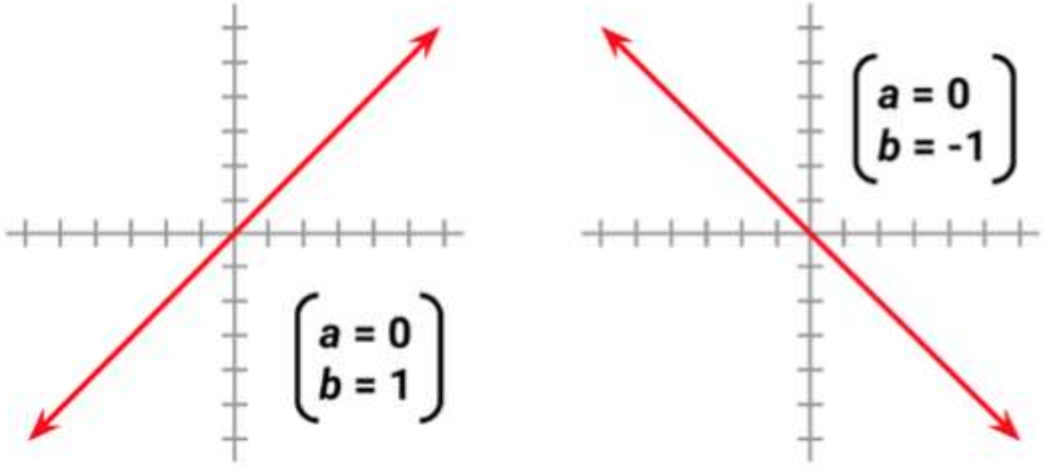
Understanding regression

- Regression is concerned with specifying the relationship between a single numeric dependent variable (the value to be predicted) and one or more numeric independent variables (the predictors).
- As the name implies, the dependent variable depends upon the value of the independent variable or variables.
- The simplest forms of regression assume that the relationship between the independent and dependent variables follows a straight line.

From basic algebra that lines can be defined in a slope-intercept form similar to y = a + bx.

In this form, the letter y indicates the dependent variable and x indicates the independent variable.

The slope term b specifies how much the line rises for each increase in x. Positive values define lines that slope upward while negative values define lines that slope downward.



- The term a is known as the intercept because it specifies the point where the line crosses, or intercepts, the vertical y axis. It indicates the value of y when x = 0.

- When there is only a single independent variable it is known as simple linear regression. In the case of two or more independent variables, this is known as multiple linear regression, or simply "multiple regression".

- Regression can also be used for other types of dependent variables and even for some classification tasks.
- For instance, **logistic regression** is used to model a binary categorical outcome.
- Poisson regression—named after the French mathematician Siméon Poisson—models integer count data. The method known as **multinomial logistic regression** models a categorical outcome; thus, it can be used for classification.

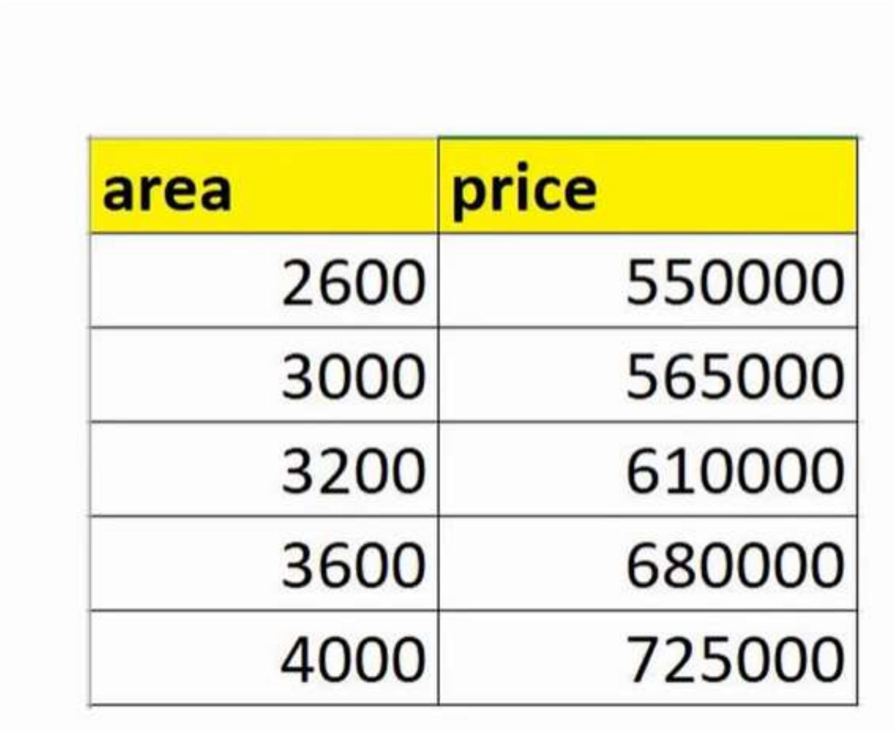
Simple linear Regression

- A simple linear regression model defines the relationship between a dependent variable and a single independent predictor variable using a line defined by an equation in the following form: $y=a+bx$

The intercept, a describes where the line crosses the y axis,

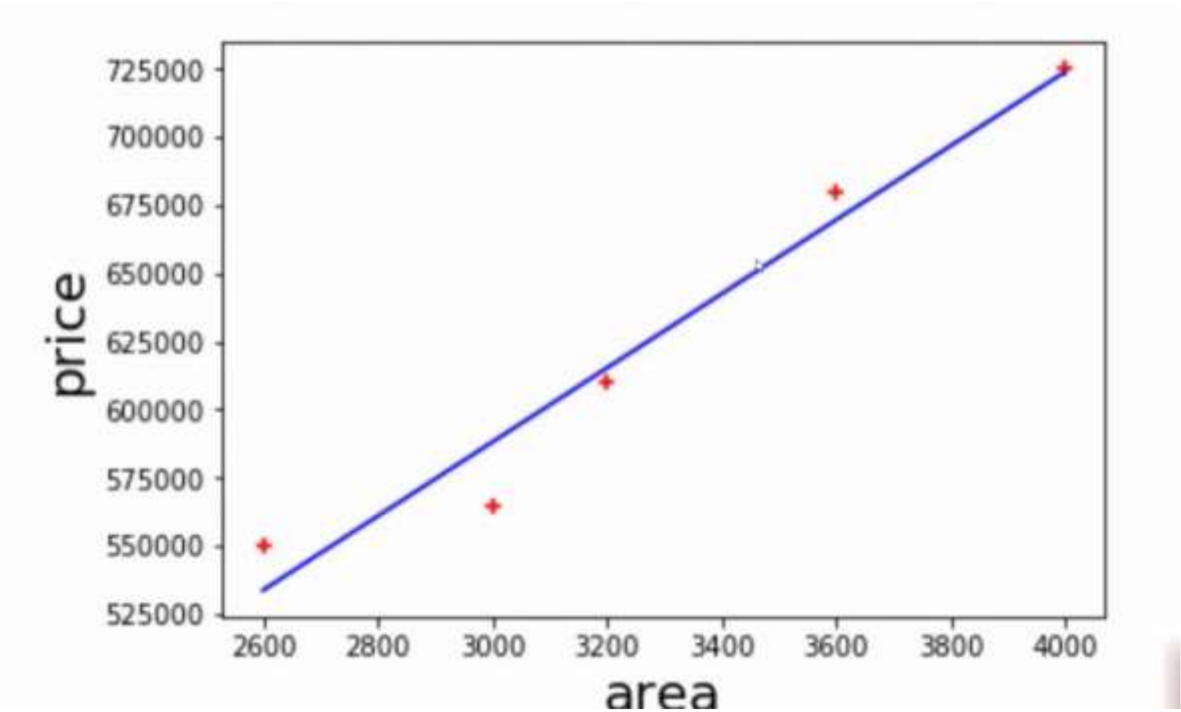
slope, b describes the change in y given an increase of x

In the given example ‘area’ is the independent variable and price is the ‘dependent’ variable.

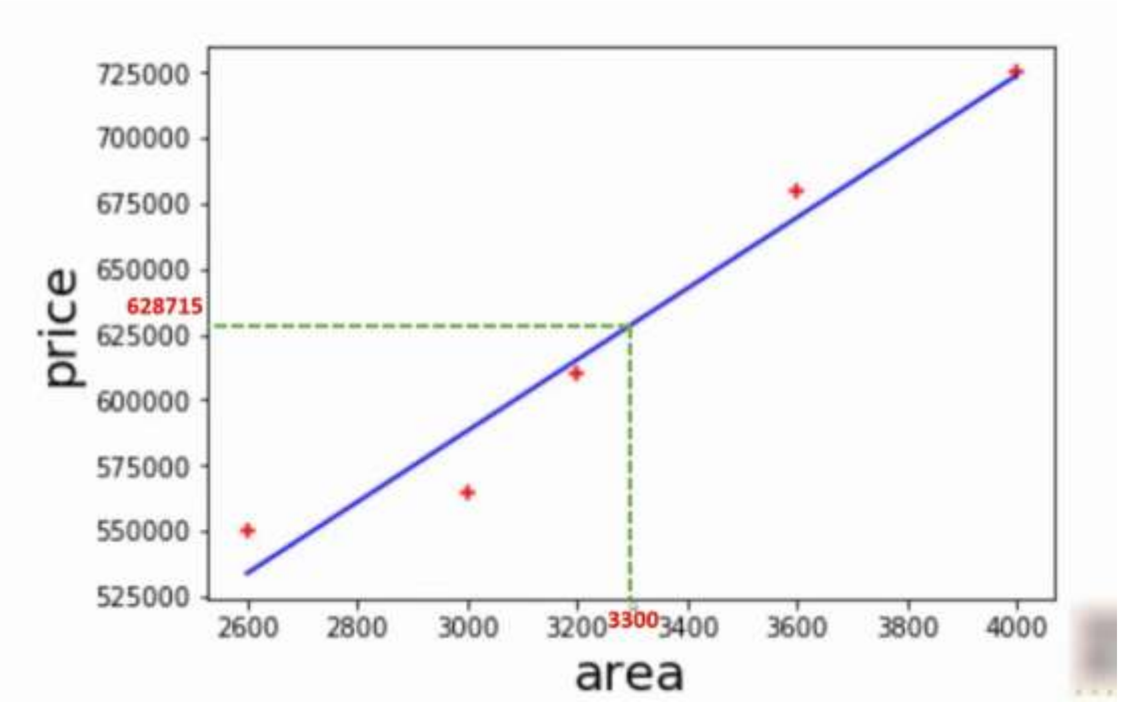


area	price
2600	550000
3000	565000
3200	610000
3600	680000
4000	725000

- The regression line corresponds to the dataset



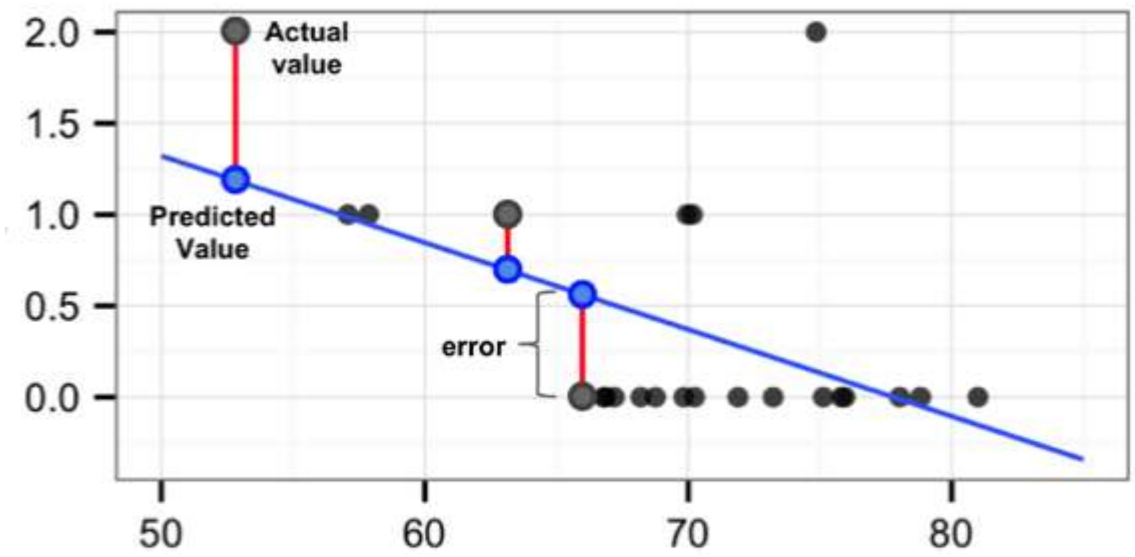
- To find the ‘price’ of new ‘area’ it can be find out from the regression line



Ordinary least squares estimation

- In order to determine the optimal estimates of a and b, an estimation method known as Ordinary Least Squares (OLS) was used.

- In OLS regression, the slope and intercept are chosen so that they minimize the sum of the squared errors, that is, the vertical distance between the predicted y value and the actual y value.
- These errors are known as residuals, and are illustrated for several points in the following diagram:



In mathematical terms, the goal of OLS regression can be expressed as the task of minimizing the following equation:

$$\sum(y_i - \hat{y}_i)^2 = \sum e_i^2$$

In plain language, this equation defines e (the error) as the difference between the actual y value and the predicted y value.

In general equation of line is Y=bx+a

The solution for a depends on the value of b. It can be obtained using the following formula $a = \bar{y} - b\bar{x}$

The value of b that results in the minimum squared error is $$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

The variance involves finding the average squared deviation from the mean of x. This can be expressed as

$$\mathrm{Var}(x) = \frac{\sum(x_i - \bar{x})^2}{n}$$

Covariance function for x and y, denoted as Cov(x, y). The covariance formula is $\mathrm{Cov}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$

Divide the covariance function by the variance function, the n terms get cancelled and can rewrite the formula for b as

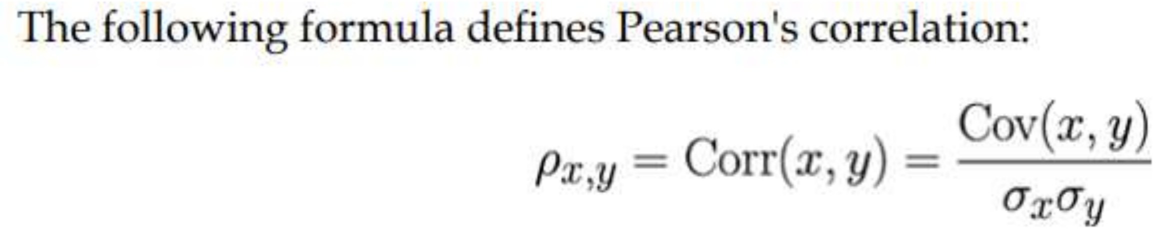
$$b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

Correlations

The correlation between two variables is a number that indicates how closely their relationship follows a straight line.

Correlation typically refers to Pearson's correlation coefficient, which was developed by the 20th century mathematician Karl Pearson.

The correlation ranges between -1 and +1. The extreme values indicate a perfectly linear relationship, while a correlation close to zero indicates the absence of a linear relationship.

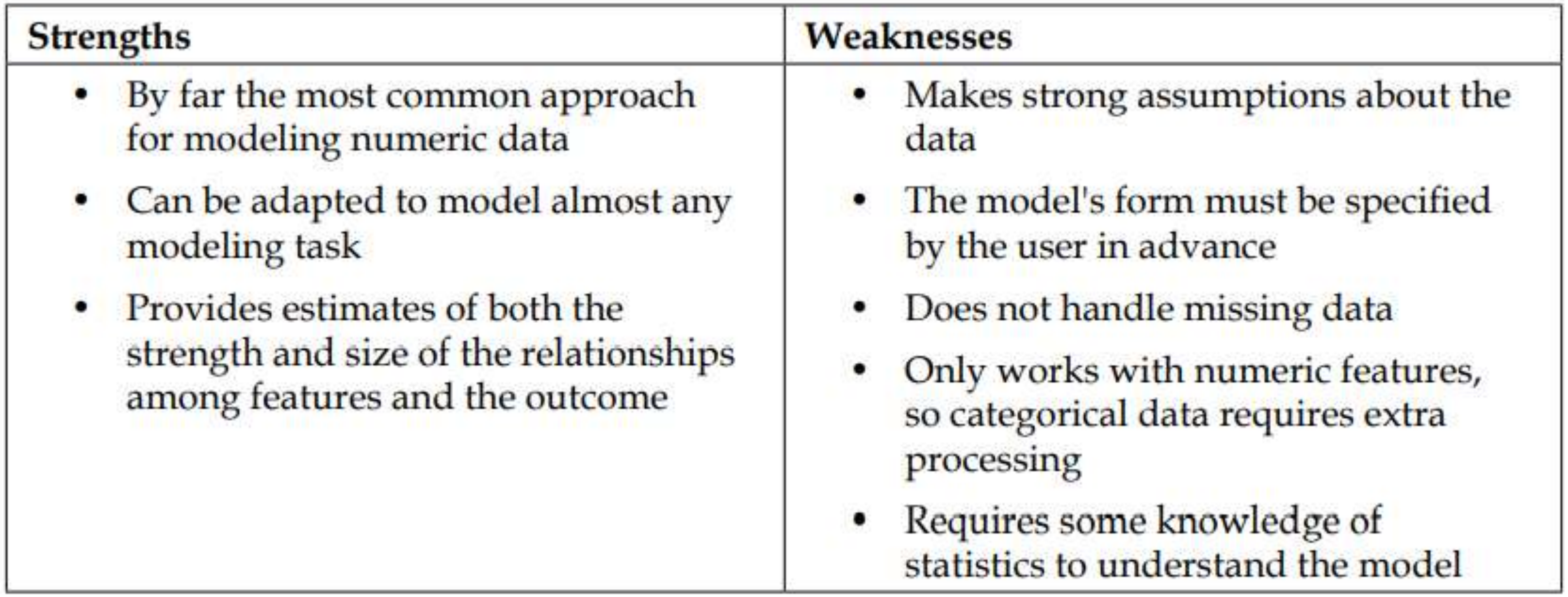


The following formula defines Pearson's correlation:

$$\rho_{x,y} = \mathrm{Corr}(x, y) = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

Multiple linear regression

Most real-world analyses have more than one independent variable. Therefore, it is likely that you will be using multiple linear regression for most numeric prediction tasks.



Strengths	Weaknesses
• By far the most common approach for modeling numeric data • Can be adapted to model almost any modeling task • Provides estimates of both the strength and size of the relationships among features and the outcome	• Makes strong assumptions about the data • The model's form must be specified by the user in advance • Does not handle missing data • Only works with numeric features, so categorical data requires extra processing • Requires some knowledge of statistics to understand the model

Multiple regression as an extension of simple linear regression. The goal in both cases is similar—find values of beta coefficients that minimize the prediction error of a linear equation. The key difference is that there are additional terms for additional independent variables.

Multiple regression equations generally follow the form of the following equation. The dependent variable y is specified as the sum of an intercept term α plus the product of the estimated β value and the x values for each of the i features. An error term (denoted by the Greek letter epsilon) has been added here as a reminder that the predictions are not perfect.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

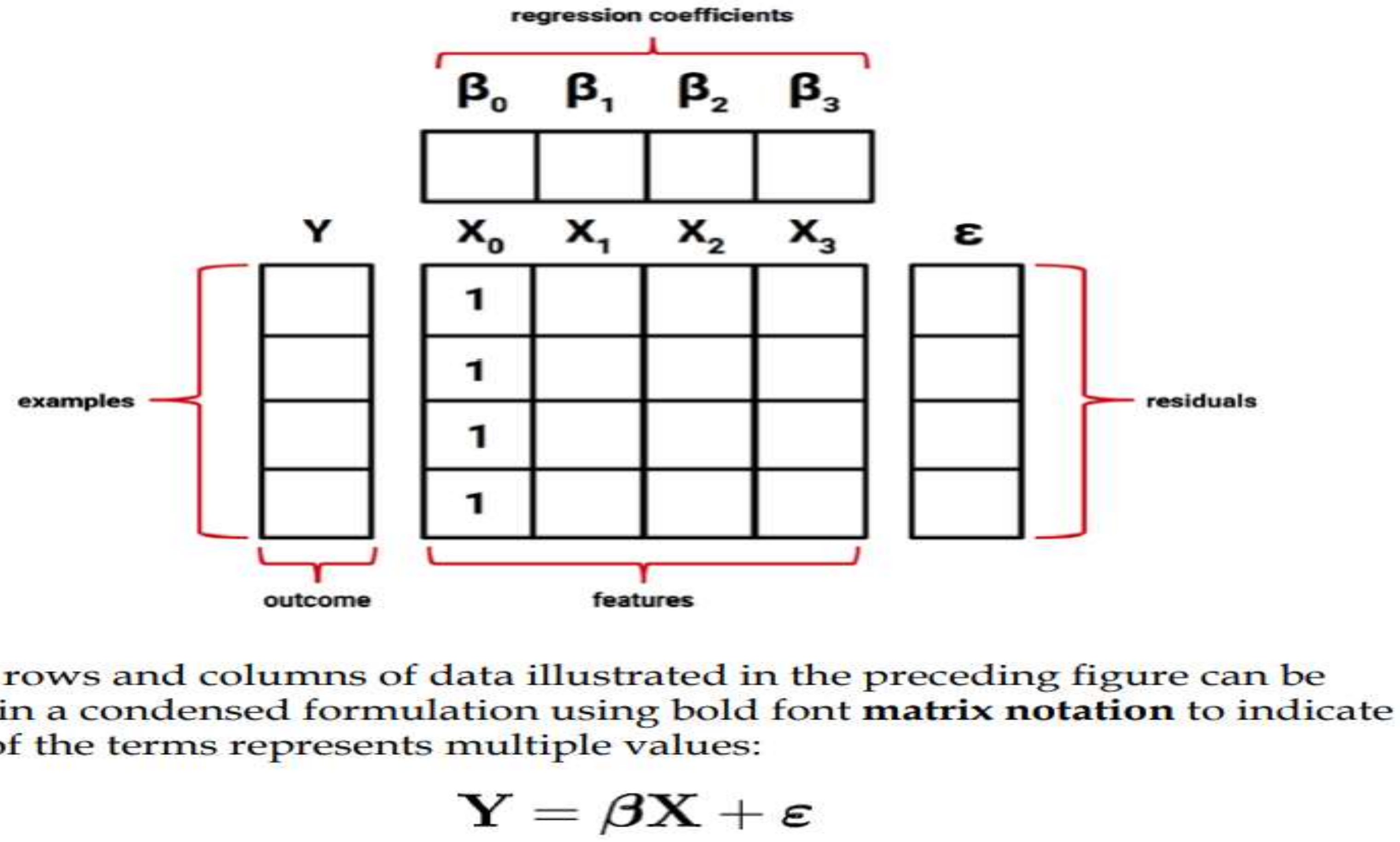
- Since the intercept term α is really no different than any other regression parameter, it is also sometimes denoted as β0 (pronounced beta-naught), as shown in the following equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i + \varepsilon$$

- e, the intercept is unrelated to any of the independent x variables. However, for reasons that will become clear shortly, it helps to imagine β0 as if it were being multiplied by a term x0 , which is a constant with the value 1

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i + \varepsilon$$

In order to estimate the values of the regression parameters, each observed value of the dependent variable y must be related to the observed values of the independent x variables using the regression equation in the previous form. The following figure illustrates this structure



The many rows and columns of data illustrated in the preceding figure can be described in a condensed formulation using bold font **matrix notation** to indicate that each of the terms represents multiple values:

$$\mathbf{Y} = \beta\mathbf{X} + \varepsilon$$

The goal is now to solve for β, the vector of regression coefficients that minimizes the sum of the squared errors between the predicted and actual Y values. Finding the optimal solution requires the use of matrix algebra; The best estimate of the vector β can be computed as:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

This solution uses a pair of matrix operations—the T indicates the transpose of matrix X, while the negative exponent indicates the matrix inverse.