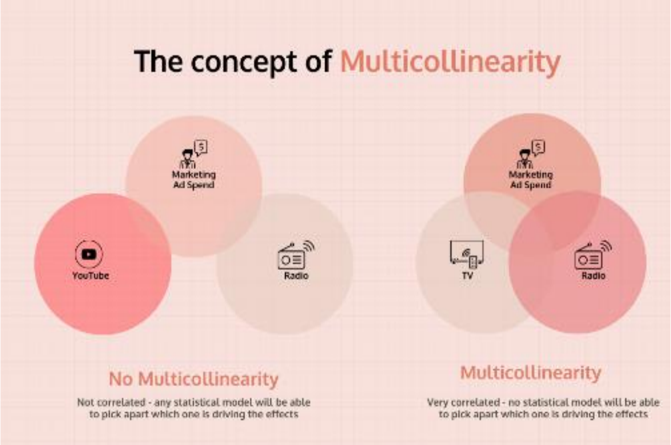
WAYS TO HANDLE MULTICOLLINEARITY


The concept of Multicollinearity
Marketing Ad Spend
YouTube
Radio
No Multicollinearity
Not correlated - any statistical model will be able to pick apart which one is driving the effects
Marketing Ad Spend
TV
Radio
Multicollinearity
Very correlated - no statistical model will be able to pick apart which one is driving the effects

Remove one or more of the independent variables from the model, use a different statistical method, such as ridge regression or LASSO regression, to fix multicollinearity, one can remove one of the highly correlated variables, combine them into a single variable, or use a dimensionality reduction technique such as principal component analysis to reduce the number of variables while retaining most of the information.

Variance inflation factor

- Fit a Multiple Linear Regression Model: Start by fitting a multiple linear regression model using all predictor variable

- Check for High VIF Values: Examine the VIF values for each predictor variable. Typically, a VIF value above 10 or 5 indicates a high degree of multicollinearity.

- If high VIF values are found, consider the following approaches to address multicollinearity:

- Drop predictor variables with high VIF values from the model, especially those with the highest VIF values. This reduces multicollinearity as these variables are likely highly correlated with other predictors.

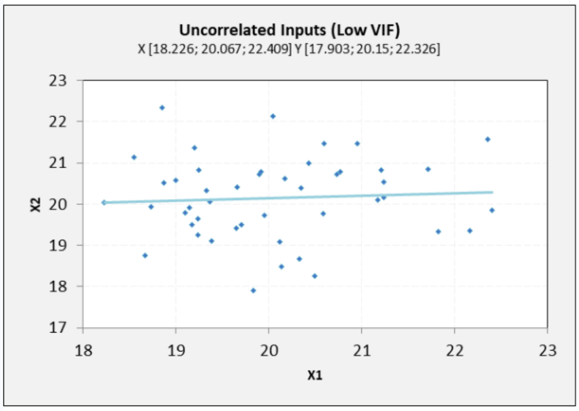
- Sometimes, highly correlated variables can be combined into composite variables. For instance, if two variables are highly correlated, you could create a new variable that represents the average or sum of the two.

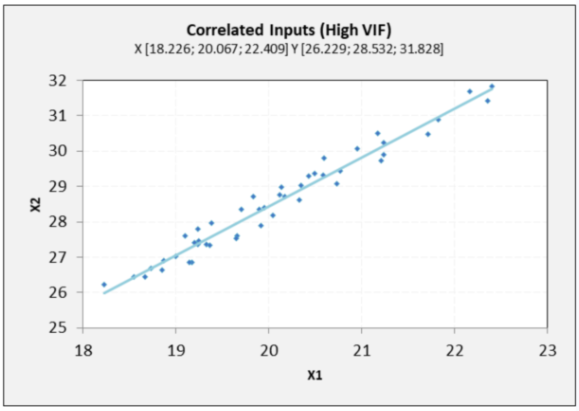
- Use regularization techniques such as Ridge Regression or Lasso Regression, which introduce a penalty term to the regression equation, helping to reduce the impact of multicollinearity.

In VIF method, we pick each feature and regress it against all of the other features. For each regression, the factor is calculated as :

$$VIF = \frac{1}{1-R^2}$$

- $VIF = \frac{1}{1-R_{sqr}}$
- VIF = 1 means no collinearity (orthogonal)
- VIF between 5-10 or higher indicates collinearity
- **Solution**: Find VIF, remove the redundant term





Real-life example where the Variance Inflation Factor (VIF) can be applied.

- Imagine you're working on a project to predict housing prices based on various features of the houses, such as square footage, number of bedrooms, number of bathrooms, and distance from the city center. You suspect that there might be multicollinearity among some of these features, particularly between square footage and number of bedrooms.

  Here's how you can use VIF to assess multicollinearity in this scenario:
- Calculate VIF: Calculate the VIF for each predictor variable (feature) in your regression model. To calculate the VIF for a particular variable, you regress that variable against all the other predictor variables.

- Interpret VIF Values: Examine the VIF values for each predictor variable. VIF values greater than a certain threshold (commonly 5 or 10) indicate multicollinearity. High VIF values suggest that the variance of the regression coefficients for that variable is inflated due to multicollinearity with other predictor variables.

- Address Multicollinearity: If you find high VIF values indicating multicollinearity, you can take appropriate actions to address it. This might involve removing one of the highly correlated variables, combining them into a single composite variable, or using regularization techniques like ridge regression or Lasso regression to mitigate multicollinearity.

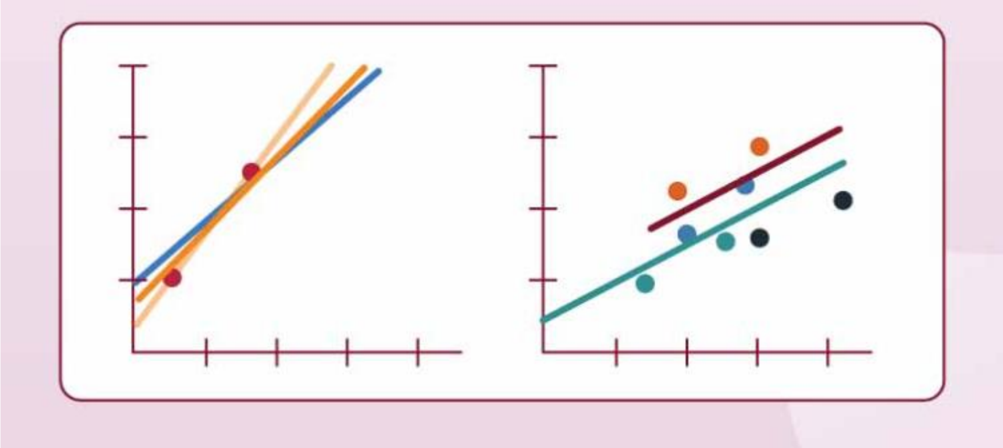
Lasso regression

- In the context of multicollinearity, this can effectively eliminate redundant predictors from the model, as highly correlated predictors tend to have similar coefficients.

- When one of these predictors has its coefficient forced to zero by the lasso penalty, the other correlated predictors can compensate for its influence, resulting in a more stable and interpretable model.

- Suppose you're a data scientist working for an e-commerce company, and your task is to build a model to predict customer purchase behavior based on various features such as customer demographics, browsing history, and purchase history.

$$J(\theta) = \mathrm{MSE}(\theta) + \alpha \sum_{i=1}^{n} |\theta_i|$$



Real-life example where the Lasso Regression can be applied.

- Gather data on customer demographics (age, gender, location), browsing behavior (pages visited, time spent on site), purchase history (products purchased, purchase frequency), and any other relevant features.
- Use Lasso regression to select the most relevant features for predicting customer purchase behavior. Since Lasso regression penalizes the absolute size of the coefficients, it automatically performs feature selection by setting less important coefficients to zero.

- After training the Lasso regression model, examine the coefficients to understand which features have the most significant impact on customer purchase behavior. Since Lasso regression shrinks less important coefficients towards zero, it helps in identifying the most relevant features for predicting customer behavior.

Principal Component Analysis (PCA):

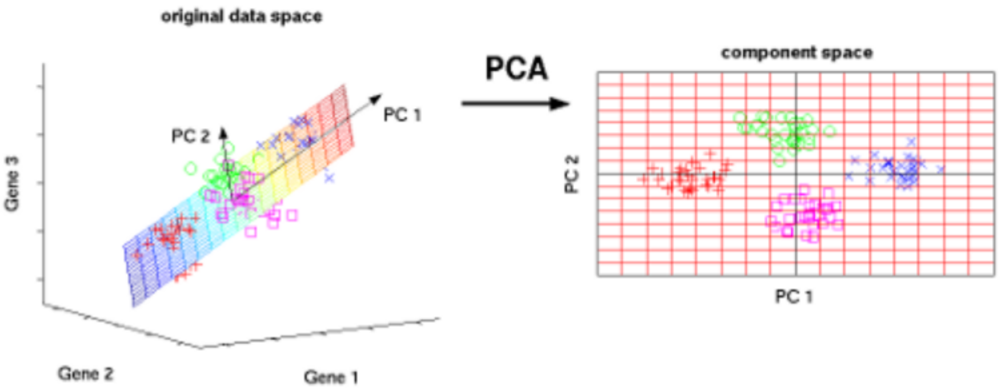
- Principal Component Analysis (PCA) can indirectly handle multicollinearity by transforming the original correlated variables into a set of uncorrelated variables called principal components.

- When multicollinearity exists among the original variables, the principal components extracted by PCA are orthogonal to each other, meaning they are linearly independent and do not exhibit multicollinearity.

- In the context of Principal Component Analysis (PCA), a principal component is a linear combination of the original variables in a dataset.

- These components are derived in such a way that they are orthogonal to each other, meaning they are uncorrelated.

- Each principal component captures a certain amount of variance in the data, with the first component capturing the maximum variance, the second component capturing the maximum remaining variance orthogonal to the first, and so on.



Real time Exampe for PCA can be applied

- Imagine you work for a real estate agency, and your task is to analyze a dataset containing information about various houses, including features such as square footage, number of bedrooms, number of bathrooms, lot size, and distance from the city center. You want to understand the underlying structure of the dataset and identify the most significant factors that influence house prices.

Here's how you can apply PCA in this scenario:

- Dimensionality Reduction with PCA: Apply PCA to the standardized dataset to reduce its dimensionality. PCA will identify the principal components (linear combinations matrix of the original features) that capture the most significant sources of variation in house characteristics.

- Interpretation: Examine the principal components to understand what they represent in terms of housing characteristics. For example, one principal component might capture overall house size (e.g., square footage, number of bedrooms, number of bathrooms), while another might represent location-related factors (e.g., distance from the city center).

- Feature Transformation: Use the principal components as new features to represent each house's characteristics. Instead of using the original features, you can represent each house using a much smaller number of principal components, effectively reducing the dimensionality of the dataset.

- Analysis and Prediction: Analyze the relationship between the principal components and house prices using regression analysis or other predictive modeling techniques. By using the reduced-dimensional representation of housing characteristics, you can build models that predict house prices more efficiently and accurately.

- Insights and Decision Making: Use the insights gained from the analysis to make informed decisions in the real estate market. For example, if a particular principal component consistently has a strong influence on house prices, it may indicate areas where the company can focus its efforts to maximize property value.