Linear Regression

- Sonal Ghanshani

Introduction to Regression Analysis

- It is a technique used to:
- Estimate a relationship between variables
- Predict the value of one variable (dependent variable) on the basis of other variables (independent variables)
- **Example**:
  - $Y = \beta_0 + \beta_1 x + \varepsilon$
  - Here, Y is a dependent variable, whereas β_0, $\beta_{1,}$ x, and ε are independent variables.

Why Linear Regression?

Why Linear Regression?

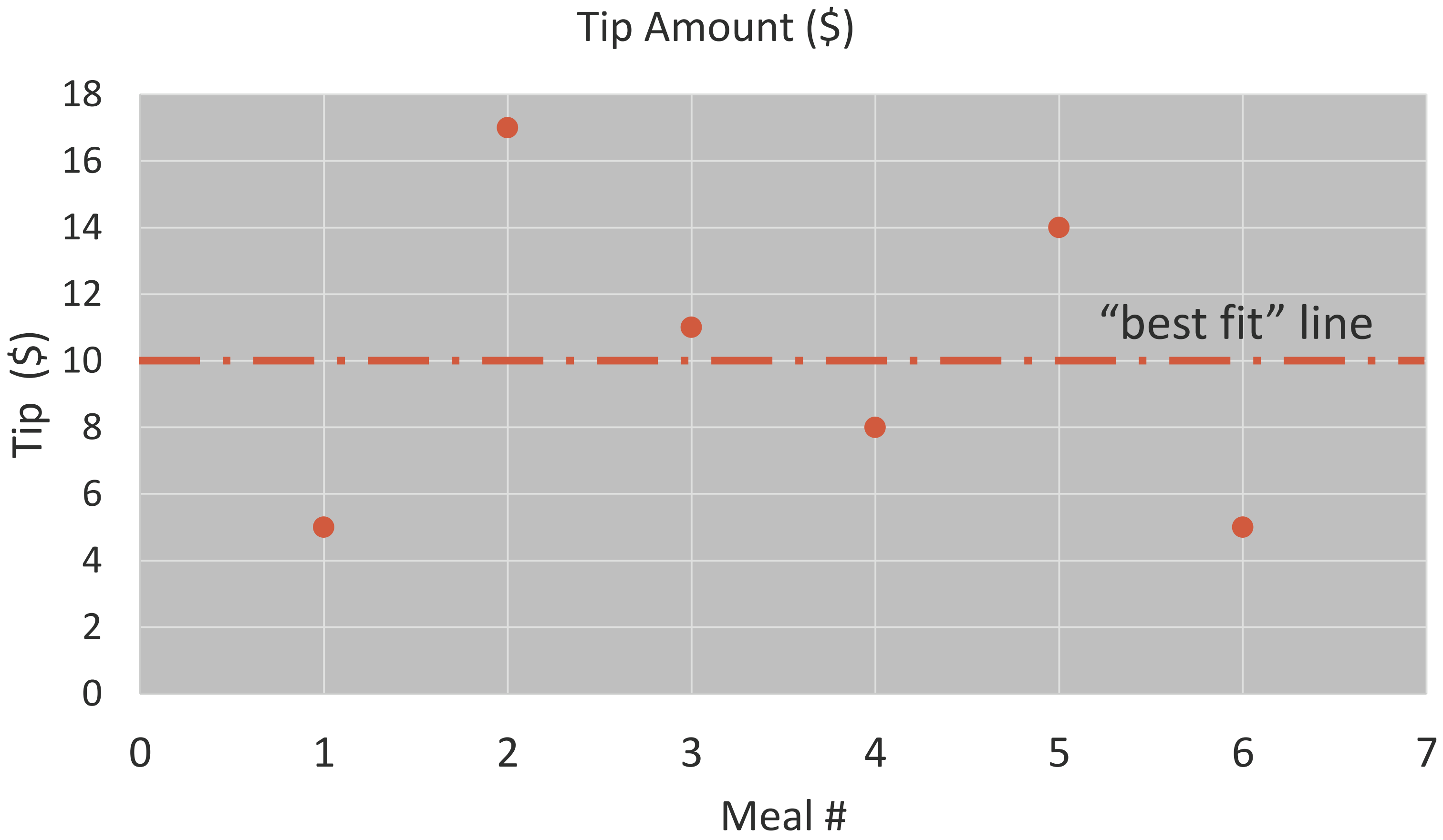
- Consider

- Let us assume there is a student at a university and he works in a small restaurant after the classes. “Tips” are very important part of the waiter’s pay. Most of the time dollar amount of the tip is related to the dollar amount of the total bill.

- As, a data science student, he would like to develop a model that will allow him to make a prediction about what amount of tip to expect for any bill amount. Therefore, one evening he collects the data for six meals.

Let’s see…

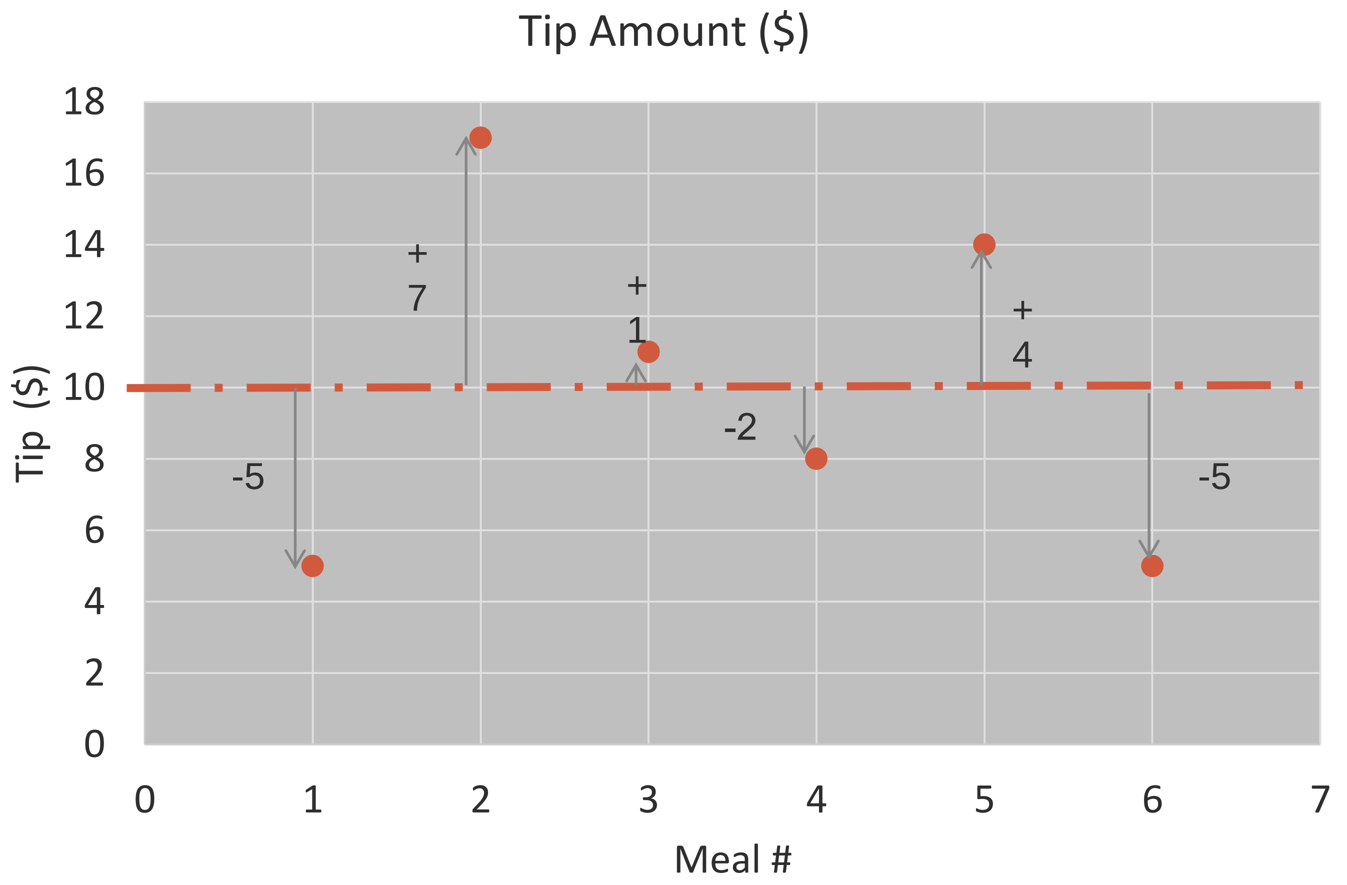
Meal #	Tip Amount
1	5
2	17
3	11
4	8
5	14
6	5



With only one variable, and no other information, the best prediction for the next tip amount is mean of the sample itself.

Let’s see…

Meal #	Error	(Error)^2
1	-5	25
2	7	49
3	1	1
4	-2	4
5	4	16
6	-5	25



Sum of Errors = 0

Sum of Square of Errors = 120

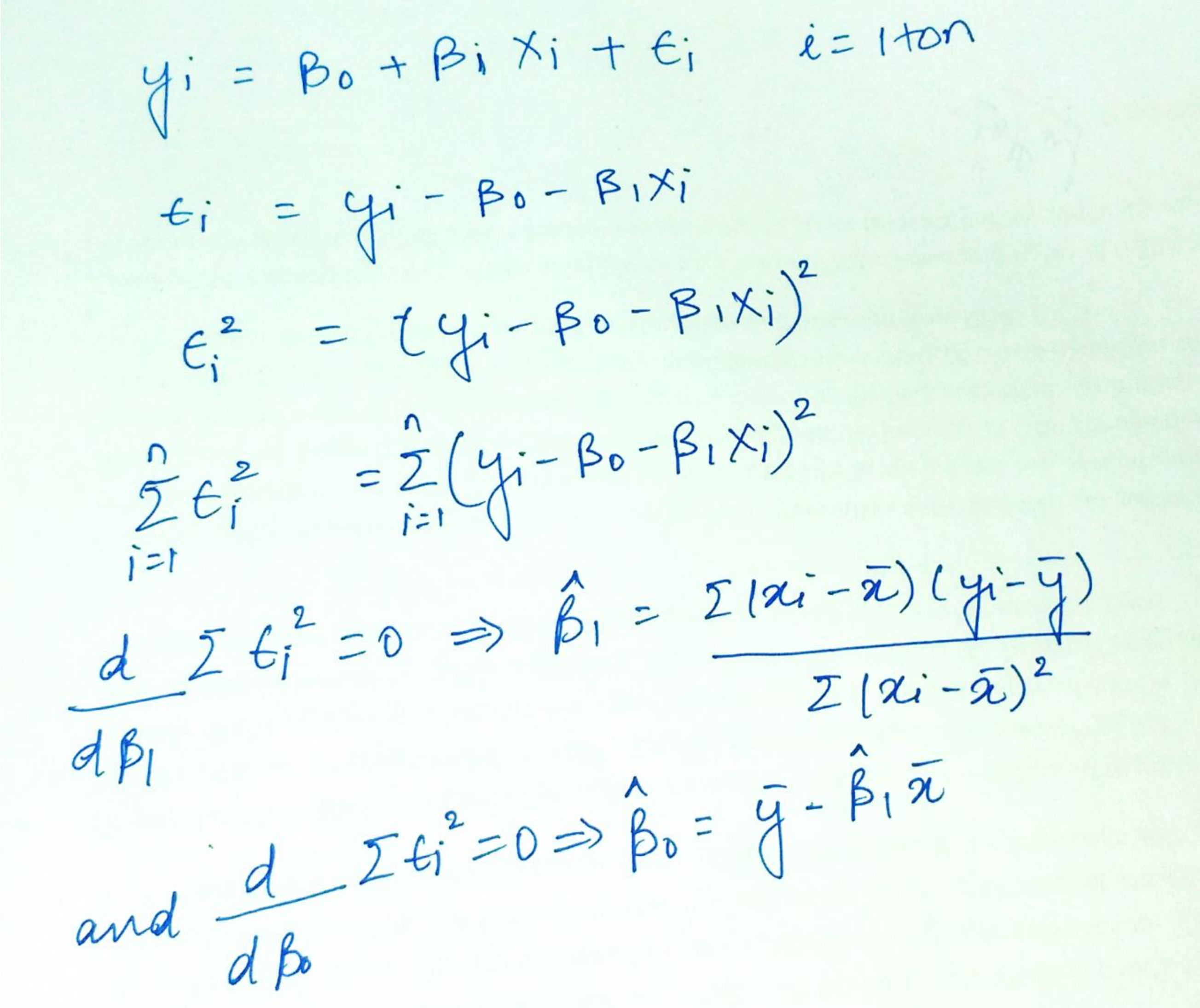
Why square the residuals?
To make them positive and emphasize larger deviations

Why?

- Linear regression is applied when there exists a linear relationship between the dependent and independent variable. (also known as response and features)
- The goal of the simple linear regression is to find the best fitting line through the data that minimizes the SSE.

Ordinary Least Squares

Ordinary Least Squares



$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1 \text{ to } n$

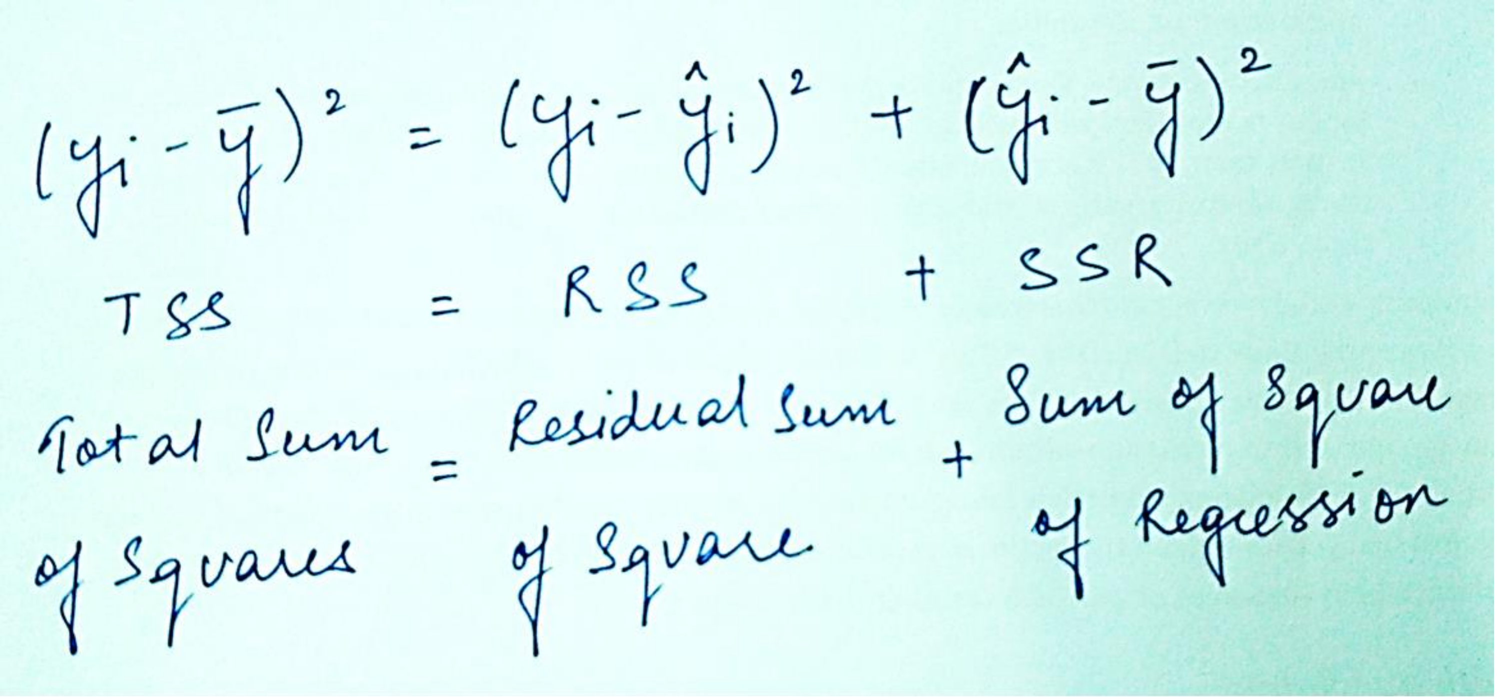
$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$

$\epsilon_i^2 = (y_i - \beta_0 - \beta_1 x_i)^2$

$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

$\frac{d}{d\beta_1} \sum \epsilon_i^2 = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

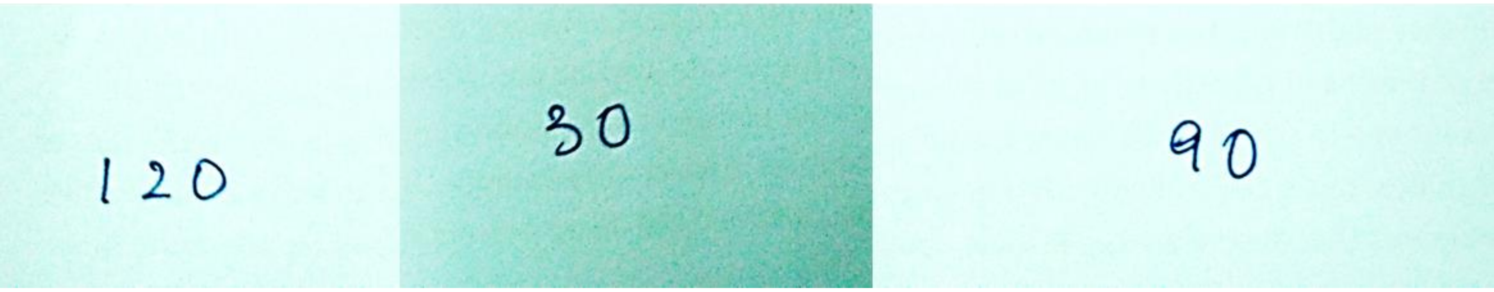
and $\frac{d}{d\beta_0} \sum \epsilon_i^2 = 0 \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$



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

TSS = RSS + SSR

Total Sum of Squares = Residual Sum of Square + Sum of Square of Regression



120 30 90

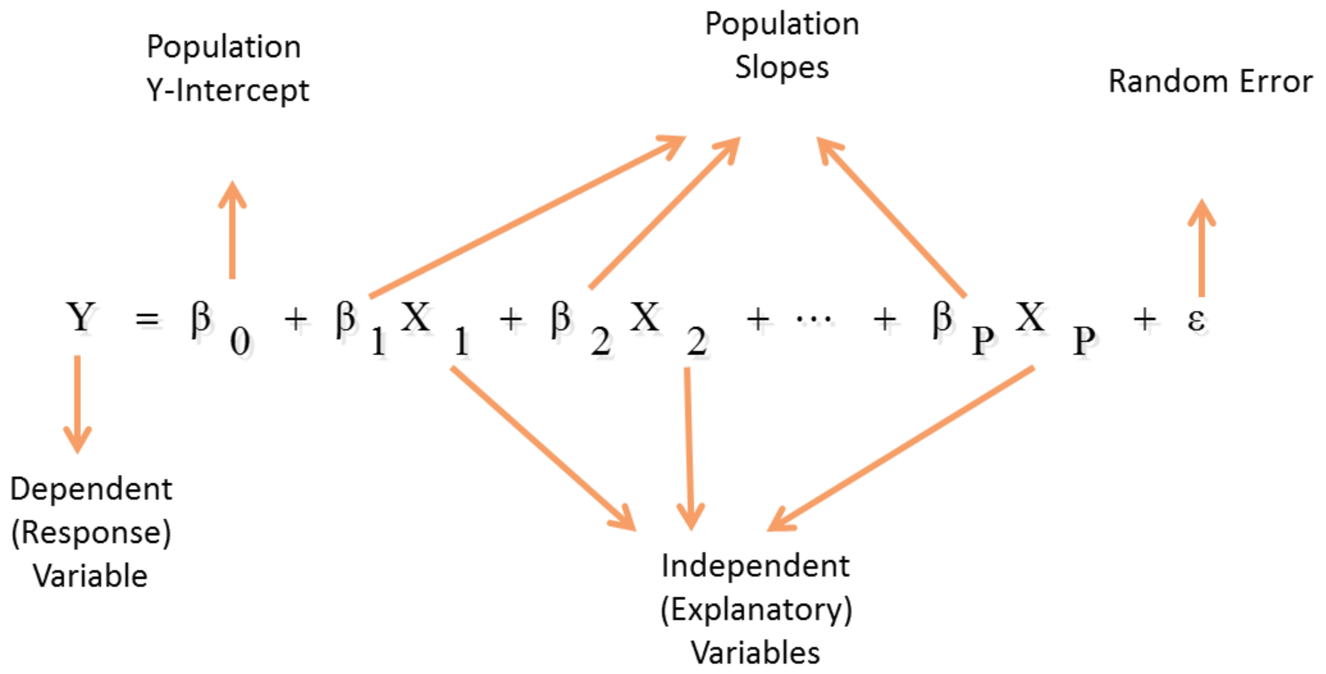
Simple Regression Analysis

Simple Regression Analysis

- In this analysis:
  - A relationship between a dependent variable and an independent variable is modeled
  - The output is a function to predict the dependent variable on the basis of the values of independent variables
  - A straight line is fit to the data

Simple Linear Regression Model

- It depicts the relationship between one dependent and two or more independent variables. An example and its components are explained below:



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \varepsilon$$

Population Y-Intercept

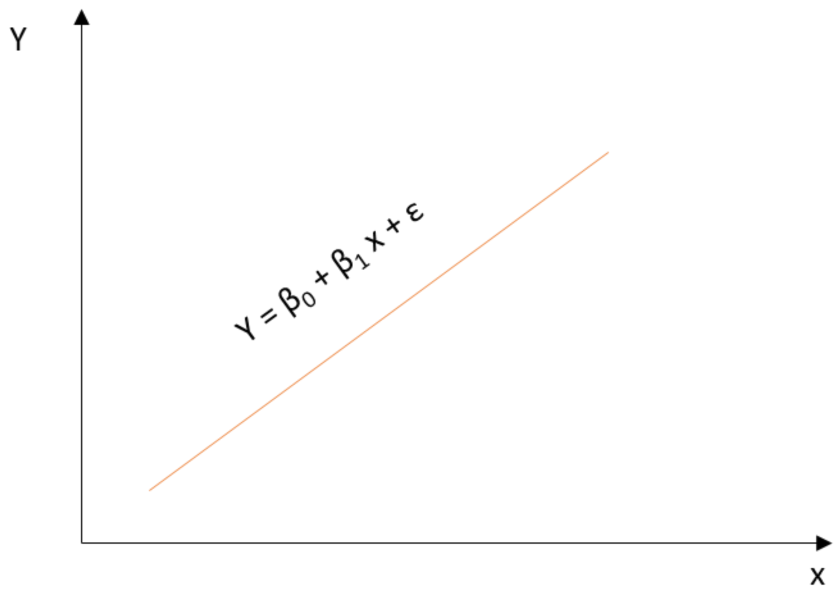
Population Slopes

Random Error

Dependent (Response) Variable

Independent (Explanatory) Variables

Example of Linear Regression Model Relationships



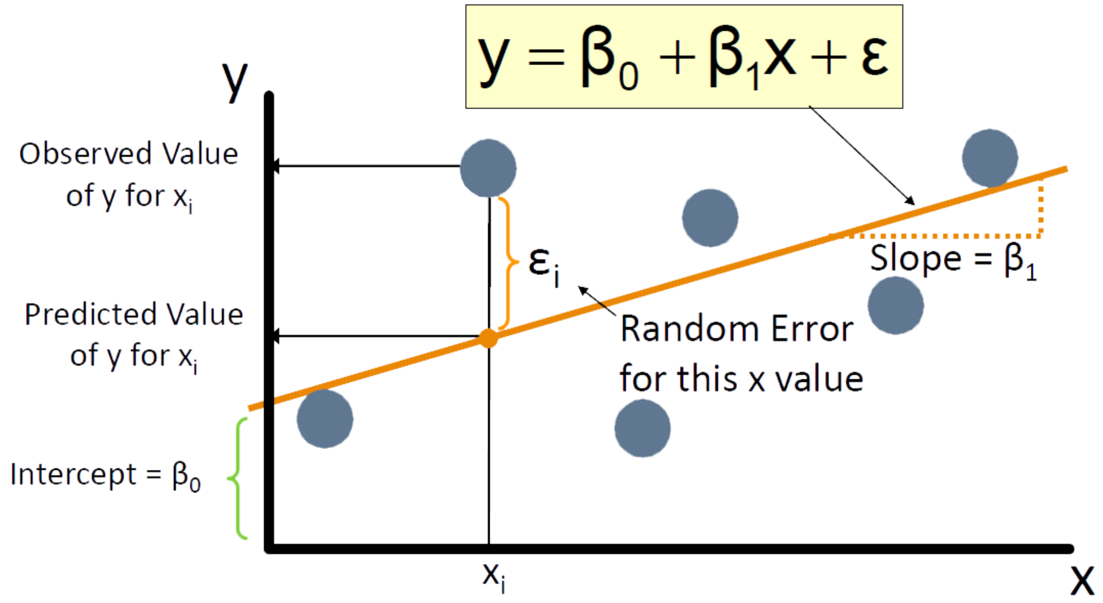
Simple Linear Regression Model Explained

- A more descriptive graphical representation of simple linear regression is given below:

- β1 represents the slope. A slope of two variables implies that each one-unit change in x results in a two-unit change in y.

- β1 represents the estimated change in the average value of y as a result of a one-unit change in x.

- β0 represents the estimated average value of y when the value of x is zero.


$y = \beta_0 + \beta_1 x + \varepsilon$
y
Observed Value of y for x_i
ε_i
Slope = β_1
Predicted Value of y for x_i
Random Error for this x value
Intercept = β_0
x_i
x

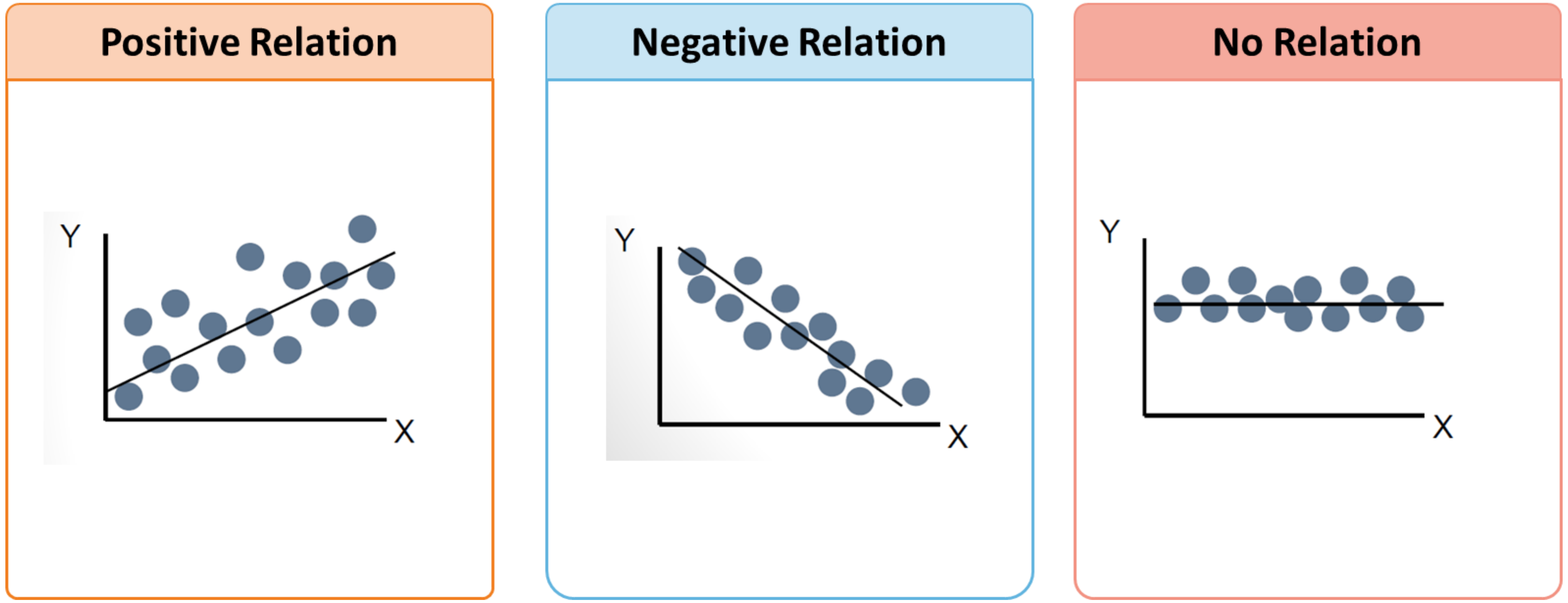
Method of Least Squares Regression Model

- It selects the line with the lowest total sum of squared prediction errors (Sum of Squares of Error, or SSE).
- Mathematically,
  - $SSR = \sum (y - yhat)^2$ (measure of an explained variation)
  - $SSE = \sum (y - y)^2$ (measure of an unexplained variation)
  - $SST = SSR + SSE = \sum (y - \bar{y})2$ (measure of the total variation in y)

Correlation - How to determine relationship between variables?

Correlation

- X and Y can exist in three different types of relations



Positive Relation

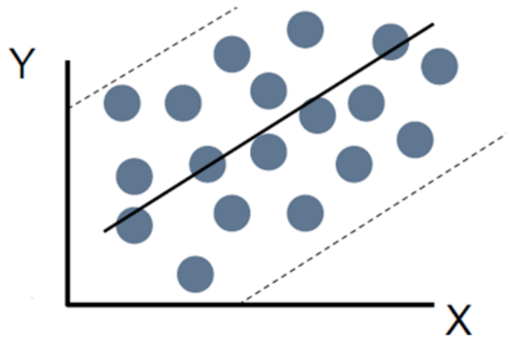
Negative Relation

No Relation

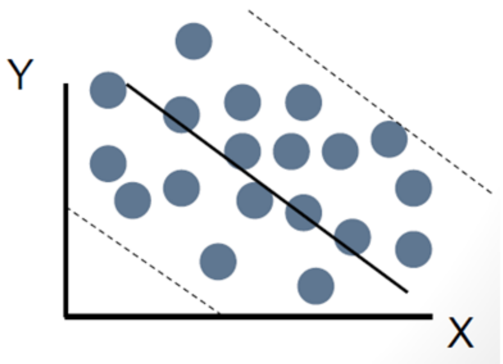
Correlation

- They can also exist in a weak relation

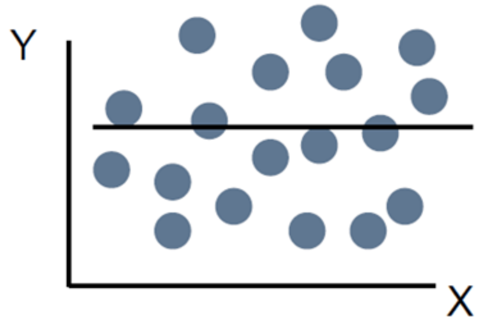
Weak Relation Type I



Weak Relation Type II



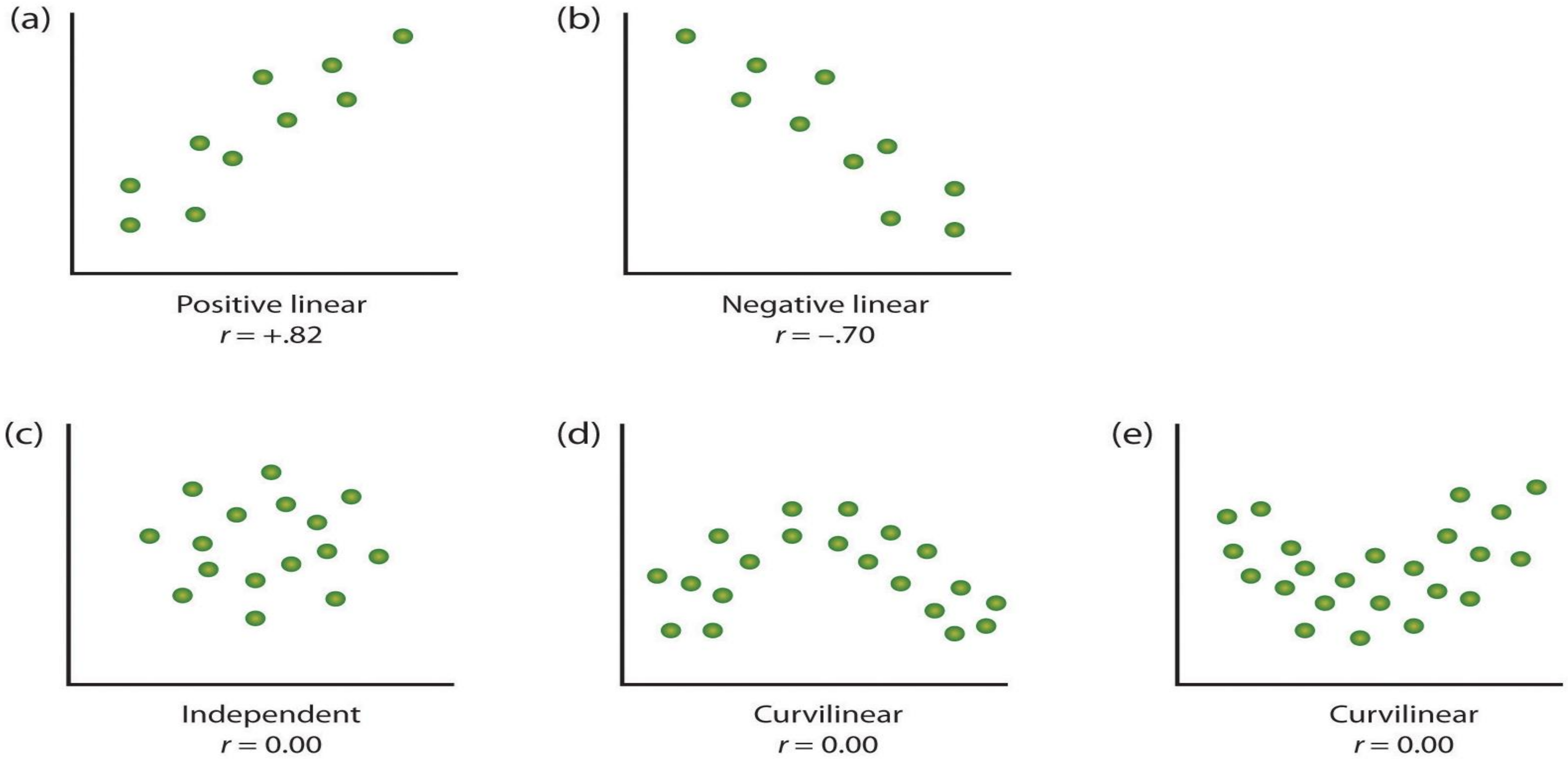
Weak Relation Type III



Correlation

- Correlation is a statistical technique that predicts whether and how strongly pairs of variables are related.
  - The main result of a correlation is called the correlation coefficient (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.
  - If r is close to 0, it means there is no relationship between the variables
  - If r is positive, it means that as one variable gets larger the other gets larger
  - If r is negative, it means that as one gets larger the other gets smaller (often called an "inverse" correlation)

Correlation



Positive linear
r = +.82

Negative linear
r = −.70

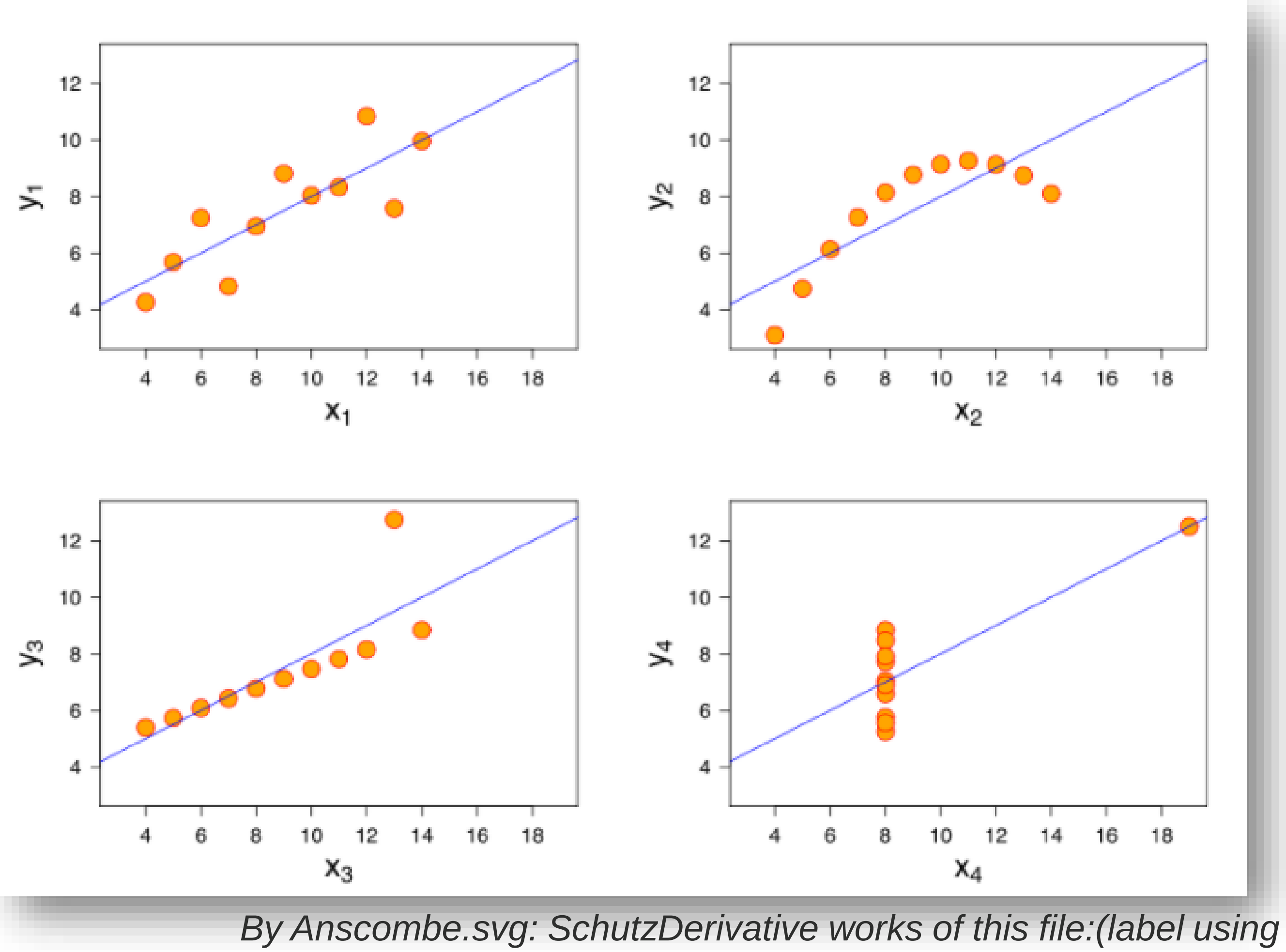
Independent
r = 0.00

Curvilinear
r = 0.00

Curvilinear
r = 0.00

Corelation

- Four sets of data with the same correlation of 0.816



By Anscombe.svg: SchutzDerivative works of this file:(label using subscripts): Avenue - Anscombe.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=9838454

Model Validation

Coefficient of Multiple Determination Regression Model

It:

- Determines the relation between X and Y
- Is often referred by R
- R2 = SSR/SST = (SSR/(SSR+SSE))
- 0 < R2 <1
- The higher the value, the more accurate the regression model

Mean Squared Error

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y_i} - Y_i)^2$$

Application

Predictive Analytics

- Predictive analytics i.e. forecasting future opportunities and risks is the most prominent application of regression analysis in business.
- Demand analysis, for instance, predicts the number of items which a consumer will probably purchase. However, demand is not the only dependent variable when it comes to business. Regression analysis can go far beyond forecasting impact on direct revenue.
- For example, we can forecast the number of shoppers who will pass in front of a particular billboard and use that data to estimate the maximum to bid for an advertisement.
- Insurance companies heavily rely on regression analysis to estimate the credit standing of policyholders and a possible number of claims in a given time period.

Operation Efficiency

- Regression models can also be used to optimize business processes.
- A factory manager, for example, can create a statistical model to understand the impact of oven temperature on the shelf life of the cookies baked in those ovens.
- In a call centre, we can analyse the relationship between wait times of callers and number of complaints.
- Data-driven decision making eliminates guesswork, hypothesis and corporate politics from decision making.
- This improves the business performance by highlighting the areas that have the maximum impact on the operational efficiency and revenues.

Supporting Decisions

- Businesses today are overloaded with data on finances, operations and customer purchases.
- Increasingly, executives are now leaning on data analytics to make informed business decisions thus eliminating the intuition and gut feel.
- Regression analysis can bring a scientific angle to the management of any businesses. By reducing the tremendous amount of raw data into actionable information, regression analysis leads the way to smarter and more accurate decisions.
- This does not mean that regression analysis is an end to managers creative thinking. This technique acts as a perfect tool to test a hypothesis before diving into execution.

Correcting Errors

- Regression is not only great for lending empirical support to management decisions but also for identifying errors in judgment.
- For example, a retail store manager may believe that extending shopping hours will greatly increase sales.
- Regression analysis, however, may indicate that the increase in revenue might not be sufficient to support the rise in operating expenses due to longer working hours (such as additional employee labour charges).
- Hence, regression analysis can provide quantitative support for decisions and prevent mistakes due to manager's intuitions.

New Insights

- Over time businesses have gathered a large volume of unorganized data that has the potential to yield valuable insights.
- However, this data is useless without proper analysis. Regression analysis techniques can find a relationship between different variables by uncovering patterns that were previously unnoticed.
- For example, analysis of data from point of sales systems and purchase accounts may highlight market patterns like increase in demand on certain days of the week or at certain times of the year.
- You can maintain optimal stock and personnel before a spike in demand arises by acknowledging these insights.