Scientia Africana, Vol. 14 (No.1), June 2015. Pp 189-196
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

MICRONUMEROSITY IN CLASSICAL LINEAR REGRESSION

*** G. M. Oyeyemi, *A. Bolakale, **A. I. Folorunsho and * M. K. Garba**

**Department of Statistics,*
University of Ilorin, Ilorin, Nigeria.

***Department of Mathematics & Statistics,*
Osun State Polytechnic, Iree, Nigeria

Received: 22-12-14
Accepted:16-06-15

ABSTRACT

This study studied the problem of micronumerosity in CLR in other to prescribe appropriate remedy to the problem if encountered at any CLR analysis. The study is aimed at determining an optimum sample size n, such that when the number of observations of variables in CLR is greater than (i.e. n > n*) then micronumerosity is not a problem. It also suggests means of correcting micronumerosity in CLR. The optimum minimum sample size (n) for a given number of independent variables (p) and level of correlation between the dependent and independent variable(s) were determined. Also, Factor Analysis served as the best method of overcoming problem of micronumerosity.*

Key Words: Micronumerosity, Multicollinearity, Linear Regression, Principal Component Analysis, Factor Analysis

INTRODUCTION

When estimating parameters of one sample, two samples, simple or multiple linear regression equations, a researcher is poised to have at the back of his mind the precision of estimation and least standard error. These conditions are as important as carrying out the research itself. In many situations, varieties of conditions such as wrong use of estimator, incorrect formulas, multicollinearity, insufficient sample size, etc., could make the precision of estimation and least standard error unrealizable. In multiple linear regressions, virtually all econometrics literatures blame the failure of an estimate to be non-precise or having large standard error on independent variables being significantly correlated with one another, that is, multicollinearity. This situation is the same if the sample size barely exceeds the number of independent variables in the regression equation. Goldberger (1964) has christened this scenario as the problem of micronumerosity, which simply means small sample size (Gujarati, 2004).

Thus, micronumerosity is a situation whereby the sample size is not sufficient to obtain a precise (unbiased) estimate with relatively least standard errors. We cannot estimate a regression model with Ordinary Least Squares (OLS) method in a case of exact micronumerosity, or having fewer observations than parameters to be estimated. Also, we have relatively large standard errors with near micronumerosity, which means the number of observations barely exceeds the number of parameters to be estimated.

To drive home the importance of sample size, Goldberger (1964) coined the term micronumerosity, to counter the exotic polysyllabic name *multicollinearity*. According to Goldberger, exact micronumerosity (the counterpart of exact multicollinearity) arises when *n*, the sample size, is zero, in which case any kind of estimation is impossible. Near micronumerosity, like near multicollinearity, arises when the number of observations barely exceeds the number of parameters to be estimated.

Regression analysis is concerned with the study of the dependent variable on one or more explanatory variables, with a view of estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter. The number of observations n must be greater than the number of parameters to be estimated. Alternatively, the number of observations n must be greater than the number of explanatory variables in linear regression modeling (Gujarati, 2004).

Goldberger (1989) argues that if the sample size is less or equal to the number of predictors in a CLR equations, it is impossible to estimate the regression parameters or fit an appropriate model to the data. He furthered stated that if the sample size barely exceeds the number of predictors, there is lack of fit in the regression equation even if all other basic assumptions of CLS holds. He discussed the consequences of micronumerosity as similar to that of multicollinearity, and finally suggested the development of a critical value for the sample size, such that when taking observations, if the number of observations is greater than the sample size obtained, then micronumerosity is not a problem. Monogan (2011) also described the numerous consequences of micronumerosity as Goldberger (1990) did and furthered that the methods of correcting micronumerosity are the same as correcting multicollinearity in multiple linear regression. He further suggested the use of multivariate measurement in solving micronumerosity problems as used in multicollinearity problem.

Application of principal component analysis (PCA) in regression has long been introduced by Kendall (1957). Jeffers (1967) suggested obtaining a new set of uncorrelated ordered variables (known as principal components, PC) from the original variables, to achieve an easier and more stable model. The most important of these conditions is that the transformed variables are uncorrelated. Correlation of variables is basically an indication of the strength and direction of a linear relationship between two variables (Weisberg 1980) and it must be considered if redundant data is to be acknowledged and eliminated.

Exploratory factor analysis (EFA) is a causal modeling technique that attempts to “explain” correlations among a set of observed (manifest) variables through the linear combination of a few unknown number of latent (unobserved) random factors. The procedure was originated by the psychologist Charles Spearman in the early

Scientia Africana, Vol. 14 (No.1), June 2015. Pp 189-196
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria ISSN 1118 – 1931

1900's to model human intelligence. Spearman's (1904) single factor model was later generalized by Thurstone (1933, 1947) to multiple factors (Timm, 2002). Factor analysis regression (FAR) provides a model-based estimation method that is particularly tailored to cope with multicollinearity in variables setting. Scott (1966, 1969) was the first to address this issue by deriving "factor analysis regression equations" from a factor model of both the dependent and the explanatory variables. The theoretical deficiencies of Scott's approach are criticized for the most part by King (1969). He showed that Scott's FAR estimator is biased and that the bias still exists asymptotically. Scott's FAR approach has been reconsidered by Lawley and Maxwell (1973), Chan (1977) and Isogawa and Okamoto (1980).

MATERIALS AND METHODS

Here, we start by defining the vector matrix of means μ and variance-covariance matrix ∑ for the dependent variable and p-independent variable(s) such that all the assumptions of multiple linear regression stated in earlier holds. A random sample, starting from size n= p+1, is then simulated using the vector matrix of means μ and variance-covariance matrix ∑ earlier defined from a normal population. The correlation among the variables in the data is then checked to ensure that the underlying assumptions of linearity between dependent variable and the independent variable(s) and no multicollinearity between independent variable(s) hold.

Further, we regress the dependent variable on the independent variable(s), and the model diagnosis using Analysis of Variance, ANOVA. If the F-statistic computed is significant at 0.01 level of significance, then we accept the sample size n used, as the minimum sample size required to avert micronumerosity, otherwise, the sample size is rejected and another sample is taken by increasing the sample size until a significant model is obtained.

At the end of varying the sample size, the correlation between the dependent and independent variable(s) was also varied to see the effect of correlation on the sample size required. This procedure was repeated for p-independent variables ranging from 1 to 10.

A multiple linear regression of the minimum sample size required on the number of predictors p and the correlation between the dependent and independent variable(s) were then obtained to establish the relationship of the likelihood minimum sample size required to avoid micronumerosity.

A principal component regression and factor analysis regression were also performed to check if the problem of micronunerosity could be solved with the two methods. The R^2 obtained was compared to see which of the methods performs better in treating micronumerosity.

RESULTS

The simulation for the research and analysis of variance is conducted to determine which sample size is sufficient enough to avoid micronumerosity. At the different stages of the simulation and analysis, it would be shown how the change in sample size affects the change in F-statistic for the fitness of the regression model, and the t-values for the regressors.

Lastly, using the SPSS package, factor analysis and principal component analysis regression were performed to determine if the methods of correcting multicollinearity is suitable for remedying micronumerosity, and also to determine which of factor analysis or principal component analysis regression performs better in fitting an alternative model, using their respective adjusted R^2 values.

Linear regression model

Starting with one dependent variable, Y and one independent variable, X, Let the vector u = (6 4) and $\mathcal{E} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ be the vector of means and matrix of variance-covariance for Y and X respectively, observations was sampled from a normal population, with n = p + 1 (i.e. n = 2).

A linear regression equation Y = 2 + X was obtained but the model diagnosis could not be obtained because of the singularity of the matrix $X'X$.

By increasing the sample size by 1, i.e, n = 3, the following analysis of variance was obtained.

Table 1: ANOVA for 3 sample size (n = 3) with a dependent and one independent variable.

Coefficient	Estimate	Std. error	t-value	Pr (>|t|
Intercept	2.000	2.581 x 10^{-15}	7.749 x 10^{14}	< 0.000
X	1.000	5.924 x 10^{-16}	1.688 x 10^{15}	< 0.000
F (1, 1) = 2.849 x 10^{30} p-value < 0.000 R^2 = 1.000 $R^2_{adj} = 1.000$ Residual std. error = 9.553 x 10^{-16}				

From the results obtained in Table 1, the p-value (3.772e-16) for the F-statistic is highly significant, and thus we conclude that the model is suitable and appropriate for predicting. Thus, the required sample size (n) when the correlation between the dependent and the independent variable is 1 is 3, i.e., n = 3.

Now, by setting the variance-covariance matrix as $\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$, the following model was obtained, starting with n = 3

Table 2: ANOVA for 3 sample size (n = 3) with a dependent and an independent variable (r = 0.8).

Coefficient	Estimate	Std. error	t-value	Pr (>|t|
Intercept	3.49581	0.17577	19.89	0.032
X	0.57264	0.03834	14.94	0.053
F (1, 1) = 223.1 p-value = 0.053 R^2 = 0.9855 $R^2_{adj} = 0.981$ Residual std. error = 0.075				

From the results in table 2, the p-value (0.053) for the F-statistic is not significant at 0.01 level of significance, and thus we conclude that the model is not appropriate for predicting. We continue to increase the sample size and fit the regression model at each step. The model was found to be significant when n = 11.

Scientia Africana, Vol. 14 (No.1), June 2015. Pp 189-196
© College of Natural and Applied Sciences, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

Table 3: ANOVA for 11 sample size (n = 3) with a dependent and an independent variable (r = 0.8).

Coefficient	Estimate	Std. error	t-value	Pr (>|t|
Intercept	3.49581	0.17577	19.89	0.032
X	0.57264	0.03834	14.94	0.053
F (1, 1) = 223.1 p-value = 0.053 R^2 = 0.9855 R^2_{adj} = 0.981 Residual std. error = 0.075				

From the results obtained in table 3, the p-value (0.003223) for the F-statistic is highly significant, and thus we conclude that the model is suitable and appropriate for predicting. Thus, sample size n required when the correlation r between the dependent and the independent variable is 0.8 is 11.

This process was continued for other p, the number of predictors, i.e., p = 2,3,4,5,…,10, while the correlation r between the dependent and independent variables was varied but zero (0) correlation was maintained between the independent variables. The result is presented in table 4.

Table 4: Minimum required sample size (n) for a given number of predictor(s) p, for different level of correlation between dependent and independent variables.

S/N	Number of predictors (p)	Correlation (r)	Number of observations (n)
1	1	1	3
2	1	0.8	11
3	1	0.6	21
4	2	1	4
5	2	0.8	14
6	2	0.6	28
7	3	1	5
8	3	0.8	17
9	3	0.6	33
10	4	1	6
11	4	0.8	19
12	4	0.6	38
13	5	1	7
14	5	0.8	22
15	5	0.6	43
16	6	1	8
17	6	0.8	24
18	6	0.6	38
19	7	1	9
20	7	0.8	29
21	7	0.6	45

By regressing the sample size (n) on the number of predictors (p) and correlation between dependent and independent variables (r), as presented in table 4, the results of the regression model are presented in Table 5.

Table 5: ANOVA for regression of sample size on number of predictor variables and correlation coefficient

Coefficient	Estimate	Std. error	t-value	Pr (>|t|
Intercept	71.0976	3.5197	20.200	< 0.000
P	1.6991	0.2923	5.813	< 0.000
R	-73.150	4.4143		< 0.000
F (2, 19) = 137.30 p-value < 0.000 $R^2 = 0.935$ $R^2_{adj} = 0.929$ Residual std. error = 3.226				

$$n = 71.1 + 1.7p - 73.15r \dots\dots\dots\dots\dots\dots\dots\dots.(***)$$

Where $0 \leq | r | \leq 1$

Factor analysis and principal component regression

In this section, both factor analysis (FA) and principal component analysis (PCA) are used as a way of solving the problem of Micronumerosity. A p- variate multivariate data set (p=10) with problem of micronumerosity is sampled using R statistical package. PCA is used to obtain the p- components with their respective eigen-values. Scree plot of the eigen values was used to determine the appropriate number of components to be used in the principal component regression. The Scree plot as shown in figure 1 suggested three (3) components for the regression model. The results of the principal component regression model are presented in table 5

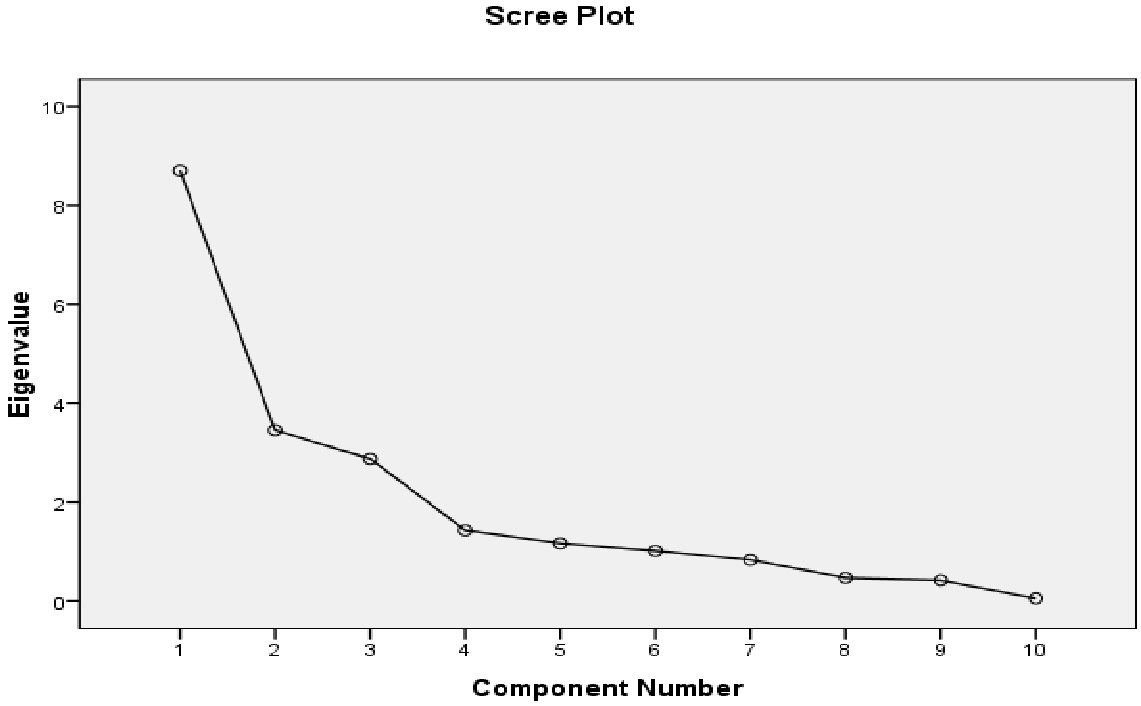

Figure 1: Scree plot of components derived from ten independent variables

Table 5: Model Summary and ANOVA Table for PCA regression

Source	Sum of Squares	df	Mean Square	F	Sig.	R	R Square	Adjusted R Square	Std. Error of the Estimate
Regression	1399.130	3	466.377	151.392	0.000	.986	.972	.966	1.75516
Residual	40.048	13	3.081						
Total	1439.178	16							

The same simulated data was used to obtain factor analysis regression retaining three factors (as the independent variable) to be consistent with principal component regression model obtained earlier. The results of the factor analysis regression are presented in table 6.

Table 6: Model Summary and ANOVA for factor analysis regression

Source	Sum of Squares	df	Mean Square	F	Sig.	R	R Square	Adjusted R Square	Std. Error of the Estimate
Regression	1438.970	3	479.657	29978.56	0.000	1.000	1.000	1.000	0.127
Residual	0.208	13	0.016						
Total	1439.178	16							

DISCUSSION

The simulation study on Micronumerosity problem in classical linear regression shows that the sample size (n) should at least be greater than the relationship:

$$n = 71.1 + 1.7p - 73.15r$$

to avoid Micronumerosity problem. Where p is number of independent variables and r is the minimum correlation coefficient between the dependent variables and independent variable(s).

Also, both principal component and factor analysis regression model have been demonstrated as means of solving Micronumerosity problem. This was achieved by reducing the independent variables into a fewer components or factors, although, factor analysis model is better than principal components regression because it provided higher R^2.

In conclusion, the presence of micronumerosity is prevalent and might not be avoidable in many econometrics researches due to reasons beyond the control of the econometrician. However, the problem of micronumerosity should not be confused with multicollinearity (except both occurs in the same model) since both show nearly the same symptoms and some of the ways of treating multicollinearity are applicable to micronumerosity. Multicollinearity most times does not affect the predictability of a regression equation whereas the first thing micronumerosity affects is the predictability of Classical Linear Regression. It is advisable though to check first multicollinearity in a dataset that has any of the symptoms stated earlier to ascertain which of the two problems to tackle.

Therefore, if there is presence of Micronumerosity in a data set, then additional data should be obtained (increase the sample size n). If it is not possible to increase the sample size,then the best method of remedying Micronumerosity is to use factor analysis regression or principal component regression.

REFERENCES

Chan, N. N. (1977). On an unbiased predictor in factor analysis. Biometrica 64, pp. 642 – 644.

Goldberger, A. S. (1964). *Econometric Theory*, John Wiley & Sons, New York.

Goldberger, A. S. (1989). *Introductory Econometrics*, Harvard University Press, Cambridge, Mass.

Goldberger, A. S. (1990). *A Course in Econometrics*, Harvard University Press, Cambridge, Mass.

Gujarati, D. N. (2004). *Essentials of Econometrics*, 2d ed., McGraw-Hill, New York.

Isogawa, Y. and Okamato, M. (1980). *Linear Prediction in the Factor Analysis Model*.

Jeffers, J.N.R. (*1967*). Two case studies in The application principal component analysis. Applied Statistics, 16, pp. 225-236.

Kendall, M. G. (1957). A Course in Multivariate Analysis. London: Griffin.

King, B. (1969). Comment on ‘Factor Analysis and Regression’. Econometrica 37, pp. 538 – 540.

Lawley, D. N. and Maxwell, A. E. (1973). Regression and Factor Analysis. Biometrica, 60, pp. 331 – 336.

Monogan, J. (2011). *Multicollinearity & Micronumerosity*,University of Georgia.

Scott, J. T. (1966). Factor Analysis and Regression. Econometrica 34, pp. 552 – 562.

Scott, J. T. (1969). Factor Analysis Regression revisited. Econometrica 37, pp. 719 – 722.

Spearman, C. (1904). General Intelligence, “Objectively determined and measured”.American Journal of Psychology, 15, pp. 201-293.

Timm, N. H. (2002). Applied Multivariate Analysis; prentice-Hill, New Jersey, Inc. Englewood Cliff.

Thurstone, L. L. (1933). The theory of multiple factors. Ann Arbor, MI: Edwards Brothers, Inc.

Thurstone, L. L. (1947). Multiple-factor analysis. Chicago: University of Chicago Press.

Weisberg, S. (1980). *Applied Linear Regression*, 2nd edition, Wiley, New York.