LINEAR REGRESSION ANALYSIS

MODULE – II

Lecture - 3

Simple Linear Regression Analysis

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Properties of the direct regression estimators

Unbiased property

Note that $b_1 = \dfrac{s_{xy}}{s_{xx}}$ 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 $\sum_{i=1}^{n} k_i x_i = 1$,

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

Thus b_1 is an unbiased estimator of β_1 . Next

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

Thus b_0 is an unbiased estimators of β_0.

Variances

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

$$Var(b_1) = \sum_{i=1}^{n} k_i^2 Var(y_i) + \sum_i \sum_{j \neq i} k_i k_j Cov(y_i, y_j)$$

$$= \sigma^2 \frac{\sum_i (x_i - \bar{x})^2}{s_{xx}^2} \text{ (since } y_1, ..., y_n \text{ are independent)}$$

$$= \frac{\sigma^2 s_{xx}}{s_{xx}^2}$$

$$= \frac{\sigma^2}{s_{xx}}.$$

Similarly, 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

$$Cov(\bar{y}, b_1) = E\left[\{\bar{y} - E(\bar{y})\}\{b_1 - E(b_1)\}\right]$$

$$= E\left[\bar{\varepsilon}(\sum_i k_i y_i - \beta_1)\right]$$

$$= \frac{1}{n} E\left[(\sum_i \varepsilon_i)(\beta_0 \sum_i k_i + \beta_1 \sum_i k_i x_i + \sum_i k_i \varepsilon_i) - \beta_1 \sum_i \varepsilon_i\right]$$

$$= \frac{1}{n}[0 + 0 + 0 + 0]$$

$$= 0$$

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

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

$$= -\frac{\bar{x}}{s_{xx}}\sigma^2.$$

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

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

where

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

Estimation of σ^2

The estimator of σ^2 is obtained from residual sum of squares as follows. Assuming that Since y_i is normally distributed, so 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 dependent estimate** of σ^2 .

Estimate 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 $\hat{\sigma}^2 = s^2$ 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 $\varepsilon_i (i = 1,...,n)$. 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, model $y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ has a centered version as follows:

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

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

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

and

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

respectively.

Thus the form of the estimate of slope parameter β_1 remains same in 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

$$E(b_0^*) = \beta_0^*, \quad E(b_1) = \beta_1,$$

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

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

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

and the predicted values are

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

Note that in 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

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

For example, in analyzing the relationship between illumination of bulb (*y*) and electric current (*X*), the illumination of bulb is zero when current 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 ensures 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:

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

$$= \frac{\sum_{i=1}^{n} x_i^2 \beta_1}{\sum_{i=1}^{n} x_i^2}$$

$$= \beta_1.$$

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