1.2. Multiple Linear Regression

In practice, it is not usual that there is only one independent (input) variable. Multiple independent variables together can influence the target (output) variable and in different directions.

Suppose you have to estimate the price of a certain house you want to buy. You know the floor area, the age of the house, its distance from your workplace, the crime rate of the place, etc. Now, some of these factors will affect the price of the house positively. For example, more the area, the more the price. On the other hand, factors like distance from the workplace, and the crime rate can influence your estimate of the house negatively

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

Multiple regression also allows you to determine the overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained.

In Multiple Linear Regression, there is one quantitative response i.e., dependent variable (Y) and more than one predictor or independent variable such as $X_1, X_2, \dots, X_k$.

A population model for a multiple linear regression model that relates a y-variable to k x-variables is written as

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon}$$

Where β_0 is the Y intercept of the regression surface and each $\beta_1, \beta_2, \dots, \beta_k$, is the slope of the regression surface with respect to each variable $X_1, X_2, \dots, X_k$.

The word "linear" in "multiple linear regression" refers to the fact that the model is linear in the parameters, $\beta_1, \beta_2, \dots, \beta_k$. This simply means that each parameter multiplies an X-variable, while the regression function is a sum of these "parameter times X-variable" terms. Each X-variable can be a predictor variable or a transformation of predictor variables (such as the square of a predictor variable or two predictor variables multiplied together).

Note that *even β_0 represents a "parameter times X-variable" term if you think of the X-variable that is multiplied by β_0 as being the constant function "1."*

Interpretation of the Model Parameters:

Each β coefficient represents the change in the mean response, $E(Y)$, per unit increase in the associated predictor variable when all the other predictors are held constant. For example, β_1 represents the change in the mean response, $E(Y)$, per unit increase in X_1 when X_2, X_3, ..., X_k are held constant.

The intercept term, β_0, represents the mean response, $E(Y)$, when all the predictors $X_1, X_2, \ldots, X_k$, are all zero (which may or may not have any practical meaning).The estimates of the β coefficients are the values that minimize the sum of squared errors for the sample. The letter b is used to represent a sample estimate of a β coefficient. Thus b_0 is the sample estimate of β_0, b_1 is the sample estimate of β_1, and so on.

Thus, a **fitted (or predicted) value** is calculated as:

$$\boldsymbol{\hat{y}_i = b_0 + b_1 x_{i\,1} + b_2 x_{i\,2} + \ldots + b_k x_{i\,k}}$$

The subscript i refers to the i^{th} individual or unit in the data. In the notation for the x-variables, the subscript following i simply denotes which x-variable it is.

A residual (error) term is calculated as $e_i = y_i - \hat{y}_i$, the difference between an actual and a predicted value of y.

Finding the intercept and coefficients:

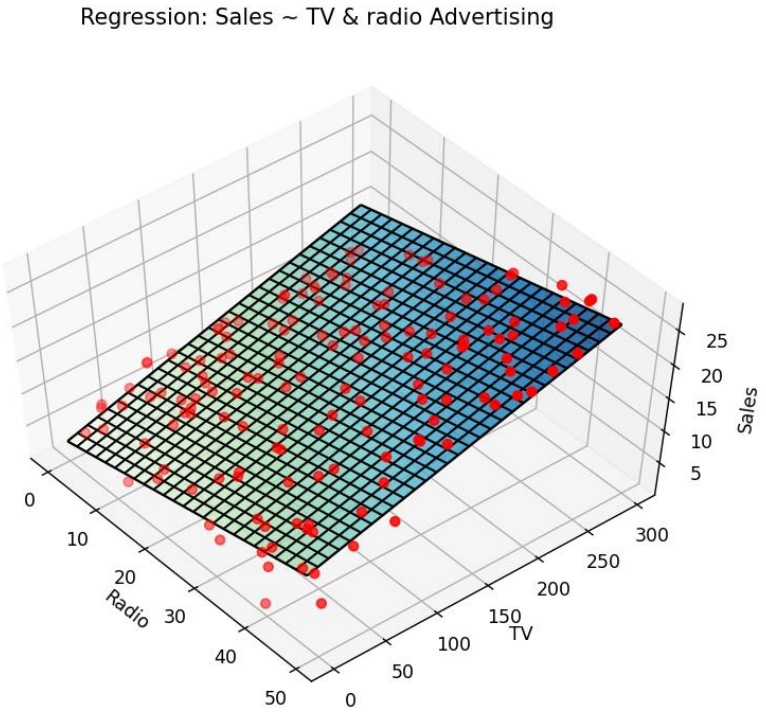
Finding the values of these constants(β) is what regression model does by using Ordinary Least Squares (OLS) procedure.

This method seeks to minimize the sum of the squared residuals (errors) and fitting the best line, plane or hyperplane (depending on the number of input variables**).

This means that given a regression line (line of best fit) through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together.

$$Sum\ of\ Square\ of\ Residual\ or\ SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

This is the quantity that ordinary least squares seeks to minimize.



***In the case of two predictors, the estimated regression equation yields a plane (as opposed to a line in the simple linear regression setting). For more than two predictors, the estimated regression equation yields a hyperplane.*

OLS approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients. It means that all of the data must be available and you must have enough memory to fit the data and perform matrix operations.

Matrix Formulation of the Multiple Regression Model

In the multiple regression setting, because of the potentially large number of predictors, it is more efficient to use matrices to define the regression model.

let's start with the simple case first. Consider the following simple linear regression function:

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

Consider now writing an equation for each observation:

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$

$$\begin{array}{cccc} . & . & . & . \\ . & . & . & . \end{array}$$

$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$$

Well, that's a pretty inefficient way of writing it all out! As you can see, there is a pattern that emerges. By taking advantage of this pattern, we can instead formulate the above **simple linear regression function in matrix notation**:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\boldsymbol{Y = X\beta + \varepsilon}$$

That is, instead of writing out the *n* equations, using matrix notation, our simple linear regression function reduces to a short and simple statement:

$$Y = X\,\beta + \varepsilon$$

- $\boldsymbol{X}$ is an $n \times 2$ matrix called as the **design matrix**
- $\boldsymbol{Y}$ is an $n \times 1$ column vector called as **the response vector**
- $\boldsymbol{\beta}$ is a 2×1 column vector called as the **vector of parameters**
- and $\boldsymbol{\varepsilon}$ is an $n \times 1$ column vector called as the **error vector**

Least squares estimates in matrix notation:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (X'X)^{-1}X'Y$$

where, $(X'X)^{-1}$ is the inverse of the $X'X$ matrix, and X' is the transpose of the X matrix.

In case of more than one independent variable i.e., multiple linear regression:

$$Y_i = \beta_0 + \beta_1 X_{i\,1} + \beta_2 X_{i\,2} + \ldots + \beta_k X_{i\,k} + \varepsilon_i$$

Design Matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix}$$

Coefficient matrix β:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

Least Squares Solution

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Fitted (predicted) values for the mean of Y are

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

One important matrix that appears here is the so-called "hat matrix," $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, since it puts the hat on $\boldsymbol{Y}$.

Model Assumptions:

The four assumptions that comprise the multiple linear regression model generalize the simple linear regression model conditions to take account of the fact that we now have multiple predictors:

1. **Linearity**

   First, multiple linear regression requires the relationship between the independent and dependent variables to be linear. The linearity assumption can best be tested with scatterplots. Scatterplots can show whether there is a linear or curvilinear relationship.

   If the relationship displayed in your scatterplots and partial regression plots are not linear, you will have to either run a non-linear regression analysis or "transform" your data.

2. **No (perfect) Collinearity**

   Multiple Linear Regression assumes that there is NO multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. This leads to problems with understanding which independent variable contributes to the variance explained in the dependent variable, as well as technical issues in calculating a multiple regression model. *Multicollinearity can be present in the model, as long as it is not 'perfect'*. We

can detect pairs of highly correlated variables by examining the correlation matrix for high absolute values.

Try removing one of the correlated predictors from the model, or combining them into a single predictor if any pair of independent variables are found to be highly correlated.

3. **Residual must be Normally Distributed with mean zero**
   The multiple linear regression analysis requires that the errors between observed and predicted values (i.e., the residuals of the regression or e_i) should be approximately normally distributed with mean zero. A histogram (with a superimposed normal curve) can be used for visually checking this assumption. The histogram must symmetric and centred at zero.

4. **Homoscedasticity**

   This assumption states that the variance of error terms is similar (constant) across the values of the independent variables.

   A scatter plot of residuals (e_i) versus predicted values $(\hat{y}_i)$ is good way to check for homoscedasticity. Thus, we make a plot of residual values on the y-axis and the predicted values on the x-axis. There should be no clear pattern in the distribution; if there is a cone-shaped pattern or funnel shape in residual plot (as shown below), the data is heteroscedastic.

Feature Selection:

How do I decide which features to include in a linear model? Which one or which two are important? Are all of them important?

To find this out, we will perform Feature Selection or variable selection. Now one way of doing this is trying all possible combinations of independent variables. Another idea is strategically creating a model by checking whether the R-squared as you add new features.

This idea is called Forward Selection and the steps to follow are:

1. We start with a model without any predictor and just the intercept term.

2. We then perform simple linear regression for each predictor to find the best performer (greatest R^2).

3. We then add another variable to it and check for the best 2-variable combination again by checking greatest R^2 .

4. After that the best 3-variable combination is checked, and so on. The approach is stopped when some stopping rule is satisfied.

Example: Sales prediction from advertising expenditure

We will again consider the Advertising Data, which we used in Simple Linear Regression.

The advertising data set consists of the sales of a product in 200 different markets, along with advertising budgets for three different media: TV, radio, and newspaper. Here's how it looks like:

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

**Sales (in '000 units) & Advertising budget (in '000 $)

The first row of the data says that the advertising budgets for TV, radio, and newspaper were $230.1k, $37.8k, and $69.2k respectively, and the corresponding number of units that were sold was 22.1k (or 22,100).

In Simple Linear Regression, we can saw how each advertising medium affects sales when applied without the other two media. However, in practice, all three might be working together to impact net sales. We did not consider the combined effect of these media on sales.

Multiple Linear Regression solves the problem by taking account of all the variables in a single expression. Hence, our Linear Regression model can now be expressed as:

$$\boldsymbol{sales} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} * \boldsymbol{TV} + \boldsymbol{\beta_2} * \boldsymbol{radio} + \boldsymbol{\beta_3} * \boldsymbol{newspaper}$$

Finding the values of these constants (β) is what regression model does by minimizing the error function and fitting the best line (or plane or hyperplane). We get the following result after using OLS:

$$\boldsymbol{sales} = 2.89 + 0.044 * TV + 0.199 * radio + 0.001 * newspaper$$

Interpretation of $\boldsymbol{\beta}$ coefficients :

If we fix the budget for TV & Newspaper, then increasing the radio budget by $1000 will lead to an increase in sales by around 199 units (0.199*1000).

Similarly, by fixing the radio & newspaper, we infer an approximate rise of 44 units of products per $1000 increase in the TV budget.

However, for the newspaper budget, since the coefficient is quite negligible (close to zero), it's evident that the newspaper is not affecting the sales.