Statistics Part

Zhenyu(Daniel) Bai

Linearity in Parameters

The first assumption is that there is a linear relationship between dependent and independent variables. It can be examined by plotting scatter plots and checking if there are linear relationships. This assumption will be violated if the relationship between explanatory variables and dependent variables is not linear. If it happens, we can do data transformation on the variables to achieve a linear relationship.

When the relationship between variables is non-linear, while the model is linear regression, it will lead to biased and inefficient estimations, because the estimates cannot be explained by one single linear coefficient. As a result, the model will fail to capture the linear relationship between variables correctly and efficiently.

Random Sampling

In the linear regression, the data are assumed to be drawn randomly from the population. They should be representative and unbiased. The unbiased data can help models to capture the linear relationship between explanatory and dependent variables, ensuring the estimated result is close to the real underlying relationships between data.

If the data is not drawn randomly from the population, it will cause overfitting for the model. For the training dataset, the model might show good results. However, the model still couldn’t capture the correct movement in the validation dataset due to the overfitting in the training phase, causing the failure in model estimations.

No Perfect Multicollinearity

The independent variables should not be perfectly collinear. In other words the variable cannot be a linear combination of one another. It will make sure that each parameter is identifiable and estimable. This error will occur when there are redundant variables that can explain each other, leading to intermediate variables and causing the failure in model estimations. We can use correlation heat maps to identify the multicollinearity between variables and use feature engineering to combine variables together to reduce the redundancy.

When there is multicollinearity between variables, it will introduce bias into the estimation since the model cannot distinguish the individual effects of collinear variables. Sometimes variables will cancel each other out and cause inefficiency in the model estimation.

Zero Conditional Mean of Error Terms:

The expected value of the error term, given any value of the independent variables, is zero. This assumption is crucial for unbiasedness in the OLS estimators. On the other hand, if the expected value of the error term is not zero, it often indicates that there are omitted variables, reverse causality, and etc., because the given variables cannot explain the dependent variables holistically.

If the conditional mean doesn't equal zero, it will cause the estimation of the model to be systematically off the target. In such cases, the OLS estimates will be biased, as they systematically overestimate or underestimate the true effect of the independent variables on the dependent variable. This bias makes it difficult to trust the estimated coefficients' magnitudes or even their signs, causing models failure in capturing the effects of each variable on the target.

Homoscedasticity

Heteroskedasticity occurs when the variance of the error terms differs across the levels of the independent variables, violating the linear regression assumption of constant error variance. This phenomenon can be visually inspected through residual plots, where the spread of residuals varies with the fitted values, indicating the presence of heteroskedasticity. Applying transformations to the dependent variable by using measures such as square root or logarithm can mitigate the effects of heteroskedasticity and stabilize the variance of the error terms.

When heteroskedasticity is present, it undermines the efficiency of the OLS estimators, making them no longer the Best Linear Unbiased Estimators. Although the OLS estimates remain unbiased, the standard errors are affected, which misleads the inference about the coefficients such as hypothesis testing.

No Autocorrelation of Error Terms

The assumption of no autocorrelation states that the error terms of the regression model are uncorrelated with each other. When the error terms are correlated to one another, it will further cause inconsistency in the model estimation. The effect is particularly significant in time series data, where the value of a variable at one time point could be correlated with its value at another time point. This problem can often be solved by incorporating lagged variables of dependent variables, which helps to capture the dynamics in time series data.

Violation of this assumption will compromise the efficiency of the OLS estimates. While the estimates remain unbiased, the standard errors become unreliable, affecting hypothesis tests and confidence intervals.

Normality of Error Terms

The assumption of normally distributed error terms primarily affects the inference in regression analysis. When the error term is not normally distributed, this assumption will be violated. While OLS does not require normality for estimation, the assumption ensures the validity of many hypothesis tests, such as the t-test and F-test, especially in small samples.

Violation of this assumption does not bias the OLS estimates but may affect the accuracy of confidence intervals and the power of hypothesis tests. As a result, we will fail to identify redundant variables and improve the model.