Asymptotic properties of the least squares estimator

Professor V.M. Auken, PhD

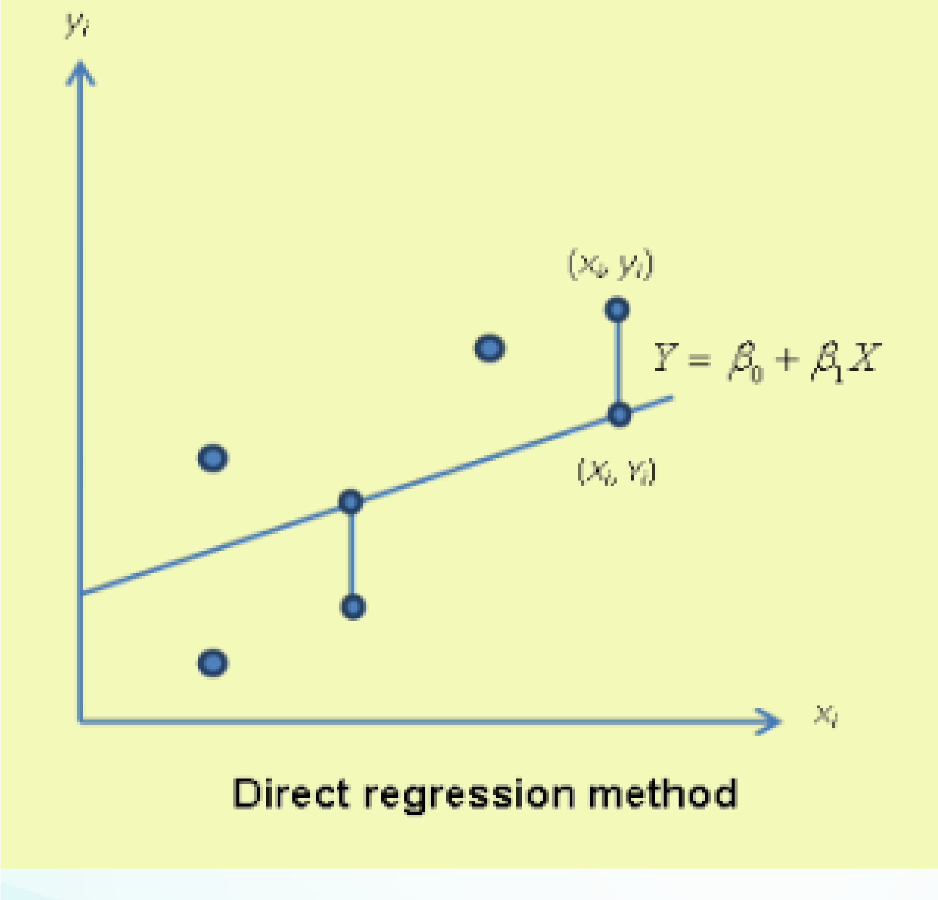
Least squares estimation

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**.

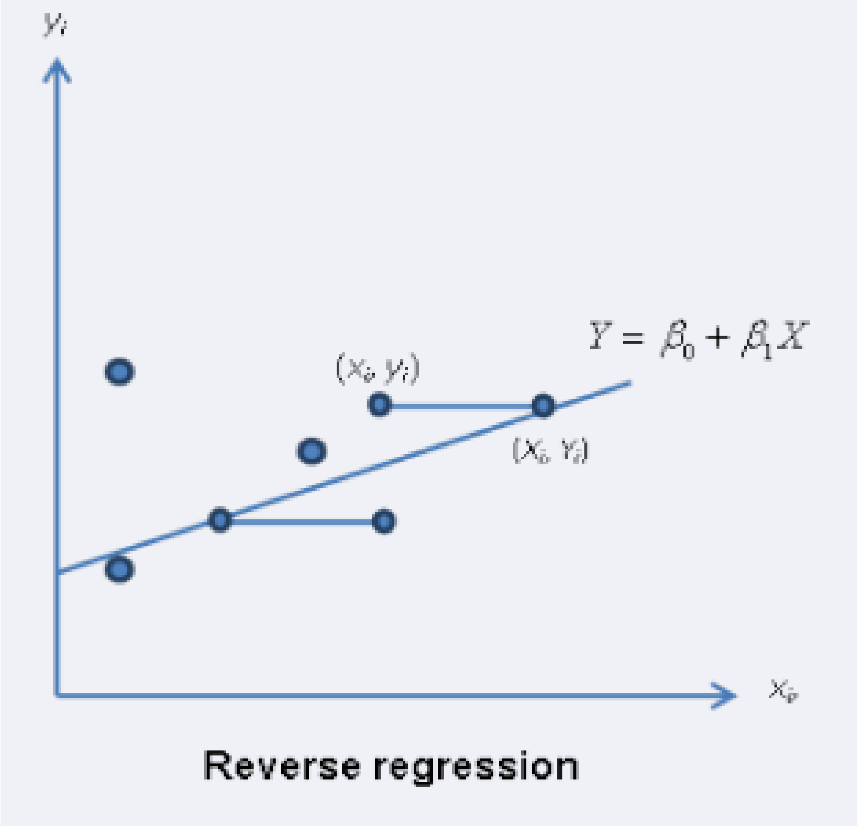
Least squares estimation



Least squares estimation

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.**

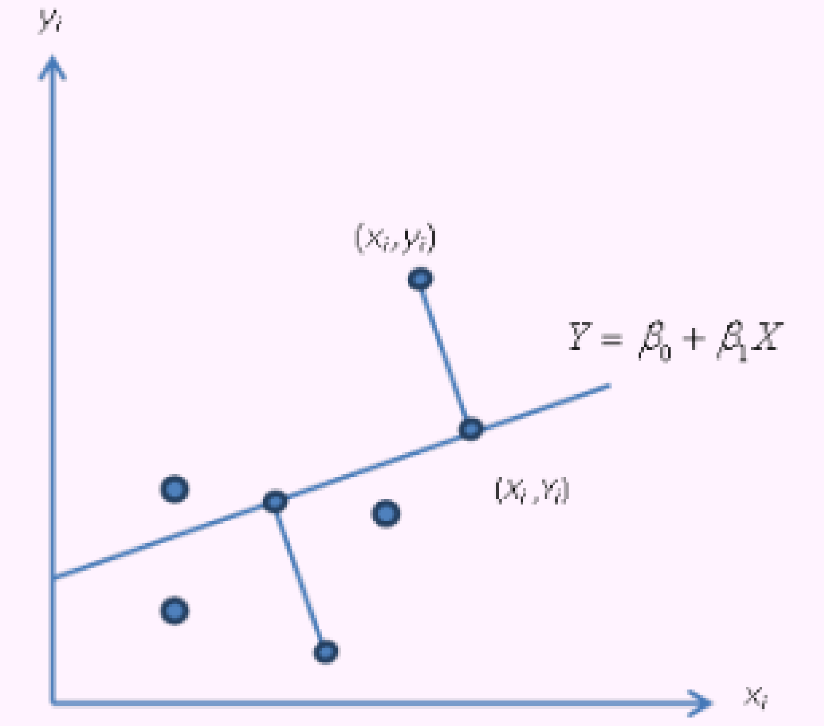
Least squares estimation



Least squares estimation

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.**

Least squares estimation

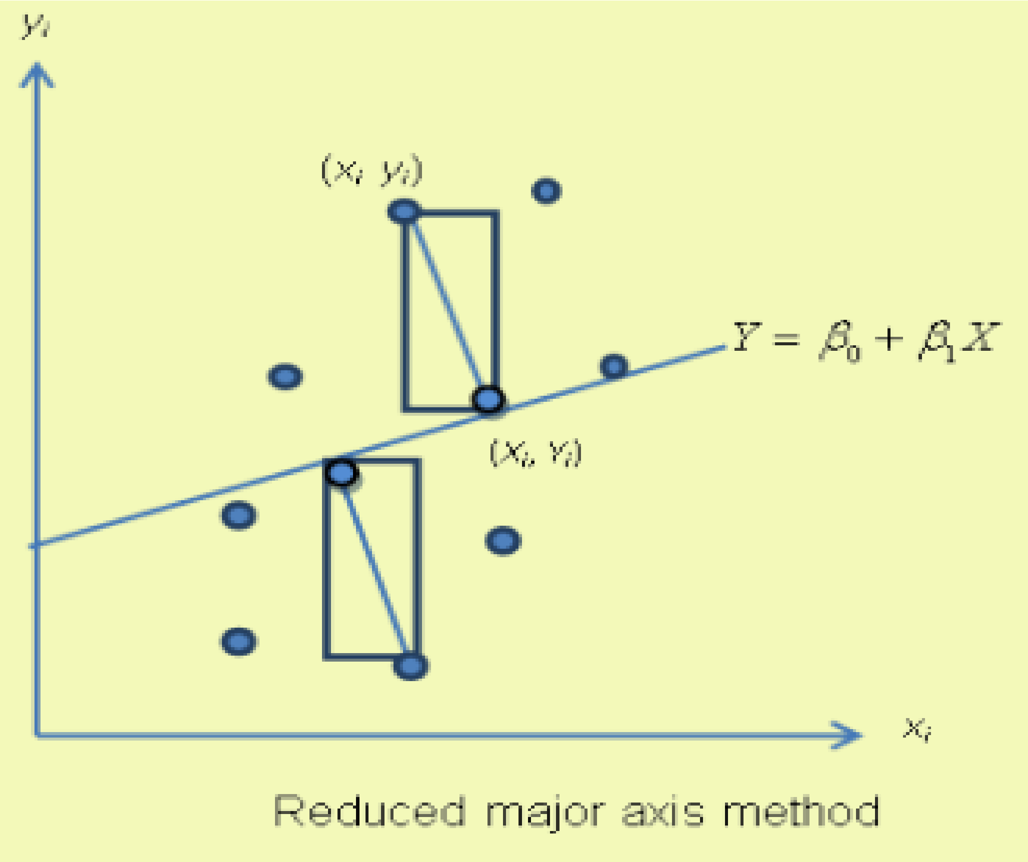


Major axis regression method

Least squares estimation

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:

Least squares estimation



Least squares estimation

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.

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.

Direct regression method

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

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

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

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

Direct regression method

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.

Direct regression method

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

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

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

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

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

Direct regression method

Further, we have

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

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

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

Direct regression method

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

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

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

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

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.

Direct regression method

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).

Direct regression method

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

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

Consistency

$$
\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y \\
&= (X'X)^{-1}X'(X\beta+\varepsilon) \\
&= \beta_0 + (X'X)^{-1}X'\varepsilon \\
&= \beta_0 + \left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n}
\end{aligned}
$$

Consider the last two terms. By assumption $\lim_{n\to\infty}\left(\frac{X'X}{n}\right) = Q_X \Rightarrow \lim_{n\to\infty}\left(\frac{X'X}{n}\right)^{-1} = Q_X^{-1}$, since the inverse of a nonsingular matrix is a continuous function of the elements of the matrix. Considering $\frac{X'\varepsilon}{n}$,

Consistency

$$\frac{X'\varepsilon}{n} = \frac{1}{n}\sum_{t=1}^{n} x_t\varepsilon_t$$

$$V(x_t\varepsilon_t) = x_t x_t' \sigma_0^2,$$

and

$$\mathcal{E}(x_t\varepsilon_t\varepsilon_s x_s') = 0, t \neq s.$$

Consistency

So the sum is a sum of independent, nonidentically distributed random variables, each with mean zero. Supposing that $V(x_t\varepsilon_t) < \infty, \forall t$, the KLLN implies

$$\frac{1}{n}\sum_{t=1}^{n} x_t\varepsilon_t \stackrel{a.s.}{\rightarrow} 0.$$

This implies that

$$\hat{\beta} \stackrel{a.s.}{\rightarrow} \beta_0.$$

Consistency

This is the property of *strong consistency:* the estimator converges almost surely to the true value. If we has used a weak LLN (defined in terms of convergence in probability), we would have *(simple, weak) consistency.*

- The consistency proof does not use the normality assumption.

QUESTIONS!

Download from
Dreamstime.com
This watermarked comp image is for previewing purposes only.
27436137
Alexander Maslennikov | Dreamstime.com