28

REGRESSION & CORRELATION

CONTENTS

REGRESSION

Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of regression is to predict the value of the dependent variable based on the values of the independent variables.

For example, regression could be used to understand whether exam performance can be predicted based on revision time and lecture attendance. As well, regression could be used to understand whether obesity can be predicted based on eating habits and income. The variable being predicted is called the **dependent variable** (or sometimes, the outcome, or target). The variable or variables being used to predict the value of the dependent variable are called the **independent variables** (or the explanatory or regressor variables).

There are several types of regression analysis, but the most common are:

Linear regression: Models the linear relationship between the dependent variable and one or more independent variables. In **simple linear regression**, there is one independent variable, while in **multiple linear regression**, there are multiple independent variables.

Non-linear regression: Models a non-linear relationship between the dependent variable and the independent variables.

SIMPLE REGRESSION

Simple regression is a linear approach to modelling the relationship between two variables. It involves one independent variable and one dependent variable. In statistical notation, y denotes the dependent variable and x denotes the independent variable. A **regression model** is an equation which describes how y is related to x and an error term.

The simple regression model is as follows:

$$\boxed{y = \beta_0 + \beta_1 x + \epsilon}$$

β_0 and β_1 are the parameters of the model, and ε is the error term. The error term accounts for the variations in y that cannot be explained by the linear relationship between x and y.

The **regression equation** is an equation that describes how the expected value of y, denoted $E(y)$, is related to x.

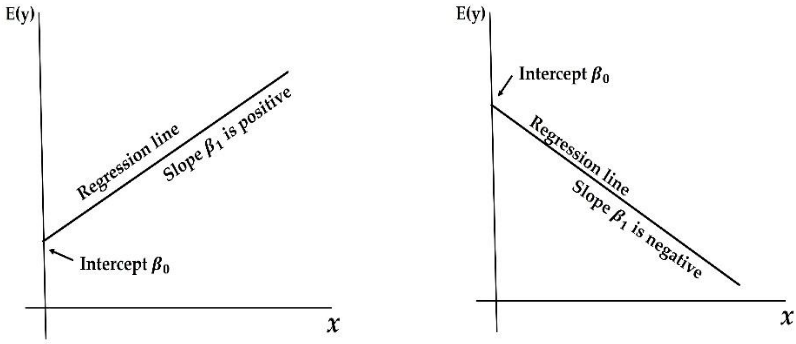
The simple regression equation is as follows.

$$\boxed{E(y) = \beta_0 + \beta_1 x}$$

The graph of the simple regression equation is a straight line. β_0 is the y-intercept of the line. β_1 is the slope. $E(y)$ is the expected value (the mean) of y for any given value of x.

Figure 28.1 Possible Regression Lines in Simple Regression

a. Positive linear relationship b. Negative linear relationship



β_0 and β_1 are population parameters which are unknown. In practice, sample statistics, b_0 and b_1, are computed as estimates. b_0 and b_1 are substituted for β_0 and β_1 in the regression equation to obtain the **estimated regression equation**.

The estimated regression equation is as follows.

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

The **least squares method** is a process for using sample data to find the estimated regression equation; the proof uses simple calculus and linear algebra. Using the least squares method, the values of b_0 and b_1 can be found using:

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

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

Where y_i = the value of the ith observation of the dependent variable; x_i = the value of the ith observation of the independent variable; $\bar{y}$ = mean of the dependent variable; and $\bar{x}$ = mean of the independent variable.

☑ EXAMPLE 28.1

Pizza Robots is a company using robots to deliver pizza. The following data were collected from a sample of 10 Pizza Robots outlets located near university campuses.

Student population (in thousands)	9	11	14	14	23	26	27	32	34	40
Monthly sales ($'000)	54	165	163	195	169	316	219	329	324	416

a) Taking monthly sales as the dependent variable, develop the estimated regression equation by computing the values of b_0 and b_1

b) What is the level of monthly sales for an outlet to be located near a campus with 20,000 students?

SOLUTION **tips**

Let x be the size of the student population (in thousands) and y the monthly sales (in thousands of dollars).

a) There are 10 outlets, we have n = 10 observations.

Begin the calculations by computing $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{230}{10} = 23 \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{2350}{10} = 235$$

Calculate the slope (b_1) as follows.

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{9463}{1018} = 9.3$$

Purchase the full book at:
[*https://unimath.5profz.com/*](https://unimath.5profz.com/)

We donate 0.5% of the book sales every year to charity, forever. When you buy **University Mathematics (I & II)** *you are helping orphans and the less privileged.*