

Chapter 2

The simple linear regression model

Introduction: regression

Regression analysis is one of the most commonly used statistical techniques in social and behavioral sciences as well as in physical sciences which involves identifying and evaluating the relationship between a dependent variable and one or more independent variables, which are also called predictor or explanatory variables. It is particularly useful for assess and adjusting for confounding. Model of the relationship is hypothesized and estimates of the parameter values are used to develop an estimated regression equation. Various tests are then employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given values for the independent variables.

Linear regression explores relationships that can be readily described by straight lines or their generalization to many dimensions. A surprisingly large number of problems can be solved by linear regression, and even more by means of transformation of the original variables that result in linear relationships among the transformed variables.

When there is a single continuous dependent variable and a single independent variable, the analysis is called a **simple linear regression analysis**. This analysis assumes that there is a linear association between the two variables. **Multiple regression** is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

Independent variables are characteristics that can be measured directly; these variables are also called predictor or explanatory variables used to predict or to explain the behavior of the dependent variable.

Dependent variable is a characteristic whose value depends on the values of independent variables.

REGRESSION VERSUS CAUSATION

Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation. In the words of Kendall and Stuart, "A statistical relationship, however strong and however suggestive, can

never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other."

In the crop-yield example cited previously, there is no *statistical reason* to assume that rainfall does not depend on crop yield. The fact that we treat crop yield as dependent on rainfall (among other things) is due to nonstatistical considerations: Common sense suggests that the relationship cannot be reversed, for we cannot control rainfall by varying crop yield.

In all the examples cited in Section 1.2 the point to note is that **a statistical relationship in itself cannot logically imply causation**. To ascribe causality, one must appeal to a priori or theoretical considerations. Thus, in the third example cited, one can invoke economic theory in saying that consumption expenditure depends on real income.

REGRESSION VERSUS CORRELATION

Correlation quantifies the degree and direction to which two variables are related. Correlation does not fit a line through the data points. But simply is computing a correlation coefficient that tells how much one variable tends to change when the other one does. When r is 0.0, there is no relationship. When r is positive, there is a trend that one variable goes up as the other one goes up. When r is negative, there is a trend that one variable goes up as the other one goes down.

With correlation, it doesn't have to think about cause and effect. It doesn't matter which of the two variables is call dependent and which is call independent, if the two variables swapped the degree of correlation coefficient will be the same.

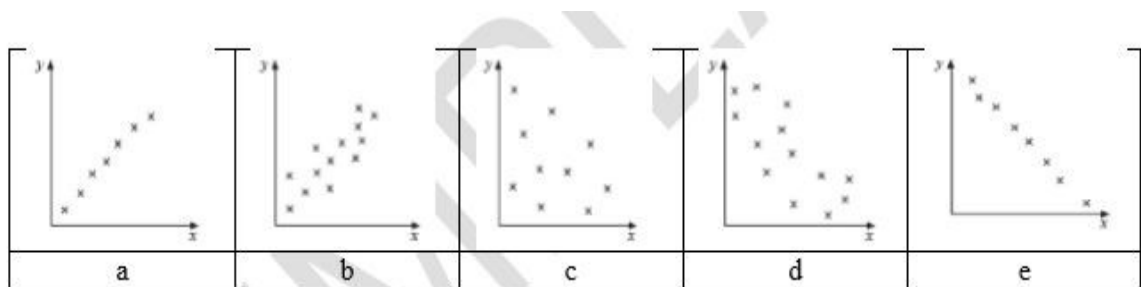
Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that

means that the change in one causes the change in the other. However, correlation does not imply causation. There may be an unknown factor that influences both variables similarly.

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. Although this correlation is fairly obvious your data may contain unsuspected correlations. You may also suspect there are correlations, but don't know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data.

- ❖ Correlation is Positive or direct when the values increase together, and
- ❖ Correlation is Negative when one value decreases as the other increases, and so called inverse or contrary correlation.



If the points plotted were all on a straight line we would have perfect correlation, but it could be positive or negative as shown in the diagrams above,

- a) Strong positive correlation between x and y . The points lie close to a straight line with y increasing as x increases.
- b) Weak, positive correlation between x and y . The trend shown is that y increases as x increases but the points are not close to a straight line
- c) No correlation between x and y ; the points are distributed randomly on the graph.
- d) Weak, negative correlation between x and y . The trend shown is that y decreases as x increases but the points do not lie close to a straight line

- e) Strong, negative correlation. The points lie close to a straight line, with y decreasing as x increases. Correlation can have a value: 3.

❖ **1 is a perfect positive correlation .**

❖ **0 is no correlation (the values don't seem linked at all) .**

❖ **-1 is a perfect negative correlation**

Linear Models and Regression Analysis

Suppose the outcome of any process is denoted by a random variable y , called as dependent (or study) variable, depends on k independent (or explanatory) variables denoted by $X_1, X_2, \dots, X_k$. Suppose the behaviour of y can be explained by a relationship given by

$$y = f(X_1, X_2, \dots, X_k, \beta_1, \beta_2, \dots, \beta_k) + \varepsilon$$

where f is some well-defined function and $\beta_1, \beta_2, \dots, \beta_k$ are the parameters which characterize the role and contribution of $X_1, X_2, \dots, X_k$, respectively. The term ε reflects the stochastic nature of the relationship between y and $X_1, X_2, \dots, X_k$ indicates that such a relationship is not exact in nature. When $\varepsilon = 0$, then the relationship is called the mathematical model otherwise the statistical model. The term "**model**" is broadly used to represent any phenomenon in a mathematical framework.

A model or relationship is termed as linear if it is linear in parameters and nonlinear if it is not linear in parameters. In other words, if all the partial derivatives of y with respect to each of the parameters $\beta_1, \beta_2, \dots, \beta_k$, are independent of the parameters, then the model is called a **linear model**. If any of the partial derivatives of y with respect to any of the $\beta_1, \beta_2, \dots, \beta_k$ is not independent of the parameters, then the model is called as nonlinear. Note that the linearity or non-linearity of the model is not described by the linearity or nonlinearity of explanatory variables in the model. Different statistical estimation procedures, e.g., method of maximum likelihood, principal of least squares, method of moments etc. can be employed to estimate the parameters of the model. The method of maximum likelihood needs further knowledge of the distribution of y . In contrast, the method of moments and the principle of least squares do not need any knowledge about the distribution of y .

The regression analysis is a tool to determine the values of the

parameters given the data on y and $X_1, X_2, \dots, X_k$. The literal meaning of regression is “to move in the backward direction”. Before discussing and understanding the meaning of “backward direction”, let us find which of the following statement is correct: S_1 : model generates data or S_2 : data generates model. Obviously S_1 is correct. It can be broadly thought that the model exists in nature but is unknown to the experimenter. When some values to the explanatory variables are provided, then the values for the output or study variable are generated accordingly, depending on the form of the function f and the nature of the phenomenon. So ideally, the pre-existing model gives rise to the data. Our objective is to determine the functional form of this model. Now we move in the backward direction. We propose to first collect the data on study and explanatory variables. Then we employ some statistical techniques and use this data to know the form of function f . Equivalently, the data from the model is recorded first and then used to determine the parameters of the model. The regression analysis is a technique which helps in determining the statistical model by using the data on study and explanatory variables. The classification of linear and nonlinear regression analysis is based on the determination of linear and nonlinear models, respectively. Consider a simple example to understand the meaning of “regression”. Suppose the yield of the crop (y) depends linearly on two explanatory variables, viz., the quality of fertilizer (X_1) and level of irrigation (X_2) as

$$y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

There exist the true values of β_1 and β_2 in nature but are unknown to the experimenter. Some values on y are recorded by providing different values to X_1 and X_2 . There exists some relationship between y and X_1, X_2 which gives rise to a systematically behaved data on y, X_1 and X_2 . Such a relationship is unknown to the experimenter. To determine the model, we move in the backward direction in the sense that the collected data is used to determine the parameters β_1 and β_2 of the model. In this sense,

such an approach is termed as regression analysis.

The theory and fundamentals of linear models lay the foundation for developing the tools for regression analysis that are based on valid statistical theory and concepts.

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 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

$$E(y) = \beta_0 + \beta_1 X \quad \text{and} \quad \text{Var}(y) = \sigma^2.$$

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

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

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

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

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) ($i = 1, \dots, n$) 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.

THE CLASSICAL LINEAR REGRESSION MODEL:

If our objective is to estimate β_1 and β_2 only, the method of OLS discussed in the preceding section will suffice. In regression analysis our objective is not only to obtain $\hat{\beta}_1$ and $\hat{\beta}_2$ but also to draw inference about the true β_1 and β_2 . For example, we would like to know how close $\hat{\beta}_1$ and $\hat{\beta}_2$ are to their counterparts in the population or how close $\hat{Y}_i$ is to the true $E(Y | X_i)$. To that end, we must not only specify the functional form of the model, but also make certain assumptions about the manner in which Y_i are generated. To see why this requirement is needed, look at the PRF: $Y_i = \beta_1 + \beta_2 X_i + u_i$. It shows that Y_i depends on both X_i and u_i . Therefore, unless we are specific about how X_i and u_i are created or generated, there is no way we can make any statistical inference about the Y_i and also, as we shall see, about β_1 and β_2 . Thus, the assumptions made about the X_i variable(s) and the error term are extremely critical to the valid interpretation of the regression estimates.

Assumption 1: Linear regression model. The regression model is linear in the parameters, $Y_i = \beta_1 + \beta_2 X_i + u_i$

We already discussed model in Chapter 1. Since linear-in-parameter regression models are the starting point of the CLRM, we will maintain this assumption throughout this. Keep in mind that the regressand and Y and the regressor X themselves may be nonlinear

Assumption 2: X values are fixed in repeated sampling. Values taken by the regressor X

are considered fixed in repeated samples. More technically, X is assumed to be *non stochastic*.

This assumption is implicit in our discussion of the PRF in Chapter 1. But it is very important to understand the concept of “fixed values in repeated sampling,” which can be explained in terms of our example given. Consider the various Y populations corresponding to the levels of income shown in that table. Keeping the value of income X fixed, say, at level

What all this means is that our regression analysis is conditional regression analysis, that is, conditional on the given values of the regressor(s) X .

Assumption 3: Zero mean value of disturbance u_i . Given the value of X , the mean, or expected, value of the random disturbance term u_i is zero. Technically, the conditional mean value of u_i is zero. Symbolically, we have

$$E(u_i | X_i) = 0$$

Assumption 3 states that the mean value of u_i , conditional upon the given X_i , is zero. Geometrically, this assumption can be pictured as in Figure 3.3, which shows a few values of the variable X and the Y populations associated with each of them. As shown, each Y population corresponding to a given X is distributed around its mean value (shown by the circled points on the PRF) with some Y values above the mean and some below it. The distances above and below the mean values are nothing but the u_i , and what requires is that the average or mean value of these deviations corresponding to any given X should be zero.

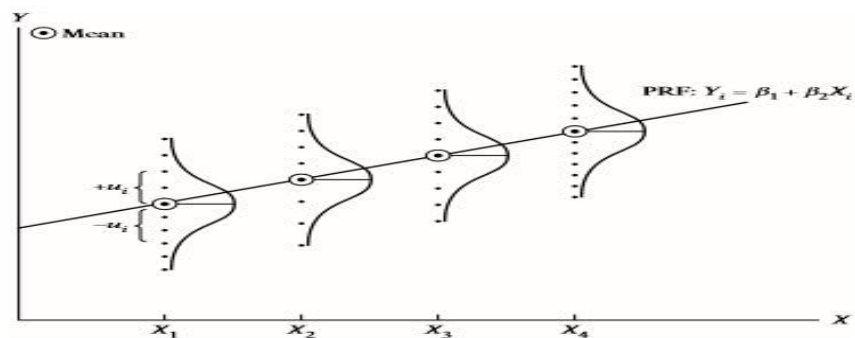


FIGURE 3.3 Conditional distribution of the disturbances u_i .

This assumption should not be difficult to comprehend in view of the discussion in Section 2.4 [see Eq. (2.4.5)]. All that this assumption says is that the factors not explicitly included in the model, and therefore subsumed in u_i , do not systematically affect the mean value of Y ; so to speak, the positive u_i values cancel out the negative u_i values so that their average or mean effect on Y is zero. In passing, note that the assumption $E(u_i | X_i) = 0$

Assumption 4: Homoscedasticity or equal variance of u_i . Given the value of X , the variance of u_i is the same for all observations. That is, the conditional variances of u_i are identical. Symbolically, we have

$$\begin{aligned} \text{var}(u_i | X_i) &= E[u_i - E(u_i | X_i)]^2 \\ &= E(u_i^2 | X_i) \quad \text{because of} \end{aligned}$$

Assumption 3

$$= \sigma^2$$

variance of u_i for each X_i (i.e., the conditional variance of u_i) is some positive constant number equal to σ^2 . Technically, represents the assumption of homoscedasticity, or *equal* (homo) *spread* (scedasticity) or *equal variance*. The word comes from the Greek verb *skedanime*, which means to disperse or scatter. Stated differently, means that the Y populations corresponding to various X values have the same variance. Put simply, the variation around the regression line (which is the line of average relationship between Y and X) is the same across the X values; it neither increases or decreases as X varies. Diagrammatically, the situation is as depicted in Fig

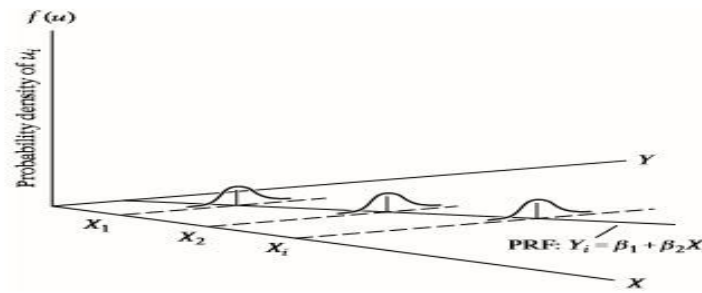


FIGURE 3.4 Homoscedasticity.

Assumption 5: No autocorrelation between the disturbances. Given any two X values,

X_i and X_j ($i \neq j$), the correlation between any two u_i and u_j ($i \neq j$) is zero. Symbolically,

$$\begin{aligned} \text{cov}(u_i, u_j | X_i, X_j) &= E \{ [u_i - E(u_i)] | X_i \} \{ [u_j - E(u_j)] | X_j \} \\ &= E(u_i | X_i)(u_j | X_j) \quad (\text{why?}) \\ &= 0 \end{aligned}$$

where i and j are two different observations and where cov means covariance.

In words, postulates that the disturbances u_i and u_j are uncorrelated. Technically, this is the assumption of no serial correlation, or no autocorrelation. This means that, given X_i , the deviations of any two Y values from their mean value do not exhibit patterns such as those shown in Figure 3.6a and b. In Figure 3.6a, we see that the u 's are positively correlated, a positive u followed by a positive u or a negative u followed by a negative u . In Figure 3.6b, the u 's are negatively correlated, a positive u followed by a negative u and vice versa.

Assumption 6: Zero covariance between u_i and X_i , or $E(u_i X_i) = 0$. Formally,

$$\begin{aligned} \text{cov}(u_i, X_i) &= E [u_i - E(u_i)][X_i - E(X_i)] \\ &= E [u_i (X_i - E(X_i))] \quad \text{since } E(u_i) = 0 \\ &= E(u_i X_i) - E(X_i)E(u_i) \quad \text{since } E(X_i) \text{ is non} \\ &\quad \text{stochastic} \\ &= E(u_i X_i) \quad \text{since } E(u_i) = 0 \end{aligned}$$

$$= 0 \quad \text{by assumption}$$

Assumption 6 states that the disturbance u and explanatory variable X are uncorrelated.

The rationale for this assumption is as follows: When we expressed the PRF as in , we assumed that X and u (which may represent the influence of all the omitted variables) have separate (and additive) influence on Y . But if X and u are correlated, it is not possible to assess their individual effects on Y . Thus, if X and u are positively correlated, X increases

when u increases and it decreases when u decreases. Similarly, if X and u are negatively correlated, X increases when u decreases and it decreases when u increases. In either case, it is difficult to isolate the influence of X and u on Y .

Assumption 6 is automatically fulfilled if X variable is nonrandom or nonstochastic and Assumption 3 holds, for in that case, $\text{cov}(u_i, X_i) = [X_i - E(X_i)] E[u_i - E(u_i)] = 0$. (Why?) But since we have assumed that our X variable not only is nonstochastic but also assumes fixed values in repeated samples, Assumption 6 is not very critical for us; it is stated here merely to point out that the regression theory presented in the sequel holds true even if the X 's are stochastic or random, provided they are independent or at least uncorrelated with the disturbances u_i .

Assumption 7: The number of observations n must be greater than the number of parameters to be estimated. Alternatively, the number of observations n must be greater than the number of explanatory variables.

This assumption is not so innocuous as it seems. In the hypothetical example of imagine that we had only the first pair of observations on Y and X (4 and 1). From this single observation there is no way to estimate the two unknowns, β_1 and β_2 . We need at least two pairs of observations to estimate the two unknowns.

Assumption 8: Variability in X values. The X values in a given sample must not all be the same. Technically, $\text{var}(X)$ must be a finite positive number.¹³

This assumption too is not so innocuous as it looks. Look at Eq. (3.1.6). If all the X values are identical, then $X_i = \bar{X}$ (Why?) and the denominator of that equation will be zero, making it impossible to estimate β_2 and therefore β_1 . Intuitively, we readily see why this assumption is important.

Assumption 9: The regression model is correctly specified. Alternatively, there is no specification bias or error in the model used in empirical analysis.

As we discussed in the Introduction, the classical econometric methodology assumes implicitly, if not explicitly, that the model used to test an economic theory is “correctly specified.” This assumption can be explained informally as follows. An econometric investigation begins with the specification of the econometric model underlying the phenomenon of interest. Some important questions that arise in the specification of the model include the following:

- (1) What variables should be included in the model?
- (2) What is the functional form of the model? Is it linear in the parameters, the variables, or both?
- (3) What are the probabilistic assumptions made about the Y_i , the X_i , and the u_i entering the model?

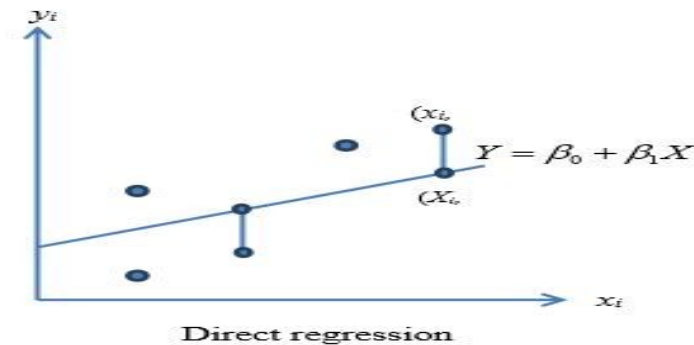
Assumption 10: There is no perfect multicollinearity. That is, there are no perfect linear relationships among the explanatory variables.

Least squares estimation

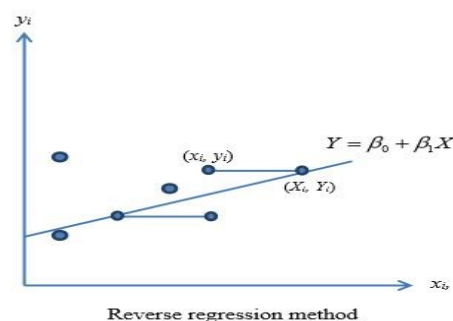
Suppose a sample of n sets of paired observations (x_i, y_i) ($i = 1, 2, \dots, n$) is available. These observations are assumed to satisfy the simple linear regression model, and so we can write

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

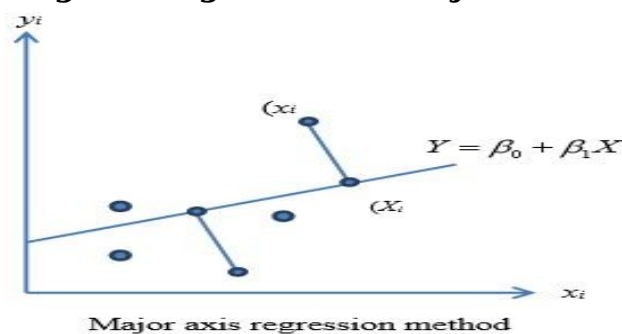
The principle of least squares estimates the parameters β_0 and β_1 by minimizing the sum of squares of the difference between the observations and the line in the scatter diagram. Such an idea is viewed from different perspectives. When the **vertical difference** between the observations and the line 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**.



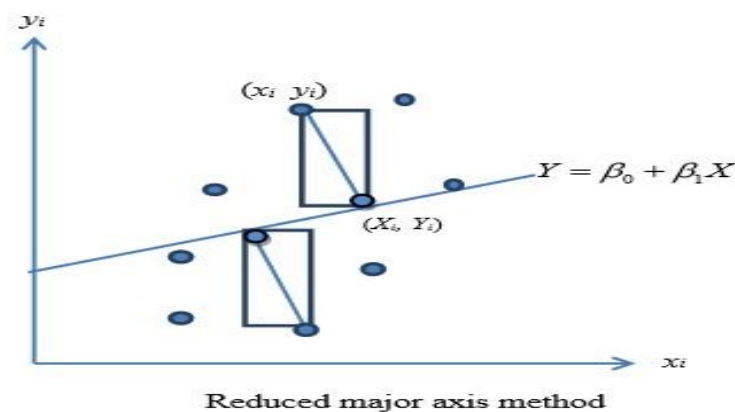
Alternatively, the sum of squares of the difference between the observations and the line in the horizontal direction in the scatter diagram can be minimized to obtain the estimates of β_0 and β_1 . This is known as a **reverse (or inverse) regression method**.



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 method is known as **orthogonal regression** or **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:



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 square 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 $i \neq j$ ($i, j = 1, 2, \dots, n$). 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**.

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

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

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

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

The solutions of β_0 and β_1 are obtained by setting

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

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

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

$$s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$