# Simple Linear Regression

Simple linear regression analysis is a statistical tool that gives us the ability to estimate the mathematical relationship between a dependent variable and an independent variable.

In simple linear regression, our objective is to use the data to position a line that best represents the relationship between the two variables.

If we are interested to study the dependency of a variable on a single variable, we use the method of simple regression.

## Some Applications of Simple Linear Regression

- A scientist wants to know at which level of sound pollution will effect human health.
- A computer scientist wants to compress an image for minimum storage.
- An economist wants to investigate the relationship between the petrol price and the inflation rate.
- A sale manager wants to predict the total sale in next year based on the number of staffs.

## Simple Linear Regression Model

Simple linear regression analysis involves one dependent variable ($y$) and one independent variable ($x$) in which the relationship between the variables is approximated by a straight line.

The general form of simple linear regression model is

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

$\beta_0$ and $\beta_1$ are the parameters of the model, and $\varepsilon$ (the Greek letter epsilon) is a random variable referred to as the error term. The error term accounts for the variability in $y$ that can not be explained by the linear relationship between $x$ and $y$.

## Simple Linear Regression Equation

Each distribution of $y$ values has its own mean or expected value. The equation that describes how the expected value of $y$, denoted by $E(y)$, is related to $x$ is called the regression equation. The regression equation for simple linear regression follows.

$$E(y) = \beta_0 + \beta_1 x \cdots\cdots (1)$$

The graph of the simple linear regression equation is a straight line; $\beta_0$ is the $y$-intercept of the regression line which measures the value of $y$ when $x = 0$, $\beta_1$ is the slope which measures the average rate of change in dependent variable per unit change in independent variable, and $E(y)$ is the mean or expected value of $y$ for a given value of $x$.

## Estimated Simple Linear Regression Equation

If the values of the population parameters $\beta_0$ and $\beta_1$ were known, we could use simple linear regression equation to compute the mean value of $y$ for a given value of $x$. In practice, the parameter values are not known, and are needed to be estimated using sample data. Sample statistics (denoted by $b_0$ and $b_1$) are computed as estimates of the population parameters $\beta_0$ and $\beta_1$. Substituting the values of the sample statistics $b_0$ and $b_1$ for $\beta_0$ and $\beta_1$ in the regression equation, we obtain the estimated regression equation.

The estimated regression equation for simple linear regression follows.

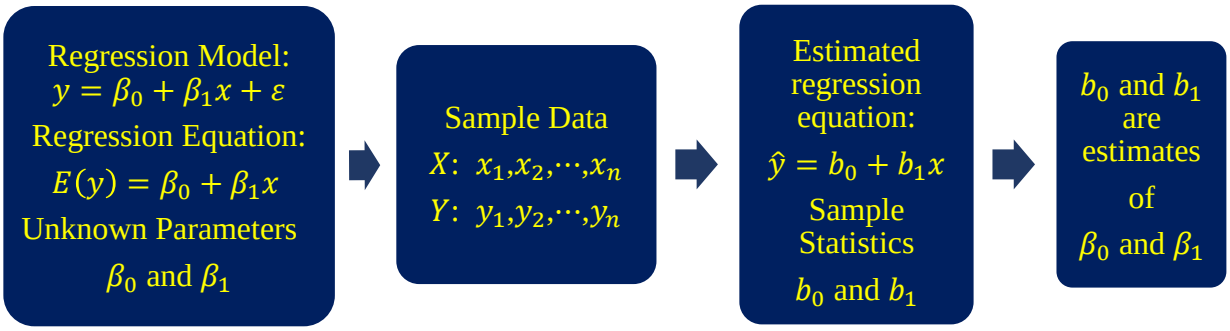
$$\hat{y} = b_0 + b_1 x$$

The graph of the estimated simple linear regression equation is called the estimated regression line; $b_0$ is the $y$ intercept and $b_1$ is the slope. In general, $\hat{y}$ is the point estimator of $E(y)$, the mean value of $y$ for a given value of $x$.

The value of $\hat{y}$ provides both a point estimate of $E(y)$ for a given value of $x$ and a point estimate of an individual value of $y$ for a given value of $x$, we will refer to simply as the estimated value of $y$.

## Estimation Process

The following figure provides a summary of the estimation process for simple linear regression.



## Assumptions of Simple Linear Regression

- The relationship between the dependent and independent variables is linear.
- The error term $\varepsilon$ is a random variable with a mean or expected value of zero; that is, E($\varepsilon$) = 0.
- The variance of $\varepsilon$, denoted by $\sigma^2$ is the same for all values of *x*.
- The values of $\varepsilon$ are independent.
- The error term $\varepsilon$ is a normally distributed random variable.

## Properties of Regression Coefficient

- Regression coefficient is independent of origin and scale.
- Correlation coefficient is the geometric mean of the regression coefficients. That is, if $b_{yx}$ and $b_{xy}$ are the regression coefficient of *y* on *x* and the regression coefficient of *x* on *y* then $r = \sqrt{b_{yx} * b_{xy}}$
- $\frac{b_{yx}+b_{xy}}{2} \geq r$, where *r* is the coefficient of correlation.
- If $b_{yx} > 1$, then $b_{xy} \leq 1$
- Regression coefficient is not a pure number.

## Estimating Regression Equation: Ordinary Least Square Method

The ordinary least squares method is the most widely used method to estimate the regression equation. Carl Friedrich Gauss (1777–1855) proposed the least squares method.

The ordinary least squares method is a procedure for using sample data to find the estimated regression equation.

The least squares method provides an estimated regression equation that minimizes the sum of squared deviations between the observed values of the dependent variable $y$ and the estimated values of the dependent variable $\hat{y}$. This least squares criterion is used to choose the equation that provides the best fit.

## Estimation of Parameters

Let the two variables x and y have n pairs of observations $(x_1, y_1)$, $(x_2, y_2)$, $\cdots$ $(x_n, y_n)$. A pair of the regression line is obtained from the bivariate variable.

The regression equation of *y* on *x* is $\hat{y}_i = b_0 + b_1 x_i$.

The sum of squared residuals is given by

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

Substituting $\hat{y}_i = b_0 + b_1 x_i$, we get $\sum e^2 = \sum (y_i - b_0 - b_1 x_i)^2$ as the expression that must be minimized. To minimize this expression, we have to take the partial derivatives with respect to $b_0$ and $b_1$, set them equal to zero, and solve. Doing so, we get

$$\frac{\partial \sum (y_i - b_0 - b_1 x_i)^2}{\partial b_0} = -2 \sum (y_i - b_0 - b_1 x_i) = 0 \cdots\cdots\cdots (2)$$

$$\frac{\partial \sum (y_i - b_0 - b_1 x_i)^2}{\partial b_1} = -2 \sum x_i (y_i - b_0 - b_1 x_i) = 0 \cdots\cdots\cdots (3)$$

Dividing equation (2) by two and summing each term individually yields

$$-\sum y_i + \sum b_0 + \sum b_1 x_i = 0$$

$$\Rightarrow n b_0 + (\sum x_i) b_1 = \sum y_i \quad [\sum b_0 = n b_0]$$

$$\Longrightarrow b_0 = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} \cdots\cdots\cdots (4)$$

Similar algebraic simplifications applied to equation (3) implies that

$$\left(\sum x_i\right) b_0 + \left(\sum x_i\right)^2 b_1 = \sum x_i y_i$$

$$\Longrightarrow \frac{\sum x_i \sum y_i}{n} + \frac{\sum {x_i}^2}{n} b_1 + \sum ({x_i}^2) b_1 = \sum x_i y_i \cdots\cdots\cdots (5)\text{(Using Equation (4))}$$

By rearranging the Equation (5), we obtain

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

$$\text{Equivalently, } b_1 = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum {x_i}^2) - (\sum x_i)^2} \cdots\cdots\cdots (6)$$

Because $\overline{y} = \sum y_i / n$ and $\overline{x} = \sum x_i / n$ , we can rewrite Equation (4) as

$$b_0 = \overline{y} - b_1 \overline{x} \cdots\cdots\cdots (7)$$

Equations (6) and (7) are used to compute the coefficients in the estimated regression equation.

## Interpretation of the Slope and the Intercept

**Interpretation of $b_1$:** $b_1$ is the estimated change in the average value of $y$ as a result of a one-unit change in $x$.

**Interpretation of $b_0$:** It is the value of the dependent variable (*y*) when the value of independent variable (*x*) is zero. It is of little significance.

## Measures of Variation

- **Total Sum of Squares (*SST*):** It is a measure of variation on the values of dependent variable (y) around their mean value ($\overline{y}$). That is

$$SST = \sum (y_i - \overline{y})^2 = \sum {y_i}^2 - \left(\sum y_i\right)^2 / n$$

- **Regression Sum of Squares (*SSR*):** It is the sum of the squared differences between the predicted value of y and the mean value of y.

$$SSR = \sum (\hat{y_i} - \overline{y})^2$$

- **Error Sum of Squares (*SSE*):** It is the sum of the squared differences between the observed value of y and the predicted value of y.

$$SSE = \sum (y_i - \hat{y_i})^2 = \sum {y_i}^2 - b_0 \sum y_i - b_1 \sum x_i y_i$$

- The relationship between *SST*, *SSR* and *SSE* is $SST = SSR + SSE$

# Index of Goodness of Fit

The goodness of fit of a statistical model describes how well it fits a set of observations. After estimating the parameters and determining the least square regression line, we need to know how "good" the fit of this line to the sample data is. We need an index of goodness of fit and in regression analysis, coefficient of determination is used for measuring the goodness of fit.

The coefficient of determination measures the strength of the association that exists between dependent variable (Y) and independent variable (X).

# Coefficient of Determination

In statistics, coefficient of determination, $r^2$ is a number that determines the proportion of variation in Y which is explained by independent variable of the regression line.

It is defined as

$$r^2 = \frac{Explained\ Variation}{Total\ Variation} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y}_i)^2}$$

The value of $r^2$ lies between 0 and 1. The greater value of $r^2$ implies better fitting of the model to the data.

For example, $r^2 = 0.74$ implies 74% of the total variation of *Y* can be explained by the regression model.

# Merits and Demerits of Simple Linear Regression

## Merits

- Simple linear regression works well irrespective of data size.
- Simple linear regression gives information about the relevance of features.

## Demerits

- If the assumptions of simple linear regression are violated then, it provides biased or inefficient estimates.
- Results are strongly affected by outliers in simple linear regression.
- Simple linear regression assumes that all the variables are normally distributed.

## Example 1

The following data shows the price and overall score (based primarily on picture quality) for ten 42-inch plasma televisions.

| Brand | Price | Score |
|---|---|---|
| Dell | 2800 | 62 |
| Hisense | 2800 | 53 |
| Hitachi | 2700 | 44 |
| JVG | 3500 | 50 |
| LG | 3300 | 54 |
| Maxent | 2000 | 39 |
| Panasonic | 4000 | 66 |
| Philips | 3000 | 55 |
| Proview | 2500 | 34 |
| Samsung | 3000 | 39 |

a) Fit a regression line on overall score for a 42-plasma television given the price.

b) What is the overall score for a television with a price of 3200 dollars?

c) Find the Coefficient of Determination and interpret?

d) What is your interpretation for the model?

## Solution of Example 1:

| Brand | Price ($x$) | Score ($y$) | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| Dell | 2800 | 62 | 173600 | 7840000 | 3844 |
| Hisense | 2800 | 53 | 148400 | 7840000 | 2809 |
| Hitachi | 2700 | 44 | 118800 | 7290000 | 1936 |
| JVG | 3500 | 50 | 175000 | 12250000 | 2500 |
| LG | 3300 | 54 | 178200 | 10890000 | 2916 |
| Maxent | 2000 | 39 | 78000 | 4000000 | 1521 |
| Panasonic | 4000 | 66 | 264000 | 16000000 | 4356 |
| Philips | 3000 | 55 | 165000 | 9000000 | 3025 |
| Proview | 2500 | 34 | 85000 | 6250000 | 1156 |
| Samsung | 3000 | 39 | 117000 | 9000000 | 1521 |
| | $\sum x = 29600$ | $\sum y = 496$ | $\sum xy = 1503000$ | $\sum x^2 = 90360000$ | $\sum y^2 = 25584$ |

$$b_1 = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum {x_i}^2) - (\sum x_i)^2}$$

$$b_1 = 0.0127$$

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

Here, $\overline{y} = 49.6$ and $\overline{x} = 2960$

$$b_0 = 12.008$$

# Interpretation

a) The fitted regression model is,

$$\hat{y} = 12.008 + 0.0127x$$

b) The overall score for a television with a price of 3200 dollars is

$$\hat{y} = 12.008 + 0.0127 * 3200 = 52.648$$

c) $r^2 = 1 - \frac{SSE}{SST}$ $\quad SST = \sum {y_i}^2 - \frac{(\sum y_i)^2}{n}$

$$SSE = \sum {y_i}^2 - b_0 \sum y_i - b_1 \sum x_i y_i$$

$$r^2 = 1 - \frac{539.932}{982.4} = 0.45$$

Interpretation: 45% variation in overall score can be explained by the price of television.

d) $b_0 = 12.008$ means overall score of television will be 12.008, when the price of the television is zero.

$b_1 = 0.0127$ means that the average overall score of television will increase by $0.0127$ unit when the price of the television will increase by one dollar.

## Example 2

The following data is extracted from an article "Reactions on Painted Steel under the Influence of Sodium Chloride and Combinations Thereof". The independent variable ($x$) is $SO_2$ deposition rate ($mg/m^2/d$) and the dependent variable ($y$) is steel weight loss ($g/m^2$).

| $x$ | $y$ |
|---|---|
| 14 | 280 |
| 18 | 350 |
| 40 | 470 |
| 43 | 500 |
| 45 | 560 |
| 112 | 1200 |

a) Develop the estimated regression equation on steel weight loss given the $SO_2$ deposition rate .

b) What is your interpretation for the model?

c) Comment on the goodness of fit of the model.

d) Predict the steel weight loss with $SO_2$ deposition rate 50 $mg/m^2/d$.

## Output from R

lm(formula = y ~x)

Residuals:

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 11.762 | 44.516 | -40.338 | -38.273 | 3.104 | 19.229 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 137.8756 | 26.3776 | 5.227 | 0.0064 ** |
| x | 9.3116 | 0.4745 | 19.622 | 3.98e-05 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.39 on 4 degrees of freedom

Multiple R-squared: 0.9897, Adjusted R-squared: 0.9871

F-statistic: 385 on 1 and 4 DF, p-value: 3.978e-05

# Interpretation

a) The fitted regression model is,

$$\hat{y} = 137.8756 + 9.3116x$$

b) Interpretation:

$b_0 = 137.8756$ means steel weight loss will be $137.8756$ g/m$^2$, when the $SO_2$ deposition rate is zero.

$b_1 = 9.3116$ means that the average steel weight loss will increase by 9.3116 g/m$^2$ when the $SO_2$ deposition rate will increase by one unit.

c) Comment on goodness of fit:

$$r^2 = 0.9897$$

Interpretation: 98.97% variation in steel weight loss can be explained by $SO_2$ deposition rate.

d) The steel weight loss with $SO_2$ deposition rate 50 mg/m$^2$/d is

$$\hat{y} = 137.8756 + 9.3116*50=603.4556$$

So, the predicted steel weight loss is approximately 603.4556 g/m$^2$.

## Practice Problems:

**Textbook:** Probability and Statistics for Engineering and the Sciences (Devore)

CHAPTER 12 Simple Linear Regression and Correlation

Page 488: 17(a, b, c), 19(a, b)

**Textbook:** Statistical Techniques in Business & Economics (LIND MARCHAL WATHEN)

Chapter 13: CORRELATION AND LINEAR REGRESSION

Page 458: 16,17 ,Page 482: 50