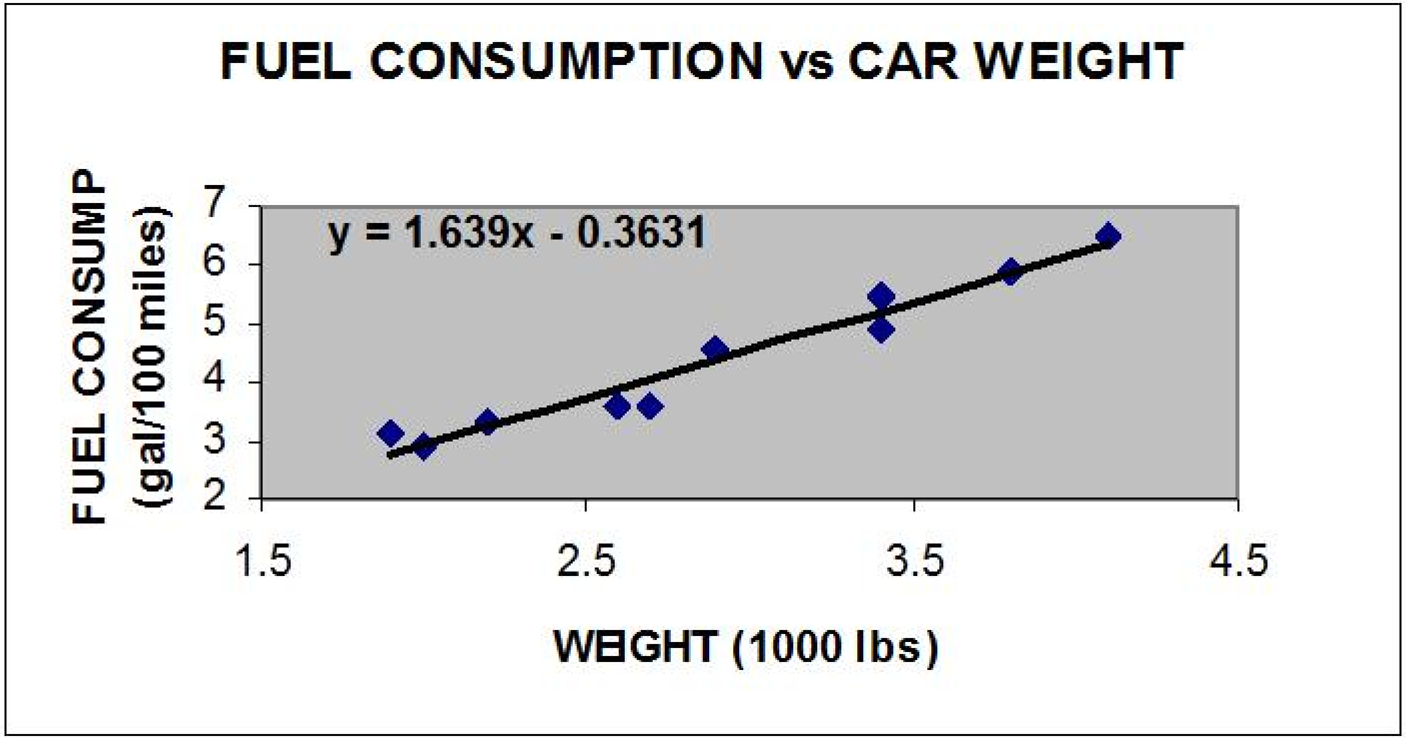
Regression

Regression

After knowing the relationship between two variables we may be interested in estimating (predicting) the value of one variable given the value of another.

Definition:

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.



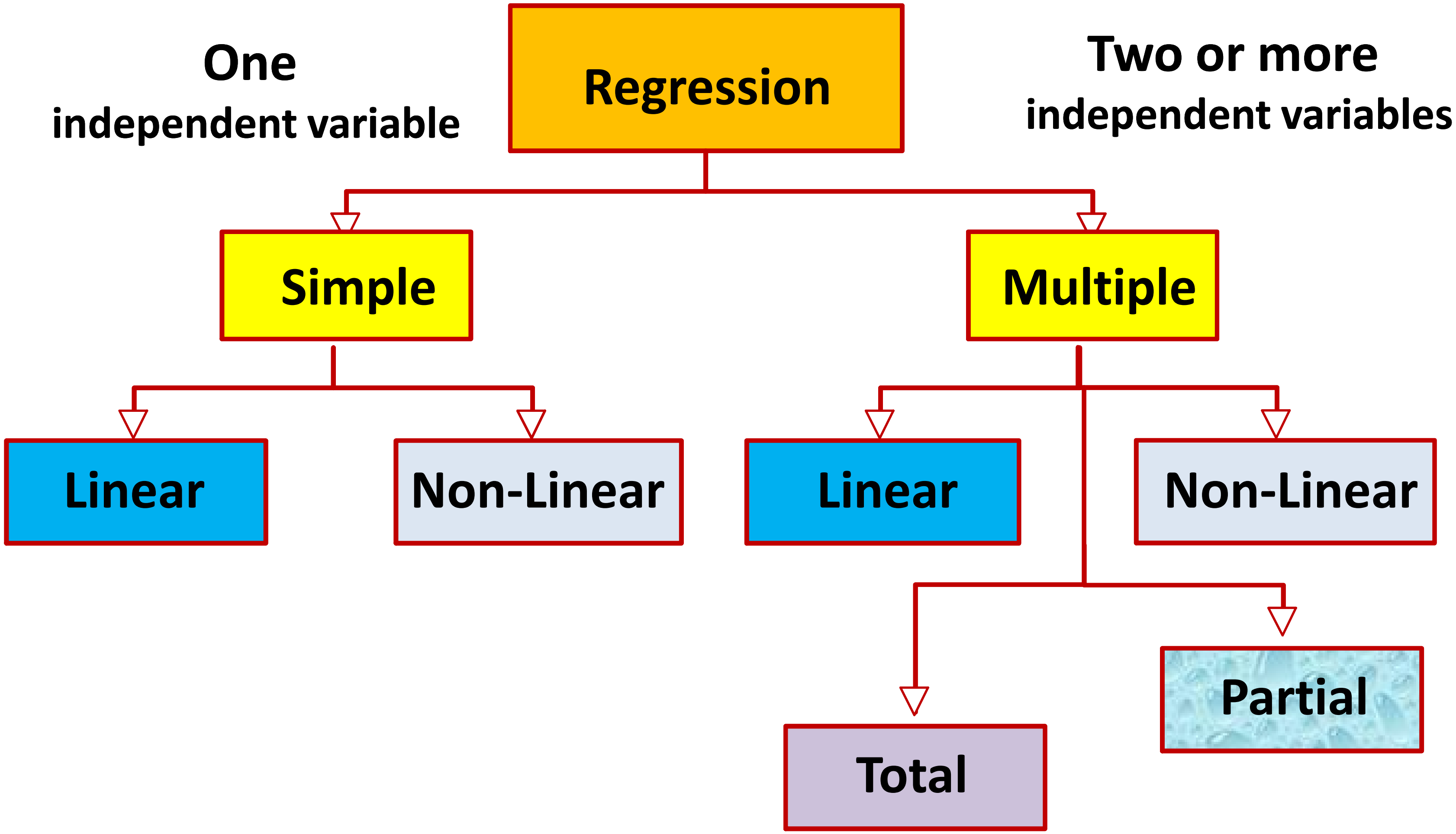
Regression Analysis

- **Regression analysis** is used to:

–Predict the value of a dependent variable based on the value of independent variable

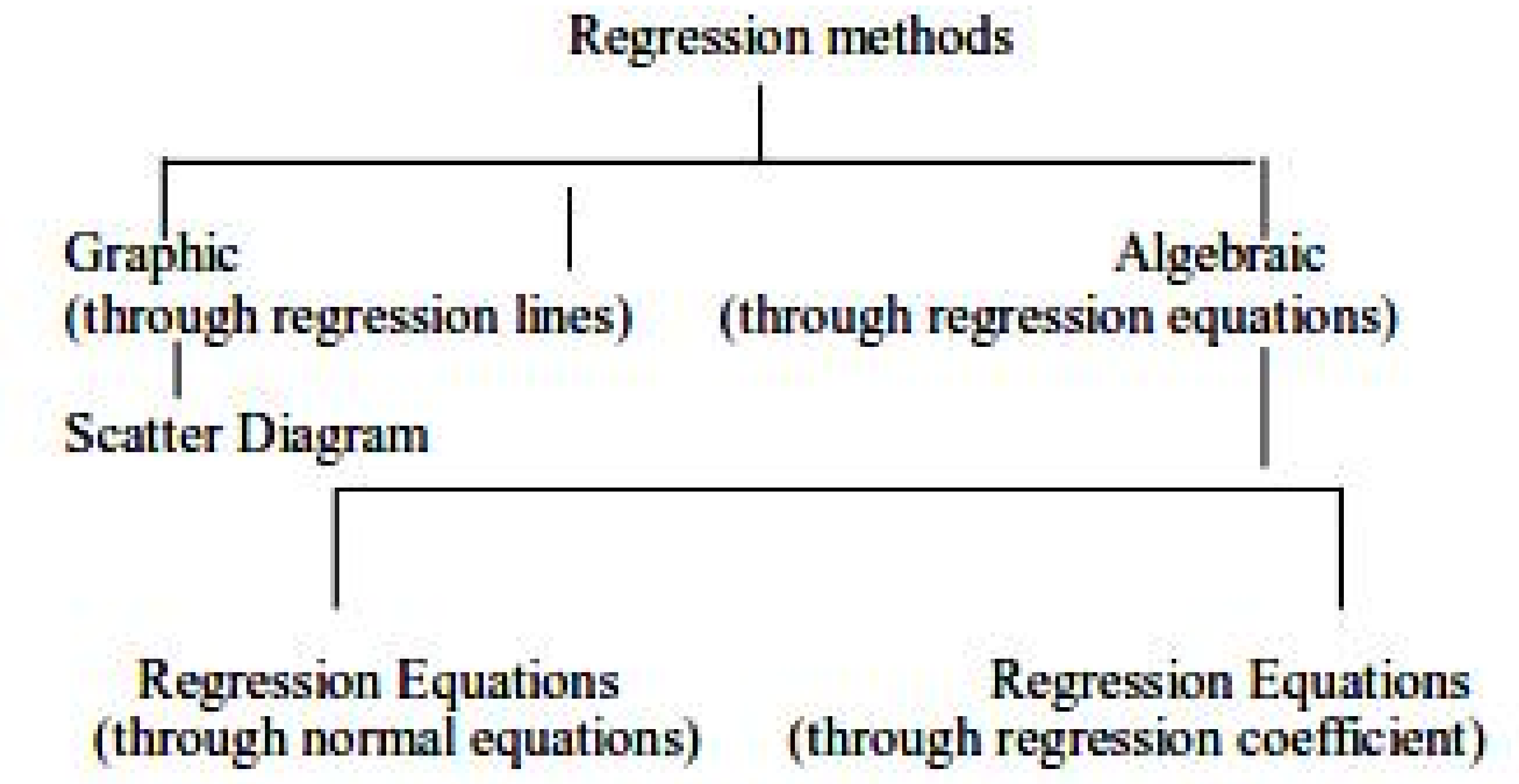
–Explain the impact of changes in an independent variable on the dependent variable

Types of Regression Models



Methods of Regression Analysis:

The various methods can be represented in the form of chart given below:



Simple Linear Regression

- The equation that describes how **y** is related to **x** and an error term ε is called the **regression**
- The **simple linear** is:

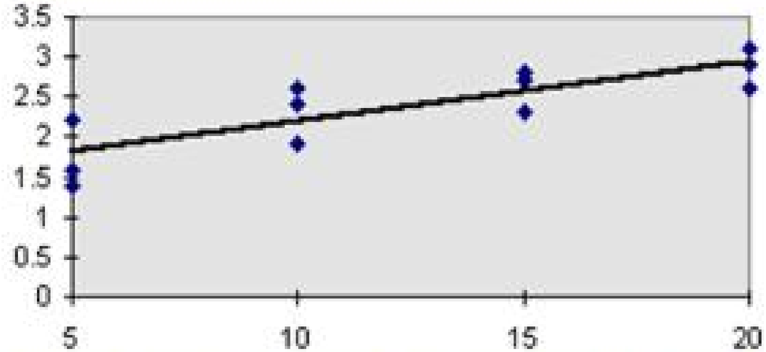
$$y = \beta_0 + \beta_1 x + \varepsilon$$

  - β_0 and β_1 are called **parameters of regression**.
  - ε is a random variable called the **error term**.

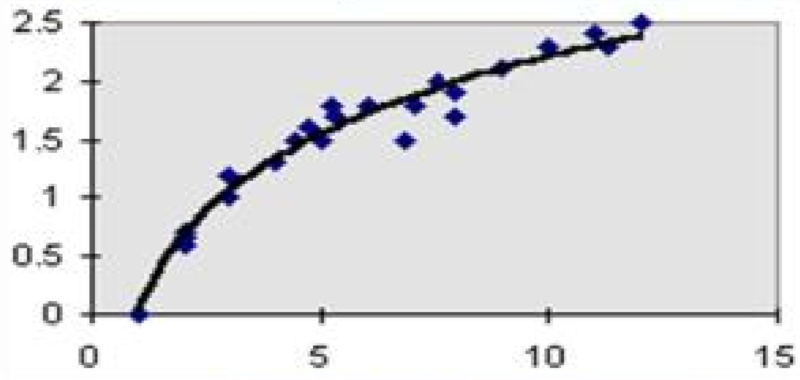
- Only **one** independent variable, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be caused by changes in x

Types of Regression

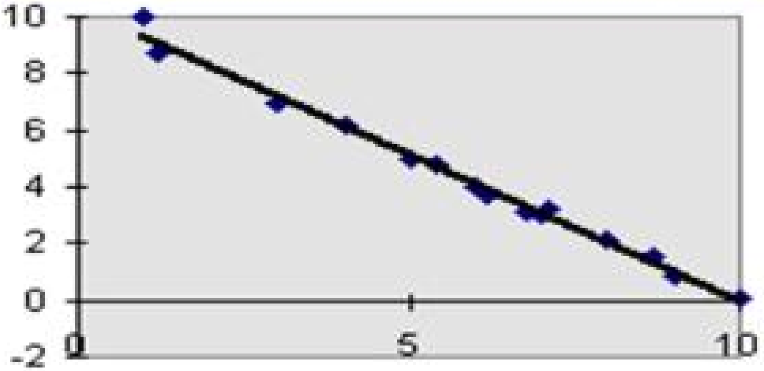
Positive Linear Relationship



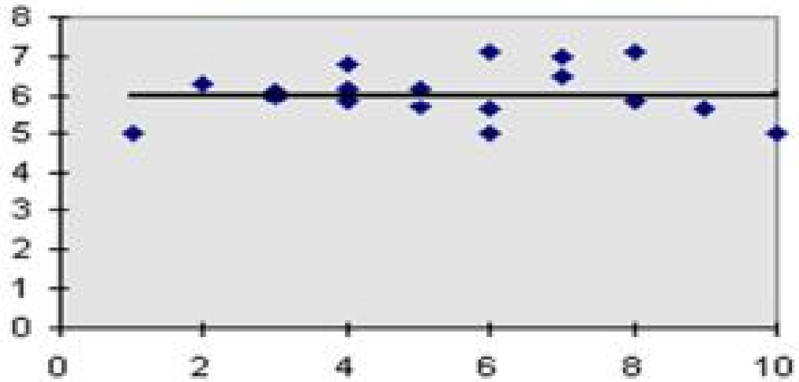
Relationship NOT Linear



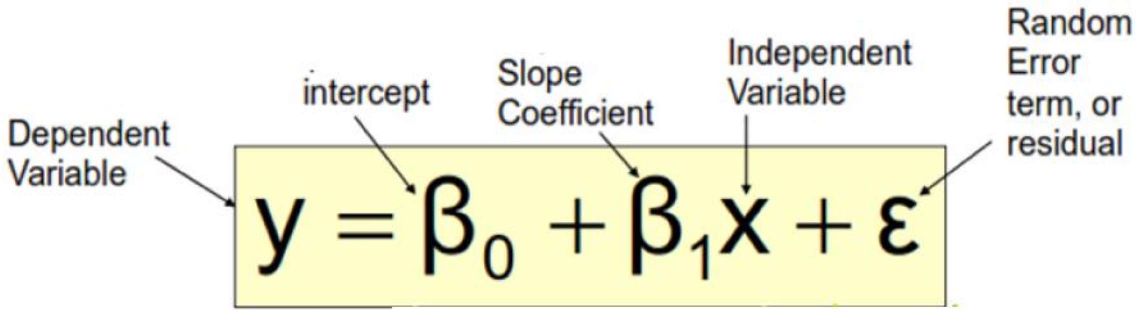
Negative Linear Relationship

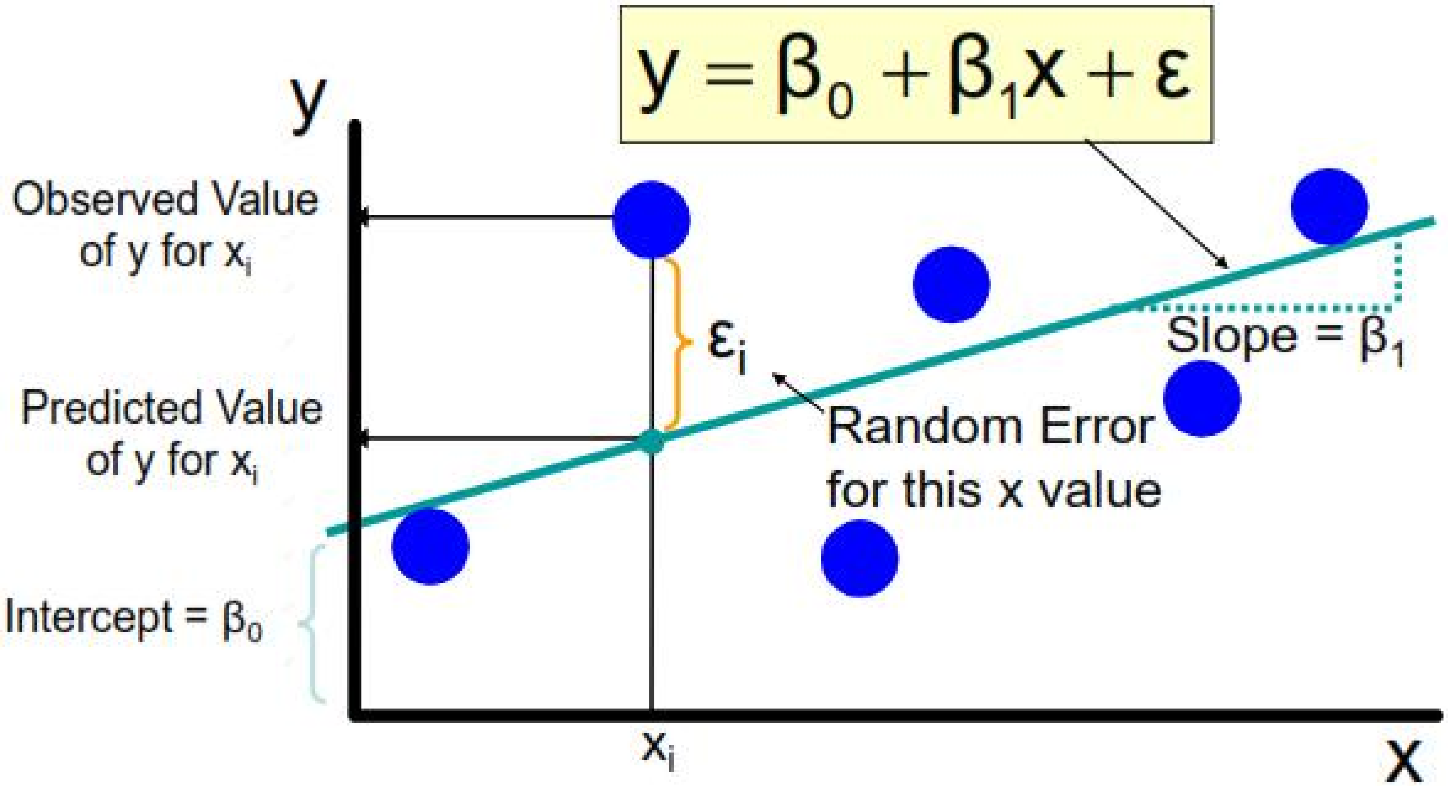


No Relationship



Linear Regression





Estimated Regression

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

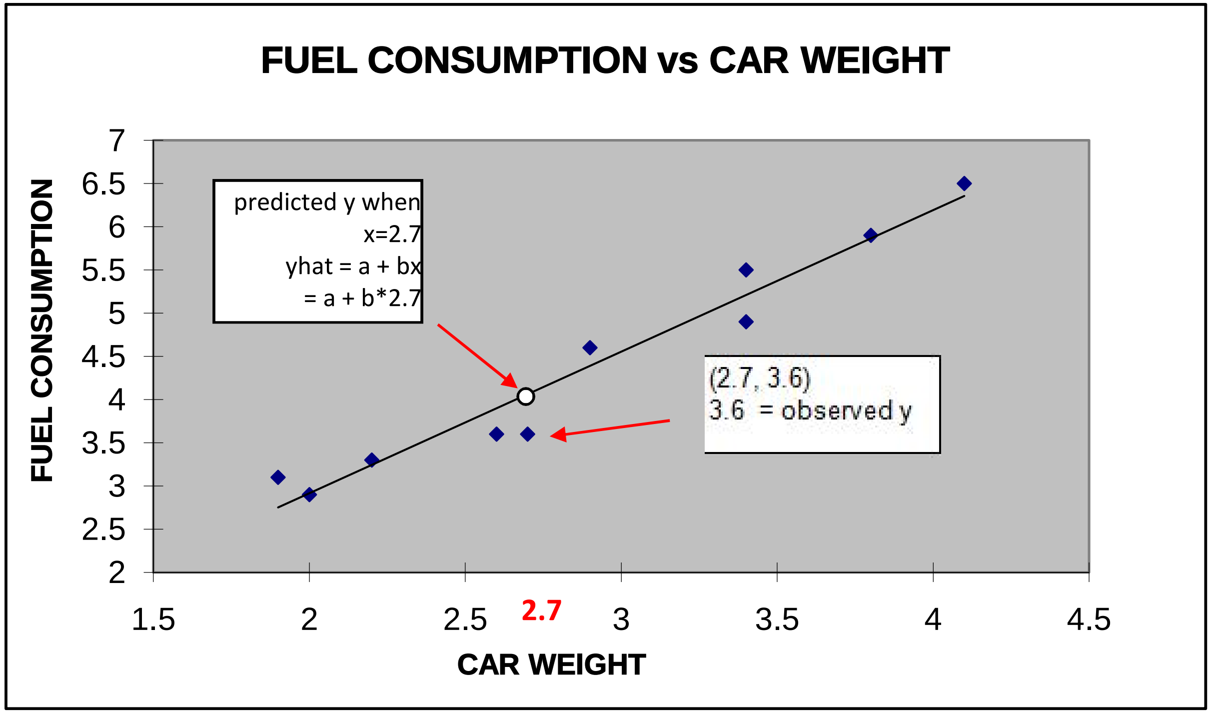
Estimated (or predicted) y value

Estimate of the regression intercept

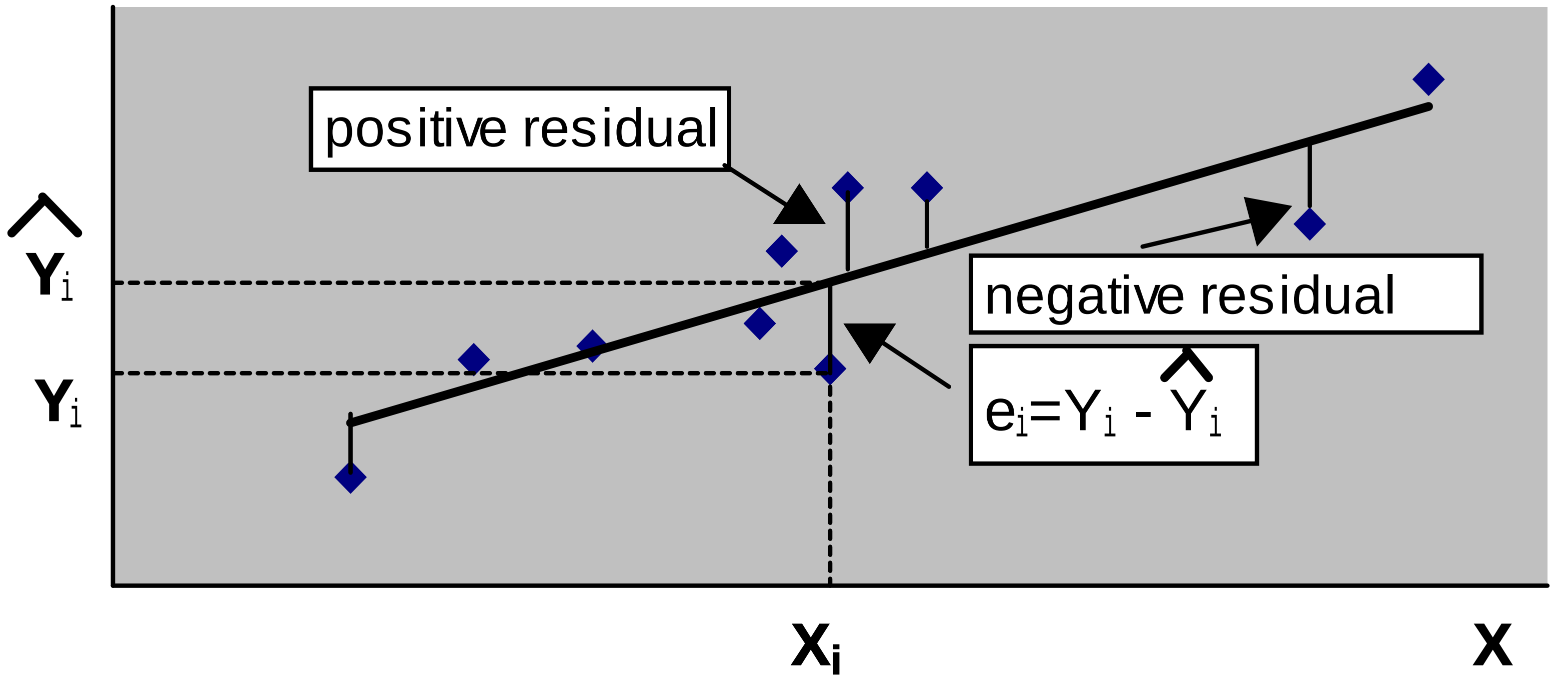
Estimate of the regression slope

Independent variable

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


FUEL CONSUMPTION vs CAR WEIGHT
FUEL CONSUMPTION
CAR WEIGHT
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
2
2.5
2.7
3
3.5
4
4.5
predicted y when x=2.7
yhat = a + bx
= a + b*2.7
(2.7, 3.6)
3.6 = observed y

Graphical Display of Residuals



The least squares method

- The method of least squares chooses the line that makes the **sum of squares of the residuals as small as possible**
- This line has slope $\mathbf{b_1}$ and intercept $\mathbf{b_0}$ that minimizes

$$\sum e^2 = \sum (y - \hat{y})^2$$
$$= \sum (y - (b_0 + b_1 x))^2$$

where:

y_i = observed value of the dependent variable

$\hat{y}_i$ = estimated value of the dependent variable

DERIVING THE LEAST SQUARES MATHOD

II. Sum of Squared Residuals:

$$\sum e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum (Y_i - (a + bX_i))^2 = \sum (Y_i - a - bX_i)^2$$

$$= \sum [(Y_i - a - bX_i)(Y_i - a - bX_i)]$$

$$= \sum [Y_i^2 - 2aY_i - 2bX_iY_i + a^2 + 2abX_i + b^2X_i^2]$$

$$\sum e_i^2 = \sum Y_i^2 - 2a\sum Y_i - 2b\sum X_iY_i + na^2 + 2ab\sum X_i + b^2\sum X_i^2$$

III. Partial Derivatives:

$$\frac{\partial \sum e_i^2}{\partial b} = -2\sum X_iY_i + 2a\sum X_i + 2b\sum X_i^2$$

$$\frac{\partial \sum e_i^2}{\partial a} = -2\sum Y_i + 2na + 2b\sum X_i$$

IV. Set Derivatives to Zero, Manipulate Terms, and Divide by Two:

$$\sum X_i Y_i = a \sum X_i + b \sum X_i^2$$

$$\sum Y_i = na + b \sum X_i$$

V. Solve the Normal Equations for the Unknowns, a and b:

$$b = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - \left(\sum X_i\right)^2}$$

$$a = \frac{\sum Y_i}{n} - b \frac{\sum X_i}{n}$$

The least squares method

Slope for the Estimated Regression Equation

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

Intercept for the Estimated Regression Equation

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

where:

x_i = value of independent variable

y_i = value of dependent variable

$\bar{x}$ = mean value for independent variable

$\bar{y}$ = mean value for dependent variable

n = total number of observations

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

Example

The relationship between the density and salinity concentration in water is given by the following formula: $\boldsymbol{\rho} = \boldsymbol{\rho_o} + (\boldsymbol{\beta} . \mathbf{C})$.

By using regression lines equations, Find the constants ($\boldsymbol{\rho_o}$, $\boldsymbol{\beta}$), by using data below.

998	998.8	1000.4	1002	1003.6	1005.2
0	0.002	0.006	0.01	0.014	0.018

n	y	x	x.y	x^2
1	998	0	0	0
1	998.8	0.002	1.9976	0.000004
1	1000.4	0.006	6.0024	0.000036
1	1002	0.01	10.02	0.0001
1	1003.6	0.014	14.0504	0.000196
1	1005.2	0.018	18.0936	0.000324
6				
	x	y	yx	x2
	6008	0.05	50.164	0.00066
	400	b		
	998	a		

Example

The relationship between the density and temperature is given by the following formula:

$\rho = \rho_o + (\beta/T)$.

By using regression lines equations, Find the constants (ρ_o, β), by using data below.

Density, **ρ (kg/m³)**	**882.5**	**880**	**878.33**	**877.14**	**876.25**
Temperature **(C°)**	**20**	**25**	**30**	**35**	**40**

N	T	p	x	y	x.y	x2
1	20	882.5	**0.05**	**882.5**	**44.125**	**0.0025**
1	25	880	**0.04**	**880**	**35.2**	**0.0016**
1	30	878.33	**0.033333333**	**878.33**	**29.27766667**	**0.001111111**
1	35	877.14	**0.028571429**	**877.14**	**25.06114286**	**0.000816327**
1	40	876.25	**0.025**	**876.25**	**21.90625**	**0.000625**
			x	y	yx	x2
5			0.176904762	4394.22	155.5700595	0.006652438
		b	250.0668089			
		a	869.9963981			

Explained and Unexplained Variation

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Sum of Squares Regression	Sum of Squares Error
$SST = \sum (y - \bar{y})^2$	$SSR = \sum (\hat{y} - \bar{y})^2$	$SSE = \sum (y - \hat{y})^2$

where:

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

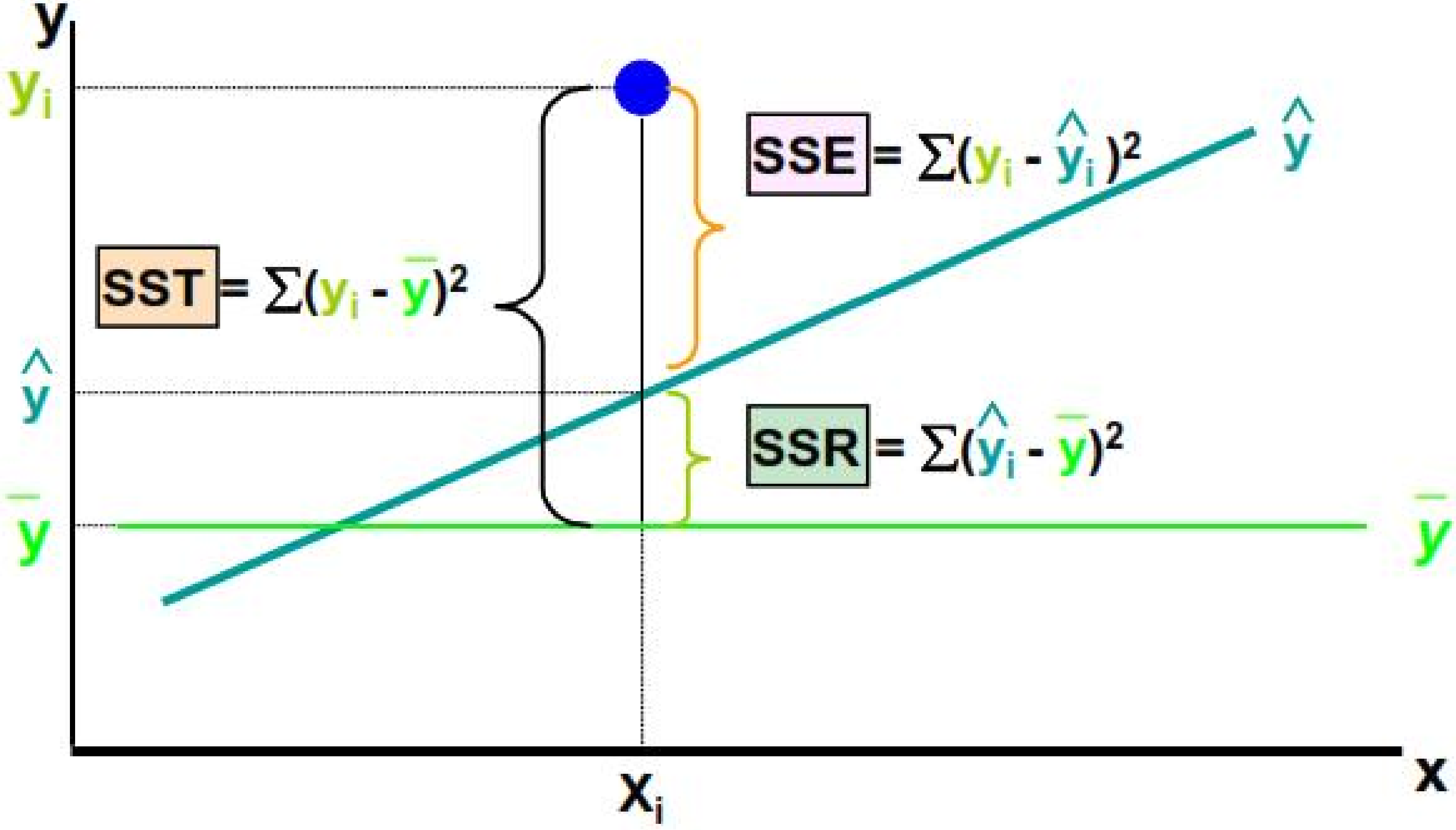
y = Observed values of the dependent variable

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

Explained and Unexplained Variation

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

Explained and Unexplained Variation



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

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

Coefficient of Determination, R^2

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

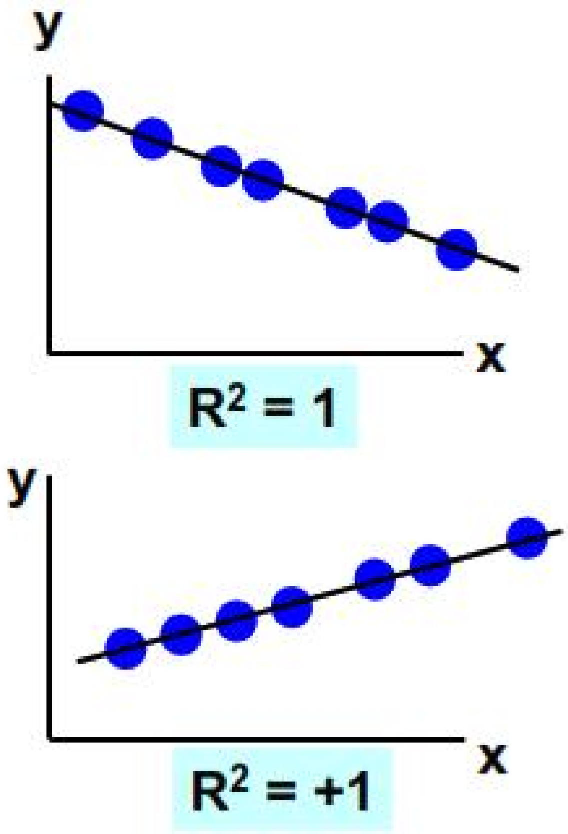
Note: In the single independent variable case, the coefficient of determination is

$$R^2 = r^2$$

where:

R^2 = Coefficient of determination
r = Simple correlation coefficient

Examples of R^2 Values

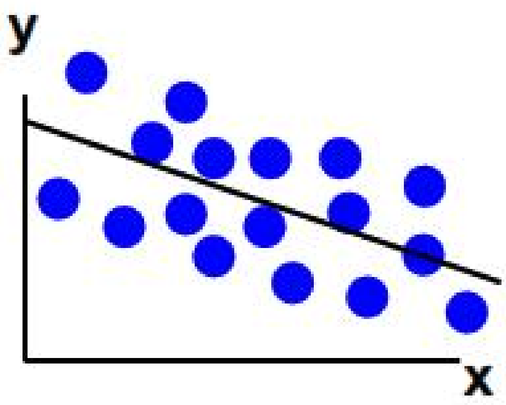


$R^2 = 1$

Perfect linear relationship between x and y:

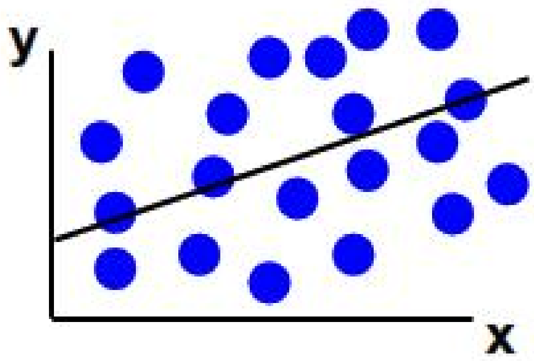
100% of the variation in y is explained by variation in x

Examples of R^2 Values



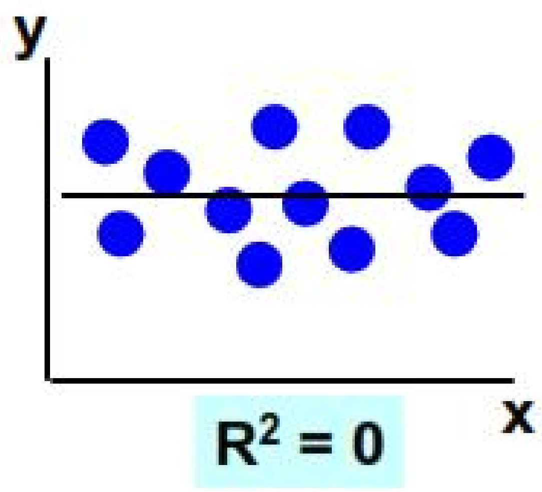
$0 < R^2 < 1$

Weaker linear relationship between x and y:



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

Examples of R^2 Values



$R^2 = 0$

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)