**Multi-collinearity**

**Multicollinearity** is the state in which there is a moderate to very high inter-correlations or inter-associations among the predictors occurs. If multicollinearity exist between two or more predictors the reliability of the regression coefficients is reduced and therefore statistical inferences are not reliable. In such situation, collinear variables do not provide new information, and it becomes difficult to separate the effect of such variables on the dependent variable (redundant information). Thus, multicollinearity is a statistical problem as it can increase the variance of the regression coefficients, making them unstable and difficult to interpret.

**Effects of multicollinearity:**

1. **Inflated standard error**: Multicollinearity increases the variance of the estimated regression coefficients. As a result, the standard errors of the coefficients become inflated. The consequence is that larger standard errors reduce the statistical significance of the predictors, making it harder to detect meaningful relationships between the independent variables and the dependent variable. Even if a predictor is important, its coefficient may not be statistically significant due to high standard errors.

**2**. **Unstable coefficient estimates**: When multicollinearity is present, small changes in the data (e.g., adding or removing observations) can lead to large changes in the estimated regression coefficients. Coefficients may have unexpected signs or magnitudes. The consequence is that the instability of the coefficients makes it difficult to interpret their magnitude and direction reliably. This undermines the trustworthiness of the model for prediction or inference.

**3. Difficulty in Identifying Individual Predictor Effects:** Multicollinearity obscures the individual contribution of each predictor because the predictors are highly interrelated.Since the correlated variables tend to move together, the model struggles to distinguish their separate impacts.

Consequence is that it becomes challenging to determine which specific predictor is driving the relationship with the dependent variable. This makes it harder to draw meaningful conclusions about the relationships between the predictor variables and the dependent variable.

**4. Reduced statistical significance**: High correlation among predictors inflates **standard errors**, which result into wider confidence interval, leading to large p-values. The hypothesis tests (e.g., t-tests) for the regression coefficients may fail to reject the null hypothesis even when a predictor is actually significant. We might fail to identify important predictors simply because of multicollinearity.

5. **Overfitting Risk**: Multicollinearity increases model complexity, making it prone to **overfitting** and poor generalization to new data. Overfitting means that the model performs well on the training data but poorly on new, unseen data.

* It reduces a model’s overall predictive power; it can produce estimates of the regression coefficients that are not statistically significant.
* The partial regression coefficient may not be estimated precisely. The standard errors of regression coefficients are likely to be high. change in the signs as well as in the magnitudes of the partial regression coefficients from one sample to another sample.
* The estimated regression coefficient of any one variable depends on which other predictors are included in the model
* The precision of the estimated regression coefficients decreases as more predictors are added to the model
* The marginal contribution of any one predictor variable in reducing the error sum of squares depends on which other predictors are already in the model
* Hypothesis tests for β*k* = 0 may yield different conclusions depending on which predictors are in the model.

In the presence of high multicollinearity, the confidence intervals of the coefficients tend to become very wide and the statistics tend to be very small. It becomes difficult to reject the null hypothesis of any study when multicollinearity is present in the data under study.

**Analysis of multicollinearity**

Multicollinearity can be tested with three criteria:

1. **Correlation Matrix**: Compute the matrix of Karl Pearson’s correlation among all independent variables. Examine whether particular pair of independent variables have high correlation.
2. **Tolerance**: The tolerance measures the influence of one independent variable on all other independent variables. The tolerance is calculated with an initial linear regression analysis. Tolerance T is defined as T = 1 - R2 for these first step regression analysis. With T < 0.01 there might be multicollinearity in the data and with T < 0.01 there certainly is.
3. **Variance Inflation Factor (VIF)**:

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

The multi-collinearity can be examined by computing the variance inflation factor (VIF) for each explanatory variable.

**Variance Inflation Factor (VIF)**

The variance inflation factor is computed as follows:

**Step 1**: We run the ordinary least square regression that has Xi as function of all other explanatory variables.

**Step 2**: Calculate the variance inflationary factor (VIF) for Xi which is given by,

= 

Where,

= Variance inflation factor of ith independent variable 

 = Coefficient of determination of explanatory variable  with all other explanatory variable.

Variance inflation factors (VIF) measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. The VIF factor lies between 1 and .

**Step 3**: Analyze the magnitude of [multicollinearity](https://en.wikipedia.org/wiki/Multicollinearity) by considering the size of the  {\displaystyle \operatorname {VIF} ({\hat {\beta }}\_{i})}.

Use the following guidelines to interpret the VIF

|  |  |
| --- | --- |
| VIF | Status of predictors |
| VIF = 1 | Not correlated |
| 1 < VIF < 5 | Moderately correlated |
| VIF > 5 to 10 | Highly correlated |

* If a set of predictors is uncorrelated then VIF = 1 for each predictor.
* If one the independent variable has a VIF factor greater than 5, we will eliminate the independent variable and perform the best-subset approach to the model building.
* If more than one independent variable has a VIF factor greater than 5, we will eliminate that independent variable that has the highest VIF, and perform the best-subset approach to the model building.
* If any independent variable has VIF > 10, that variable definitely has to be removed before modelling.

**Two explanatory variables case**:

In this case (coefficient of determination of X1 on X2) and  (coefficient of determination between X2 on X1) are equal and== . Therefore, VIF1 = VIF2. If VIF > 5, we will eliminate either X1 or X2 from the model, depending upon how they are correlated with dependent variable Y.

**Three explanatory variables case**:

In this case we run three ordinary least square regression and find coefficients of multiple determination.

1.  running regression equation of X1 on X2 and X3
2.  running regression equation of X2 on X1 and X3
3.  running regression equation of X3 on X1 and X2

We next calculate VIF1, VIF2 and VIF3 and evaluate X1, X2 and X3 for their inclusion in the model.