

Assignment 4

1. What is the difference between a univariate and a multivariate function?

Univariate involves the analysis of a single variable while multivariate analysis examines two or more variables.

- Most multivariate analysis involves a dependent variable and multiple independent variables.

Uni-variate analysis can be performed using:

- Measure of central tendency.
- Measure of dispersion.
- Frequency distribution helps us understand the distribution of data.

Multivariate analysis can be performed using:

- Multiple linear regression.
- Principle component analysis.

2. What are the various ways to check the applicability of a particular regression model on a dataset?

Evaluation of residuals help us visualize whether linear regression is applicable for our dataset or not.

For Linear Regression -The histogram should look approximately normal and the scatterplot of residuals should show random scatter.

For linear models

- Only compare linear models (**Linear, Lasso, Ridge Regression etc.**) for the same dataset.
- Find a model with a **high adjusted R²**.
- Make sure this model has equally distributed residuals around zero.
- Make sure the errors of this model are within a small bandwidth.

For Polynomial Regression: Residuals should be close to normal distribution.

3. What are the basic assumptions of the Linear Regression Algorithm?

- **Linear relationship:** The relationship between the independent and dependent variables to be linear.
- **Multivariate normality:** Linear regression analysis requires all variables to be multivariate normal.
- **No or little multi-collinearity:** The independent variables are not highly correlated with each other.
- **No auto-correlation:** The residuals are not linearly auto-correlated.
- **Homoscedasticity:** The residuals must be equal across the regression line.

4. What are the different Evaluation metrics used in Linear Regression?

There are **three error metrics** that are commonly used for evaluating and reporting the performance of a regression model; they are.

- **Mean Squared Error (MSE).**
- **Root Mean Squared Error (RMSE).**
- **Mean Absolute Error (MAE)**

5. Why do we square the residuals instead of the Modulus?

Squaring the residuals helps in penalizing the error more and contributing to maximum value of error.

6. Which evaluation metric should you prefer if you have a lot of outliers present in it?

Mean Absolute Error (MAE) is robust to outliers compared to RMSE, because it is not squaring the value of errors. So dataset having more outliers prefers Mean Absolute Error evaluation metrics.