Data Science - Multicollinearity In Regression

Gili Gutfeld (ID. 209284512)

Submitted as final project report for the TDL course, BIU, 2023

1 Abstract

The element in the data science pipeline that the project tries to improve is addressing multicollinearity in multiple regression analysis. The chosen element needs improvement because multicollinearity is a common issue in multiple regression analysis, it occurs when two or more predictor variables are highly correlated with each other. This can result in unstable and unreliable estimates of regression coefficients, making it difficult to interpret the effect of individual features on the response variable. Additionally, it can also cause problems in model selection and hypothesis testing. Some of the potential solutions include removing highly correlated independent variables, combining independent variables through linear combinations, or using partial least squares regression to create a set of uncorrelated components for inclusion in the model. The ability to effectively address multicollinearity is crucial for building accurate and interpretable predictive models. Next, we will try to develop an approach that can identify and address multicollinearity in a robust and efficient manner, and we will check it on 4 datasets: houses, cars, medical insurance and cancer.

2 Problem Description

Multicollinearity is a phenomenon that arises when the independent variables in a regression model are strongly correlated with each other. This can make it difficult to interpret the model and can also lead to overfitting. To avoid this problem, researchers usually test for multicollinearity before selecting variables for their regression model. The problem with multicollinearity is that when independent variables are highly correlated, a change in one variable will cause a change in another variable, leading to significant fluctuations in the model results. This instability can create several problems, including difficulties in choosing a list of significant variables for the model, unstable coefficient estimates, and challenges in interpreting the model. Furthermore, this instability can

lead to overfitting, where the model performs well on the training data but poorly on a new set of data.

3 Solution Overview

There there are various scenarios where dealing with multicollinearity may not be necessary. For instance, the degree of multicollinearity may not be severe enough to cause problems, or it may not affect the variables of primary interest. Alternatively, if the issue is structural multicollinearity, centering the variables can help resolve it. However, in cases where severe multicollinearity exists, it becomes crucial to address the issue. Unfortunately, this is often a challenging task, as each method to resolve the problem has its drawbacks. It is crucial to draw upon subject-area knowledge and research goals to determine the most suitable approach that balances the advantages and disadvantages of each method. Some of the potential solutions include removing highly correlated independent variables, combining independent variables through linear combinations, or using partial least squares regression to create a set of uncorrelated components for inclusion in the model. Additionally, advanced regression techniques such as LASSO and Ridge regression can also handle multicollinearity. With a little additional study, those familiar with linear least squares regression can perform these analyses with ease. The ability to effectively address multicollinearity is crucial for building accurate and interpretable predictive models. Next, we will try to develop an approach that can identify and address multicollinearity in a robust and efficient manner.

3.1 Identify Multi-Collinearity

Determining whether Multi-Collinearity is present in a regression model can be done using various methods, with one simple approach being to examine the correlation matrix of all the independent variables. For instance, we applied this method to the medical dataset from the Kaggle competition, which aims to predict the charges price based on various personal-related features. In our analysis, we chose a subset of numerical variables to include in the model, and plotted their correlation matrix to check for any high correlations between them.

	age	bmi	children	smoker	charges
age	1.000000	0.109272	0.042469	-0.025019	0.299008
bmi	0.109272	1.000000	0.012759	0.003750	0.198341
children	0.042469	0.012759	1.000000	0.007673	0.067998
smoker	-0.025019	0.003750	0.007673	1.000000	0.787251
charges	0.299008	0.198341	0.067998	0.787251	1.000000

Figure 1: Base model loss

One way to identify Multi-Collinearity is by plotting a correlation matrix and color coding it to highlight the pairwise correlation between the variables, including the dependent variable. This can be a helpful trick for selecting independent variables to include in the model. In the case of the medical data from a Kaggle competition, it revealed several variables with high correlation to each other. For example, the 'charges' and 'smoker' had a correlation of over 0.7 due to smmoker people that tend to get sick and use the insurance more. Another method to check Multi-Collinearity is by calculating the Variance Inflation Factor (VIF) for each independent variable. This measure quantifies the degree of multicollinearity among multiple regression variables, where a higher value indicates a stronger correlation with the rest of the variables.

	features	vif_Factor
0	age	9.18
1	bmi	8.06
4	charges	7.83
3	smoker	3.95
2	children	1.80

Figure 2: VIF of the base model

When we compute the VIF values for the independent variables, a value greater than 10 is often indicative of high correlation between that variable and the others. However, the acceptable range for VIF values may vary depending on specific requirements and constraints. After computing the VIF values for the features in our model, we can observe that most of the features exhibit medium correlation with the other independent variables. Only 2 of 5 features meet the criterion of having a VIF value below 5.

3.2 Variable Selection

To address the issue of multicollinearity, there are two common methods: variable selection and variable transformation. In variable selection, we remove highly correlated variables and only include the most significant ones in the model. For instance, in the medical data example, we observed that smoker and age have the two highest correlations with the dependent variable, charges. Therefore, we will include them in the model and eliminate others with low correlation with charges, for example lower than 0.1.

	features	vif_Factor
0	age	9.13
3	charges	7.81
1	bmi	7.69
2	smoker	3.94

Figure 3: Base model loss

Again, let's try to fix the collinearity. In the medical data example, we observed that smoker and age have the highest correlation with the dependent variable, charges. Therefore, we will include only that in the model and eliminate the others, in this case the correlation lower than 0.2.

	features	vif_Factor
2	charges	7.78
1	smoker	3.94
0	age	3.28

Figure 4: VIF when using variable selection

However, this method may not work in all cases, especially when the important variables still have high VIF values. In this case, we can use variable transformation to reduce the correlation between variables while maintaining their features.

3.3 Variable Transformation

As we mentioned, we can use variable transformation to reduce the correlation between variables while maintaining their features. For example, in the medical data example, we transformed the variable 'bmi' into 'new_bmi' by subtracting a minimal bmi (15) from the bmi.

	features	vif_Factor
2	charges	8.01
0	age	6.17
3	new_bmi	5.05
1	smoker	3.96

Figure 5: VIF when using variable transforamtion

3.4 Linearly combination

As we mentioned, we can use linearly combination of the independent variables to reduce the correlation between variables while maintaining their features. For example, in the medical data example, we transformed the variable 'bmi' into 'new_bmi' such that new_bmi = bmi * 2 if he smokes and new_bmi = bmi otherwise, and than we remove the 'bmi' and 'smoker' columns. The idea is to increase the bmi of the person if he smokes in thought that as the bmi increases, the charges increase and also when person smokes, therefore we can makes it linear although it's a nominal feature (smoker):

	features	vif_Factor
3	new_bmi	8.67
2	charges	7.35
0	age	3.28
1	children	1.74

Figure 6: VIF when using linear combination

By applying these methods, we can reduce the VIF values and improve the stability and interpretability of the model.

3.5 PCA - Principal Component Analysis

PCA, or Principal Component Analysis, is a widely used technique in data analysis that involves breaking down a dataset into a set of independent factors, with the aim of reducing its dimensionality. This technique is often used to simplify calculations in statistical models, by reducing the number of predictive factors. However, in the context of our current case, we will be leveraging the independence of variables provided by PCA to address the issue of multicollinearity in our model.

	features	vif_Factor
0	0	1.0
1	1	1.0
2	2	1.0

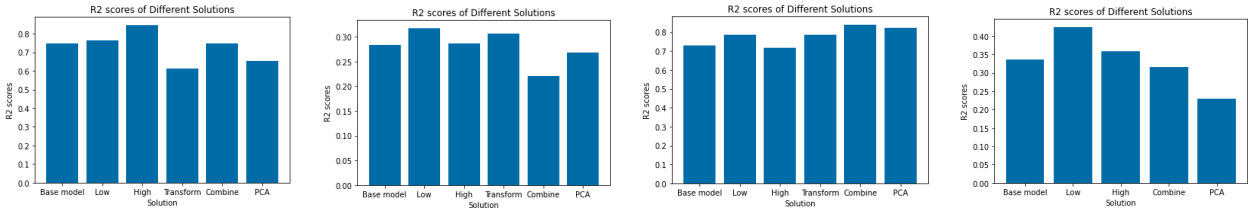
Figure 7: VIF when using PCA

Despite retaining the same number of variables as in the original data, it can be observed that the three variables are no longer correlated with each other following the PCA transformation. This enables us to use the new set of three variables as independent variables to medical charges.

However, it should be noted that this approach has a significant drawback. The PCA transformation does not maintain the identity of each variable, which makes interpreting the results difficult. By fitting a model with these variables, we can see that the structural multicollinearity has been addressed and any remaining multicollinearity is not severe enough to require further corrective measures.

4 Experimental Evaluation

In the origin dataset there are features with high correlation with the feature we want to epredict and the vif scores are relatively high. We will check each one of the solutions we mentioned by comparing between them, the R-squared score, the Mean Absolute Error, the max error and the percentage of small and big mistakes. We will check all of that in 4 different datasets: House prices, Cars, Medical insurance and Cancer and the following graphs describes the datasets by the same order we mentioned from left to right.



4.1 Variable Selection

We tested the models after removing features once with low correlation and second time with high correlation. We found that removing features that has relatively low correlation with the labels in most of the cases can improve the model and the vif scores, but we need to be careful not removing features that has too high correlation with the labels because it impairs the performance of the model although the vif scores can be better. In the house and car datasets the high correlation gave us the best results and in the cancer dataset it was the low correlation.

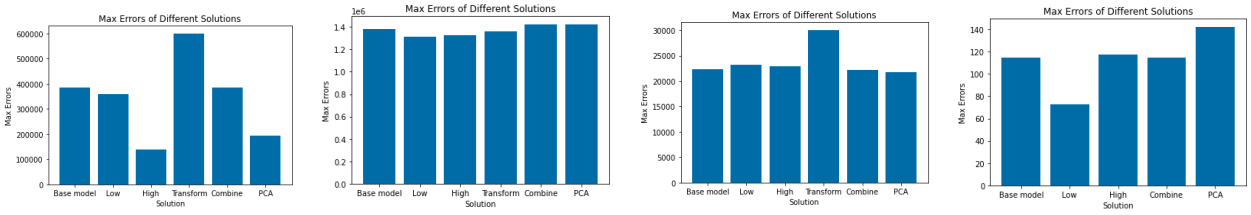
4.2 Variable Transformation and Linearly Combination

Transformations and linearly combination can sometimes improve the model but we need to find and check what is the best change for each feature and it's not always easy and actually it's not

working sometimes. In all the datasets it was a little improvement from the base model but it was not the best or the worst model, so when we want to get a small and moderate improvement we can use this solution.

4.3 PCA - Principal Component Analsys

PCA can also improve the model but sometimes it can harm, and it should be noted that this approach has a significant drawback. The PCA transformation does not maintain the identity of each variable, which makes interpreting the results difficult. And most of the times the PCA did't success to improve the base model, for example in the house and cancer datasets. But when it worked with the medical dataset it was the best model by far.



5 Related Work

"Multicollinearity in Regression Analysis" by Jim Frost, discusses the issue of multicollinearity, which occurs when independent variables in a regression model are highly correlated with each other. The article explains the problems that multicollinearity can cause and offers some solutions to mitigate it, such as centering variables, using principal component analysis (PCA), and dropping some of the correlated variables. He explains that "Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of your regression model". "Principal Component Analysis (PCA) for Machine Learning" by Muhammet Ali Kula, provides an overview of PCA, a technique that can be used to reduce the dimensionality of a dataset by identifying the most important features. The article explains the mathematical basis of PCA and provides an example of how it can be applied to a dataset. He says about PCA that "can be used on its own, or it can serve as a data cleaning or data preprocessing technique used before another machine learning algorithm". "Using Regression with Correlated Data" by Emily A. Halford, explains how correlation among independent variables in a regression model can affect the accuracy of the model and offers some solutions to mitigate the issue, such as using regularization techniques or clustering the independent variables, and she summarizes that "it is always important to check the

assumptions of a given technique and to make sure that your analytic strategy is appropriate for your data".

All of these works provide valuable insights into the problem of multicollinearity in regression analysis and offer various techniques to mitigate it, and I got the ideas of solving the problem from these articles, But they only discuss it. My goal is also to develop an approach that can identify and address multicollinearity in a robust and efficient manner and to identify the most appropriate technique or combination of techniques for addressing multicollinearity based on the characteristics of the data and the goals of the analysis.

6 Conclusion

It is important to check for the issue of Multi-Collinearity prior to building a regression model. VIF can be used to easily assess each independent variable to determine whether they have high correlations with each other. A correlation matrix is also useful in selecting important factors when unsure of which variables to choose for the model. Additionally, the correlation matrix can help to understand why certain variables may have high VIF values. We found that removing features that has relatively low correlation with the labels in most of the cases can improve the model and the vif scores, but we need to be careful not removing features that has too high correlation with the labels because it impairs the performance of the model although the vif scores can be better. Transformations and linearly combination can sometimes improve the model but we need to find and check what is the best change for each feature and it's not always work. PCA can also improve the model but sometimes it can harm, and it should be noted that this approach has a significant drawback. The PCA transformation does not maintain the identity of each variable, which makes interpreting the results difficult. While there are methods to address multi-collinearity, such as PCA, this approach can lead to the loss of model interpretability and the need to re-transform the data for application to a new set. It is best to reduce correlation by carefully selecting the appropriate variables and transforming them as needed. In cases where a variable has a relatively high VIF value but is important in predicting the result, it is up to the individual to decide whether to keep or discard the variable. Trial and error is often necessary, including testing different sets of variables, building the model, and checking for overfitting against test data. However, even after performing these checks, it is important to remember that there are other statistical assumptions that should be considered before constructing a model.