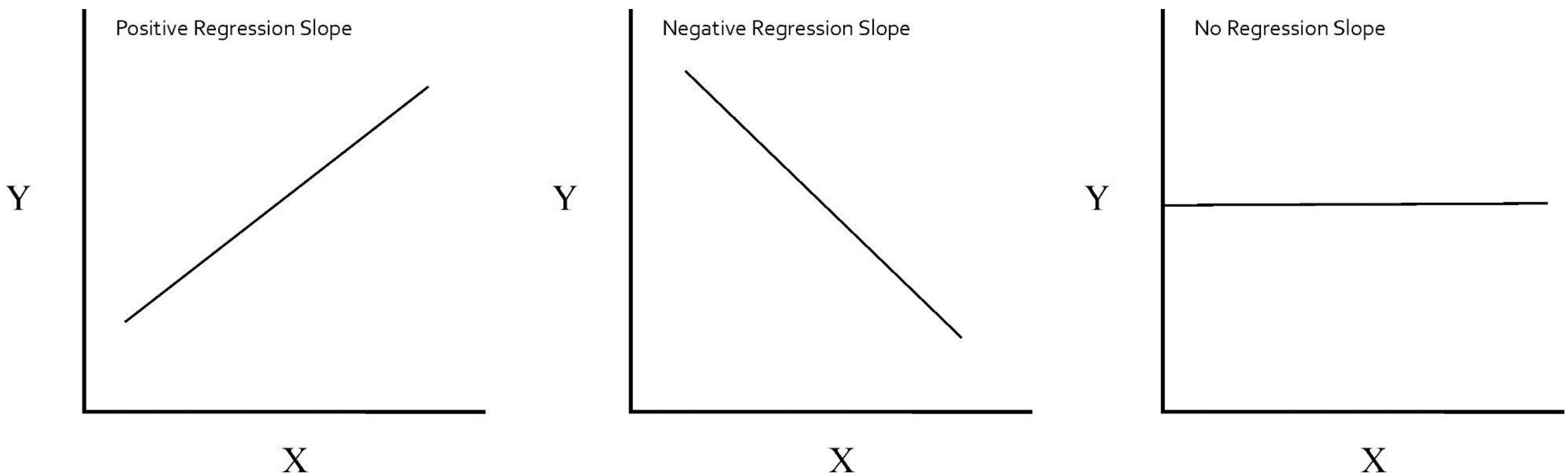
Introduction to Regression

Regression – a means of predicting a dependent variable based one or more independent variables.

-This is done by fitting a line or surface to the data points that minimizes the total error.

- The line or surface is called the regression model or equation.

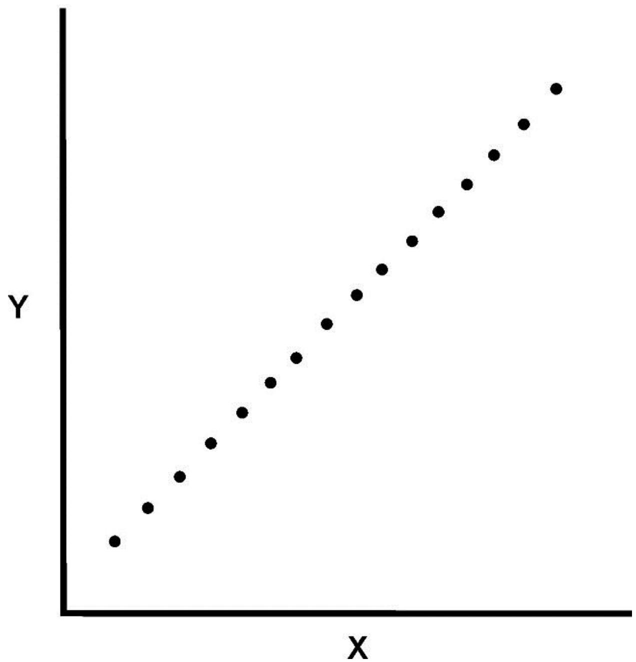
- In this first section we will only work with simple linear bivariate regression (lines).


Positive Regression Slope
Y
X
Negative Regression Slope
Y
X
No Regression Slope
Y
X

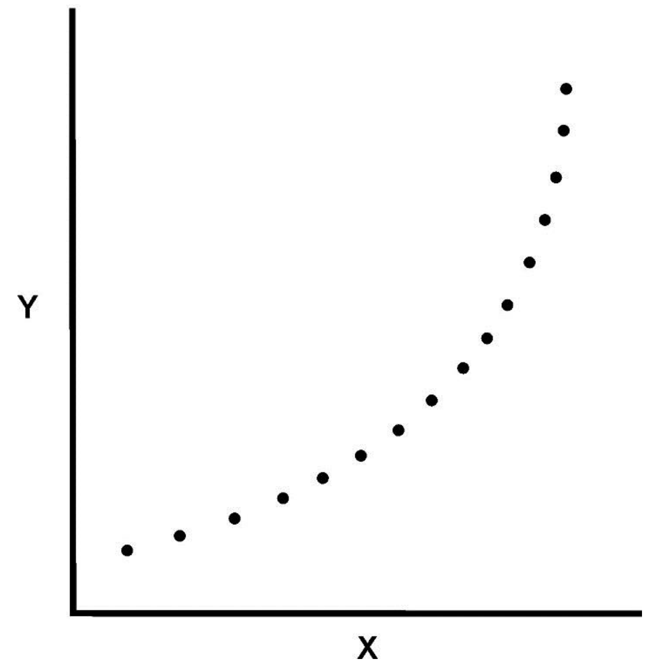
Homoscedasticity – equal variances.

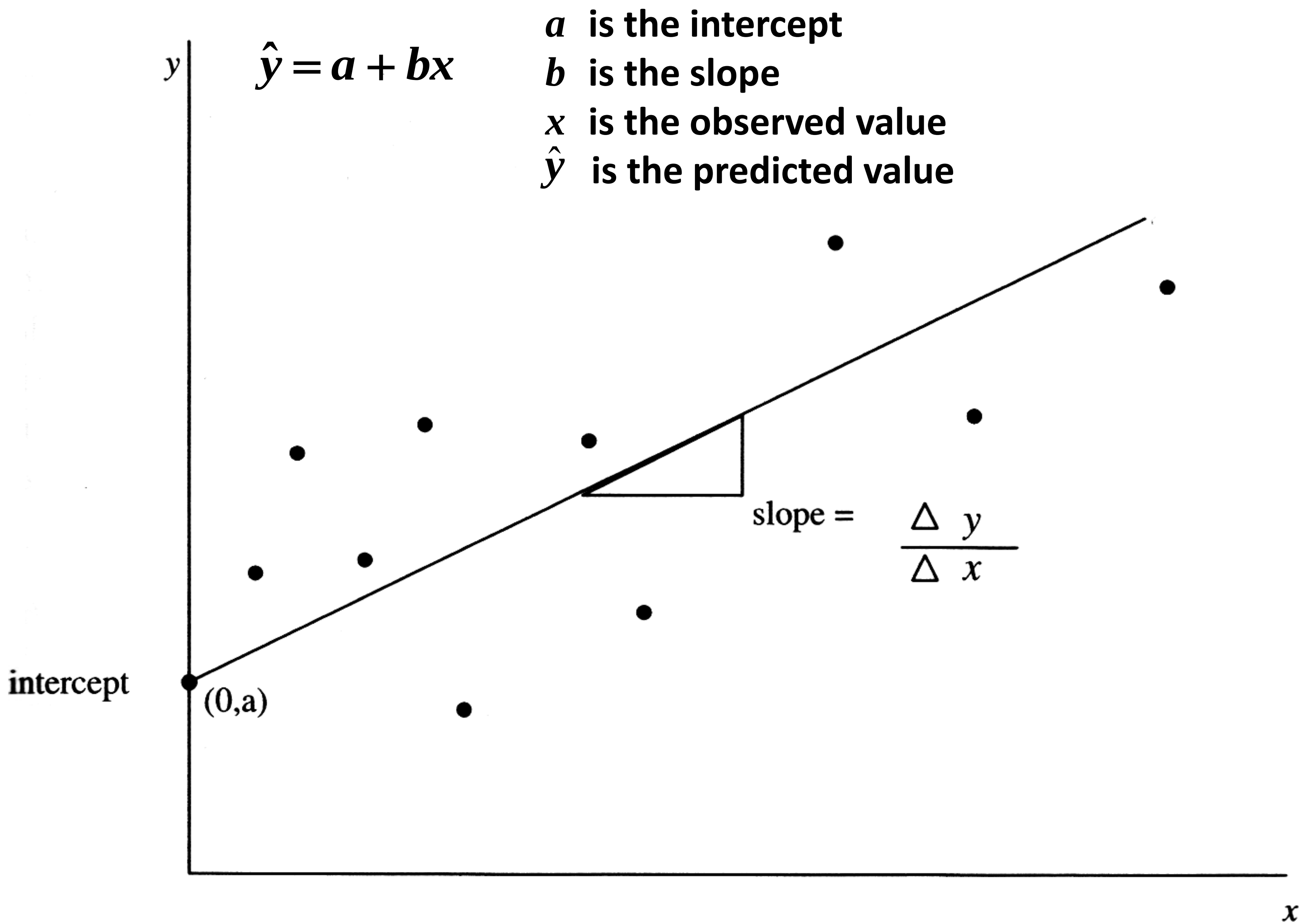
Heteroscedasticity – unequal variances. The value of one variable increases at an increasing or decreases at a decreasing (non-linear) rate. These data may be transformed to meet the linearity assumption.

Homoscedastic Data



Heteroscedastic Data




$\hat{y} = a + bx$
a is the intercept
b is the slope
x is the observed value
$\hat{y}$ is the predicted value
y
x
slope = $\frac{\Delta y}{\Delta x}$
intercept
(0,a)

More formally:

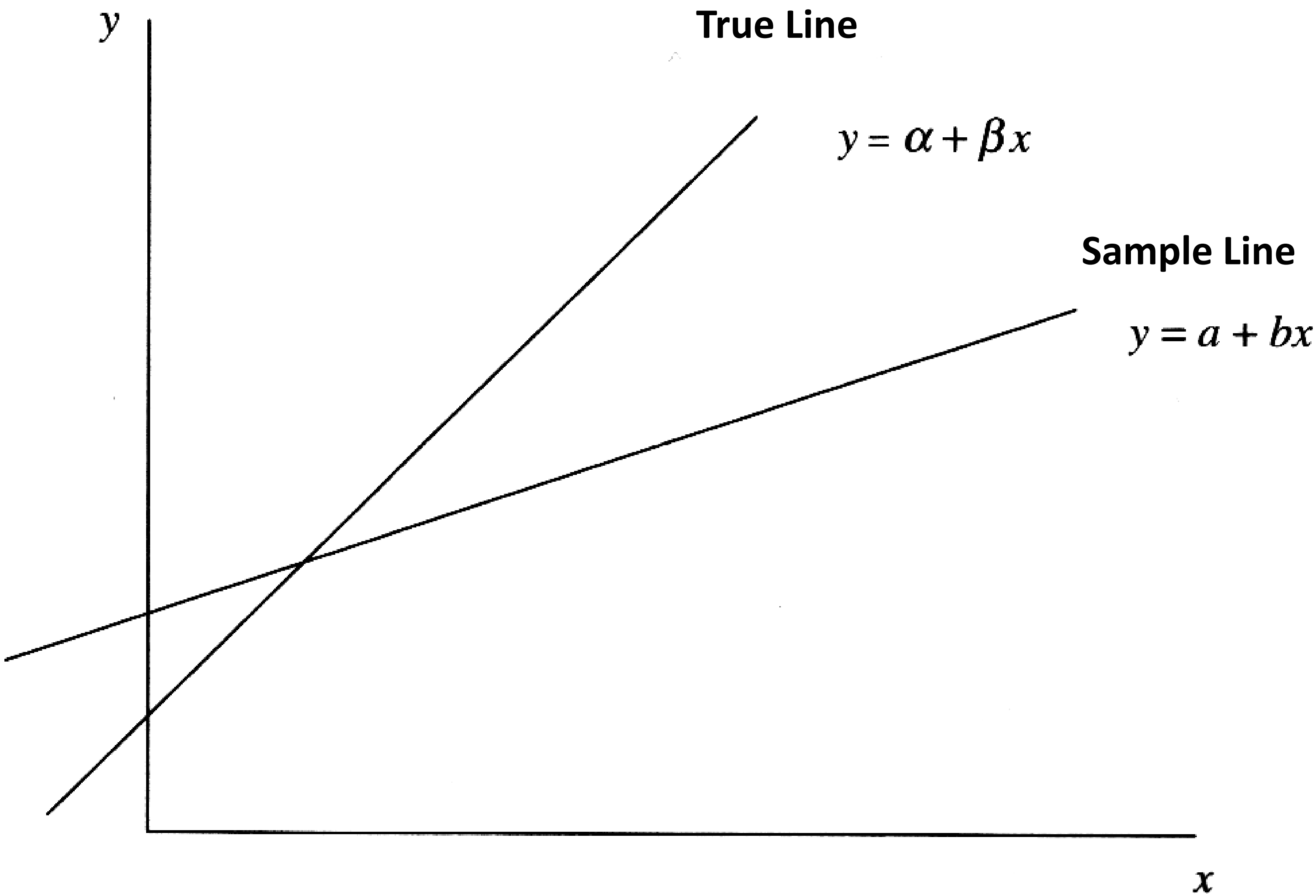
In simple bivariate linear regression there are the following parameters:

$\hat{y}$: the predicted value.

a : the y-axis intercept (also called the *constant*).

b: the slope of the line (the change in the dependent or y variable for each unit change in the independent or x variable).

x : the observed value of the independent variable.



True Line
$y = \alpha + \beta x$
Sample Line
$y = a + bx$
y
x

Each observation of the dependent (y) variable may be expressed as the predicted value + a residual (error).

$$y = a + bx + e = \hat{y} + e$$

where y is the actual value, $\hat{y}$ is the predicted value, and e is the residual (error).

Residual – the difference between the true value and the predicted value.

$$e = y - \hat{y}$$

Regression Residuals:

Unless the r^2 is 100%, there will be some amount of variation in *y* which remains unexplained by *x*.

The unexplained variation is the error component of the regression equation.

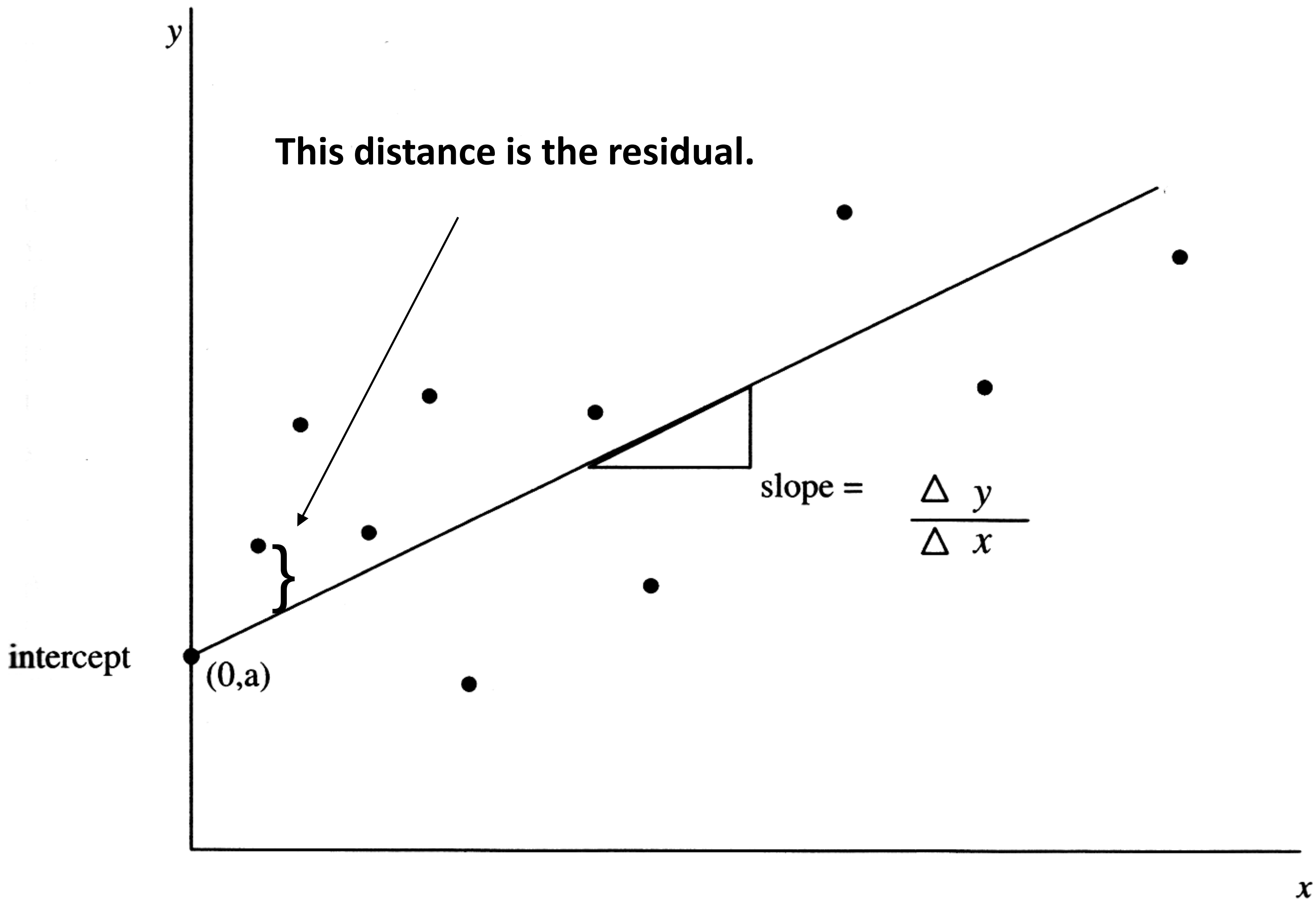
That error is the sum of the differences between each observed value and its value as predicted by the regression equation.

Key Points:

- We are trying to fit a line through a set of plotted points that minimizes the residuals (errors).

- This line is called the *line of best fit*.

- We fit this line in such a way that the *sum of the squared residuals* is minimized.



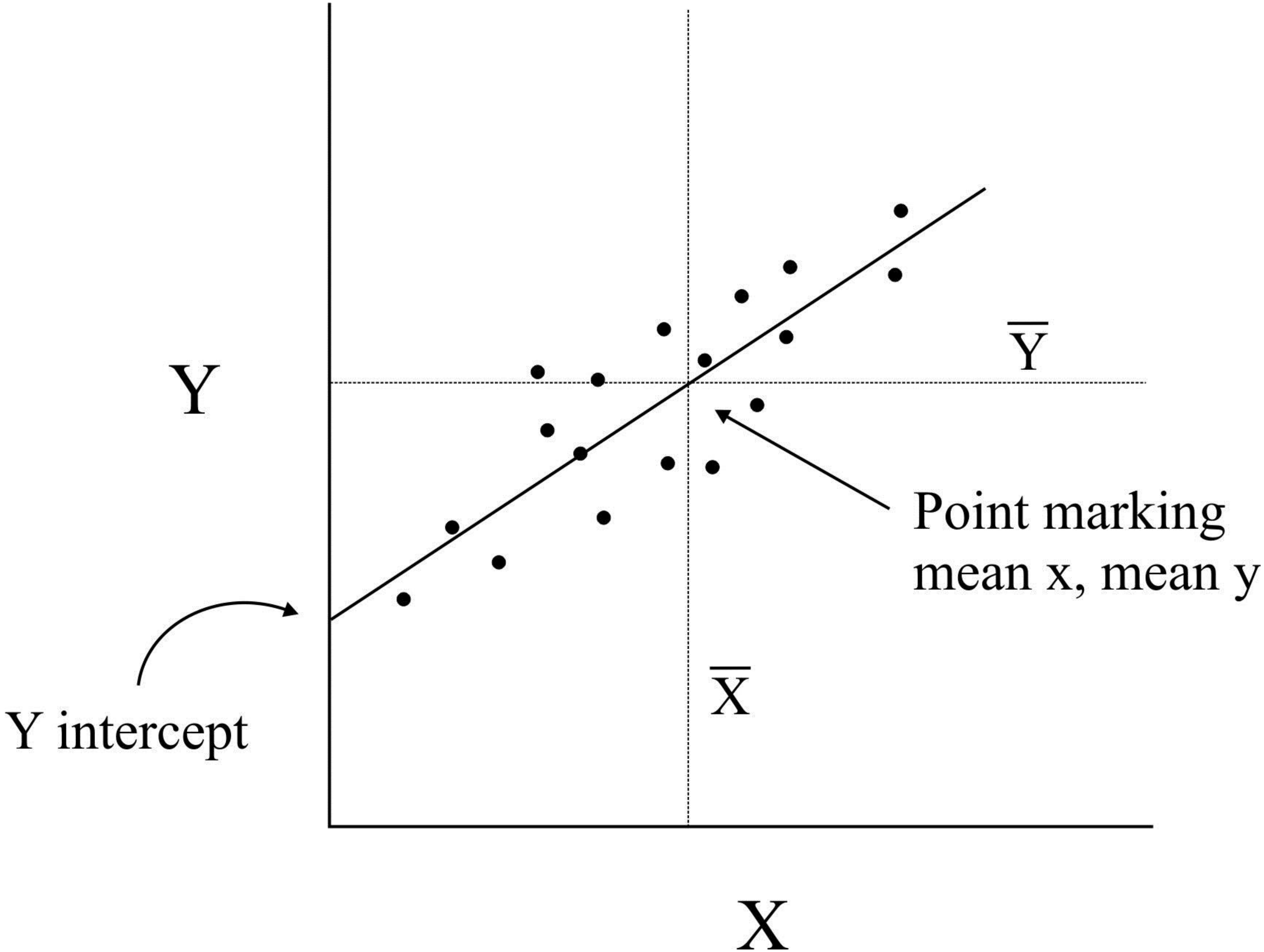
y
This distance is the residual.
slope = $\frac{\Delta y}{\Delta x}$
intercept
(0,a)
x

The way we determine which line (there are an infinite number of potential lines) is the *best fit* is easy…

- We need to define a line that passes through the point determined by the *mean x* value and the *mean y* value.

- The slope of this line needs to minimize the residual error.

Notice that not all of the points fall on the line.

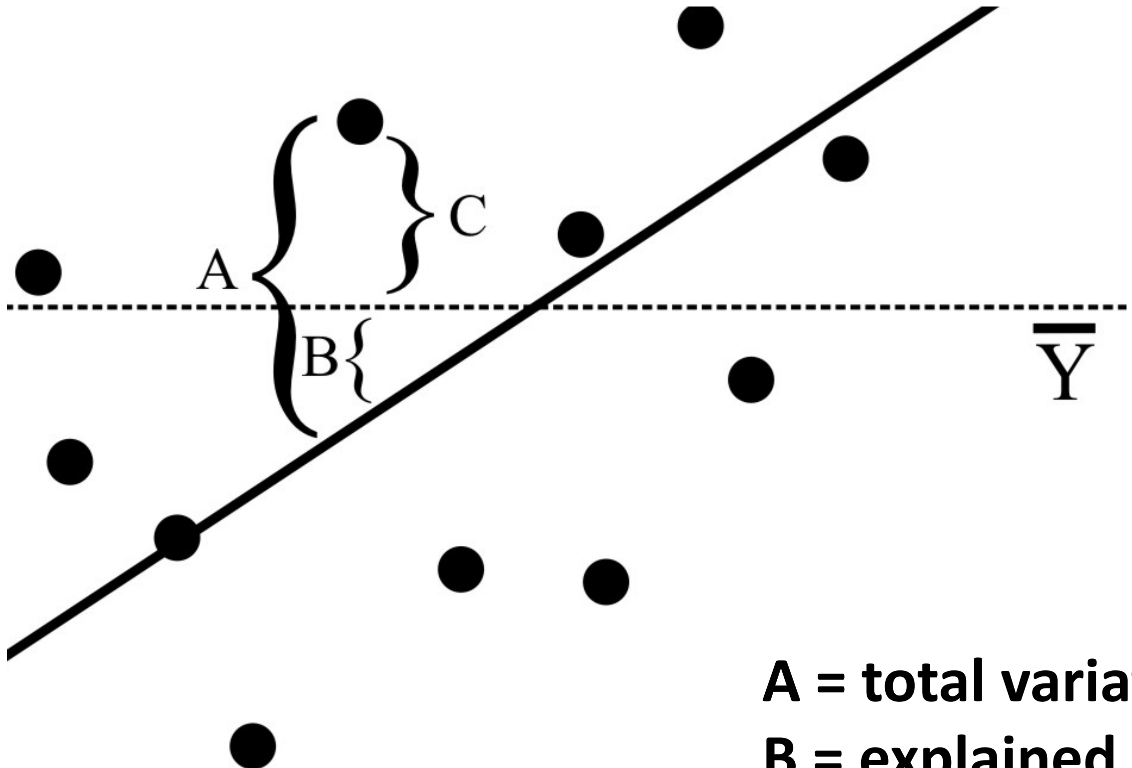


Explained and Unexplained Variation

- The variation in the dependent (*y*) variable can be "partitioned".

- This is similar to the TSS, BSS, and WSS terms in AOV.

  - Total variation in the dependent (*y*) variable.
  - Variation in the dependent (*y*) variable explained by the independent (*x*) variable.
  - The variation in the dependent (*y*) variable NOT explained by the independent (*x*) variable (residual).



A = total variation in y
B = explained variation in y
C = residual

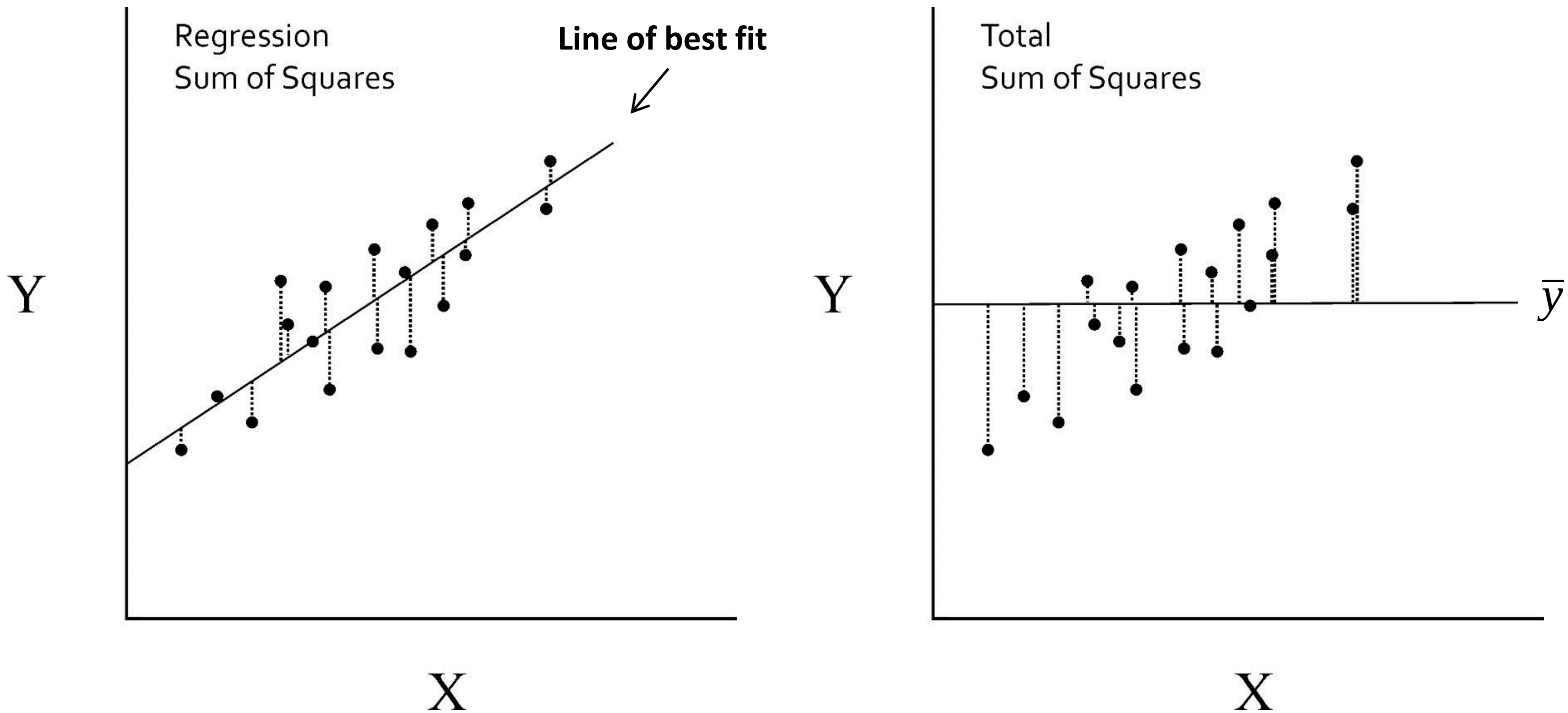
In the form of an equation:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2$$

Total sum of squares

Regression (explained) sum of squares

Residual (unexplained) sum of squares


Regression
Sum of Squares
Line of best fit
Y
X
Total
Sum of Squares
Y
$\bar{y}$
X

The proportion of the total explained variation in y is called the *coefficient of determination or* r^2:

Regression (explained) sum of squares

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Total sum of squares

Key Points

The coefficient of determination (r^2) is equal to the square of the correlation coefficient.

The r^2 is equal to the explained sum of squares divided by the total sum of squares.

r^2 is also equal to 1 minus the ratio of the residual sum of squares divided by the total sum of squares.

What are the units of r^2? What is the range of r^2?

Assumptions of Regression

1. The relationship between y (dependent) and x (independent) is linear.

2. The errors (residuals) do not vary with x.

3. The residuals are independent, meaning that the value of one residual does not influence the value of another.

4. The residuals are normally distributed.

Machine calculation of *a* and *b* (intercept and slope)

$$b = \frac{\sum xy}{\sum x^2}$$

where

$$\sum xy = \sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}$$

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

$$a = \overline{Y} - b\overline{X}$$

Machine calculation of r^2 (coefficient of determination)

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

$$RSS = \frac{\left(\sum xy\right)^2}{\sum x^2}$$

$$ESS = TSS - RSS$$

$$r^2 = \frac{RSS}{TSS}$$

Significance Testing in Regression

There are several hypotheses that are tested in regression:

1. That the variation explained by the model is not due to chance (F test).

2. That the slope of the regression line is significantly different than zero (t test of the β parameter).

3. That the y intercept is significantly different than zero (t test of the constant parameter).
    - This test result can be ignored unless there is some reason to believe that the y intercept should be zero.

We can test the significance of model using the F-statistic in the following form:

$$F = \frac{RSS/(v-1)}{ESS/(n-2)} \qquad df = v-1, n-2$$

where v = the number of parameters and n is the sample size. Since in bivariate linear regression we are estimating only 2 parameters (a and b) v will always be 2.

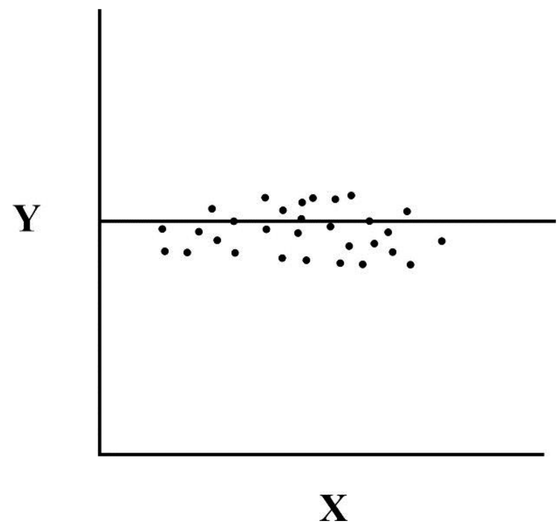
We can test the null hypothesis that β = 0 using the t test in the following form:

$$t = \frac{b}{s_b}$$

…where s_b is the standard deviation of the slope..

$$s_b = \sqrt{\frac{ESS/(n-2)}{\sum x^2}}$$

$$df = n - 2$$



When β=0, for each unit change in x there is no change in y.

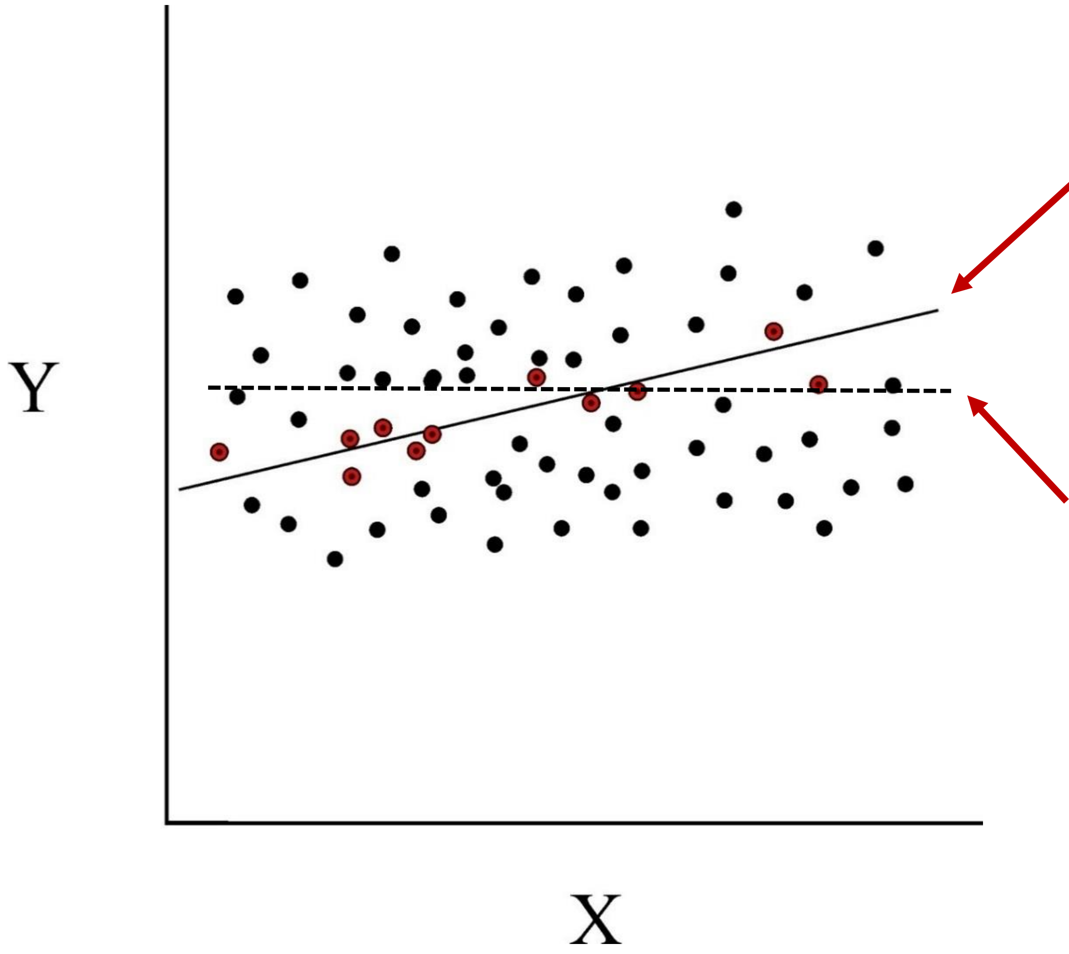
Key points regarding the F and t statistics:

- The F statistic tells you whether the association is due to chance.
- The t statistic tells you whether a specific independent variable is contributing to the model.

A few points concerning regression analysis:

- Be sure to specify the correct dependent variable since the procedure makes an assumption concerning causality.
- Data may be transformed to meet the linearity requirement.
- Do not predict y values beyond the data range.
- If there is no causality, correlation is the correct method.
  - For example, human leg and arm lengths are linearly associated, but one does not *cause* the other.

Sampling for regression should span the entire range of values whenever possible. Here the red dots mark a sample from a much larger hypothetical population.



If only the red observations are used there appears to be a causal association.

If the entire range of data are used the causal association disappears.... Y does not change as X increases.

Residual plot examination:

Normally distributed residuals appear scattered randomly about the mean residual line.

Heteroscedastic residuals fan out from the residual mean line.

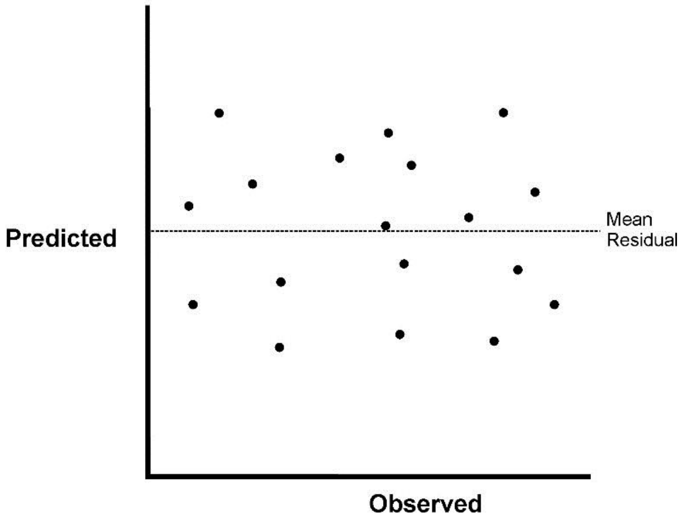
If an important explanatory variable is missing the predicted values increase as the observed values increase.

Non-linear association between the variables appears as an arc running through the mean residual line.

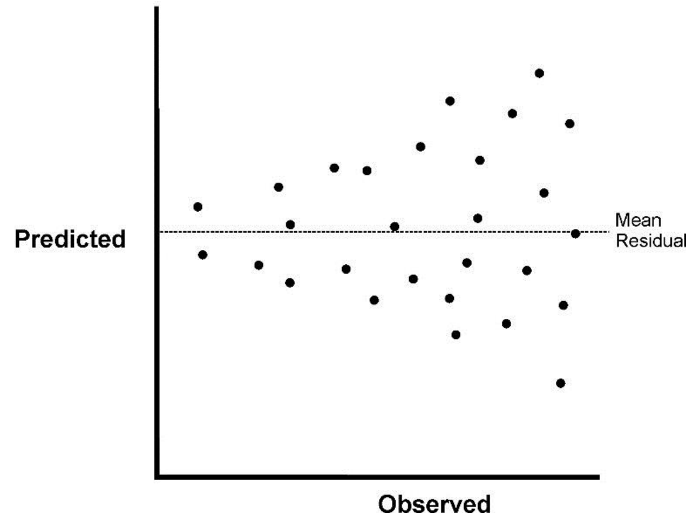
The last three of the above (heteroscedasticity, missing variable, non-linear relastionship) point to data problems.

Residual Plots

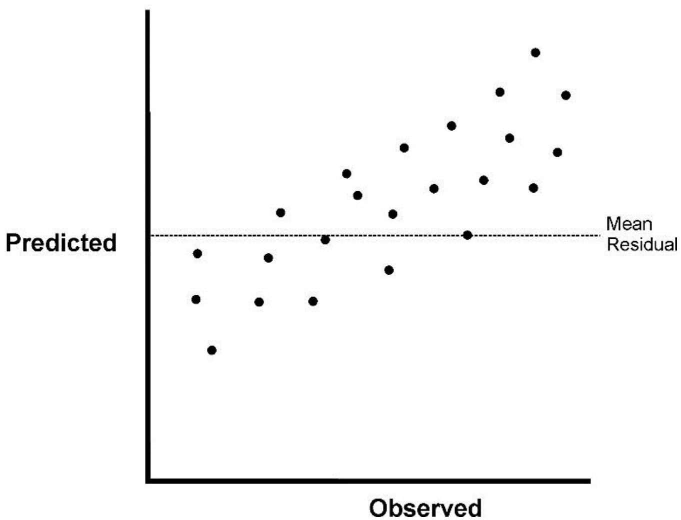
Normally distributed residuals.



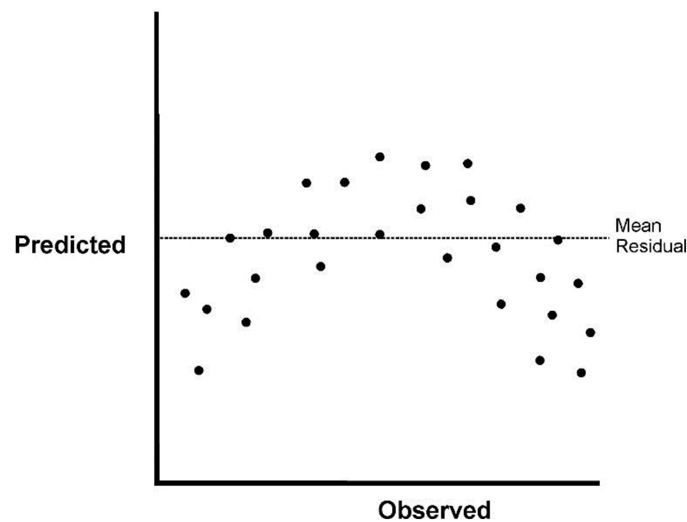
Heteroscedastic data.



Missing explanatory variable.



Original data have a non-linear relationship.



In-Class Example
Isle of Mann TT

Fatalities by Slope per Location

$$a = \bar{Y} - b\bar{X}$$

$$b = \frac{\sum xy}{\sum x^2}$$

$$\sum xy = \sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}$$

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

Location	*Fatalities*	*Slope (Deg)*
Alpine Cottage	10	0.652
Waterworks Corner	2	22.1
Quarry Bends	6	7.9
Greeba Castle	4	7.7
Mountain Box	5	13.6
Ballaugh Bridge	7	1.7
Glentramman	8	4.9
Stonebreaker's Hut	4	13.3
Appledene	5	6.0
Handley's Corner	3	11.9
Glen Helen	8	2.1
Vernadah	4	18.1

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

$$RSS = \frac{\left(\sum xy\right)^2}{\sum x^2}$$

$$ESS = TSS - RSS$$

$$r^2 = \frac{RSS}{TSS}$$

$$F = \frac{RSS/(v-1)}{ESS/(n-2)} \quad df = v-1, n-2$$

$$t = \frac{b}{s_b} \qquad s_b = \sqrt{\frac{ESS/(n-2)}{\sum x^2}}$$

$$df = n-2$$