← Go Back to Machine Learning

Course Content

The Assumptions of Linear Regression

When applying linear regression to solve real-life regression problems, we need to understand that the algorithm operates under some assumptions. These assumptions allow us to correctly conclude from the results of our analysis. In case one or more of these assumptions are violated, the results and the performance of the model might not be reliable enough to extract inferences from the model and apply those to the unseen data. This is why it is necessary to check whether the below-specified assumptions are satisfied by the linear regression model or not.

- **Linearity:** The first and foremost assumption is related to the type of relation between the dependent and independent variables. Since the linear regression model fits a line to the data, the **relation between the independent variables and the dependent variable must be linear;** otherwise, it might fail to identify the pattern in the data and result in a higher error.
- **Multicollinearity:** This is a phenomenon related to the **correlation among the independent variables in the data**. Linear regression assumes that there is **no multicollinearity** between the independent variables. Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of the regression model. This is because correlated variables will provide repeated information to the model and hence the significance of one variable might decrease than the actual value. To avoid this, it is necessary to include only non-correlated independent variables in the model.
- **Homoscedasticity:** According to this assumption, the error associated with each data point should be **equally spread** (meaning “constant variance”) along the best fit line. The derivation of the equation of the linear regression model assumes that the error terms have a constant variance. Hence, the results from the linear regression model might not be reliable in case the assumption of homoscedasticity is violated.
- **Normal Distribution of error terms:** If error terms are non-normally distributed, confidence intervals may become too wide or narrow. Once the confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on the **minimization of least squares**. The presence of a non–normal distribution suggests that there are a **few unusual data points** that must be studied closely to make a better model.
- **Endogeneity:** This is the phenomenon of independent variables being correlated to the error terms of the model. With endogeneity, **the optimization process will lead to biased parameters of the model**. This will **adversely affect** the performance of the model. To avoid this, it is assumed that the independent variables are not correlated to the error terms.

Previous Next