Journal of Modern Mathematics and Statistics 2 (3): 120-122, 2008
ISSN: 1994-5388

Empirical Investigation of Effect of Multicollinearity on Type 1 Error Rates of the Ordinary Least Squares Estimators

[1]O.O. Alabi, [2]Kayode Ayinde and [3]B.A. Oyejola
[1]Department of Mathematical Science, Olabisi Onabanjo University, P.M.B. 2002, Ago-Iwoye, Ogun State, Nigeria
[2]Department of Pure and Applied Mathematics, Ladoke Akintola University of Technology, P.M.B. 4000, Ogbomoso, Oyo State, Nigeria
[3]Department of Statistics, University of Ilorin, P.M.B. 1515, Ilorin, Kwara State, Nigeria

Abstract: The effect of multicollinearity on the parameters of regression model using the Ordinary Least Squares (OLS) estimator is not only on estimation but also on inference. Large standard errors of the regression coefficients result in very low values of the t-statistic. Consequently, this study attempts to investigate empirically the effect of multicollinearity on the type 1 error rates of the OLS estimator. A regression model with constant term (β_0) and two independent variables (with β_1 and β_2 as their respective regression coefficients) that exhibit multicollinearity was considered. A Monte Carlo study of 1000 trials was conducted at 8 levels of multicollinearity (0, 0.25, 0.5, 0.7, 0.75, 0.8, 0.9 and 0.99) and sample sizes (10, 20, 40, 80, 100, 150, 250 and 500). At each specification, the true regression coefficients were set at unity. Results show that multicollinearity effect on the OLS estimator is not serious in that the type 1 error rates of β_0 is not significantly different from the preselected level of significance (0.05), in all the levels of multicollinearity and samples sizes and that that of β_1 and β_2 only exhibits significant difference from 0.05 in very few levels of multicollinearity and sample sizes. Even at these levels the significant level different from 0.06.

Key words: Regression model, OLS estimator, multicollinearity, type 1 error rates

INRODUCTION

Regression theory postulates that there exists a stochastic relationship between a variable Y and a set of other variables (X_1, X_2, …, X_n). In other words, Y (called the dependent, endogenous or explained variable) depends on other observed variables, X_1, X_2, …X_n (called independent, exogeneous or explanatory variables). However, one of the assumptions of this model is that the explanatory variables are independent. This is not often the case in economic variables; variables like age and year of experience do exhibit a form of linear relationship. When this assumption is violated, it results into multicollinearity problem (Chatterjee *et al*., 2000).

Multicollinearity could be perfect or imperfect. When it is perfect, estimates obtained are not unique (Searle, 1971). So far multicollinearity is not perfect; the OLS estimator has been shown to be unbiased but inefficient. Other consequences or indications of multicollinearity problem include:

- Small changes in the data can produce significant changes in the parameter estimates (regression coefficients).
- The regression coefficients may have wrong signs and/or unreasonable magnitudes.
- Regression coefficients have high standard errors which result in very low values of the t-statistic and thus affect the significance of the parameters (Chatterjee *et al*., 2000; Fomby *et al*., 1984).

Thus, the presence of multicollinearity in a data set does not only affect parameter estimation but also inferences on the parameters of the model. Consequently, with generated collinear data, this study attempts to investigate empirically type 1 error rates of the OLS estimator.

MATERIALS AND METHODS

Consider the regression model of the form:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + e_t \quad (1)$$

Corresponding Author: Mr. O.O. Alabi, Department of Mathematical Science, Olabisi Onabanjo University, P.M.B. 2002, Ago-Iwoye, Ogun State, Nigeria

Where,

$$\varepsilon_t \sim N(1, \sigma^2)\ t = 1, 2, ..., n$$

y : Is the dependent variable
x_1 and x_2 : Are regressors which exhibit
ρ : Correlation (multicollinearity)
β_0, β_1 and β_2 : Are the regression coefficient (parameters) of the model.

Now, suppose

$$Xi \sim n\left(\mu_i, \sigma_i^2\right) i = 1, 2$$

If these variables are correlated, then X_1 and X_2 can be generated with the equations

$$X_1 = \mu_1 + \sigma_1 z_1$$
$$X_2 = \mu_2 + \rho\sigma_2 z_1 + \sigma_2 z_2\sqrt{1-\rho^2} \quad (2)$$

Where:

$$Z_i \sim N(1, 0) \text{ and } i = 1, 2$$

is the value of correlation between the two variables (Ayinde, 2006; Ayinde and Oyejola, 2007).

Monte Carlo experiments were performed 1000 times for 8 sample sizes (n =10, 20, 40, 80, 100, 150, 250 and 500) and 8 levels of multicollinearity (ρ = 0, 0.25, 0.5, 0.7, 0.75, 0.8, 0.9 and 0.99) with stochastic regressors that are normally distributed. At a particular specification of n and ρ (a scenario), the first replication was obtained by generating $e_i \sim N(1, 0)$. Next, $X_{1i} \sim N(1, 0)$ and $X_{2i} \sim N(1, 0)$ were generated using Eq. 2 such that they exhibit ρ correlation. The values y_i in Eq. 1 were obtained by taking the true regression coefficients as unity. This process is continued until all the 1000 replications had been done. Another scenario is then started until all the scenarios were completed. For each replication in the scenario, the OLS method of parameter estimation was used to obtain estimates of the regression coefficients; and using the t-statistic, hypothesis about the true regression coefficient was tested at 0.05 level of significance to examine the type 1 error rates of each of the regression coefficients. All these were done by writing a computer program using the Time Series Processor (TSP) software. Furthermore, the type 1 error rates were obtained after 1000 replications and these were tested against the pre-selected level of significance (0.05) using the z-statistic to find out whether or not the proportion exhibit difference. Where they were significantly different, the hypothesis was tested against 0.06.

RESULTS AND DISCUSSION

The summary of the type 1 error rates of β_0, β_1 and β_2 at different levels of multicollinearity and sample size are shown in Table 1-3, respectively.

From Table 1, it can be seen that at a specified sample size the type 1 error rate of β_0 is not only the same at all the levels of multicollinearity but also not statistically significant from 0.05, the pre-selected level of significance.

From Table 2, the type 1 error rate of β_1 is not significantly different from 0.05 except in few cases. At these instances, however, the error rates are not significantly different from 0.06.

From Table 3 except when the sample size is very small (n = 10), the type 1 error rate of β_2 is not significantly different from 0.05 and it diminishes, though not

Table 1: Type 1 error rate of β_0 at different levels of multicollinearity (ρ) and sample sizes

	Sample size							
ρ	10	20	40	80	100	150	250	500
0	0.059	0.052	0.05	0.042	0.046	0.047	0.052	0.045
0.25	0.059	0.052	0.05	0.043	0.046	0.047	0.052	0.045
0.5	0.059	0.052	0.052	0.042	0.046	0.047	0.052	0.045
0.7	0.059	0.053	0.052	0.043	0.046	0.048	0.052	0.045
0.75	0.059	0.053	0.052	0.043	0.046	0.048	0.052	0.045
0.8	0.059	0.053	0.052	0.042	0.046	0.048	0.052	0.045
0.9	0.06	0.053	0.051	0.042	0.046	0.048	0.052	0.045
0.99	0.06	0.053	0.051	0.042	0.046	0.048	0.052	0.045

Table 2: Type 1 error rate of β_1 at different levels of multicollinearity (ρ) and sample sizes

	Sample size							
ρ	10	20	40	80	100	150	250	500
0	0.06	0.058	0.056	0.056	0.063	0.056	0.054	0.058
0.25	0.061	0.059	0.059	0.053	0.07+	0.057	0.056	0.05
0.5	0.069+	0.058	0.065+	0.062	0.072+	0.065+	0.063	0.042
0.7	0.061	0.065+	0.063	0.063	0.065	0.065+	0.055	0.053
0.75	0.059	0.064+	0.061	0.057	0.063	0.062	0.053	0.055
0.8	0.056	0.056	0.063	0.056	0.056	0.061	0.057	0.056
0.9	0.057	0.056	0.052	0.05	0.05	0.065+	0.052	0.049
0.99	0.058	0.051	0.047	0.049	0.046	0.057	0.037	0.052

Table 3: Type 1 error rate of β_2 at different levels of multicollinearity (ρ) and sample sizes

	Sample size							
ρ	10	20	40	80	100	150	250	500
0	0.067+	0.058	0.061	0.053	0.053	0.058	0.046	0.058
0.25	0.067+	0.058	0.061	0.054	0.053	0.058	0.046	0.058
0.5	0.063	0.057	0.06	0.054	0.053	0.059	0.046	0.058
0.7	0.063	0.057	0.054	0.053	0.052	0.058	0.046	0.057
0.75	0.063	0.057	0.054	0.053	0.052	0.058	0.046	0.056
0.8	0.064+	0.052	0.054	0.051	0.052	0.058	0.046	0.056
0.9	0.064+	0.053	0.052	0.046	0.049	0.054	0.045	0.056
0.99	0.062	0.051	0.052	0.045	0.045	0.052	0.038	0.052

+ - Significantly different from 0.05 but not 0.06

substantially, as levels of multicollinearity increases. When n = 10, the type 1 error rates exhibit significant different from 0.05 and not 0.06 at most levels of multicollinearity.

Consequently, it can be seen that the effect of multicollinearity on the type 1 error rates of the OLS estimator is trivial in all the levels of multicollineariy and sample size.

CONCLUSION

In spite of the level of multicollinearity and sample size, this study has revealed that the effect of multicollinearity on the type 1 error rates of the OLS estimator is trivial and not serious in that the error rates exhibit no or little significant difference from the pre-selected level of significance.

REFERENCES

Ayinde, K., 2006. A comparative study of the performances of the OLS and Some GLS estimators when regressors are both stochastic and collinear. West Afr. J. Biophy. Biomath., 2: 54-67.

Ayinde, K. and B.A. Oyejola, 2007. A comparative study of the peformances of the OLS and some GLS estimators when regressors are correlated with error terms. Res. J. Applied Sci., 2 (3): 215-220.

Chatterjee, S., A.S. Hadi and B. Price, 2000. Regression analysis by example. 3rd Edn. A Wiley-Interscience Pubication. John Wiley and Sons.

Fomby, T.B., R.C. Hill and S.R. Johnson, 1984. Advanced econometric methods. Springer-Verlag, New York Berlin Heidelberg London Paris Tokyo.

Searle, S.R., 1971. Linear models. New York, John Wiley and Sons.