Simple Linear Regression Model

A simple linear regression model contains only one explanatory (or independent) variable, denoted by X_i, for $i = 1,2,\ldots,n$, and corresponding response variable Y_i. The simple linear regression model is given by the equation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \textbf{Model (3.1)}$$

Where:

Y_i is the value of the response variable in the ith element.

β_0 is the regression coefficient that gives the y-intercept of the regression line.

β_1 is the regression coefficient that gives the slope of the regression line.

X_i is the value of the independent variable in the ith element.

ε_i is a random error term or disturbance or stochastic term.

Model (3.1) is said to be ***simple***, ***linear in the parameters***, and ***linear in the independent variable***.

- It is *simple* in that there is only one independent variable.
- It is *linear in the parameters* because no parameters appears as an exponent or is multiplied or divided by another parameters.
- It is *linear in the independent variable* because this variable appears only in the first power.

A model which is linear in the parameters and the independent variable is also called a ***first-order model***.

Take Note!!!

The term *"linear"* regression model will always mean a regression that is linear in the parameters, the $\beta's$ (that is, the parameter are raised to the first power only). It may or may not in the explanatory variables, the $X's$.

Exercise:

Which of the following equations may be analyzed using linear regression method and explain why.

1. $Y_i = \beta_0 + \beta_1 \left(\frac{1}{X_i} \right) + \varepsilon_i$
2. $Y_i = \beta_0 + \beta_1 ln(X_i) + \varepsilon_i$
3. $ln(Y_i) = ln(\beta_0) + \beta_1 ln(X_i) + \varepsilon_i$
4. $ln(Y_i) = ln(\beta_0) + \beta_1 ln(X_i) + \varepsilon_i$, if we let $\alpha = ln(\beta_0)$
5. $ln(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$
6. $Y_i = \beta_0 + \beta_1 ln(X_i) - \beta_2 X_i^2 + \varepsilon_i$
7. $Y_i = \left(\frac{1}{1 + e^{\beta_0 + \beta_1 X_i + \varepsilon}} \right)$

Take Note!!!

A model that can be made linear in the parameters is called an ***intrinsically linear regression model***. It can be linearized by an appropriate transformation on the dependent variable.

Exercise:

Show that this model is intrinsically linear regression model.

$$Y_i = \left(\frac{1}{1 + e^{\beta_0 + \beta_1 X_i + \varepsilon}} \right)$$

Alternative Version of Model

Another modification sometimes helpful is to use for the independent variable the deviation $X_i - \bar{X}$ rather than X_i.

Exercise:

Show that

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i \quad \textbf{(3.2)}$$

is an alternative model version of $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.

Conditional Expected Value and Unconditional Expected Value

It is important to distinguish these conditional expected value from the unconditional expected value.

Sample questions:

- What is the expected value of weekly consumption on expenditure of a family? ***This is unconditional expected value.***
- What is the expected value of weekly consumption on expenditure of a family whose monthly income is say, Php 25,000? ***This is conditional expected value.***

The expected value of the distribution of Y given X_i is functional related to X_i. In simple terms, it tells how the mean or average response of Y varies with X.

Conditional mean $E(X_i)$ is a function of X_i, where X_i is a given value of X. Symbolically,

$$E(Y \mid X_i) = f(X_i) \text{ or } E(Y_i) = f(X_i) \quad \textbf{(3.3)}$$

(1.3) is the ***population regression function (PRF) or conditional expectation function (CEF)***, where $f(X_i)$ denotes some function of the explanatory variable X.

$$E(Y_i) = \beta_0 + \beta_1 X_i \quad \textbf{(3.4)}$$

(1.4) is the ***linear population regression model***.

Construction of Regression Models

1. Selection of independent variables.
2. Functional form of regression equation.
3. Scope of model.

Meaning of Regression Parameters

- β_1 indicates the change in the mean of the probability distribution of Y per unit increase in X.

  Sample Interpretation:

  The height coefficient in the regression equations is 115.6. This coefficient represents the mean increase of weight in kilograms for every additional one meter in height.

- If the scope of the model includes $x = 0$, β_0 gives the mean of probability distribution of Y at $x = 0$.

  Sample Interpretation:

  We would expect an average height of 42 cm for scrub in partial sun with no bacterial in the soil.

- When the scope of the model does not cover $x = 0$, β_0 does not have any particular meaning as a separate term in the regression model.

Important Features of Model

1. The observed value of Y in the ith elements is the sum of two components:

   1.1. The constant term: $\beta_0 + \beta_1 X_i$

   1.2. The random term or stochastic term: ε_i.

   Hence, Y_i is a random variable.

2. Since we deal with stochastic variable, we all know that ε_i has a probability distribution. ε_i have a probabilistic property:

   2.1. It has a mean: $E(\varepsilon_i) = 0$

   2.2. It has a variance: $\sigma^2(\varepsilon_i) = \sigma^2$

3. ε_i and ε_j are uncorrelated/independent so that the covariance $\sigma(\varepsilon_i, \varepsilon_j) = 0$ for all $i, j : i \neq j$.

4. The error terms are normally distributed.

5. Since, $E(\varepsilon_i) = 0$, it follows from $E(a + Y) = a + E(Y)$ that:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

   Thus, the response Y_i, when the level of X existing in the ith elements is X_i, comes from a probability distribution whose mean is $E(Y_i) = \beta_0 + \beta_i X_i$.

6. The observed value of Y_i in the ith elements exceeds or falls short of the value of the regression function by the error term amount ε_i.

7. The error terms ε_i are assumed to have constant variance σ^2. It therefore follows that the variance of the response Y_i is $\sigma^2(Y_i) = \sigma^2$.

   Since, using theorem $\sigma^2(a + Y) = \sigma^2(Y)$,

   we have: $\sigma^2(\beta_0 + \beta_i X_i + \varepsilon_i) = \sigma^2(\varepsilon_i) = \sigma^2$

   Thus, model 3.1 assumes that the probability distributions of Y have the same variance σ^2, regardless of the level of the independent variables X.

In summary, model 3.1 implies that the response variables observations Y_i come from probability distribution whose mean are $E(Y_i) = \beta_0 + \beta_i X_i$ and whose variance are σ^2, the same for all levels of X. Further, any two observations Y_i and Y_j are uncorrelated.

Significance of Random Error Term/Stochastic Term

The ε_i is a surrogate for all those variables that are omitted from the model but that collectively affect Y.

1. Vagueness of theory
2. Unavailability of data
3. Corevariables versus peripheral variables
4. Intrinsic randomness in human behaviors
5. Poor proxy variables
6. Principle of parsimony
7. Wrong functional form.

Take Note!!!

We should not exclude relevant and important variables just to keep the regression model simple.

Sample Regression Function (SRF)

Our task now is to estimate the population regression function on the basis of the sample information. Since we cannot estimate or in real situations we do not have the entire population for examination, we well estimate the values of the unknown β_0 and β_1 on the basis of observations on Y and X.

To estimate the population regression function, we will use the sample counter part, we can express the SRF in its stochastic form as follows:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad \textbf{(3.5)}$$

Where:

$\hat{Y}_i$ is estimator of $E(Y \mid X_i)$ or $E(Y_i)$

$\hat{\beta}_0$ is estimator of β_0

$\hat{\beta}_1$ is estimator of β_1

To sum up, then, we find our primary objective in regression analysis is to estimate the population regression function

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

on the basis of the sample regression function

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

In terms of the sample regression function, the observed Y_i can be expressed as:

$$Y_i = \hat{Y} + e_i$$

and in terms of the population regression function, it can be expressed as

$$Y_i = E(Y \mid X_i) + \varepsilon_i \text{ or } Y_i = E(Y_i) + \varepsilon_i$$

Residuals

The *i*th residual is the difference between the observed value Y_i and the corresponding fitted value $\hat{Y}_i$. This residual is denoted by e_i and is defines in general as follows:

$$\begin{aligned} e_i &= Y_i - \hat{Y}_1 \\ &= Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \\ &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \end{aligned}$$

We need to distinguish between the model random error term value $\varepsilon_i = Y_i - E(Y_i)$ and residual $e_i = Y_i - \hat{Y}_i$.

- $\varepsilon_i = Y_i - E(Y_i)$ involves the vertical deviation of Y_i from the unknown true regression line and hence in unknown.
- $e_i = Y_i - \hat{Y}_i$ is the vertical deviation of Y_i from the fitted value $\hat{Y}_i$ on the estimated regression line, and it is known.

Residuals are highly useful for studying whether a given regression model is appropriate for the data at hand.

Method of Least Squares

To find *"good"* estimators of the regression parameters β_0 and β_1, we shall employ the method of least squares.

The objective of the method of least squares is to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1, respectively, for which Q is a minimum.

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Least Squares Estimators

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \qquad \textbf{(3.6)}$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \qquad \textbf{(3.7)}$$

The equation 3.6 and 3.7 are called ***normal equations***. $\hat{\beta}_0$ and $\hat{\beta}_1$ are called point estimators of β_0 and β_1, respectively.

The least squares principle chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squares of the residuals

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

The estimators for β_0 and β_1 are obtained by using calculus to find the values that minimize the sum of squares of the residuals.

Least Square Estimator of β_0***:***

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \textbf{(3.8)}$$

Least Square Estimator of β_1***:***

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \qquad \textbf{(3.9)}$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \qquad \textbf{(3.10)}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2} \qquad \textbf{(3.11)}$$

Exercise:

Derive the least squares estimators (LSEs) of the parameters in the simple linear regression model.

Take Note!!!

The estimated regression function $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ have alternative model which can be written as:

$$\hat{Y}_i = \bar{Y} + \hat{\beta}_1 (X_i - \bar{X}) \qquad \textbf{(3.12)}$$

Exercise:

Show that $\hat{Y}_i = \bar{Y} + \hat{\beta}_1 (X_i - \bar{X})$ is an alternative model version of $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

Properties of Fitted Regression Line

The estimated regression line $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ fitted by the method of least squares has a number of properties worth noting.

1. The sum of the residual is zero.

$$\sum_{i=1}^{n} e_i = 0$$

2. The sum of squared residuals $\sum_{i=1}^{n} e_i^2$ is a minimum. This was the requirement to be satisfied in deriving the least squares estimators of the regression parameters.
3. The sum of observed values Y_i equals the same of the fitted values $\hat{Y}_i$.

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

4. The sum of the weighted residuals is zero when the residual in the ith element is weighted by the level of the predictor variable in the ith element.

$$\sum_{i=1}^{n} X_i e_i = 0$$

5. A consequence of properties 1 and 4 is the same of the weighted residual is zero when the residual in the ith element is weighted by the fitted of the response variable for the ith element.

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$$

6. The regression line always through the point $(\bar{x}, \bar{y})$. When $x = \bar{x}$, we have $\hat{Y} = \bar{Y}$.

Assignment

Derive the estimators of the parameters in the simple linear regression model using maximum likelihood method.

Prepared by:

KATRINA D. ELIZON
Department of Mathematics and Statistics
College of Science