MULTICOLLINEARITY

Multicollinearity may not affect the accuracy of the model as much but we might lose reliability in determining the effects of individual independent features on the dependent feature in your model and that can be a problem when we want to interpret your model.

Detecting Multicollinearity using VIF

Let's try detecting multicollinearity in a dataset to give you a flavor of what can go wrong.

Although correlation matrix and scatter plots can also be used to find multicollinearity, their findings only show the bivariate relationship between the independent variables. VIF is preferred as it can show the correlation of a variable with a group of other variables.

" VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable. "
or
VIF score of an independent variable represents how well the variable is explained by other independent variables.

R^2 value is determined to find out how well an independent variable is described by the other independent variables. A high value of **R^2** means that the variable is highly correlated with the other variables. This is captured by the **VIF** which is denoted below:

$$VIF=1/(1-R^2)$$

So, the closer the **R^2** value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable.

- VIF starts at 1(when R^2=0, VIF=1 – minimum value for VIF) and has no upper limit.
- VIF = 1, no correlation between the independent variable and the other variables.
- VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others.
- Some researchers assume VIF>5 as a serious issue for our model while some researchers assume VIF>10 as serious, it varies from person to person.

Solutions for Multicollinearity

1. Drop the variables causing the problem.
– If using a large number of X-variables, a stepwise regression could be used to determine which of the variables to drop.
– Removing collinear X-variables is the simplest method of solving the multicollinearity problem.

2. If all the X-variables are retained, then avoid making inferences about the individual parameters. Also, restrict inferences about the mean value of Y of values to X that lie in the experimental region.

3. Re-code the form of the independent variables.
For example, if x1 and x2 are collinear, you might try using x1 and the ratio x2/x1 instead.

4. Ridge and Lasso Regression– This is an alternative estimation procedure to ordinary least squares. Penalizes for the duplicate information and shrinks or drops to zero the parameters of a regression model.

5. By standardizing the variables i.e, by subtracting the mean value or taking the deviated forms of the variables (xi=Xi-mean(X))

7. Increase in sample size may sometimes solve the problem of multicollinearity.

What Is a Correlation Matrix?

A correlation matrix is a statistical technique used to evaluate the relationship between two variables in a data set. The matrix is a table in which every cell contains a correlation coefficient, where 1 is considered a strong relationship between variables, 0 a neutral relationship and -1 a not strong relationship. It's most commonly used in building regression models.

In multiple linear regression, the correlation matrix determines the correlation coefficients between the independent variables of a model.

When Do You Use the Correlation Matrix?

The correlation matrix gives you an idea about your data set.

For example, let's say you want to predict the price of a car on the basis of fuel type, transmission type or age, etc. A correlation matrix would be very useful.

Using a correlation matrix, we can evaluate the relationship between two variables:

- If the relationship is 1, then the relationship is strong.
- If the relationship is 0, then it means the relationship is neutral.
- If the relationship is -1, then it means the relationship is negative or not strong.
- By using a correlation matrix, you can better understand your data set, analyze it and visualize the result.
- The correlation matrix is a statistical technique that gives you the values between -1 to 1 which you can determine the relationship between variables.

EIGEN VALUE DECOMPOSITION

How Eigenvalue Decomposition Relates to Multicollinearity

1. **Understanding the Covariance or Correlation Matrix**:
    - Multicollinearity is tied to the linear dependencies among features in the dataset.
    - These dependencies are captured in the covariance matrix (or correlation matrix, if features are standardized).

2. **Eigenvalue Decomposition**:
    - Eigenvalue decomposition of the covariance matrix expresses it in terms of its eigenvalues and eigenvectors: Covariance Matrix=QΛQ−1\text{Covariance Matrix} = Q \Lambda Q^{-1}Covariance Matrix=QΛQ−1
        - QQQ: Matrix of eigenvectors.
        - Λ\LambdaΛ: Diagonal matrix of eigenvalues.
3. **Eigenvalues and Multicollinearity**:
    - Eigenvalues (λi) represent the variance along the directions defined by the eigenvectors.
    - **Multicollinearity occurs when one or more eigenvalues are very small or close to zero.** This indicates that the corresponding eigenvectors define directions in the feature space with near-zero variance, suggesting redundancy or linear dependence among features.

Steps for Using Eigenvalue Decomposition to Detect Multicollinearity

1. **Compute the Correlation Matrix**:
    - Use the standardized version of your data to compute the correlation matrix.
2. **Perform Eigenvalue Decomposition**:
    - Decompose the correlation matrix to obtain eigenvalues and eigenvectors.
3. **Analyze Eigenvalues**:
    - Look for **small eigenvalues** (close to zero). The smaller the eigenvalue, the higher the dependency among features in that direction.
4. **Condition Number**:
    - Calculate the **condition number**, which is the ratio of the largest eigenvalue to the smallest eigenvalue:
      Condition Number=λmax/λmin
    - A high condition number (e.g., > 30) indicates severe multicollinearity.

How to Address Multicollinearity Using Eigenvalue Insights

1. **Remove Redundant Features**:
    - Examine the eigenvectors corresponding to small eigenvalues to identify which features contribute most to collinearity.
2. **Dimensionality Reduction**:

- Use techniques like **Principal Component Analysis (PCA)** to combine correlated features into uncorrelated principal components.

3. **Regularization**:
    - Apply methods like Ridge Regression to handle multicollinearity without feature elimination.

Summary

Eigenvalue decomposition provides a mathematically robust way to detect and quantify multicollinearity by analyzing the variance structure of features. Small eigenvalues, large condition numbers, and their corresponding eigenvectors help pinpoint issues and guide solutions like feature elimination or transformation.