# A MODEL TO MINIMIZE MULTICOLLINEARITY EFFECTS

**Baciu Olivia, Parpucea Ilie**
*The department of Statistics,Mathematics and Forecasting, Faculty of Economics and Business Administration, Babes-Bolyai University, Cluj-Napoca, Romania*
*oli_baciu@yahoo.com*

***Abstract**. Multicollinearity implies near-linear dependence among regressors and is one of the diagnostics that harms enough the quality and the estimation of the regression models. Among the effects of multicollinearity can be mentioned that parameter estimates could lead to opposite signs or the variables turn out to having insignificant coefficients although it is known from theory or reality that the relationship exists. Also, when other variables are included or removed from the model this can affect the parameter estimates. Usually, multicollinearity is measured with the help of Variance Inflation Factor. A value greater than ten indicates severe multicollinearity in the model. Different approaches are known to reduce or eliminate multicollinearity effects but some of them are not always applicable due to data. The most used methods include addition of more data or elimination of the variable that is highly correlated with other independent variables or the use of the Ridge Regression. In addition to the well known and used models it is proposed here a new approach for the multicollinearity reduction. This method implies creating an index variable as a linear combination of the highly correlated ones. The index coefficients are selected under specific constraints imposed on the variables such that the new variable becomes highly correlated with the response variable but not with the independent ones. The best coefficients can be chosen out of the solution domain using an optimization program. In the new model, the highly correlated variables are replaced by the index one. The quality of the new model is improved by reducing or even eliminating the effects of multicollinearity. The regression model is expected to yield proper estimates. Also, VIF returns appropriate values, lower than ten. The method is exemplified on the BRD stock portfolio. Multicollinearity was eliminated, as showed by a value of one of the VIF and the model is expected to improve.*

***Keywords**: multicollinearity, regression model, index variable*

***JEL classification**: C20, C51, C58*

## 1.Introduction

Reality has to be modeled using statistical tools by creating models that include those variables that can express it the most accurate. The question that arises is which variables should be included to benefit the most for the purposes of the analysis. What happens when these variables help in explaining the reality but they influence each other?

The purpose of this article is to propose a remedial measure method that fixes the problem of multicollinearity without excluding any explanatory variables from the model. The originality of the method is that it allows keeping in the model all the variables that are highly correlated as a new variable, which is a linear combination of them. This single index can be constructed when two or more predictor variables are highly correlated and are revealing properties of a common feature. In the new model, the multicollinearity will be reduced or even eliminated.

In practice it is difficult to find data that does not contain extreme observations and is perfectly related. Also, the included variables may help in reducing the remaining variation of the dependent one but they can create near-linear dependence among the regressors. These two approaches are not always easy to use. Other techniques used to reduce the effects of multicollinearity are Ridge Regression or Principal Component Analysis.

## 2.Literature review

Multicollinearity and how to overcome the deficiencies of a model being affected by it has been widely discussed by the theoreticians. Different approaches were proposed over time, some of them being used and applied today. In practice, there are several used methods to fix its effects. Among these methods, can be mentioned data elimination or addition, Ridge Regression or techniques related to Pricipal Component Analisys. In most cases, economical practice imposes to keep all the variables in the model for a better transposition of the reality. In other cases, there is no more data to add in order to balance the sample lack of information.

A large discussion about multicollinearity detection and how to approach it depending on the point of view of an econometrician or a computer programmer is presented in Farrar & Glauber (1967).

A different approach from the generally used methods is proposed by Conklin & Lipovetsky (2003). This method assumes a change in the correlation matrix by creating a new matrix of the same structure. The negative and large in absolute value correlations are replaced in the new matrix by the opposite sign values. In such way, the only negative values will be the ones with small absolute values.

O'Brien (2007), suggest combining variables into a single index as a measure of multicollinearity reduction but it does not indicate how to estimate the coefficients of the linear combination.

Chen (2012) is taking care of the "implausible estimates" caused by the presence of multicollinearity in a modern approach. It is using external informations, given by the theory and reality, combined with the statistical confidence region. "Based on a priori knowledge" it can be chosen" the most reasonable set of coefficients inside the confidence region". These coefficients will be significant because "they are highly consistent with the data".

## 3.Multicollinearity

Construction of regression models has been a well researched theme over time. The key for a well done study is a well done constructed model. Each step in the strategy of building a model should be carefully inspected. A sensitive issue is the model selection. According to Kutner et al. (2005), the selection of a regression model depends mostly on the diagnostic results and multicollinearity is one diagnostic that can harm enough the model.

Highly correlated predictor variables do not damage the prediction but they do have an impact on the parameter estimates of the regression models. Due to multicollinearity, variables can have statistically insignificant coefficients though there is a relation between the dependent and the set of independent ones. It can also lead to parameter estimates with opposite sign than expected from theory or reality. When predictor variables are added or removed there are important changes in the estimated parameters.