LINEAR REGRESSION ANALYSIS

MODULE – II

Lecture - 2

Simple Linear Regression Analysis

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

The simple linear regression model

We consider the modeling between the dependent and one independent variable. When there is only one independent variable in the linear regression model, the model is generally termed as simple linear regression model. When there are more than one independent variables in the model, then the linear model is termed as the multiple linear regression model.

Consider a simple linear regression model

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

where y is termed as the **dependent** or **study variable** and X is termed as **independent** or **explanatory variable**. The terms β_0 and β_1 are the parameters of the model. The parameter β_0 is termed as intercept term and the parameter β_1 is termed as slope parameter. These parameters are usually called as **regression coefficients**. The unobservable error component ε accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of y. This is termed as **disturbance or error term**. There can be several reasons for such difference, e.g., the effect of all deleted variables in the model, variables may be qualitative, inherit randomness in the observations etc. We assume that ε is observed as independent and identically distributed random variable with mean zero and constant variance σ^2. Later, we will additionally assume that ε is normally distributed.

The independent variable is viewed as controlled by the experimenter, so it is considered as non-stochastic whereas y is viewed as a random variable with

$$E(y) = \beta_0 + \beta_1 X$$

And

$$Var(y) = \sigma^2.$$

Sometimes X can also be a random variable. In such a case, instead of simple mean and simple variance of y, we consider the conditional mean of y given $X = x$ as

$$E(y \mid x) = \beta_0 + \beta_1 x$$

and the conditional variance of y given $X = x$ as

$$Var(y \mid x) = \sigma^2.$$

When the values of β_0, β_1 and σ^2 are known, the model is completely described.

The parameters β_0, β_1 and σ^2 are generally unknown and ε is unobserved. The determination of the statistical model $y = \beta_0 + \beta_1 X + \varepsilon$ depends on the determination (i.e., estimation) of β_0, β_1 and σ^2.

In order to know the value of the parameters, n pairs of observations $(x_i, y_i)(i = 1,...,n)$ on (X, y) are observed/collected and are used to determine these unknown parameters.

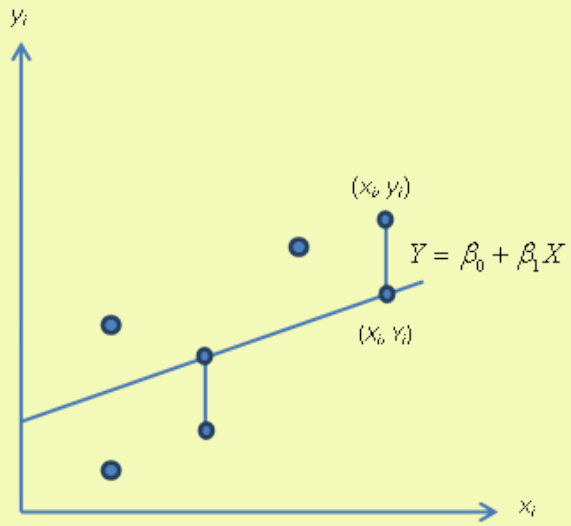
Various methods of estimation can be used to determine the estimates of the parameters. Among them, the least squares and maximum likelihood principles are the popular methods of estimation.

Least squares estimation

Suppose a sample of *n* sets of paired observations $(x_i, y_i)(i = 1, 2, ..., n)$ are available. These observations are assumed to satisfy the simple linear regression model and so we can write

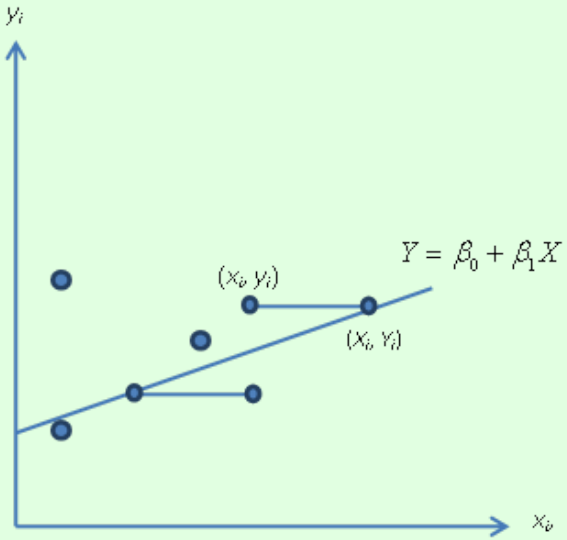
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i (i = 1, 2, ..., n).$$

The method of least squares estimates the parameters β_0 and β_1 by minimizing the sum of squares of difference between the observations and the line in the scatter diagram. Such an idea is viewed from different perspectives. When the **vertical difference** between the observations and the line in the scatter diagram is considered and its sum of squares is minimized to obtain the estimates of β_0 and β_1, the method is known as **direct regression**.



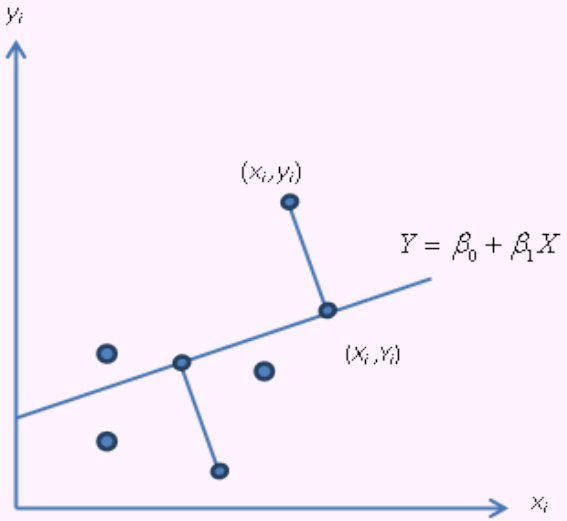
Direct regression method

Alternatively, the sum of squares of difference between the observations and the line in horizontal direction in the scatter diagram can be minimized to obtain the estimates of β_0 and β_1 . This is known as **reverse (or inverse) regression method.**



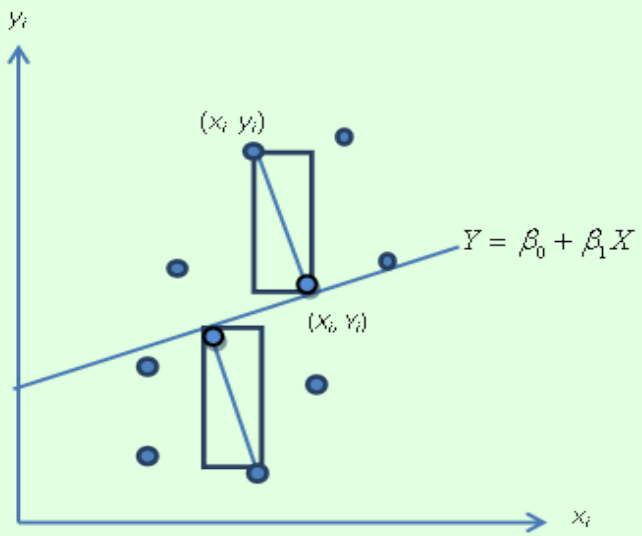
Reverse regression

Instead of horizontal or vertical errors, if the sum of squares of perpendicular distances between the observations and the line in the scatter diagram is minimized to obtain the estimates of β_0 and β_1, the method is known as **orthogonal regression** or **major axis regression method.**



Major axis regression method

Instead of minimizing the distance, the area can also be minimized. The **reduced major axis regression method** minimizes the sum of the areas of rectangles defined between the observed data points and the nearest point on the line in the scatter diagram to obtain the estimates of regression coefficients. This is shown in the following figure:



Reduced major axis method

The method of **least absolute deviation regression** considers the sum of the absolute deviation of the observations from the line in the vertical direction in the scatter diagram as in the case of direct regression to obtain the estimates of β_0 and β_1

No assumption is required about the form of probability distribution of ε_i in deriving the least squares estimates. For the purpose of deriving the statistical inferences only, we assume that ε_i's are observed as random variable with

$$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2 \text{ and } Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ for all } i \neq j (i, j = 1, 2, ..., n).$$

This assumption is needed to find the mean, variance and other properties of the least squares estimates. The assumption that ε_i's are normally distributed is utilized while constructing the tests of hypotheses and confidence intervals of the parameters.

Based on these approaches, different estimates of β_0 and β_1 are obtained which have different statistical properties. Among them the direct regression approach is more popular. Generally, the direct regression estimates are referred as the **least squares estimates** or **ordinary least squares estimates.**

Direct regression method

This method is also known as the **ordinary least squares estimation**. Assuming that a set of n paired observations on $(x_i, y_i),\ i = 1, 2, ..., n$ are available which satisfy the linear regression model $y = \beta_0 + \beta_1 X + \varepsilon$. So we can write the model for each observation as $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ ,(i = 1, 2, ..., n)$.

The direct regression approach minimizes the sum of squares due to errors given by

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

with respect to β_0 and β_1.

The partial derivatives of $S(\beta_0, \beta_1)$ with respect to β_0 are

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_t - \beta_0 - \beta_1 x_i)$$

and the partial derivative of $S(\beta_0, \beta_1)$ with respect to β_1 is

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i.$$

The solution of β_0 and β_1 is obtained by setting

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0.$$

The solutions of these two equations are called the **direct regression estimators**, or usually called as the **ordinary least squares (OLS)** estimators of β_0 and β_1.

This gives the ordinary least squares estimates b_0 of β_0 and b_1 of β_1 as

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

$$b_1 = \frac{s_{xy}}{s_{xx}}$$

where

$$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

$$s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Further, we have

$$\frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_0^2} = -2\sum_{i=1}^{n}(-1) = 2n,$$

$$\frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_1^2} = 2\sum_{i=1}^{n} x_i^2$$

$$\frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_0 \partial \beta_1} = 2\sum_{i=1}^{n} x_t = 2n\bar{x}.$$

The Hessian matrix which is the matrix of second order partial derivatives in this case is given as

$$H^* = \begin{pmatrix} \frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_0^2} & \frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_1^2} \end{pmatrix}$$

$$= 2\begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

$$= 2\begin{pmatrix} \ell' \\ x' \end{pmatrix}(\ell,\ x)$$

where $\ell = (1,1,...,1)'$ is a *n*-vector of elements unity and $x = (x_1,...,x_n)'$ is a *n*-vector of observations on *X*. The matrix *H** is positive definite if its determinant and the element in the first row and column of *H** are positive.

The determinant of *H* is given by

$$|H^*| = 2\left(n\sum_{i=1}^{n} x_i^2 - n^2\bar{x}^2 \right)$$

$$= 2n\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\geq 0.$$

The case when $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$ is not interesting because then all the observations are identical, i.e. $x_i = c$ (some constant). In such a case there is no relationship between *x* and *y* in the context of regression analysis. Since $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0,$ therefore $|H^*| > 0.$ So *H* is positive definite for any (β_0,β_1); therefore $S(\beta_0,\beta_1)$ has a global minimum at (b_0,b_1).

The **fitted line** or the **fitted linear regression model** is

$$y = b_0 + b_1 x$$

and the predicted values are

$$\hat{y}_i = b_0 + b_1 x_i \ (i = 1, 2, ..., n).$$

The difference between the observed value y_i and the fitted (or predicted) value $\hat{y}_i$ is called as a **residual**.

The i^{th} residual is

$$e_i = y_i \sim \hat{y}_i (i = 1, 2, ..., n).$$

We consider it as

$$e_i = y_i - \hat{y}_i$$

$$= y_i - (b_0 + b_1 x_i).$$