Chapter 2
Simple Linear Regression Analysis

The simple linear regression model

We consider the modelling 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 a 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.

The linear 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 the independent or explanatory variable. The terms β_0 and β_1 are the parameters of the model. The parameter β_0 is termed as an intercept term, and the parameter β_1 is termed as the 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. There can be several reasons for such difference, e.g., the effect of all deleted variables in the model, variables may be qualitative, inherent 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 variables are 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 the sample mean and sample 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

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

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 in practice 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 values of these 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.

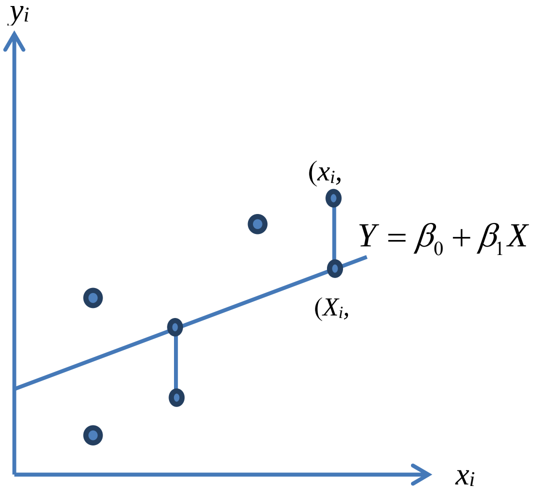
Various methods of estimation can be used to determine the estimates of the parameters. Among them, the methods of least squares and maximum likelihood are the popular methods of estimation.

Least squares estimation

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

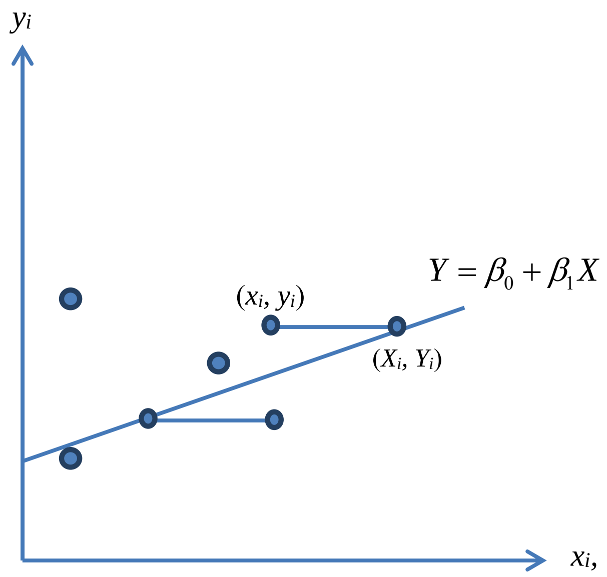
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i (i = 1,2,...,n).$$

The principle of least squares estimates the parameters β_0 and β_1 by minimizing the sum of squares of the 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**.



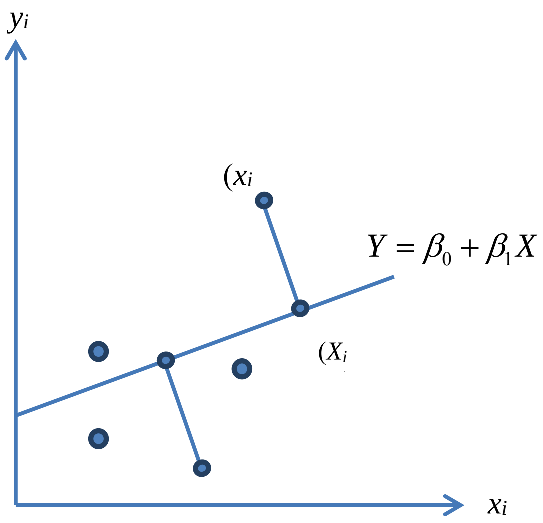
Direct regression

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



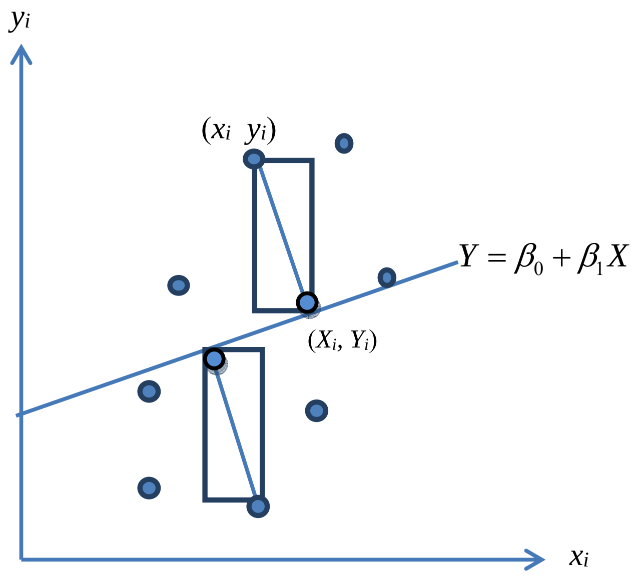
Reverse regression method

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 the 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 random variable with $E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$ 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 to 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

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

with respect to β_0 and β_1.

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

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

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 solutions of β_0 and β_1 are 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

$$\begin{aligned}
\frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_0^2} &= -2\sum_{i=1}^{n}(-1) = 2n, \\
\frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_1^2} &= 2\sum_{i=1}^{n} x_i^2 \\
\frac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} &= 2\sum_{i=1}^{n} x_t = 2n\overline{x}.
\end{aligned}$$

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

$$\begin{aligned}
H^* &= \begin{pmatrix} \dfrac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_0^2} & \dfrac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} \\ \dfrac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} & \dfrac{\partial^2 S(\beta_0,\beta_1)}{\partial \beta_1^2} \end{pmatrix} \\
&= 2\begin{pmatrix} n & n\overline{x} \\ n\overline{x} & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \\
&= 2\begin{pmatrix} \ell' \\ x' \end{pmatrix}(\ell, x)
\end{aligned}$$

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

$$\begin{aligned}
|H^*| &= 4\left(n\sum_{i=1}^{n} x_i^2 - n^2\overline{x}^2\right) \\
&= 4n\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&\geq 0.
\end{aligned}$$

The case when $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0$ is not interesting because all the observations, in this case, 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 $\sum_{i=1}^{n}(x_i - \overline{x})^2 > 0,$ 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

$$y = b_0 + b_1 x.$$

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 a **residual.** The i^{th} residual is defined as

$$\begin{aligned} e_i &= y_i \sim \hat{y}_i (i = 1, 2, ..., n) \\ &= y_i - \hat{y}_i \\ &= y_i - (b_0 + b_1 x_i). \end{aligned}$$

Properties of the direct regression estimators:

Unbiased property:

Note that $b_1 = \frac{s_{xy}}{s_{xx}}$ and $b_0 = \bar{y} - b_1 \bar{x}$ are the linear combinations of $y_i (i = 1, ..., n).$

Therefore

$$b_1 = \sum_{i=1}^{n} k_i y_i$$

where $k_i = (x_i - \bar{x}) / s_{xx}$. Note that $\sum_{i=1}^{n} k_i = 0$ and $\sum_{i=1}^{n} k_i x_i = 1,$ so

$$\begin{aligned} E(b_1) &= \sum_{i=1}^{n} k_i E(y_i) \\ &= \sum_{i=1}^{n} k_i (\beta_0 + \beta_1 x_i). \\ &= \beta_1. \end{aligned}$$

This b_1 is an unbiased estimator of β_1. Next

$$\begin{aligned} E(b_0) &= E\left[\bar{y} - b_1 \bar{x}\right] \\ &= E\left[\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} - b_1 \bar{x}\right] \\ &= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\ &= \beta_0. \end{aligned}$$

Thus b_0 is an unbiased estimator of β_0.

Variances:

Using the assumption that y_i's are independently distributed, the variance of b_1 is

$$
\begin{aligned}
Var(b_1) &= \sum_{i=1}^{n} k_i^2 Var(y_i) + \sum_i \sum_{j \neq i} k_i k_j Cov(y_i, y_j) \\
&= \sigma^2 \frac{\sum_i (x_i - \bar{x})^2}{s_{xx}^2} \qquad (Cov(y_i, y_j) = 0 \text{ as } y_1, ..., y_n \text{ are independent}) \\
&= \frac{\sigma^2 s_{xx}}{s_{xx}^2} \\
&= \frac{\sigma^2}{s_{xx}}.
\end{aligned}
$$

The variance of b_0 is

$$Var(b_0) = Var(\bar{y}) + \bar{x}^2 Var(b_1) - 2\bar{x} Cov(\bar{y}, b_1).$$

First, we find that

$$
\begin{aligned}
Cov(\bar{y}, b_1) &= E\left[\{\bar{y} - E(\bar{y})\}\{b_1 - E(b_1)\}\right] \\
&= E\left[\bar{\varepsilon}(\sum_i c_i y_i - \beta_1)\right] \\
&= \frac{1}{n} E\left[(\sum_i \varepsilon_i)(\beta_0 \sum_i c_i + \beta_1 \sum_i c_i x_i + \sum_i c_i \varepsilon_i) - \beta_1 \sum_i \varepsilon_i\right] \\
&= \frac{1}{n}\left[0 + 0 + 0 + 0\right] \\
&= 0
\end{aligned}
$$

So

$$Var(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right).$$

Covariance:

The covariance between b_0 and b_1 is

$$
\begin{aligned}
Cov(b_0, b_1) &= Cov(\bar{y}, b_1) - \bar{x} Var(b_1) \\
&= -\frac{\bar{x}}{s_{xx}} \sigma^2.
\end{aligned}
$$

It can further be shown that the ordinary least squares estimators b_0 and b_1 possess the minimum variance in the class of linear and unbiased estimators. So they are termed as the Best Linear Unbiased Estimators (BLUE). Such a property is known as the **Gauss-Markov theorem,** which is discussed later in multiple linear regression model.

Residual sum of squares:

The residual sum of squares is given as

$$
\begin{aligned}
SS_{res} &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \\
&= \sum_{i=1}^{n} \left[y_i - \bar{y} + b_1 \bar{x} - b_1 x_i \right]^2 \\
&= \sum_{i=1}^{n} \left[(y_i - \bar{y}) - b_1 (x_i - \bar{x}) \right]^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 + b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2 b_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
&= s_{yy} + b_1^2 s_{xx} - 2 b_1^2 s_{xx} \\
&= s_{yy} - b_1^2 s_{xx} \\
&= s_{yy} - \left(\frac{s_{xy}}{s_{xx}} \right)^2 s_{xx} \\
&= s_{yy} - \frac{s_{xy}^2}{s_{xx}} \\
&= s_{yy} - b_1 s_{xy}.
\end{aligned}
$$

where $s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$

Estimation of σ^2

The estimator of σ^2 is obtained from the residual sum of squares as follows. Assuming that y_i is normally distributed, it follows that SS_{res} has a χ^2 distribution with $(n-2)$ degrees of freedom, so

$$\frac{SS_{res}}{\sigma^2} \sim \chi^2 (n-2).$$

Thus using the result about the expectation of a chi-square random variable, we have

$$E(SS_{res}) = (n-2)\sigma^2.$$

Thus an unbiased estimator of σ^2 is

$$s^2 = \frac{SS_{res}}{n-2}.$$

Note that SS_{res} has only $(n-2)$ degrees of freedom. The two degrees of freedom are lost due to estimation of b_0 and b_1. Since s^2 depends on the estimates b_0 and b_1, so it is a **model-dependentt estimate** of σ^2.

Estimates of variances of b_0 and b_1:

The estimators of variances of b_0 and b_1 are obtained by replacing σ^2 by its estimate $\hat{\sigma}^2 = s^2$ as follows:

$$\widehat{Var}(b_0) = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)$$

and

$$\widehat{Var}(b_1) = \frac{s^2}{s_{xx}}.$$

It is observed that since $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0,$ so $\sum_{i=1}^{n} e_i = 0.$ In the light of this property, e_i can be regarded as an estimate of unknown ε_i $(i = 1,...,n)$. This helps in verifying the different model assumptions on the basis of the given sample $(x_i, y_i),\ i = 1,2,...,n.$

Further, note that

(i) $\sum_{i=1}^{n} x_i e_i = 0,$

(ii) $\sum_{i=1}^{n} \hat{y}_i e_i = 0,$

(iii) $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$ and

(iv) the fitted line always passes through $(\bar{x}, \bar{y}).$

Centered Model:

Sometimes it is useful to measure the independent variable around its mean. In such a case, the model $y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ has a centred version as follows:

$$\begin{aligned} y_i &= \beta_0 + \beta_1(x_i - \bar{x}) + \beta_1\bar{x} + \varepsilon \quad (i = 1,2,...,n) \\ &= \beta_0^* + \beta_1(x_i - \bar{x}) + \varepsilon_i \end{aligned}$$

where $\beta_0^* = \beta_0 + \beta_1\bar{x}$. The sum of squares due to error is given by

$$S(\beta_0^*, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}\left[y_i - \beta_0^* - \beta_1(x_i - \bar{x})\right]^2.$$

Now solving

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

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

we get the direct regression least squares estimates of β_0^* and β_1 as

$$b_0^* = \overline{y}$$

and

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

respectively.

Thus the form of the estimate of slope parameter β_1 remains the same in the usual and centered model whereas the form of the estimate of intercept term changes in the usual and centered models.

Further, the Hessian matrix of the second order partial derivatives of $S(\beta_0^*, \beta_1)$ with respect to β_0^* and β_1 is positive definite at $\beta_0^* = b_0^*$ and $\beta_1 = b_1$ which ensures that $S(\beta_0^*, \beta_1)$ is minimized at $\beta_0^* = b_0^*$ and $\beta_1 = b_1$.

Under the assumption that $E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$ and $Cov(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j = 1, 2, ..., n$, it follows that

$$E(b_0^*) = \beta_0^*, \; E(b_1) = \beta_1,$$
$$Var(b_0^*) = \frac{\sigma^2}{n}, \; Var(b_1) = \frac{\sigma^2}{s_{xx}}.$$

In this case, the fitted model of $y_i = \beta_0^* + \beta_1(x_i - \overline{x}) + \varepsilon_i$ is

$$y = \overline{y} + b_1(x - \overline{x}),$$

and the predicted values are

$$\hat{y}_i = \overline{y} + b_1(x_i - \overline{x}) \quad (i = 1, ..., n).$$

Note that in the centered model

$$Cov(b_0^*, b_1) = 0.$$

No intercept term model:

Sometimes in practice, a model without an intercept term is used in those situations when $x_i = 0 \Rightarrow y_i = 0$ for all $i = 1, 2, ..., n$. A no-intercept model is

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

For example, in analyzing the relationship between the velocity (y) of a car and its acceleration (X), the velocity is zero when acceleration is zero.

Using the data $(x_i, y_i),\ i = 1, 2, ..., n,$ the direct regression least-squares estimate of β_1 is obtained by minimizing $S(\beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$ and solving

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

gives the estimator of β_1 as

$$b_1^* = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}.$$

The second-order partial derivative of $S(\beta_1)$ with respect to β_1 at $\beta_1 = b_1$ is positive which insures that b_1 minimizes $S(\beta_1)$.

Using the assumption that $E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$ and $Cov(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j = 1, 2, ..., n$, the properties of b_1^* can be derived as follows:

$$\begin{aligned} E(b_1^*) &= \frac{\sum_{i=1}^{n} x_i E(y_i)}{\sum_{i=1}^{n} x_i^2} \\ &= \frac{\sum_{i=1}^{n} x_i^2 \beta_1}{\sum_{i=1}^{n} x_i^2} \\ &= \beta_1 \end{aligned}$$

This b_1^* is an unbiased estimator of β_1. The variance of b_1^* is obtained as follows:

$$Var(b_1^*) = \frac{\sum_{i=1}^{n} x_i^2 Var(y_i)}{\left(\sum_{i=1}^{n} x_i^2\right)^2}$$

$$= \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

and an unbiased estimator of σ^2 is obtained as

$$\frac{\sum_{i=1}^{n} y_i^2 - b_1 \sum_{i=1}^{n} y_i x_i}{n-1}.$$

Maximum likelihood estimation

We assume that ε_i's $(i = 1, 2, ..., n)$ are independent and identically distributed following a normal distribution $N(0, \sigma^2)$. Now we use the method of maximum likelihood to estimate the parameters of the linear regression model

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

the observations y_i $(i = 1, 2, ..., n)$ are independently distributed with $N(\beta_0 + \beta_1 x_i, \sigma^2)$ for all $i = 1, 2, ..., n.$

The likelihood function of the given observations (x_i, y_i) and unknown parameters β_0, β_1 and σ^2 is

$$L(x_i, y_i; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right].$$

The maximum likelihood estimates of β_0, β_1 and σ^2 can be obtained by maximizing $L(x_i, y_i; \beta_0, \beta_1, \sigma^2)$ or equivalently in $\ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)$ where

$$\ln\, L(x_i, y_i; \beta_0, \beta_1, \sigma^2) = -\left(\frac{n}{2}\right)\ln 2\pi - \left(\frac{n}{2}\right)\ln \sigma^2 - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

The normal equations are obtained by partial differentiation of log-likelihood with respect to β_0, β_1 and σ^2 and equating them to zero as follows:

$$\frac{\partial \ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial \ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

and

$$\frac{\partial \ln L(x_i, y_i; \beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0.$$

The solution of these normal equations give the maximum likelihood estimates of β_0, β_1 and σ^2 as

$$\tilde{b}_0 = \bar{y} - \tilde{b}_1 \bar{x}$$

$$\tilde{b}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$

and

$$\tilde{s}^2 = \frac{\sum_{i=1}^{n} (y_i - \tilde{b}_0 - \tilde{b}_1 x_i)^2}{n}$$

respectively.

It can be verified that the Hessian matrix of second-order partial derivation of $\ln L$ with respect to β_0, β_1, and σ^2 is negative definite at $\beta_0 = \tilde{b}_0$, $\beta_1 = \tilde{b}_1$, and $\sigma^2 = \tilde{s}^2$ which ensures that the likelihood function is maximized at these values.

Note that the least-squares and maximum likelihood estimates of β_0 and β_1 are identical. The least-squares and maximum likelihood estimates of σ^2 are different. In fact, the least-squares estimate of σ^2 is

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

so that it is related to the maximum likelihood estimate as

$$\tilde{s}^2 = \frac{n-2}{n} s^2.$$

Thus $\tilde{b}_0$ and $\tilde{b}_1$ are unbiased estimators of β_0 and β_1 whereas $\tilde{s}^2$ is a biased estimate of σ^2, but it is asymptotically unbiased. The variances of $\tilde{b}_0$ and $\tilde{b}_1$ are same as of b_0 and b_1 respectively but $Var(\tilde{s}^2) < Var(s^2)$.

Testing of hypotheses and confidence interval estimation for slope parameter:

Now we consider the tests of hypothesis and confidence interval estimation for the slope parameter of the model under two cases, viz., when σ^2 is known and when σ^2 is unknown.

Case 1: When σ^2 is known:

Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, ..., n)$. It is assumed that $\varepsilon_i 's$ are independent and identically distributed and follow $N(0, \sigma^2)$.

First, we develop a test for the null hypothesis related to the slope parameter

$$H_0 : \beta_1 = \beta_{10}$$

where β_{10} is some given constant.

Assuming σ^2 to be known, we know that $E(b_1) = \beta_1, Var(b_1) = \dfrac{\sigma^2}{s_{xx}}$ and b_1 is a linear combination of normally distributed $y_i 's$. So

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{xx}}\right)$$

and so the following statistic can be constructed

$$Z_1 = \frac{b_1 - \beta_{10}}{\sqrt{\dfrac{\sigma^2}{s_{xx}}}}$$

which is distributed as $N(0,1)$ when H_0 is true.

A decision rule to test $H_1 : \beta_1 \neq \beta_{10}$ can be framed as follows:

Reject H_0 if $|Z_1| > Z_{\alpha/2}$

where $Z_{\alpha/2}$ is the $\alpha/2$ percent points on the normal distribution.

Similarly, the decision rule for one-sided alternative hypothesis can also be framed.

The $100(1-\alpha)\%$ confidence interval for β_1 can be obtained using the Z_1 statistic as follows:

$$P\left[-z_{\alpha/2} \le Z_1 \le z_{\alpha/2}\right] = 1-\alpha$$

$$P\left[-z_{\alpha/2} \le \frac{b_1-\beta_1}{\sqrt{\frac{\sigma^2}{s_{xx}}}} \le z_{\alpha/2}\right] = 1-\alpha$$

$$P\left[b_1 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{s_{xx}}} \le \beta_1 \le b_1 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{s_{xx}}}\right] = 1-\alpha.$$

So $100(1-\alpha)\%$ confidence interval for β_1 is

$$\left[b_1 - z_{\alpha/2}\sqrt{\frac{\sigma^2}{s_{xx}}}, b_1 + z_{\alpha/2}\sqrt{\frac{\sigma^2}{s_{xx}}}\right]$$

where $z_{\alpha/2}$ is the $\alpha/2$ percentage point of the $N(0,1)$ distribution.

Case 2: When σ^2 is unknown:

When σ^2 is unknown then we proceed as follows. We know that

$$\frac{SS_{res}}{\sigma^2} \sim \chi^2(n-2)$$

and

$$E\left(\frac{SS_{res}}{n-2}\right) = \sigma^2.$$

Further, SS_{res}/σ^2 and b_1 are independently distributed. This result will be proved formally later in the next module on multiple linear regression. This result also follows from the result that under normal distribution, the maximum likelihood estimates, viz., the sample mean (estimator of population mean) and the sample variance (estimator of population variance) are independently distributed, so b_1 and s^2 are also independently distributed.

Thus the following statistic can be constructed:

$$t_0 = \frac{b_1-\beta_1}{\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}}$$

$$= \frac{b_1-\beta_1}{\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}}$$

which follows a t-distribution with $(n-2)$ degrees of freedom, denoted as t_{n-2}, when H_0 is true.

A decision rule to test $H_1 : \beta_1 \neq \beta_{10}$ is to

reject H_0 if $|t_0| > t_{n-2,\alpha/2}$

where $t_{n-2,\alpha/2}$ is the $\alpha/2$ percent point of the t-distribution with $(n-2)$ degrees of freedom. Similarly, the decision rule for the one-sided alternative hypothesis can also be framed.

The $100(1-\alpha)\%$ confidence interval of β_1 can be obtained using the t_0 statistic as follows:

Consider

$$P\left[-t_{\alpha/2} \leq t_0 \leq t_{\alpha/2}\right] = 1-\alpha$$

$$P\left[-t_{\alpha/2} \leq \frac{b_1-\beta_1}{\sqrt{\dfrac{\hat{\sigma}^2}{s_{xx}}}} \leq t_{\alpha/2}\right] = 1-\alpha$$

$$P\left[b_1 - t_{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}} \leq \beta_1 \leq b_1 + t\alpha/2\sqrt{\frac{\hat{\sigma}^2}{s_{xx}}}\right] = 1-\alpha.$$

So the $100(1-\alpha)\%$ confidence interval β_1 is

$$\left[b_1 - t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}, b_1 + t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}\right].$$

Testing of hypotheses and confidence interval estimation for intercept term:

Now, we consider the tests of hypothesis and confidence interval estimation for intercept term under two cases, viz., when σ^2 is known and when σ^2 is unknown.

Case 1: When σ^2 is known:

Suppose the null hypothesis under consideration is

$$H_0 : \beta_0 = \beta_{00},$$

where σ^2 is known, then using the result that $E(b_0) = \beta_0, Var(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{s_x}\right)$ and b_0 is a linear combination of normally distributed random variables, the following statistic

$$Z_0 = \frac{b_0 - \beta_{00}}{\sqrt{\sigma^2\left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{s_{xx}}\right)}}$$

has a $N(0,1)$ distribution when H_0 is true.

A decision rule to test $H_1 : \beta_0 \neq \beta_{00}$ can be framed as follows:

Reject H_0 if $|Z_0| > Z_{\alpha/2}$

where $Z_{\alpha/2}$ is the $\alpha/2$ percentage points on the normal distribution. Similarly, the decision rule for one-sided alternative hypothesis can also be framed.

The $100(1-\alpha)\%$ confidence intervals for β_0 when σ^2 is known can be derived using the Z_0 statistic as follows:

$$P\left[-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}\right] = 1-\alpha$$

$$P\left[-z_{\alpha/2} \leq \frac{b_0 - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}} \leq z_{\alpha/2}\right] = 1-\alpha$$

$$P\left[b_0 - z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)} \leq \beta_0 \leq b_0 + z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}\right] = 1-\alpha.$$

So the $100(1-\alpha)\%$ of confidential interval of β_0 is

$$\left[b_0 - z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}, b_0 + z_{\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}\right].$$

Case 2: When σ^2 is unknown:

When σ^2 is unknown, then the following statistic is constructed

$$t_0 = \frac{b_0 - \beta_{00}}{\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}}$$

which follows a t-distribution with $(n-2)$ degrees of freedom, i.e., t_{n-2} when H_0 is true.

A decision rule to test $H_1 : \beta_0 \neq \beta_{00}$ is as follows:

Reject H_0 whenever $|t_0| > t_{n-2,\alpha/2}$

where $t_{n-2,\alpha/2}$ is the $\alpha/2$ percentage point of the t-distribution with $(n-2)$ degrees of freedom. Similarly, the decision rule for one-sided alternative hypothesis can also be framed.

The $100(1-\alpha)\%$ confidence interval of β_0 can be obtained as follows:

Consider

$$P\left[t_{n-2,\alpha/2} \le t_0 \le t_{n-2,\alpha/2}\right] = 1-\alpha$$

$$P\left[t_{n-2,\alpha/2} \le \frac{b_0-\beta_0}{\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)}} \le t_{n-2,\alpha/2}\right] = 1-\alpha$$

$$P\left[b_0 - t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)} \le \beta_0 \le b_0 + t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)}\right] = 1-\alpha.$$

So $100(1-\alpha)\%$ confidence interval for β_0 is

$$\left[b_0 - t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)}, b_0 + t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\overline{x}^2}{s_{xx}}\right)}\right].$$

Test of hypothesis for σ^2

We have considered two types of test statistics for testing the hypothesis about the intercept term and slope parameter- when σ^2 is known and when σ^2 is unknown. While dealing with the case of known σ^2, the value of σ^2 is known from some external sources like past experience, long association of the experimenter with the experiment, past studies etc. In such situations, the experimenter would like to test the hypothesis like $H_0:\sigma^2=\sigma_0^2$ against $H_0:\sigma^2\neq\sigma_0^2$ where σ_0^2 is specified. The test statistic is based on the result $\frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-2}$. So the test statistic is

$$C_0 = \frac{SS_{res}}{\sigma_0^2} \sim \chi^2_{n-2} \text{ under } H_0.$$

The decision rule is to reject H_0 if $C_0 < \chi^2_{n-2,\alpha/2}$ or $C_0 > \chi^2_{n-2,1-\alpha/2}$.

Confidence interval for σ^2

A confidence interval for σ^2 can also be derived as follows. Since $SS_{res} / \sigma^2 \sim \chi^2_{n-2}$, thus consider

$$P\left[\chi^2_{n-2,\alpha/2} \le \frac{SS_{res}}{\sigma^2} \le \chi^2_{n-2,1-\alpha/2}\right] = 1-\alpha$$

$$P\left[\frac{SS_{res}}{\chi^2_{n-2,1-\alpha/2}} \le \sigma^2 \le \frac{SS_{res}}{\chi^2_{n-2,\alpha/2}}\right] = 1-\alpha\,.$$

The corresponding $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left[\frac{SS_{res}}{\chi^2_{n-2,1-\alpha/2}}, \frac{SS_{res}}{\chi^2_{n-2,\alpha/2}}\right].$$

Joint confidence region for β_0 and β_1:

A joint confidence region for β_0 and β_1 can also be found. Such a region will provide a $100(1-\alpha)\%$ confidence that both the estimates of β_0 and β_1 are correct. Consider the centered version of the linear regression model

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \varepsilon_i$$

where $\beta_0^* = \beta_0 + \beta_1 \bar{x}$. The least squares estimators of β_0^* and β_1 are

$$b_0^* = \bar{y} \text{ and } b_1 = \frac{s_{xy}}{s_{xx}},$$

respectively.

Using the results that

$$E(b_0^*) = \beta_0^*,$$

$$E(b_1) = \beta_1,$$

$$Var(b_0^*) = \frac{\sigma^2}{n},$$

$$Var(b_1) = \frac{\sigma^2}{s_{xx}}.$$

When σ^2 is known, then the statistic

$$\frac{b_0^* - \beta_0^*}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \quad \text{and} \quad \frac{b_1 - \beta_1}{\sqrt{\frac{\sigma^2}{s_{xx}}}} \sim N(0,1).$$

Moreover, both statistics are independently distributed. Thus

$$\left(\frac{b_0^* - \beta_0^*}{\sqrt{\frac{\sigma^2}{n}}} \right)^2 \sim \chi_1^2 \quad \text{and} \quad \left(\frac{b_1 - \beta_1}{\sqrt{\frac{\sigma^2}{s_{xx}}}} \right)^2 \sim \chi_1^2$$

are also independently distributed because b_0^* and b_1 are independently distributed. Consequently, the sum of these two

$$\frac{n(b_0^* - \beta_o^*)^2}{\sigma^2} + \frac{s_{xx}(b_1 - \beta_1)^2}{\sigma^2} \sim \chi_2^2.$$

Since

$$\frac{SS_{res}}{\sigma^2} \sim \chi_{n-2}^2$$

and SS_{res} is independently distributed of b_0^* and b_1, so the ratio

$$\frac{\left(\frac{n(b_0^* - \beta_0^*)^2}{\sigma^2} + \frac{s_{xx}(b_1 - \beta_1)^2}{\sigma^2} \right) \Big/ 2}{\left(\frac{SS_{res}}{\sigma^2} \right) \Big/ (n-2)} \sim F_{2,n-2}.$$

Substituting $b_0^* = b_0 + b_1 \bar{x}$ and $\beta_0^* = \beta_0 + \beta_1 \bar{x}$, we get

$$\left(\frac{n-2}{2} \right) \left[\frac{Q_f}{SS_{res}} \right]$$

where

$$Q_f = n(b_0 - \beta_0)^2 + 2\sum_{i=1}^{n} x_t (b_0 - \beta_1)(b_1 - \beta_1) + \sum_{i=1}^{n} x_i^2 (b_1 - \beta_1)^2.$$

Since

$$P\left[\left(\frac{n-2}{2} \right) \frac{Q_f}{SS_{res}} \le F_{2,n-2} \right] = 1 - \alpha$$

holds true for all values of β_0 and β_1, so the $100(1-\alpha)\%$ confidence region for β_0 and β_1 is

$$\left(\frac{n-2}{2} \right) . \frac{Q_f}{SS_{res}} \le F_{2,n-2;1-\alpha.}.$$

This confidence region is an ellipse which gives the $100(1-\alpha)\%$ probability that β_0 and β_1 are contained simultaneously in this ellipse.

Analysis of variance:

The technique of analysis of variance is usually used for testing the hypothesis related to equality of more than one parameters, like population means or slope parameters. It is more meaningful in case of multiple regression model when there are more than one slope parameters. This technique is discussed and illustrated here to understand the related basic concepts and fundamentals which will be used in developing the analysis of variance in the next module in multiple linear regression model where the explanatory variables are more than two.

A test statistic for testing $H_0 : \beta_1 = 0$ can also be formulated using the analysis of variance technique as follows.

On the basis of the identity

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}),$$

the sum of squared residuals is

$$\begin{aligned} S(b) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2 - 2\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y}). \end{aligned}$$

Further, consider

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n} (y_i - \bar{y}) b_1 (x_i - \bar{x}) \\ &= b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \end{aligned}$$

Thus we have

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

The term $\sum_{i=1}^{n} (y_i - \bar{y})^2$ is called the **sum of squares about the mean, corrected sum of squares** of y (i.e., *SScorrected*), **total sum of squares**, or s_{yy}.

The term $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ describes the deviation: observation minus predicted value, viz., the residual sum of squares, i.e., $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

whereas the term $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ describes the proportion of variability explained by the regression, $SS_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$.

If all observations y_i are located on a straight line, then in this case $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$ and thus $SS_{corrected} = SS_{reg}$.

Note that SS_{reg} is completely determined by b_1 and so has only one degree of freedom. The total sum of squares $s_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ has $(n-1)$ degrees of freedom due to constraint $\sum_{i=1}^{n}(y_i - \overline{y}) = 0$ and SS_{res} has $(n-2)$ degrees of freedom as it depends on the determination of b_0 and b_1.

All sums of squares are mutually independent and distributed as χ^2_{df} with df degrees of freedom if the errors are normally distributed.

The mean square due to regression is

$$MS_{reg} = \frac{SS_{reg}}{1}$$

and mean square due to residuals is

$$MSE = \frac{SS_{res}}{n-2}.$$

The test statistic for testing $H_0 : \beta_1 = 0$ is

$$F_0 = \frac{MS_{reg}}{MSE}.$$

If $H_0 : \beta_1 = 0$ is true, then MS_{reg} and MSE are independently distributed and thus

$$F_0 \sim F_{1,n-2}.$$

The decision rule for $H_1: \beta_1 \neq 0$ is to reject H_0 if

$$F_0 > F_{1,n-2;1-\alpha}$$

at α level of significance. The test procedure can be described in an Analysis of variance table.

Analysis of variance for testing $H_0: \beta_1 = 0$

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Regression	SS_{reg}	1	MS_{reg}	MS_{reg} / MSE
Residual	SS_{res}	$n-2$	MSE	
Total	s_{yy}	$n-1$		

Some other forms of SS_{reg}, SS_{res} and s_{yy} can be derived as follows:

The sample correlation coefficient then may be written as

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}.$$

Moreover, we have

$$b_1 = \frac{s_{xy}}{s_{xx}} = r_{xy}\sqrt{\frac{s_{yy}}{s_{xx}}}.$$

The estimator of σ^2 in this case may be expressed as

$$\begin{aligned} s^2 &= \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 \\ &= \frac{1}{n-2} SS_{res}. \end{aligned}$$

Various alternative formulations for SS_{res} are in use as well:

$$\begin{aligned} SS_{res} &= \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 \\ &= \sum_{i=1}^{n} [(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2 \\ &= s_{yy} + b_1^2 s_{xx} - 2b_1 s_{xy} \\ &= s_{yy} - b_1^2 s_{xx} \\ &= s_{yy} - \frac{(s_{xy})^2}{s_{xx}}. \end{aligned}$$

Using this result, we find that

$$SS_{corrected} = s_{yy}$$

and

$$\begin{aligned} SS_{reg} &= s_{yy} - SS_{res} \\ &= \frac{(s_{xy})^2}{s_{xx}} \\ &= b_1^2 s_{xx} \\ &= b_1 s_{xy}. \end{aligned}$$

Goodness of fit of regression

It can be noted that a fitted model can be said to be good when residuals are small. Since SS_{res} is based on residuals, so a measure of the quality of a fitted model can be based on SS_{res}. When the intercept term is present in the model, a measure of goodness of fit of the model is given by

$$R^2 = 1 - \frac{SS_{res}}{s_{yy}} = \frac{SS_{reg}}{s_{yy}}.$$

This is known as the **coefficient of determination**. This measure is based on the concept that how much variation in y's stated by s_{yy} is explainable by SS_{reg} and how much unexplainable part is contained in SS_{res}. The ratio SS_{reg}/s_{yy} describes the proportion of variability that is explained by regression in relation to the total variability of y. The ratio SS_{res}/s_{yy} describes the proportion of variability that is not covered by the regression.

It can be seen that

$$R^2 = r_{xy}^2$$

where r_{xy} is the simple correlation coefficient between *x* and *y*. Clearly $0 \le R^2 \le 1$, so a value of R^2 closer to one indicates the better fit and value of R^2 closer to zero indicates the poor fit.

Prediction of values of study variable

An important use of linear regression modeling is to predict the average and actual values of the study variable. The term prediction of the value of study variable corresponds to knowing the value of $E(y)$ (in case of average value) and value of y (in case of actual value) for a given value of the explanatory variable. We consider both cases.

Case 1: Prediction of average value

Under the linear regression model, $y = \beta_0 + \beta_1 x + \varepsilon,$ the fitted model is $y = b_0 + b_1 x$ where b_0 and b_1 are the OLS estimators of β_0 and β_1 respectively.

Suppose we want to predict the value of $E(y)$ for a given value of $x = x_0$. Then the predictor is given by

$$\widehat{E(y \mid x_0)} = \hat{\mu}_{y/x_0} = b_0 + b_1 x_0 .$$

Predictive bias

Then the prediction error is given as

$$\begin{aligned}\hat{\mu}_{y|x_0} - E(y) &= b_0 + b_1 x_0 - E(\beta_0 + \beta_1 x_0 + \varepsilon)\\ &= b_0 + b_1 x_0 - (\beta_0 + \beta_1 x_0)\\ &= (b_0 - \beta_0) + (b_1 - \beta_1)x_0.\end{aligned}$$

Then

$$\begin{aligned}E\left[\hat{\mu}_{y|x_0} - E(y)\right] &= E(b_0 - \beta_0) + E(b_1 - \beta_1)x_0\\ &= 0 + 0 = 0\end{aligned}$$

Thus the predictor μ_{y/x_0} is an unbiased predictor of $E(y)$.

Predictive variance:

The predictive variance of $\hat{\mu}_{y|x_0}$ is

$$\begin{aligned}PV(\hat{\mu}_{y|x_0}) &= Var(b_0 + b_1 x_0)\\ &= Var\left[\bar{y} + b_1(x_0 - \bar{x})\right]\\ &= Var(\bar{y}) + (x_0 - \bar{x})^2 Var(b_1) + 2(x_0 - \bar{x})Cov(\bar{y}, b_1)\\ &= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{s_{xx}} + 0\\ &= \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right].\end{aligned}$$

Estimate of predictive variance

The predictive variance can be estimated by substituting σ^2 by $\hat{\sigma}^2 = MSE$ as

$$\begin{aligned}\widehat{PV}(\hat{\mu}_{y|x_0}) &= \hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right]\\ &= MSE\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right].\end{aligned}$$

Prediction interval estimation:

The 100(1-α)% prediction interval for $E(y/x_0)$ is obtained as follows:

The predictor $\hat{\mu}_{y|x_0}$ is a linear combination of normally distributed random variables, so it is also normally distributed as

$$\hat{\mu}_{y|x_0} \sim N\left(\beta_0 + \beta_1 x_0, PV\left(\hat{\mu}_{y|x_0}\right)\right).$$

So if σ^2 is known, then the distribution of

$$\frac{\hat{\mu}_{y|x_0} - E(y \mid x_0)}{\sqrt{PV(\hat{\mu}_{y|x_0})}}$$

is $N(0,1)$. So the 100(1-α)% prediction interval is obtained as

$$P\left[-z_{\alpha/2} \le \frac{\hat{\mu}_{y|x_0} - E(y \mid x_0)}{\sqrt{PV(\hat{\mu}_{y|x_0})}} \le z_{\alpha/2}\right] = 1 - \alpha$$

which gives the prediction interval for $E(y/x_0)$ as

$$\left[\hat{\mu}_{y|x_0} - z_{\alpha/2}\sqrt{\sigma^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}}\right]},\ \hat{\mu}_{y|x_0} + z_{\alpha/2}\sqrt{\sigma^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}}\right]}\right].$$

When σ^2 is unknown, it is replaced by $\hat{\sigma}^2 = MSE$ and in this case the sampling distribution of

$$\frac{\hat{\mu}_{y|x_0} - E(y \mid x_0)}{\sqrt{MSE\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}}\right]}}$$

is t-distribution with $(n-2)$ degrees of freedom, i.e., t_{n-2}.

The 100(1-α)% prediction interval in this case is

$$P\left[-t_{\alpha/2,n-2} \le \frac{\hat{\mu}_{y|x_0} - E(y \mid x_0)}{\sqrt{MSE\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}}\right]}} \le t_{\frac{\alpha}{2},n-2}\right] = 1 - \alpha.$$

which gives the prediction interval as

$$\left[\hat{\mu}_{y|x_0} - t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{(x_0-\overline{x})^2}{s_{xx}}\right)},\ \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{(x_0-\overline{x})^2}{s_{xx}}\right)}\right].$$

Note that the width of the prediction interval $E(y \mid x_0)$ is a function of x_0. The interval width is minimum for $x_0 = \overline{x}$ and widens as $|x_0 - \overline{x}|$ increases. This is also expected as the best estimates of y to be made at x-values lie near the center of the data and the precision of estimation to deteriorate as we move to the boundary of the x-space.

Case 2: Prediction of actual value

If x_0 is the value of the explanatory variable, then the actual value predictor for y is

$$\hat{y}_0 = b_0 + b_1 x_0 .$$

The true value of y in the prediction period is given by $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$ where ε_0 indicates the value that would be drawn from the distribution of random error in the prediction period. Note that the form of predictor is the same as of average value predictor, but its predictive error and other properties are different. This is the **dual nature of predictor**.

Predictive bias:

The predictive error of $\hat{y}_0$ is given by

$$\begin{aligned}\hat{y}_0 - y_0 &= b_0 + b_1 x_0 - (\beta_0 + \beta_1 x_0 + \varepsilon_0)\\ &= (b_0 - \beta_0) + (b_1 - \beta_1)x_0 - \varepsilon.\end{aligned}$$

Thus, we find that

$$\begin{aligned}E(\hat{y}_0 - y_0) &= E(b_0 - \beta_0) + E(b_1 - \beta_1)x_0 - E(\varepsilon_0)\\ &= 0 + 0 + 0 = 0\end{aligned}$$

which implies that $\hat{y}_0$ is an unbiased predictor of y_0.

Predictive variance

Because the future observation y_0 is independent of $\hat{y}_0$, the predictive variance of $\hat{y}_0$ is

$$
\begin{aligned}
PV(\hat{y}_0) &= E(\hat{y}_0 - y_0)^2 \\
&= E[(b_0-\beta_0)+(x_0-\bar{x})(b_1-\beta_1)+(b_1-\beta_1)\bar{x}-\varepsilon_0]^2 \\
&= Var(b_0)+(x_0-\bar{x})^2Var(b_1)+\bar{x}^2Var(b_1)+Var(\varepsilon_0)+2(x_0-\bar{x})Cov(b_0,b_1)+2\bar{x}Cov(b_0,b_1)+2(x_0-\bar{x})Var(b_1) \\
&\qquad\qquad \text{[rest of the terms are 0 assuming the independence of } \varepsilon_0 \text{ with } \varepsilon_1,\varepsilon_2,...,\varepsilon_n] \\
&= Var(b_0)+[(x_0-\bar{x})^2+\bar{x}^2+2(x_0-\bar{x})]Var(b_1)+Var(\varepsilon)+2[(x_0-\bar{x})+2\bar{x}]Cov(b_0,b_1) \\
&= Var(b_0)+x_0^2Var(b_1)+Var(\varepsilon_0)+2x_0Cov(b_0,b_1) \\
&= \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right]+x_0^2\frac{\sigma^2}{s_{xx}}+\sigma^2-2x_0\frac{\bar{x}\sigma^2}{s_{xx}} \\
&= \sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right].
\end{aligned}
$$

Estimate of predictive variance

The estimate of predictive variance can be obtained by replacing σ^2 by its estimate $\hat{\sigma}^2 = MSE$ as

$$
\begin{aligned}
\widehat{PV(\hat{y}_0)} &= \hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right] \\
&= MSE\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right].
\end{aligned}
$$

Prediction interval:

If σ^2 is known, then the distribution of

$$\frac{\hat{y}_0-y_0}{\sqrt{PV(\hat{y}_0)}}$$

is $N(0,1)$. So the 100(1-α)% prediction interval is obtained as

$$P\left[-z_{\alpha/2} \le \frac{\hat{y}_0-y_0}{\sqrt{PV(\hat{y}_0)}} \le z_{\alpha/2}\right] = 1-\alpha$$

which gives the prediction interval for y_0 as

$$\left[\hat{y}_0 - z_{\alpha/2}\sqrt{\sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)},\ \hat{y}_0 + z_{\alpha/2}\sqrt{\sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)}\right].$$

When σ^2 is unknown, then

$$\frac{\hat{y}_0 - y_0}{\sqrt{\widehat{PV}(\hat{y}_0)}}$$

follows a t-distribution with $(n-2)$ degrees of freedom. The $100(1\text{-}\alpha)\%$ prediction interval for $\hat{y}_0$ in this case is obtained as

$$P\left[-t_{\alpha/2,n-2} \le \frac{\hat{y}_0 - y_0}{\sqrt{\widehat{PV}(\hat{y}_0)}} \le t_{\alpha/2,n-2}\right] = 1-\alpha$$

which gives the prediction interval

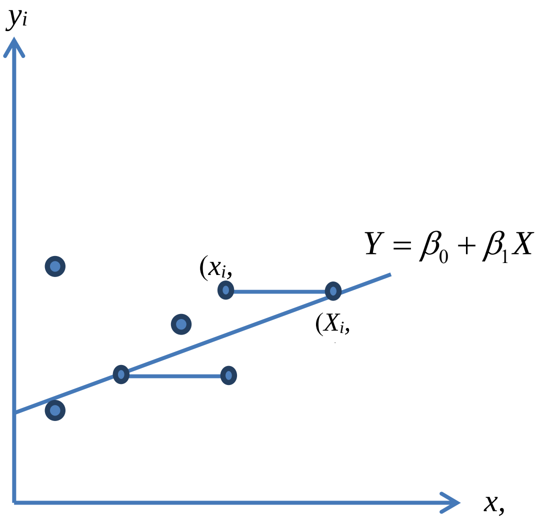
$$\left[\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)},\ \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{s_{xx}}\right)}\right].$$

The prediction interval is of minimum width at $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases.

The prediction interval for $\hat{y}_0$ is wider than the prediction interval for $\hat{\mu}_{y/x_0}$ because the prediction interval for $\hat{y}_0$ depends on both the error from the fitted model as well as the error associated with the future observations.

Reverse regression method

The reverse (or inverse) regression approach minimizes the sum of squares of horizontal distances between the observed data points and the line in the following scatter diagram to obtain the estimates of regression parameters.



Reverse regression

The reverse regression has been advocated in the analysis of gender (or race) discrimination in salaries. For example, if y denotes salary and x denotes qualifications, and we are interested in determining if there is gender discrimination in salaries, we can ask:

> "Whether men and women with the same qualifications (value of x) are getting the same salaries (value of y). This question is answered by the direct regression."

Alternatively, we can ask:

> "Whether men and women with the same salaries (value of y) have the same qualifications (value of x). This question is answered by the reverse regression, i.e., regression of x on y."

The regression equation in case of reverse regression can be written as

$$x_i = \beta_0^* + \beta_1^* y_i + \delta_i \quad (i = 1, 2, ..., n)$$

where δ_i's are the associated random error components and satisfy the assumptions as in the case of the usual simple linear regression model. The reverse regression estimates $\hat{\beta}_{OR}$ of β_0^* and $\hat{\beta}_{1R}$ of β_1^* for the model are obtained by interchanging the x and y in the direct regression estimators of β_0 and β_1. The estimates are obtained as

$$\hat{\beta}_{OR} = \bar{x} - \hat{\beta}_{1R}\bar{y}$$

and

$$\hat{\beta}_{1R} = \frac{s_{yy}}{s_{xy}}$$

for β_0 and β_1 respectively. The residual sum of squares in this case is

$$SS_{res}^* = s_{xx} - \frac{s_{xy}^2}{s_{yy}}.$$

Note that

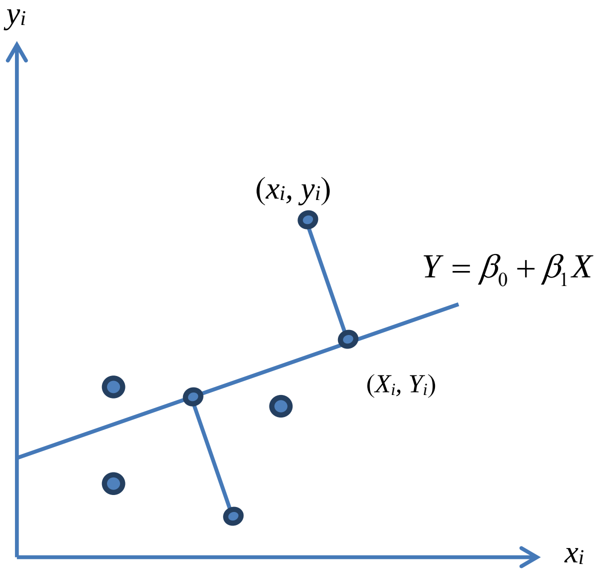
$$\hat{\beta}_{1R} b_1 = \frac{s_{xy}^2}{s_{xx}s_{yy}} = r_{xy}^2$$

where b_1 is the direct regression estimator of the slope parameter and r_{xy} is the correlation coefficient between x and y. Hence if r_{xy}^2 is close to 1, the two regression lines will be close to each other.

An important application of the reverse regression method is in solving the calibration problem.

Orthogonal regression method (or major axis regression method)

The direct and reverse regression methods of estimation assume that the errors in the observations are either in x-direction or y-direction. In other words, the errors can be either in the dependent variable or independent variable. There can be situations when uncertainties are involved in dependent and independent variables both. In such situations, the orthogonal regression is more appropriate. In order to take care of errors in both the directions, the least-squares principle in orthogonal regression minimizes the squared perpendicular distance between the observed data points and the line in the following scatter diagram to obtain the estimates of regression coefficients. This is also known as the **major axis regression method**. The estimates obtained are called **orthogonal regression estimates** or **major axis regression estimates** of regression coefficients.



Orthogonal or major axis regression

If we assume that the regression line to be fitted is $Y_i = \beta_0 + \beta_1 X_i$, then it is expected that all the observations $(x_i, y_i),\ i = 1, 2, ..., n$ lie on this line. But these points deviate from the line, and in such a case, the squared perpendicular distance of observed data $(x_i, y_i)\ \ (i = 1, 2, ..., n)$ from the line is given by

$$d_i^2 = (X_i - x_i)^2 + (Y_i - y_i)^2$$

where (X_i, Y_i) denotes the i^{th} pair of observation without any error which lies on the line.

The objective is to minimize the sum of squared perpendicular distances given by $\sum_{i=1}^{n} d_i^2$ to obtain the estimates of β_0 and β_1. The observations (x_i, y_i) $(i = 1, 2, ..., n)$ are expected to lie on the line

$$Y_i = \beta_0 + \beta_1 X_i ,$$

so let

$$E_i = Y_i - \beta_0 - \beta_1 X_i = 0.$$

The regression coefficients are obtained by minimizing $\sum_{i=1}^{n} d_i^2$ under the constraints E_i's using the Lagrangian's multiplier method. The Lagrangian function is

$$L_0 = \sum_{i=1}^{n} d_i^2 - 2\sum_{i=1}^{n} \lambda_i E_i$$

where $\lambda_1, ..., \lambda_n$ are the Lagrangian multipliers. The set of equations are obtained by setting

$$\frac{\partial L_0}{\partial X_i} = 0, \frac{\partial L_0}{\partial Y_i} = 0, \frac{\partial L_0}{\partial \beta_0} = 0 \text{ and } \frac{\partial L_0}{\partial \beta_1} = 0 \ (i = 1, 2, ..., n).$$

Thus we find

$$\frac{\partial L_0}{\partial X_i} = (X_i - x_i) + \lambda_i \beta_1 = 0$$

$$\frac{\partial L_0}{\partial Y_i} = (Y_i - y_i) - \lambda_i = 0$$

$$\frac{\partial L_0}{\partial \beta_0} = \sum_{i=1}^{n} \lambda_i = 0$$

$$\frac{\partial L_0}{\partial \beta_1} = \sum_{i=1}^{n} \lambda_i X_i = 0.$$

Since

$$X_i = x_i - \lambda_i \beta_1$$
$$Y_i = y_i + \lambda_i ,$$

so substituting these values is ε_i , we obtain

$$E_i = (y_i + \lambda_i) - \beta_0 - \beta_1 (x_i - \lambda_i \beta_1) = 0$$

$$\Rightarrow \lambda_i = \frac{\beta_0 + \beta_1 x_i - y_i}{1 + \beta_1^2}.$$

Also using this λ_i in the equation $\sum_{i=1}^{n}\lambda_i = 0$, we get

$$\frac{\sum_{i=1}^{n}(\beta_0+\beta_1 x_i - y_i)}{1+\beta_1^2} = 0$$

and using $(X_i - x_i)+\lambda_i\beta_1 = 0$ and $\sum_{i=1}^{n}\lambda_i X_i = 0$, we get

$$\sum_{i=1}^{n}\lambda_i(x_i - \lambda_i\beta_1) = 0.$$

Substituting λ_i in this equation, we get

$$\frac{\sum_{i=1}^{n}(\beta_0 x_i+\beta_1 x_i^2 - y_i x_i)}{(1+\beta_i^2)} - \frac{\beta_1(\beta_0+\beta_1 x_i - y_i)^2}{(1+\beta_1^2)^2} = 0. \qquad (1)$$

Using λ_i in the equation and using the equation $\sum_{i=1}^{n}\lambda_i = 0$, we solve

$$\frac{\sum_{i=1}^{n}(\beta_0+\beta_1 x_i - y_i)}{1+\beta_1^2} = 0.$$

The solution provides an orthogonal regression estimate of β_0 as

$$\hat{\beta}_{0OR} = \bar{y} - \hat{\beta}_{1OR}\bar{x}$$

where $\hat{\beta}_{1OR}$ is an orthogonal regression estimate of β_1.

Now, substituting β_{0OR} in equation (1), we get

$$\sum_{i=1}^{n}(1+\beta_1^2)\left[\bar{y}x_i - \beta_1\bar{x}x_i + \beta_1 x_i^2 - x_i y_i\right] - \beta_1\sum_{i=1}^{n}\left(\bar{y}-\beta_1\,\bar{x}+\beta_1 x_i - y_i\right)^2 = 0$$

or

$$(1+\beta_1^2)\sum_{i=1}^{n}x_i\left[y_i - \bar{y} - \beta_1(x_i-\bar{x})\right] + \beta_1\sum_{i=1}^{n}\left[-(y_i-\bar{y})+\beta_1(x_i-\bar{x})\right]^2 = 0$$

or

$$(1+\beta_1^2)\sum_{i=1}^{n}(u_i+\bar{x})(v_i-\beta_1 u_i) + \beta_1\sum_{i=1}^{n}(-v_i+\beta_1 u_i)^2 = 0$$

where

$$u_i = x_i - \bar{x},$$
$$v_i = y_i - \bar{y}.$$

Since $\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} u_i = 0$, so

$$\sum_{i=1}^{n} \left[\beta_1^2 u_i v_i + \beta_1 (u_i^2 - v_i^2) - u_i v_i \right] = 0$$

or

$$\beta_1^2 s_{xy} + \beta_1 (s_{xx} - s_{yy}) - s_{xy} = 0.$$

Solving this quadratic equation provides the orthogonal regression estimate of β_1 as

$$\hat{\beta}_{1OR} = \frac{\left(s_{yy} - s_{xx}\right) + sign\left(s_{xy}\right)\sqrt{(s_{xx} - s_{yy})^2 + 4s_{xy}^2}}{2s_{xy}}$$

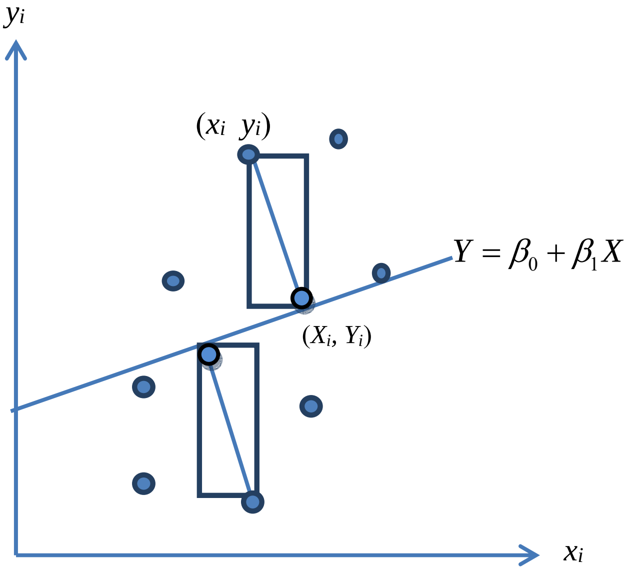
where $sign(s_{xy})$ denotes the sign of s_{xy} which can be positive or negative. So

$$sign(s_{xy}) = \begin{cases} 1 & \text{if } s_{xy} > 0 \\ -1 & \text{if } s_{xy} > 0. \end{cases}.$$

Notice that this gives two solutions for $\hat{\beta}_{1OR}$. We choose the solution which minimizes $\sum_{i=1}^{n} d_i^2$. The other solution maximizes $\sum_{i=1}^{n} d_i^2$ and is in the direction perpendicular to the optimal solution. The optimal solution can be chosen with the sign of s_{xy}.

Reduced major axis regression method:

The direct, reverse and orthogonal methods of estimation minimize the errors in a particular direction which is usually the distance between the observed data points and the line in the scatter diagram. Alternatively, one can consider the area extended by the data points in a certain neighbourhood and instead of distances, the area of rectangles defined between the corresponding observed data point and the nearest point on the line in the following scatter diagram can also be minimized. Such an approach is more appropriate when the uncertainties are present in the study and explanatory variables both. This approach is termed as reduced major axis regression.



Reduced major axis method

Suppose the regression line is $Y_i = \beta_0 + \beta_1 X_i$ on which all the observed points are expected to lie. Suppose the points $(x_i, y_i),\ i = 1, 2, ..., n$ are observed which lie away from the line. The area of rectangle extended between the i^{th} observed data point and the line is

$$A_i = (X_i \sim x_i)(Y_i \sim y_i)\ \ (i = 1, 2, ..., n)$$

where (X_i, Y_i) denotes the i^{th} pair of observation without any error which lies on the line.

The total area extended by n data points is

$$\sum_{i=1}^{n} A_i = \sum_{i=1}^{n} (X_i \sim x_i)(Y_i \sim y_i).$$

All observed data points $(x_i, y_i), (i = 1,2,...,n)$ are expected to lie on the line

$$Y_i = \beta_0 + \beta_1 X_i$$

and let

$$E_i^* = Y_i - \beta_0 - \beta_1 X_i = 0.$$

So now the objective is to minimize the sum of areas under the constraints E_i^* to obtain the reduced major axis estimates of regression coefficients. Using the Lagrangian multiplier method, the Lagrangian function is

$$\begin{aligned} L_R &= \sum_{i=1}^n A_i - \sum_{i=1}^n \mu_i E_i^* \\ &= \sum_{i=1}^n (X_i - x_i)(Y_i - y_i) - \sum_{i=1}^n \mu_i E_i^* \end{aligned}$$

where $\mu_1, ..., \mu_n$ are the Lagrangian multipliers. The set of equations are obtained by setting

$$\frac{\partial L_R}{\partial X_i} = 0, \frac{\partial L_R}{\partial Y_i} = 0, \frac{\partial L_R}{\partial \beta_0} = 0, \frac{\partial L_R}{\partial \beta_1} = 0 \quad (i = 1,2,...,n).$$

Thus

$$\begin{aligned} \frac{\partial L_R}{\partial X_i} &= (Y_i - y_i) + \beta_1 \mu_i = 0 \\ \frac{\partial L_R}{\partial Y_i} &= (X_i - x_i) - \mu_i = 0 \\ \frac{\partial L_R}{\partial \beta_0} &= \sum_{i=1}^n \mu_i = 0 \\ \frac{\partial L_R}{\partial \beta_1} &= \sum_{i=1}^n \mu_i X_i = 0. \end{aligned}$$

Now

$$\begin{aligned} X_i &= x_i + \mu_i \\ Y_i &= y_i - \beta_1 \mu_i \\ \beta_0 + \beta_1 X_i &= y_i - \beta_1 \mu_i \\ \beta_0 + \beta_1 (x_i + \mu_i) &= y_i - \beta_1 \mu_i \\ \Rightarrow \mu_i &= \frac{y_i - \beta_0 - \beta_1 x_i}{2\beta_1}. \end{aligned}$$

Substituting μ_i in $\sum_{i=1}^{n}\mu_i = 0$, the reduced major axis regression estimate of β_0 is obtained as

$$\hat{\beta}_{0RM} = \bar{y} - \hat{\beta}_{1RM}\bar{x}$$

where $\hat{\beta}_{1RM}$ is the reduced major axis regression estimate of β_1. Using $X_i = x_i + \mu_i$, μ_i and $\hat{\beta}_{0RM}$ in $\sum_{i=1}^{n}\mu_i X_i = 0$, we get

$$\sum_{i=1}^{n}\left(\frac{y_i - \bar{y} + \beta_1\bar{x} - \beta_1 x_i}{2\beta_1}\right)\left(x_i - \frac{y_i - \bar{y} + \beta_1\bar{x} - \beta_1 x_i}{2\beta_1}\right) = 0.$$

Let $u_i = x_i - \bar{x}$ and $v_i = y_i - \bar{y}$, then this equation can be re-expressed as

$$\sum_{i=1}^{n}(v_i - \beta_1 u_i)(v_i + \beta_1 u_i + 2\beta_1\bar{x}) = 0.$$

Using $\sum_{i=1}^{n}u_i = \sum_{i=1}^{n}u_i = 0$, we get

$$\sum_{i=1}^{n}v_i^2 - \beta_1^2\sum_{i=1}^{n}u_i^2 = 0.$$

Solving this equation, the reduced major axis regression estimate of β_1 is obtained as

$$\hat{\beta}_{1RM} = sign(s_{xy})\sqrt{\frac{s_{yy}}{s_{xx}}}$$

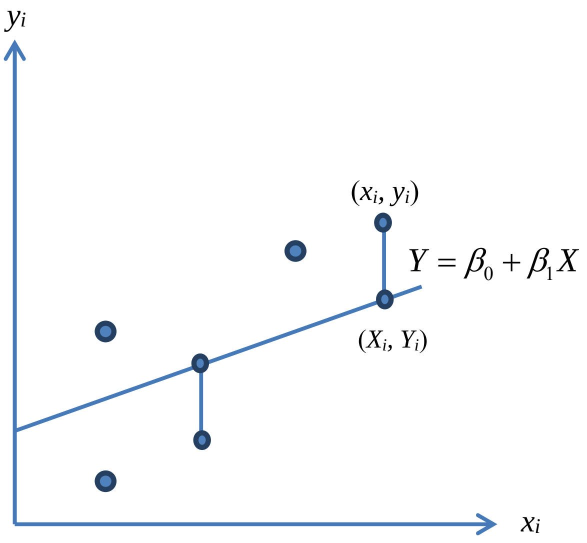
where $sign(s_{xy}) = \begin{cases} 1 & \text{if } s_{xy} > 0 \\ -1 & \text{if } s_{xy} < 0. \end{cases}$

We choose the regression estimator which has same sign as of s_{xy}.

Least absolute deviation regression method

The least-squares principle advocates the minimization of the sum of squared errors. The idea of squaring the errors is useful in place of simple errors because random errors can be positive as well as negative. So consequently their sum can be close to zero indicating that there is no error in the model and which can be misleading. Instead of the sum of random errors, the sum of absolute random errors can be considered which avoids the problem due to positive and negative random errors.

In the method of least squares, the estimates of the parameters β_0 and β_1 in the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. $(i = 1, 2, ..., n)$ are chosen such that the sum of squares of deviations $\sum_{i=1}^{n} \varepsilon_i^2$ is minimum. In the method of least absolute deviation (LAD) regression, the parameters β_0 and β_1 are estimated such that the sum of absolute deviations $\sum_{i=1}^{n} |\varepsilon_i|$ is minimum. It minimizes the absolute vertical sum of errors as in the following scatter diagram:



Least absolute deviation regression

The LAD estimates $\hat{\beta}_{0L}$ and $\hat{\beta}_{1L}$ are the estimates of β_0 and β_1, respectively which minimize

$$LAD(\beta_0, \beta_1) = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|$$

for the given observations $(x_i, y_i)(i = 1, 2, ..., n)$.

Conceptually, LAD procedure is more straightforward than OLS procedure because $|e|$ (absolute residuals) is a more straightforward measure of the size of the residual than e^2 (squared residuals). The LAD regression estimates of β_0 and β_1 are not available in closed form. Instead, they can be obtained numerically based on algorithms. Moreover, this creates the problems of non-uniqueness and degeneracy in the estimates. The concept of non-uniqueness relates to that more than one best line pass through a data point. The degeneracy concept describes that the best line through a data point also passes through more than one other data points. The non-uniqueness and degeneracy concepts are used in algorithms to judge the quality of the estimates. The algorithm for finding the estimators generally proceeds in steps. At each step, the best line is found that passes through a given data point. The best line always passes through another data point, and this data point is used in the next step. When there is non-uniqueness, then there is more than one best line. When there is degeneracy, then the best line passes through more than one other data point. When either of the problems is present, then there is more than one choice for the data point to be used in the next step and the algorithm may go around in circles or make a wrong choice of the LAD regression line. The exact tests of hypothesis and confidence intervals for the LAD regression estimates can not be derived analytically. Instead, they are derived analogously to the tests of hypothesis and confidence intervals related to ordinary least squares estimates.

Estimation of parameters when X is stochastic

In a usual linear regression model, the study variable is supped to be random and explanatory variables are assumed to be fixed. In practice, there may be situations in which the explanatory variable also becomes random.

Suppose both dependent and independent variables are stochastic in the simple linear regression model

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

where ε is the associated random error component. The observations $(x_i, y_i), i = 1, 2, ..., n$ are assumed to be jointly distributed. Then the statistical inferences can be drawn in such cases which are conditional on X.

Assume the joint distribution of X and y to be bivariate normal $N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ where μ_x and μ_y are the means of X and $y; \sigma_x^2$ and σ_y^2 are the variances of X and y; and ρ is the correlation coefficient between X and y. Then the conditional distribution of y given $X = x$ is the univariate normal conditional mean

$$E(y \mid X = x) = \mu_{y|x} = \beta_0 + \beta_1 x$$

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

$$Var(y \mid X = x) = \sigma_{y|x}^2 = \sigma_y^2(1 - \rho^2)$$

where

$$\beta_0 = \mu_y - \mu_x \beta_1$$

and

$$\beta_1 = \frac{\sigma_y}{\sigma_x}\rho.$$

When both X and y are stochastic, then the problem of estimation of parameters can be reformulated as follows. Consider a conditional random variable $y \mid X = x$ having a normal distribution with mean as conditional mean $\mu_{y|x}$ and variance as conditional variance $Var(y \mid X = x) = \sigma^2_{y|x}$. Obtain n independently distributed observation $y_i \mid x_i,\ i = 1, 2, ..., n$ from $N(\mu_{y|x}, \sigma^2_{y|x})$ with nonstochastic X. Now the method of maximum likelihood can be used to estimate the parameters which yield the estimates of β_0 and β_1 as earlier in the case of nonstochastic X as

$$\tilde{b} = \bar{y} - \tilde{b}_1 \bar{x}$$

and

$$\tilde{b}_1 = \frac{s_{xy}}{s_{xx}},$$

respectively.

Moreover, the correlation coefficient

$$\rho = \frac{E(y - \mu_y)(X - \mu_x)}{\sigma_y \sigma_x}$$

can be estimated by the sample correlation coefficient

$$\begin{aligned}
\hat{\rho} &= \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \\
&= \frac{s_{xy}}{\sqrt{s_{xx}} \sqrt{s_{yy}}} \\
&= \tilde{b}_1 \sqrt{\frac{s_{xx}}{s_{yy}}}.
\end{aligned}$$

Thus

$$\begin{aligned}
\hat{\rho}^2 &= \tilde{b}_1^2 \frac{s_{xx}}{s_{yy}} \\
&= \tilde{b}_1 \frac{s_{xy}}{s_{yy}} \\
&= \frac{s_{yy} - \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{s_{yy}} \\
&= R^2
\end{aligned}$$

which is same as the coefficient of determination. Thus R^2 has the same expression as in the case when X is fixed. Thus R^2 again measures the goodness of the fitted model even when X is stochastic.