Multiple linear regression model

A **multiple linear regression model** is a statistical technique used to predict a dependent variable based on multiple independent variables. It extends simple linear regression by considering multiple predictors.

Formulation

The multiple linear regression model with two predictors can be written as:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

Where:

- $\mathbf{y}$ is the vector of dependent variable observations, with size $n \times 1$, where n is the number of observations (or samples).
- $\mathbf{X}$ is the matrix of independent variables (or regressors), with size $n \times p$, where p is the number of regressors. In this case, with two regressors (including the intercept term), the matrix will have one column for the intercept (1), one for x_1, and one for x_2, so $\mathbf{X}$ will have size $n \times 3$.
- β is the vector of coefficients (parameters to estimate), with size $p \times 1$. In this case, β will be a vector of size 3×1 (including the intercept term).
- ε is the vector of errors (or residuals), with size $n \times 1$, representing the difference between the observed and predicted values of the model.

For the case with two regressors (including the intercept), the matrix $\mathbf{X}$ will look like this:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}_{(n\times 3)}$$

And the vector β will be:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{(3\times 1)}$$

Thus, the model becomes:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{(n\times 1)} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}_{(n\times 3)} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{(3\times 1)} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{(n\times 1)}$$

OLS estimation

The goal is to find the line (or hyperplane in higher dimensions) that best fits the data by minimizing the **sum of squared errors** (cost function).

$$\varepsilon = \mathbf{y} - \mathbf{X}\beta$$

The cost function to minimize is the sum of squared errors:

$$J(\beta) = \varepsilon^\top \varepsilon = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$$

To find the minimum of the cost function, it is necessary to calculate the derivative of $J(\beta)$ with respect to β and set it to zero.

Expand the cost function:

$$J(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\beta - \beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta$$

Since $\mathbf{y}^\top \mathbf{X}\beta$ is a scalar, it follows that $\mathbf{y}^\top \mathbf{X}\beta = \beta^\top \mathbf{X}^\top \mathbf{y}$. So the cost function becomes:

$$J(\beta) = \mathbf{y}^\top \mathbf{y} - 2\beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta$$

Now calculate the derivative of $J(\beta)$ with respect to β:

$$\frac{\partial J(\beta)}{\partial \beta} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\beta$$

Set the derivative equal to zero to find the estimated coefficients:

$$-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\beta = 0$$

Solve for β:

$$\mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top \mathbf{y}$$

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

This is the **Ordinary Least Squares (OLS)** formula for estimating the coefficients β. This vector contains the estimates of β_0, β_1, and β_2, which are the coefficients of the multiple linear regression model.

Goodness of fit measure

The coefficient of determination R^2 is a measure of how well the model fits the data. It is calculated as:

$$R^2 = 1 - \frac{\text{Sum of Squared Errors (SSE)}}{\text{Total Sum of Squares (SST)}}$$

Where:

- The **Total Sum of Squares** (SST) measures the total variability in the data relative to the mean of y:

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

  with $\bar{y}$ being the mean of y.

- The **Sum of Squared Errors** (SSE) measures the variability not explained by the model, i.e., the sum of the squared differences between the observed values and the predicted values:

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  where $\hat{y}_i$ is the predicted value for y_i.

Interpretation of R^2

The value of R^2 ranges from 0 to 1:

- $R^2 = 1$ means the model perfectly explains the variability in the data.
- $R^2 = 0$ means the model explains none of the variability in the data.

A high R^2 value indicates that a large portion of the variability in y is explained by the independent variables in the model, while a low value suggests the model has limited predictive power.

Gauss–Markov assumptions

The Gauss-Markov assumptions are a set of conditions under which the Ordinary Least Squares (OLS) estimator $\hat{\beta}_{\mathrm{OLS}}$ is the Best Linear Unbiased Estimator (BLUE).

Zero conditional mean

The expected value of the residuals conditional on the independent variables should be zero:

$$\mathbb{E}(\varepsilon|\mathbf{X}) = 0$$

This assumption ensures the residuals are random and not linked to the independent variables. In other words, the regressors must be uncorrelated with the residuals i.e.

$$\mathrm{Cov}(x_i, \varepsilon_i) = 0 \quad \forall i$$

Homoscedasticity

The residuals should have constant variance across all observations.

$$\mathrm{Var}(\varepsilon|\mathbf{X}) = \sigma\mathbf{I}$$

where $\mathbf{I}$ is the identity matrix. This implies that the residuals are equally spread out across all levels of the independent variables.

No autocorrelation

The residuals should be uncorrelated with each other. This means that for all $i \neq j$, the covariance between the residuals ε_i and ε_j should be zero:

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j | \mathbf{X}) = 0 \quad i \neq j$$

Normality of residuals (optional)

The residuals should be normally distributed:

$$\varepsilon \sim N(0, \sigma \mathbf{I})$$

While this assumption is not necessary for the OLS estimator to be BLUE, it is often assumed for the purpose of hypothesis testing.