What is Linear Regression?

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting straight line through the data points that minimizes the sum of the squared differences (residuals) between the observed values and the values predicted by the line.

What is Simple Linear Regression ?

- Simpler linear regression is a statistical technique used for finding the existence of an association relationship between a dependent variable(outcome or response var) and independent variable(predictor var or feature).
- We can only establish that the change in the value of outcome variable(Y) is associated with the change in the value of feature (X), that is, regression cannot be used for establish the relationship between two variables.

Steps to Build Regression Model :

- Building a regression model is a iterative process and several iteration may be required before finalizing the appropriate model.

1) Collect Data 2) Pre-Process the Data 3) Dividing Data into Training & Test Data 4) Perform Descriptive Analytics or Data Exploration 5) Build the Model 6) Perform Model Diagnostics 7) Validate the Model and Measure the Model Accuracy 8) Decide on Model Deployment

Building Simple Linear Regression Model

Linear Relationship: The assumption that there is a linear relationship between the independent variables and the dependent variable. This relationship can be expressed with the equation of a line:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

where:

- Y is the dependent variable.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables $X_1, X_2, \ldots, X_n$.
- ϵ is the error term (residual).

Residual

1. **Prediction (Predicted Value)**: This is the value that your linear regression model estimates for a given data point. It's found using the equation of the line you fit to the data.
2. **Actual Value (Observed Value)**: This is the actual value of the dependent variable for that data point, as recorded in your data.
3. **Residual**: The residual is the difference between the actual value and the predicted value. It tells you how far off your model's prediction is from the actual data point.

The formula for the residual e_i for a specific data point i is:

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

where:

- y_i is the actual value of the dependent variable for the ith data point.
- $\hat{y}_i$ is the predicted value from the linear regression model for the ith data point.

In other words:

- If the residual is positive, it means the actual value is higher than the predicted value.
- If the residual is negative, it means the actual value is lower than the predicted value.
- If the residual is zero, it means the actual value matches the predicted value exactly.

Residuals are important because they give you insight into how well your model is performing. By examining the residuals, you can identify patterns or trends that your model might be missing, check if the assumptions of linear regression are met, and improve your model if necessary.

Sum Of Squared Errors

1. **Residuals**: As mentioned earlier, residuals are the differences between the actual values and the predicted values for each data point.
2. **Squaring the Residuals**: To avoid the issue of positive and negative residuals canceling each other out, we square each residual. This ensures that all differences are treated as positive values.
3. **Summing the Squared Residuals**: We then add up all these squared residuals to get a single number, which represents the total error in the model's predictions.

The formula for SSE is:

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

where:

- y_i is the actual value of the dependent variable for the ith data point.
- $\hat{y}_i$ is the predicted value from the linear regression model for the ith data point.
- n is the total number of data points.

In simpler terms:

- For each data point, calculate the difference between the actual value and the predicted value (this is the residual).
- Square this difference.
- Add up all these squared differences for all data points.

The resulting sum is the SSE. A lower SSE indicates that the model's predictions are closer to the actual values, meaning the model is performing well. Conversely, a higher SSE indicates greater discrepancies between the predicted and actual values, suggesting the model may not be fitting the data well.

In [ ]: