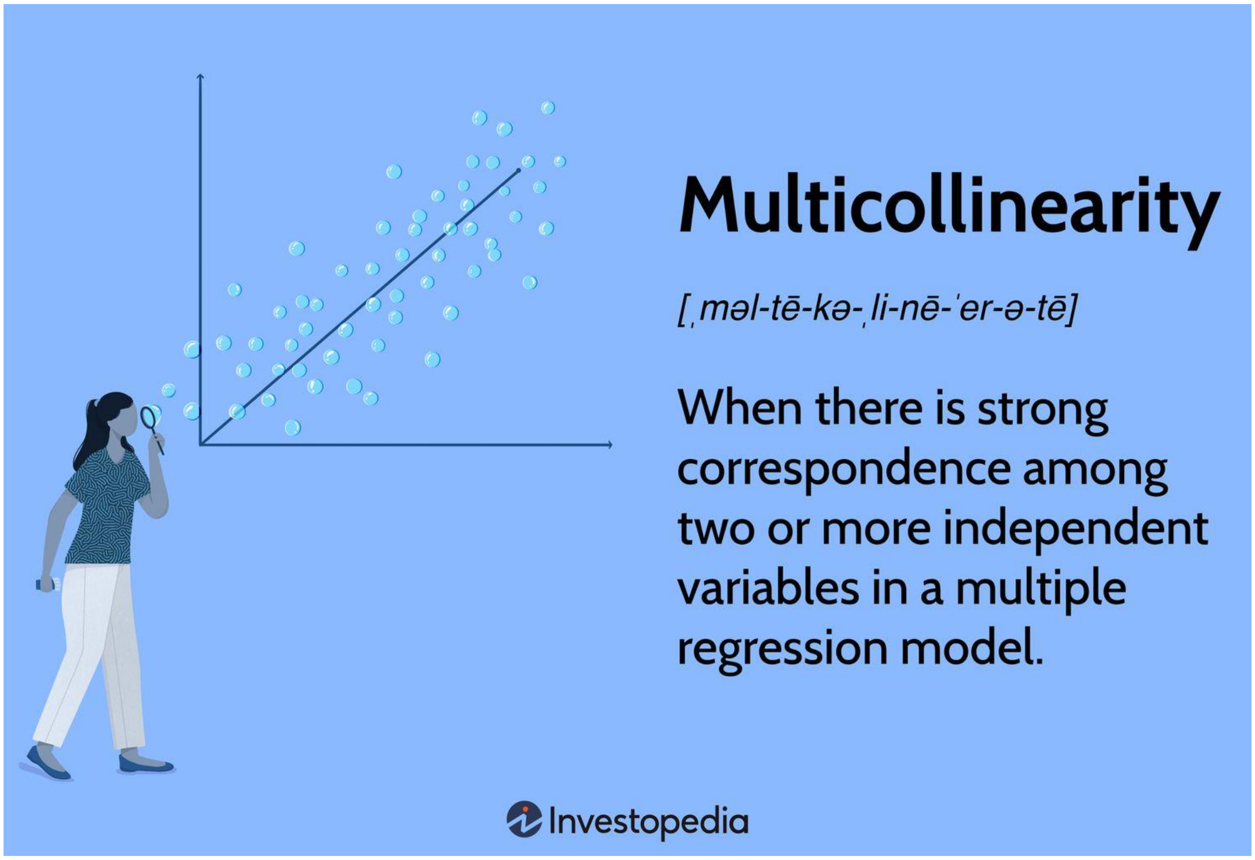
Multicollinearity

What is Multicollinearity?

Multicollinearity is a statistical concept where several independent variables in a model are correlated. Two variables are considered perfectly collinear if their correlation coefficient is +/- 1.0. Multicollinearity among independent variables will result in less reliable statistical inferences.



How to handle multicollinearity?

Centering the variables is a simple way to reduce structural multicollinearity. Centering the variables is also known as standardizing the variables by subtracting the mean.

one way of reducing **data-based multicollinearity** is to **remove one or more of the violating predictors from the regression model.** Another way is to collect additional data under different experimental or observational conditions.

To handle multicollinearity: 1) **Increase sample size** to strengthen the statistical power.

2) Remove highly correlated predictors **by checking the Variance Inflation Factor (VIF).**

3) **Combine correlated variables into a single predictor through Principal Component Analysis** (PCA) or factor analysis.

Thumb Rule: most cases, there will be some amount of multicollinearity. As a rule of thumb, **a VIF of 5 or 10 indicates that the multicollinearity might be problematic.**

VIF **less than 5 indicates** a **low correlation** of that predictor with other predictors. A **value between 5 and 10 indicates** a **moderate correlation,** while **VIF values larger than 10** are a sign for high, **not tolerable correlation** of model predictors