A model to minimize multicollinearity effects economics essay

Economics

Multicollinearity implies near-linear dependence among regressors and is one of the diagnostics that harms enough the quality of the regression model. Different approaches are known to reduce or eliminate its effects. In addition to the well known and used models 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 variables. The index coefficients are selected under specific constraints imposed on the variables. The quality of the new model is improved by reducing or even eliminating the effects of multicollinearity. The method is exemplified on the BIO stock portfolio. Keywords: multicollinearity, econometric model, regression, VIFJEL classification: C51, C30

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.

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 Lipovetsky & Conklin (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".

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. When predictor variables are highly correlated, the interpretation of the regression parameters is affected. The usual interpretation that an increase by one unit of a parameter when holding the others constant leads to a change of the

expected value of the dependant variable does not apply in the presence of multicollinearity. The most used method to detect multicollinearity is the Variance Inflation Factor. A large value of VIF is used as an indicator of a severe multicollinearity. According to Kutner et al. (2005) a VIF value exceeding 10 suggest severe multicollinearity and the same is considered according to Hair et al. (1995), Mason et al. (1989). But the choice of a significant threshold has to be made regarding other factors that influence the stability of the estimates of the ith regression coefficients and VIF values should not be fixed a priori to some specific values, according to O'Brien (2007). To overcome the effects of multicollinearity, some of the measures that are taken are either to add more data or to eliminate the variable that creates dependence among the regressors. Often, the predictor variables cannot be excluded from the model because they highlight the essence of the real situations. Elimination of key explanatory variables can lead to biased estimators of the regression coefficients. Addition of more data is used to diminish the standard errors of the parameters but this is not always possible. Another technique used is the Ridge regression. It allows biased estimators of the regression coefficients but reduces the level of multicollinearity. Considering the reasons why one should be careful when adding or eliminating explanatory variables, a method of multicollinearity reduction without variable elimination is introduced in this article.

Model description and estimation

The following linear model is considered:(1)where and are stationary time series. Otherwise, the series have to be made stationary. In most cases one variable is not enough to explain the variation of the dependant variable and

more predictor variables should be included. If the predictor variables are highly correlated it leads to the appearance of multicollinearity in the model. Considering that one predictor variable does not explain well the model (1), let introduce in the model two new time series variables and , characterized by: The fact that the covariance between the new variables is positive implies collinearity in the model. This can affect the model by reducing the stability of the parameter estimates or by inflating the standard errors. The covariances between the dependant variable and the two new variables are also greater than zero, which implies that the new variables influence the explained one. In the new created model,(2)if is relatively high is a first hint that the multicollinearity phenomena is present in the model and could bring damages to its quality. In such conditions, theory suggests that a very easy to work with method is to eliminate one of the highly correlated variables or or add more data if possible. The proposed method allows keeping all the variables that are significant in the model. In most cases the variables kept in the model are the ones that explain much of the variation of the response variable. Naturally, the question that arises is if there are plausible conditions under which a combination of the two explanatory variables could eliminate or reduce the collinearity in the regression model. It is important that explanatory variables to be correlated with the response variable but the response ones not highly correlated among themselves. The following linear combination of the two variables is considered: , where and are stationary time series. Now, are there any real values for the parameters a and b such that to be relatively small and and to be significantly correlated? The assumption used is that the covariance among the regressors has to be

lower than the minimum of the covariances among the pairs of the explanatory variable introduced in model (1) and each of the new introduced variables in model (2). These covariances has to be low. If covariances are negative values, in the below inequalities, their values will be considered in absolute value. The minimum of the covariances is denoted by m. Simple computations lead to: The above inequality becomes then(3)By solving inequality (3) it can be chosen from the solution domain those values for a and b such that to be small. In order for to be a factor influencing the variation of the dependent variable , has to kept those values of a and b such that the influence of the new variable on the dependent one to be high. This means that the following inequality must hold: The maximum of the two covariances is denoted by M. The above inequality is a condition that allows to keep as a variable in the model. After straightforward computations, the above inequality becomes:(4)By solving (4) it can be chosen those values for a and b such that to be high. Choosing the values for a and b is reduced to solving the system of inequalities (3) and (4) which allows to keep those values for a and b that helps reduce the multicollinearity from the model by reducing the covariance between the predictor variables and increasing the covariance between the dependent and independent ones. Introducing in the model as the new variable it gives: Finally, can return to the initial variables by multiplying them with the corresponding a and b:(5)Model (5) is in fact a model that allows to keep all the needed explanatory variables without having highly correlated explanatory variables in the model. One model building process step, multicollinearity diagnostics is now exceeded.

Example

To exemplify the proposed method, it is considered the case of the stock market portfolio Biofarm (BIO) over the first 18th weeks of 2007. Data is the daily closing price of each Monday in the mentioned period. This period is right before the economical crisis began and is characterized by high returns in the Romanian stock market. The estimated linear model of the price is: (6)where: price is the stock daily closing price (in RON)BET is the Bucharest Exchange Trading Index (x100 RON Index Points)The stationary constraint holds for both variables. Usually, the market index explains in a low proportion the variation of the price, up to 30%. In this particular case we have that BET explains 62% of the variation of the stock price. The high proportion can be due to the profitable chosen period. The response variable variation can be better explained if more appropriate variables are introduced in the model. In practice, the stock price is influenced by several factors, among which can be mentioned: size, earnings/price, cash flow/price, dividend/price, book-to-market equity and so on. Since the model is explained in a medium proportion by only one variable, two new variables are introduced. These variables are Price Earnings Ratio (PER) and Price to Book Value (P/BV). The simple linear model becomes:(7)where: price is the stock daily closing price (in RON)BET is the Bucharest Exchange Trading Index (x100 RON Index Points)PER is the price earnings ratio, computed as the price per share divided by the annual earnings per sharePBV is the price to book value, computed as current share price divided by the book value per shareAll the model parameters are significant at a 10% level of significance. For the given data, the model with three independent variables

explains better the variation of the stock price, which is 99% compared to 62% explained only by the stock index. The model is better explained using three and not only one explanatory variable. Table 1: Correlation matrixpretBETPBVPERPret1BET0. 62581PBV0. 80650. 28951PER0. 87430. 36570. 99151It can be seen from the above table that the two new introduced variables are highly correlated among themselves, which suggest the presence of multicollinearity in the model. They are also high correlated with the response variable, meaning that these two variables help in understanding the response one. Multicollinearity is confirmed by the VIF values, which exceeds by much the threshold. Table 2: VIF values

Variable

VIF

PBV93. 24PER98. 61BET1. 82The proposed method indicates that the problem of multicollinearity can be fixed by replacing the highly correlated variables with a linear combination of them, . The coefficients of the linear combination are solutions of the imposed constraints system of inequalities: (8)Table 3: Variance-Covariance matrix

Pret

BET

PBV

PER

Pret

0. 002081

BET

0. 0683775. 73664

PBV

0. 0246130. 4639240. 447552

PER

0. 1175462. 581661. 954988. 68641Replacing the corresponding covariances in (8), it becomes: From the domain of solutions, were chosen a= 240 and b= -43. The index variable becomes and the new model is: (9)The model parameters are significant at 10% level of significance. If going back to the original variables, (9) becomes: In the regression model (9), multicollinearity was removed if looking at the VIF values. In Table 5 it can be noticed that the correlation between the independent variables is lowered to zero while the set of independent variables with the dependent one are medium correlated. Table 4: VIF values

Variable

VIF

Z1. 05BET1. 05Table 5: Correlation matrixpretBETzPret1BET0. 62581Z0. 48440. 00361As mentioned before, one of the effects of multicollinearity is that parameter estimates can have opposite signs. This was the case of PER variable, which changed signs after multicollinearity was removed. Although in this model the stock price is lower explained by the explanatory variables, the problem of multicollinearity has been fixed. In this case, the predictive variables help in explaining the stock price in a proportion of 79%, which increased from 62% when explained only by the stock index. Even though, in

general, the method does not prove out an increase in the model quality, such results are expected. If we think at a comparison in the quality of the two multivariate models, in the below table it can be seen that after replacing the highly correlated variables with their linear combination, the model is improved by having constant variance of the errors. Table 6: Tests of the errorsModel 2 (7)Model 3 (9)Error heterokedasticity testH0: Errors have constant variancep-value= 0. 0005Nonconstant variancep-value= 0. 9556Constant varianceShapiro-Wilk test of normalityH0: Errors follow a normal distributionp-value= 0. 078Errors follow a normal distributionp-value= 0. 136Errors follow a normal distributionDurbin-Watson test of independencyD-W= 1. 982Errors are independentD-W= 1. 402Errors are independant

Conclusions

The presented method is used to overcome the effects of the multicollinearity phenomena. The originality of this study consists of creating a new variable as a linear combination of the highly correlated ones. The linear combination coefficients are obtained by imposing specific constrains on the system of inequalities. The constraints under which the coefficients are computed are using strictly the covariances between the variables. Any pair of solutions from the solution domain will lower down the multicollinearity phenomena but the best pair can be chosen with the help of an optimization program. The best pair of solutions will lower down enough the correlation between the independent variables and increase the correlation between the independent variables and the response one. The multicollinearity will be lowered down and even fixed. As further work, the

method can be developed to apply to a set of several highly correlated variables and an optimization program that chooses the best set of coefficients will be of interest. Also, some work can be done in proving that the quality of the model increases when the highly correlated variables are replaced by the index variable.