Estimation of regression parameters

Least Squares Method

The simple linear regression model is represented by the following equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i, i = 1,2,\ldots,n$$

One of its uses is to predict future values of the variable y based on X, so the parameters of this model must be estimated β_0, β_1.

The least squares method is one of the most widely used methods for estimating model parameters when model assumptions are available.
The basis of this method is to make the sum of squared errors as small as possible.
Based on the above model:

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i$$

So

$$e_i = y_i - \beta_0 - \beta_1 x_{i1}$$

By finding the sum of squares of the errors (or residuals), let Q be:

$$Q = \sum e_i^2 = \sum (y_i - \beta_0 - \beta_1 x_{i1})^2$$

Taking the partial derivative of the above equation with respect to (β_1, β_0), we get:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 x_{i1})$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum (y_i - \beta_0 - \beta_1 x_{i1}) x_{i1}$$

Where (β_1, β_0) are the estimated parameters.
By setting both equations equal to zero, we get:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum X_{i1} = \sum y_i \qquad \ldots \quad (1)$$

$$\hat{\beta}_0 \sum X_{i1} + \hat{\beta}_1 \sum X_{i1}^2 = \sum X_{i1} y_i \qquad \ldots \quad (2)$$

The above two equations are called the normal equations.

Through equation (1):

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum X_{i1} = \sum y_i$$

We get:

$$n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum X_{i1}$$
$$\hat{\beta}_0 = \frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum X_{i1}}{n}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

We substitute $\hat{\beta}_0$ in equation (2) to get the value of $\hat{\beta}_1$ as follows:

$$(\frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum X_{i1}}{n})\sum X_{i1} + \hat{\beta}_1 \sum X_{i1}^2 = \sum X_{i1} y_i$$
$$\frac{(\sum y_i)(\sum X_{i1}) - \hat{\beta}_1 (\sum X_{i1})^2}{n} + \hat{\beta}_1 \sum X_{i1}^2 = \sum X_{i1} y_i$$
$$(\sum y_i)(\sum X_{i1}) - \hat{\beta}_1 (\sum X_{i1})^2 + n\hat{\beta}_1 \sum X_{i1}^2 = n\sum X_{i1} y_i$$

By transferring the terms containing the estimated parameter $\hat{\beta}_1$, we get:

$$\hat{\beta}_1 \left[n\sum X_{i1}^2 - (\sum X_{i1})^2 \right] = n\sum X_{i1} y_i - (\sum y_i)(\sum X_{i1})$$

$$\therefore \hat{\beta}_1 = \frac{n\sum X_{i1} y_i - (\sum y_i)(\sum X_{i1})}{n\sum X_{i1}^2 - (\sum X_{i1})^2}$$

Dividing by n for the numerator and denominator, we get:

$$\therefore \hat{\beta}_1 = \frac{\sum X_{i1} y_i - \frac{(\sum y_i)(\sum X_{i1})}{n}}{\sum X_{i1}^2 - \frac{(\sum X_{i1})^2}{n}} = \frac{S_{Xy}}{S_{XX}}$$

were:

$$\begin{aligned} S_{XX} &= \sum X_{i1}^2 - \frac{(\sum X_{i1})^2}{n} \\ &= \sum X_{i1}^2 - n\bar{X}^2 \\ &= \sum (X_i - \bar{X})^2 \\ &= \sum (X_i - \bar{X}) X_i \end{aligned}$$

and

$$S_{Xy}=\sum X_{i1}y_i-\frac{(\sum y_i)(\sum X_{i1})}{n}$$

$$=\sum X_{i1}y_i-n\bar{X}\ \bar{y}$$

$$=\sum(X_i-\bar{X})y_i$$

$$=\sum(y_i-\bar{y})X_i$$

$$=\sum(X_i-\bar{X})(y_i-\bar{y})$$

Properties of the regression line equation

1- The sum of the observed values equals the sum of the predicted values.

Proof:

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}$$

$$\hat{y}_i=(\bar{y}-\hat{\beta}_1\bar{X})+\hat{\beta}_1x_{i1}$$

$$\hat{y}_i=\bar{y}-\hat{\beta}_1\bar{X}+\hat{\beta}_1x_{i1}$$

$$\hat{y}_i=\bar{y}+\hat{\beta}_1(x_{i1}-\bar{X})$$

Taking the sum of both sides:

$$\sum\hat{y}_i=n\bar{y}+\hat{\beta}_1\sum(x_{i1}-\bar{X})$$

Since: $\sum(x_i-\bar{X})=0$, So:

We have $\sum(x_i-\bar{X})=0$ so:

$$\sum\hat{y}_i=n\bar{y}$$

$$\sum\hat{y}_i=\cancel{n}\frac{\sum y_i}{\cancel{n}}$$

$$\therefore\sum\hat{y}_i=\sum y_i$$

2−The sum of random errors equals zero, i.e.:

$$\sum e_i = 0$$

Proof:

We have

$$e_i = y_i - \hat{y}_i$$

Taking the sum of both sides, we get:

$$\sum e_i = \sum (y_i - \hat{y}_i)$$

$$\sum e_i = \sum y_i - \sum \hat{y}_i$$

We have $\sum \hat{y}_i = \sum y_i$ So:

$$\sum e_i = 0$$

3− The sum of the errors (or residuals) weighted by the corresponding Xi values equals zero. That is:

$$\sum e_i X_i = 0$$

Proof:

$$\sum e_i X_i = \sum X_i (y_i - \hat{y}_i)$$

Since

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$$

So

$$\sum e_i X_i = \sum X_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} \quad))$$

Since

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

So

$$\sum e_i X_i = \sum X_i (y_i - ((\bar{y} - \hat{\beta}_1 \bar{X}) + \hat{\beta}_1 x_{i1}))$$

After simplifying and factoring the estimated parameter $\hat{\beta}_1$, we get:

$$\begin{aligned}\sum e_i X_i &= \sum X_i (y_i - (\bar{y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 x_{i1})) \\ &= \sum X_i (y_i - (\bar{y} + \hat{\beta}_1 (x_{i1} - \bar{X})) \\ &= \sum X_i ((y_i - \bar{y}) - \hat{\beta}_1 (x_{i1} - \bar{X}))\end{aligned}$$

After inserting $\sum X_i$ on the bracket, we get:

$$\sum e_i X_i = \sum X_i (y_i - \bar{y}) - \hat{\beta}_1 \sum X_i (x_{i1} - \bar{X}))$$

Since

$$S_{Xy} = \sum (y_i - \bar{y}) X_i$$
$$S_{XX} = \sum (X_i - \bar{X}) X_i$$

We get

$$\sum e_i X_i = S_{Xy} - \hat{\beta}_1 S_{XX}$$

Since

$$\hat{\beta}_1 = \frac{S_{Xy}}{S_{XX}}$$

So

$$\sum e_i X_i = S_{Xy} - \frac{S_{Xy}}{\cancel{S_{XX}}} \cancel{S_{XX}}$$

$$\sum e_i X_i = S_{Xy} - S_{Xy}$$

$$\therefore \sum e_i X_i = 0$$

4− Weighing the errors with the corresponding y_i values produce the sum of the squares of the errors.

Proof:

$$\sum y_i e_i = \sum y_i (y_i - \hat{y}_i)$$

$$\because \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\therefore \sum y_i e_i = \sum y_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$

$$\because \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}$$

$$\therefore \sum y_i e_i = \sum y_i (y_i - ((\bar{y} - \hat{\beta}_1 \bar{X}) + \hat{\beta}_1 x_i)$$

$$\sum y_i e_i = \sum y_i (y_i - (\bar{y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 x_i)$$

Taking a common factor of $\hat{\beta}_1$and opening the parentheses we get:

$$\sum y_i e_i = \sum y_i (y_i - (\bar{y} + \hat{\beta}_1 (x_i - \bar{X}_1))$$

$$\sum y_i e_i = \sum y_i (y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{X}))$$

$$\sum y_i e_i = \sum y_i ((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{X}))$$

Multiplying $\sum y_i$ by the brackets, we get:

$$\sum y_i e_i = \sum y_i (y_i - \bar{y}) - \hat{\beta}_1 \sum y_i (x_i - \bar{X})$$

$$\because S_{yy} = \sum y_i (y_i - \bar{y})$$

$$\because S_{Xy} = \sum y_i (x_i - \bar{X})$$

$$\therefore \sum y_i e_i = S_{yy} - \hat{\beta}_1 S_{Xy}$$

$$\therefore SSe = S_{yy} - \hat{\beta}_1 S_{Xy}$$

Thus, the sum of squares of errors (residuals) is:

SSe =Residual sum of squares