An Effective Approach to Resolve Multicollinearity in Agriculture Data

[1]M. Chandrasekhar Reddy, [2]Dr. P.Balasubramanyam, [3]Prof. M. Subbarayudu
[1]*Research Scholar, Department of Statistics, Sri Venkateswara University, Tirupati, Andhra Pradesh*
[2]*Research Scholar, Department of Statistics, Sri Venkateswara University, Tirupati, Andhra Pradesh*
[3]*Professor, Department of Statistics, Sri Venkateswara University, Tirupati, Andhra Pradesh*

Abstract- A commonly appearing problem in most of the econometric data research is the correlated input variables or problem of multicollinearity. This problem arises due to the choice of several input variables which are having high correlation among themselves and it leads to the less précised estimates and huge forecasting errors. This problem can be identified by several approaches like variance inflation factor, correlation matrix etc and it can be removed efficiently by removing or changing the set of input variables in the data. In the current publication, a stepwise procedure is adopted to remove the variables without compromising on model efficiency. The efficiency of the model can be measured or compared by using error diagnostics like RMSE, AIC, BIC etc. For empirical investigation, we used agriculture production of Rice for the past 15 years in the country.

Keywords- *Multicollinearity, VIF, Diagnostics, RMSE, Efficiency.*

I. INTRODUCTION

The major problem in any econometric model building is the selection of appropriate variables as input variables and mis-selection can cause several problems like reduction in efficiency factor, huge forecast error. The ability of any econometrician involves mostly in finding the set of assumptions that are both sufficiently specific to the current study and sufficiently realistic to allow him to take the best possible advantage of the data available to the researcher. One of the most commonly applied econometric model is the General Linear Model and the efficiency of the General Linear Model always depends on its valid assumptions about the error terms as well as independent variables. The violation of each of these assumptions may lead to different consequences in the data and one should be very careful about the violations and consequences of General Linear Model.

II. WHAT IS MULTICOLLINEARITY

Multicollinearity is a statistical scenario where there exists a perfect or exact relationship between the explanatory variables and they are moving very closely with each other. In this scenario, it is difficult to come up with reliable estimates of their individual coefficients and also we can see huge errors in the predictions. In other words, it will result in incorrect conclusions about the relationship between response variable and input variables.

Multicollinearity increases the variances of the parameter estimates and hence this may lead to lack of statistical significance of individual explanatory variables even though the fitted model is a significant model. The presence of multicollinearity can cause major problems with the estimation of β by using ordinary least squares method and the interpretation of those estimates also may goes invalid.

III. CONSEQUENCES OF MULTICOLLINEARITY

Multicollinearity commonly occurs when a large number of independent variables are used in a multiple regression model. It is because some of them may measure the same concepts or phenomena repeatedly. A perfect multicollinearity problem in the data violates the assumption that X matrix is full ranked, making OLS estimates unfeasible. When a model is not full ranked, that is, the inverse of X cannot be defined, there can be an infinite number of least squares solutions. Some of the main consequences of Multicollinearity are listed below.

- *Un-precised estimates from OLS method*

The main consequence of Multicollinearity issue is the reduction in precision of the estimates and more variance of the estimates.We can measure the precision of estimates with some error diagnostic measures like Root Mean Squared Error(RMSE), AIC, BIC etc and high values of these measures indicates less précised estimates of the parameter.

- *Correlated error terms*

In any OLS process, the errors should be independent should not display any pattern of relationship among themselves. But with the problem of multicollinearity, we can see that error terms of having correlation among themselves and it results in huge forecast error.

- *Huge sampling variances of the estimates*

One of the key concepts in deciding the model efficiency is the variance of the estimates and it should be always as small as possible. The problem of multicollinearity always produces the estimates of the parameters as large quantities and they causes the errors in prediction.

- *Testing process with low power*

While testing the significance of coefficients of the parameter estimates, we can see that the power of the test is very low. In other words, the probability of not committing type II error is very less in the testing process. This indicates that the false acceptance of the null hypothesis.

IV. IDENTIFYING MULTICOLLINEARITY

The efficiency of the prediction results are always depends on the effective identification of multicollinearity in the data before actual prediction. There are several methods to detect the problem of multicollinearity and few of them are discussed in the current paper.

A. *Correlation matrix of input variables*

Correlation matrix is the representation of relationship between variables in one single table of diagram. For example, the following correlation matrix shows diagrammatically the correlation values between three input variables. By using correlation matrix, we can identify the close relationships between the input variables and further investigate them to decide about including them in the final model. Generally, a correlation of more than 0.6 can be treated as variable that cause the multicollinearity problem. If there is high multicollinearity between any two predictor variables, then the correlation coefficient between these two variables will be near to unity.

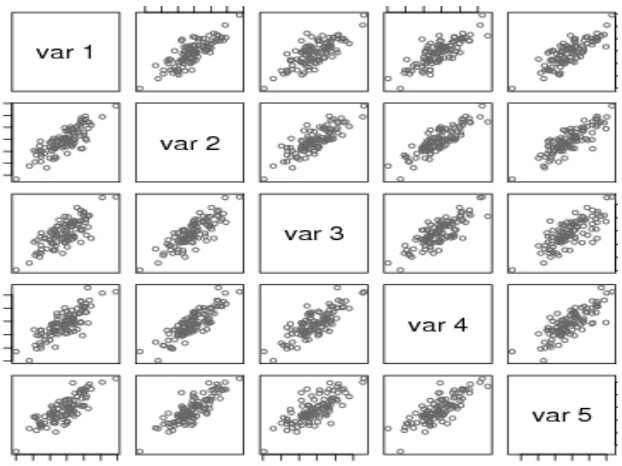

Figure 1: Correlation Coefficient

B. *Variance Inflation Factor*

The most widely applicable method of detecting the multicollinearity is Variance Inflation Factor and it is very accurate in determining the problem of multicollinearity. The expression to find the Variance Inflation factor is given by

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

Where R^2 is the coefficient of determination derived from the model. We can observe that VIF = 1 when R^2 = 0, i.e. when jth variable is not linearly related to the other predictor variables. Similarly, VIF $\rightarrow \infty$ when $R^2 \rightarrow$ 1, i.e. when jth variable is linearly related to the other predictor variables. The VIF is an index which measures how much variance of an estimated regression coefficient is increased because of multicollinearity.

The common thumb rule is if any of the VIF values exceeds 5 or 10, it implies that the associated regression coefficients are poorly estimated because of multicollinearity (Montgomery, 2001). For some practical applications, we can consider this limit up to 2 also to get more précised estimates.

C. *Approach to resolve Multicollinearity*

The best approach for the multicollinearity problem is involves three regular steps as follows.

- Identification (Scatteplots).
- VIF measure
- Removal and refitting

For identification, we can initially use scatterplots to decide the highly related variables and then use the model fitting technique to fix those variables in the first stage, SAS has an option of VIF to REG procedure and gives VIF values in any model. The common syntax for that is

```
proc reg data=data ;
model Response=explanatory variables /vif tol;
run;
quit;
```

From the output of SAS, we can identify and finalize the variable that having maximum variance inflation factor value and in the second stage we can remove it and analyze it further. This procedure is continued till all the existing variables are having VIF value less than 2 and we can say that the model is out of multicollinearity problem in the model.

D. Example from agriculture data

In this paper, we tried to explain the impacts of multicollinearity in the model performance with the help of an agriculture related data. The current numerical data is an extract of rice production in the country for the past 15 years and it includes related information like amount of fertilizers utilized, previous year price, yield per hectare, amount of pesticides used and area cultivated.

Correlation matrix plot- As described in the earlier section, we can use the correlation matrix to study the inter relationships among themselves. The matrix plot of correlations are given as

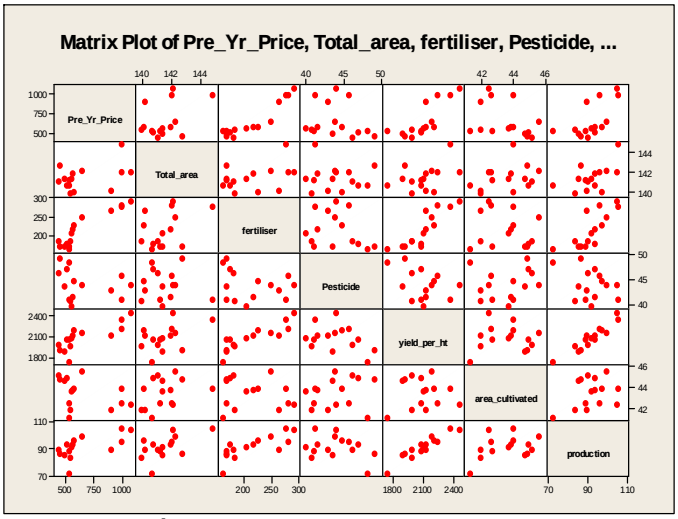

Figure 2: Matrix Plot

The above matrix plot clearly shows some of the variables are intercorrelated among themselves and causing multicollinearity in the data.

Identification of Multicollinearity in the current data:

To identify the multicollinearity in the data, we fitted the multiple regression model and the sas will produce the following output

TABLE I: PARAMETER ESTIMATE TABLE

Variable	Parameter Estimate	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	-113.63363	-8.86	<.0001	.	0
Total Area	0.21625	2.15	0.0600	0.67028	1.49192
fertiliser	0.00244	0.47	0.6468	0.18735	5.33771
Pesticide	-0.01040	-0.25	0.8102	0.75684	1.32128
Yield per Ht	0.04233	30.79	<.0001	0.17589	5.68535
Area cultivated	1.96689	20.83	<.0001	0.67227	1.48750

From the above table, we can see that the VIF values for Fertiliser and Yield per ht is greater than 2 and we can remove them one by one in multiple linar regression model. With this procedure the final estimate table after three iterations will looks as follows.

TABLE I: FINAL ESTIMATE TABLE

Variable	PE	SE	t Value	Pr > |t|	Toleran -ce	VI
Intercept	-114.84707	12.06713	-9.52	<.00 01	.	0
Total Area	0.22474	0.09506	2.36	0.03 97	0.69223	1.44461
Yield per Ht	0.04288	0.00067583	63.46	<.00 01	0.67131	1.48961
Pesticide	-0.00511	0.03892	-0.13	0.89 82	0.81426	1.22811
Area cultivated	1.94682	0.08106	24.02	<.00 01	0.84160	1.18822

Where PE and SE refers to Parameter Estimate and Standard Error respectively. So each of the VIF values in the above table are within the limit and the fitted model can be written as

Production = (-114.84) + (1.80939) Total Area + (0.03508) yield per hectare + (-0.20432) Pesticide + (1.9462) Area cultivated

V. SUMMARY AND CONCLUSIONS:

Based on the empirical analysis of the rice data, we can draw some vital conclusions as follows. We used some of the basic statistics as well as scatter plots to know the relationships between the variable and in the rice data, we can see that the input independent variables are correlated highly among themselves and it gives the initial signal of presence of multicollinearity in the current data. We adopted an approach based on Variance Inflation Factor (VIF) to decide about the variables that are causing the multicollinearity and in the current data, the previous year price is considered as vital variable with highest inflation factor. In the final model, only four independent variables are entered in the model and multicollinearity problem is completely resolved.

VI. REFERENCES

[1] G.S. Maddala, Introduction to Econometrics (2ndedn),Macmillan,New York,1992

[2] Montgomery, D.C., Peck,E.A., Vining, G.G.(2001). Introduction to Linear regression Analysis, 3rd edition, Wiley, New York.

[3] Rao,C.R. and M.B. Rao, Matrix Algebra and its Applications to Statistics and Econometrics, 1 ed., World SciientificPublishing,Singapore ,1998.

[4] Silvey, S.D.[1969], "Multicollinearity and Imprecise Estimation", Journal of the Royal Statistical Society, Series B, Vol. 31, pp. 539-552.

[5] Steward, G.W., "Collinearity and Least Square Regression", Statistical Science.1987,2(1),68-84.

[6] Wilkinson,J.H.[1965], "The Algebraic Eigenvalue Problem".

[7] Wetherill, G.B., Duncombe, P., Kenward, M., Kollerstrom, J.(1986). Regression Analysis with Applications, 1st edition, Chapman and Hall, New York.

M. Chandrasekhar Reddy is Research Scholar in the Department of Statistics, Sri Venkateswara University, Tirupati, Andhra Pradesh, India.