Applied Mathematics, 2014, 5, 1870-1877
Published Online July 2014 in SciRes. http://www.scirp.org/journal/am
http://dx.doi.org/10.4236/am.2014.513180

Scientific Research

Empirical Determination of the Tolerable Sample Size for Ols Estimator in the Presence of Multicollinearity (ρ)

O. O. Alabi[1], T. O. Olatayo[2], F. R. Afolabi[3]

[1]Department of Mathematical Sciences, Federal University of Technology, Akure, Ondo State, Nigeria
[2]Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria
[3]Department of Mathematics and Statistics, Bowen University, Bowen, Iwo Osun State, Nigeria
Email: otimtoy@yahoo.com

Received 13 April 2014; revised 19 May 2014; accepted 2 June 2014

Abstract

This paper investigates the tolerable sample size needed for Ordinary Least Square (OLS) Estimator to be used when there is presence of Multicollinearity among the exogenous variables of a linear regression model. 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 eight 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 while 1.5, 2.0 and 2.5 were taken as the hypothesized value. The power value rate was obtained at every multicollinearity level for the aforementioned sample sizes. Therefore, whether the hypothesized values highly depart from the true values or not once the multicollinearity level is very high (*i.e.* 0.99), the sample size needed to work with in order to have an error free estimation or the inference result must be greater than five hundred.

Keywords

Regression Model, OLS Estimator Multicollinearity, Power Rate Value and Tolerable Sample Size

1. Introduction

There has been a serious argument between the researchers that multicollinearity problem could be solved with the increase of the sample size while some researchers say that Multicollinearity problem will also increase with

How to cite this paper: Alabi, O.O., Olatayo, T.O. and Afolabi F.R. (2014) Empirical Determination of the Tolerable Sample Size for Ols Estimator in the Presence of Multicollinearity (ρ). *Applied Mathematics*, **5**, 1870-1877.
http://dx.doi.org/10.4236/am.2014.513180

the increase in the size of the sample. [1] stated that Multicollinearity problem could be solved by increase of the size of the sample if the presence of multicollinearity is due to errors of measurement as well as when intercorrelation happens to exist only in our original sample but not in the population [2]. Because of these arguments this paper then investigates the tolerable sample size needed for Ordinary Least Square Estimator to be used when there is presence of Multicollinearity among the exogenous variables of a linear regression model before we can say that multicollinearity problem could be solved with increase of the sample size method.

Regression theory postulates that there exists a stochastic relationship between a variable Y and a set of other variables $(X_1, X_2, \cdots, X_n)$. In other words, Y (called the dependent, endogenous or explained variable) depends on other observed variables, $X_1, X_2, \cdots, X_n$ (called independent, exogenous 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 [3].

Multicollinearity could be perfect or imperfect. When it is perfect, estimates obtained are not unique [4]. If multicollinearity is not perfect, the OLS estimator has been shown to be unbiased but inefficient. Other consequences or indications of multicollinearity problem include:

1. Small changes in the data can produce significant changes in the parameter estimates (regression coefficients).
2. The regression coefficients may have wrong signs and/or unreasonable magnitudes.
3. Regression coefficients have high standard errors which result in very low values of the t-statistic and thus affect the significance of the parameters [3] [5].

Thus, the presence of multicollinearity in a data set does not only affect parameter estimation using the OLS estimator but also inferences on the parameters of the model. Consequently, with generated collinear data, this paper attempts to investigate empirically the most tolerable sample size where power rate value of 0.99 or 1 would be obtained with ordinary least square (OLS) estimator.

2. Methodology

Consider the regression model of the form

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \mathrm{e}_t \tag{1}$$

where $\varepsilon_t \sim N(0, \sigma^2)$ $t = 1, 2, \cdots, n$

y is the dependent variable,

x_1 and x_2 are regressors which exhibit ρ correlation (multicollinearity), and β_0, β_1, and β_2 are the regression coefficient (parameters) of the model.

Now, suppose $X_i \sim N(\mu_i, \sigma_i^2)$ $i = 1, 2$. If these variables are correlated, then X_1 and X_2 can be generated with the equations

$$\begin{aligned} X_1 &= \mu_1 + \sigma_1 z_1 \\ X_2 &= \mu_2 + \rho\sigma_2 z_1 + \sigma_2 z_2 \sqrt{1-\rho^2} \end{aligned} \tag{2}$$

where $Z_i \sim N(0,1)$ $i = 1, 2$ and $|\rho| < 1$ is the value of correlation between the two variables [6]; and [7].

Monte Carlo experiments were performed 1000 times for eight sample sizes (n = 10, 20, 40, 80, 100, 150, 250 and 500) and eight 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 (ascenario), the first replication was obtained by generating $\mathrm{e}_i \sim N(0,1)$. Next, $X_{1i} \sim N(0,1)$ and $X_{2i} \sim N(0,1)$ were generated using Equation (2) such that they exhibit ρ correlation. The values y_t in Equation (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 estimator of parameter estimation was used to obtain estimate of the regression coefficients and hypothesis about the true regression coefficient was tested at 0.05 level of significance using the t-statistic to examine the type II error of the regression coefficients. All these were done by writing a computer program using the Time Series Processor (TSP) software. The result of the effect of type II error rate on OLS estimators by [8] was considered by taken the type II error rate (β) away from 1 to obtain the power rate value for every sample sizes at all levels of multicollinearity. These power rate values were then considered at all levels of multicollinearity for all

the selected sample sizes. Then the sample size with the power rate value of 0.999 or 1.0 was chosen as the most tolerable sample size at each level of multicollinearity and different parameter values, [9] on effects of multicollinearity on the power rates of the Ordinary least Squares Estimators.

3. Results and Discussion

The summary of the most tolerable sample sizes at different level of multicollinearity and different possible combination of the parameter values are shown for β_0, β_1 and β_2 in **Tables 1-8**.

When the true values of β_1 and β_2 are maintained and that of β_0 is allowed to change, The summary of the tolerable sample sizes required for the parameter β_0 to have a power rate value of 0.99 or 1 was determined at different levels of multicollinearity and hypothesized values. The results for these are shown in **Table 1**.

Table 1. The tolerable sample sizes for β_0 when the true values of β_1 and β_2 are maintained and that of β_0 are changing at different levels of multicollinearity.

Parameter values \ ρ	0	0.25	0.5	0.7	0.75	0.8	0.9	0.99
1.5, 1, 1	250	250	250	250	250	250	250	250
2, 1, 1	80	40	40	40	40	40	40	40
2.5, 1, 1	20	20	20	20	40	40	40	40

Table 2. The tolerable sample sizes for β_1 when the true values of β_0 and β_2 are maintained and that of β_1 is allowed to change at different levels of multicollinearity.

Parameter values \ ρ	0	0.25	0.5	0.7	0.75	0.8	0.9	0.99
1, 1.5, 1	150	150	150	250	250	250	250	>500
1, 2, 1	40	40	80	80	80	100	250	>500
1, 2.5, 1	40	40	40	40	40	40	40	>500

Table 3. The tolerable sample sizes for β_2 when the true values of β_0 and β_1 are maintained and that of β_2 is allowed to change, at different levels of multicollinearity.

Parameter values \ ρ	0	0.25	0.5	0.7	0.75	0.8	0.9	0.99
1, 1, 1.5	150	150	250	250	250	250	250	500
1, 1, 2	80	40	40	40	40	80	80	>500
1, 1, 2.5	20	20	20	20	40	80	80	500

Table 4. The tolerable sample sizes for β_1 when the true value for β_0 is maintained and that of β_1 and β_2 are allowed to change at different levels of multicollinearity.

Parameter values \ ρ	0	0.25	0.5	0.7	0.75	0.8	0.9	0.99
1, 1.5, 2	150	150	150	250	250	250	500	>500
1, 1.5, 2.5	150	150	150	250	250	250	500	>500
1, 2, 1.5,	40	40	40	80	80	80	100	>500
1, 2, 2.5	40	40	40	80	80	80	100	>500
1, 2.5, 1.5	40	40	40	80	80	80	100	>500
1, 2.5, 2	40	40	40	80	80	80	100	>500

Table 5. The tolerable sample sizes for β_2 when true value of is maintained and that of β_1 and β_2 are allow to change at different levels of multicollinearity.

Parameter values \ ρ	0	0.25	0.5	0.7	0.75	0.8	0.9	0.99
1, 1.5, 2	150	150	150	250	250	250	500	>500
1, 1.5, 2.5	150	150	150	250	250	250	500	>500
1, 2, 1.5,	40	40	40	80	80	80	100	>500
1, 2, 2.5	40	40	40	80	80	80	100	>500
1, 2.5, 1.5	40	40	40	80	80	80	100	>500
1, 2.5, 2	40	40	40	80	80	80	100	>500

Table 6. The tolerable sample sizes for β_0 when all the values for β_0, β_1 and β_2 are allowed to change at different levels of multicollinearity.

Parameter values \ ρ	0	0.25	0.5	0.7	0.75	0.8	0.9	0.99
1.5, 2.5, 2	100	100	100	100	100	100	100	100
2, 1.5, 2.5	40	40	40	40	40	40	40	40
2.5, 2, 1.5	20	20	20	20	20	20	20	20

Table 7. The tolerable sample sizes for β_1 when all the values for β_0, β_1 and β_2 are allowed to change at different levels of multicollinearity.

Parameter values \ ρ	0	0.25	0.5	0.7	0.75	0.8	0.9	0.99
2, 1.5, 2.5	100	100	100	150	250	250	500	500
2.5, 2, 1.5	40	40	40	40	40	40	40	500
1.5, 2.5, 2	40	40	40	40	40	40	80	80

Table 8. The tolerable sample sizes for β_2 when all the values for β_0, β_1 and β_2 are allowed to change at different levels of multicollinearity.

Parameter values \ ρ	0	0.25	0.5	0.7	0.75	0.8	0.9	0.99
2.5, 2, 1.5	150	150	250	250	250	250	500	500
1.5, 2.5, 2	40	40	40	40	40	80	80	>500
2, 1.5,2.5	40	40	80	80	80	150	250	500

Likewise, when the true values of β_0 and β_2 are maintained and that of β_1 is allowed to change, The summary of the tolerable sample sizes required for the parameter β_1 to have a power rate value of 0.99 or 1 was determined at different levels of multicollinearity and hypothesized values. The results for these are shown in **Table 2**.

When the true values of β_0 and β_1 are maintained and that of β_2 is allowed to change, The summary of the tolerable sample sizes required for the parameter β_2 to have a power rate value of 0.99 or 1 was determined at different levels of multicollinearity and hypothesized values. The results for these are shown in **Table 3**.

The summary of the tolerable sample sizes at different levels of multicollinearity and hypothesized values are shown in **Table 3**.

Also, for all other possible combinations of the parameter values similar results were obtained.

From **Table 1** to **Table 8** the tolerable sample size value decreases as the hypothesized values departed from

the true values in all lower levels of multicollinearity, whereas at higher levels of multicollinearity the required Tolerable sample sizes increases as the hypothesized values departed from the true value. But at very high level of multicollinearity (0.99) the Tolerable sample size needed must be greater than 500 before a result with.

4. Conclusion

In conclusion, at every multicollinearity level the most tolerable sample size was then obtained as the one with the highest value of power rate, which we were able to obtain at a sample size equal or greater than five hundred. This study has revealed that whether the hypothesized values highly depart from the true values or not once the multicollinearity level is very high (*i.e*. 0.99), and the sample size needed to work with in order to have an error free estimation or inference result must be greater than five hundred, if and only if, increments of the size of the sample method would be used as a measure of correction to the presence of multicollinearity.

References

[1] Stone, R. (1961) The Measurements of Consumer Expenditure and Behavior in United Kingdom. Cambridge Publishing Company.

[2] Koutsoyiannis, A. (2003) Theory of Econometrics. 2nd Edition, Palgrave.

[3] Charterjee, S., Hadi, A.S. and Price, B. (2000) Regression Analysis by Example. 3rd Edition, Wiley-Interscience Publication, John Wiley and Sons.

[4] Searle, S.R. (1971) Linear Models. John Willey and Sons, New York.

[5] Fomby, T.B., Hill, R.C. and Johnson, S.R. (1984) Advanced Econometric Methods. Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo.

[6] Ayinde, K. (2006) A Comparative Study of the Performances of OLS and Some GLS Estimator When Regressors Are Both Stochastic and Collinear. *West African Journal of Biophysics and Biomathematics*, **2**, 54-67.

[7] Ayinde, K. and Oyejola, B.A. (2007) A Comparative Study of the Performances of OLS and Some GLS Estimator When Stochastic Regressors Are Correlated with Error Terms. *Research Journal of Applied Sciences*, **2**, 215-220.

[8] Alabi, O.O. (2007) Effects of Multicolinearity on Type 1 and Type 11 Errors of Ordinary Least Squares Estimators. Unpublished M.Sc. Thesis Submitted to the Department of Statistics University of Ilorin, Ilorin.

[9] Alabi, O.O., Ayinde, K. and Olatayo, T.O. (2008) On Effects of Multicollinearity on the Power Rates of the Ordinary Least Squares Estimators. *Journal of Mathematics and Statistics*, **4**, 75-80. http://dx.doi.org/10.3844/jmssp.2008.75.80

Appendix 1. Sample of generated data when $n = 20$, $\rho = 0.8$, $\beta 0 = 2.0$, $\beta 1 = 2.0$ and $\beta 2 = 2.5$.

S/N	Y1	X1	W1	Y2	X2	W2
1	2.32161	−0.30023	0.42772	−0.85314	−1.27768	−1.47984
2	2.96364	0.92812	−0.53482	−0.11007	−0.49427	−1.07706
3	0.048508	−0.70837	−0.61115	3.07106	0.58654	0.86485
4	−0.12086	−0.65842	−0.75213	−0.12663	−0.10619	0.22869
5	1.41147	−0.0109	0.47736	3.27376	0.37188	0.474
6	3.27133	0.217	0.97012	−0.62279	−0.23245	−0.89399
7	−0.57542	−1.03278	−0.25623	3.34894	1.13592	1.49163
8	3.37155	0.89864	0.99842	−1.50461	−0.98755	−0.96275
9	−0.68431	−0.82392	−0.53171	−0.70791	−0.24765	−0.34836
10	2.71043	1.59088	1.53358	0.35442	0.57847	0.4667
11	0.60534	0.4072	−0.62504	−1.61285	−0.72042	−0.86355
12	3.47151	1.90653	2.2132	−1.4025	−1.41486	−1.18214
13	0.79256	0.025209	−0.39163	0.39862	−0.32578	−0.14075
14	3.21058	1.27114	0.6981	0.5857	−0.70043	−0.91221
15	2.40641	0.70946	−0.04596	−0.05572	0.88588	0.18286
16	3.56253	0.59337	1.3989	2.31858	−0.81237	0.27854
17	−0.55504	−0.22546	0.27806	1.72229	0.13895	−0.51126
18	1.03542	−0.13115	0.39195	5.49071	2.0146	1.53258
19	−0.11415	−0.1015	−0.78876	3.57936	0.25744	0.14164
20	−0.49305	−1.46804	−0.49454	1.65012	0.12806	0.6164
			r = 0.696			r = 0.863

S/N	Y3	X3	W3	Y4	X4	W4
1	0.97411	0.24426	1.21819	4.45046	1.27647	1.46138
2	2.46115	2.34348	1.56775	1.75265	0.74402	−0.01299
3	3.08746	−0.23536	0.029878	0.77563	0.52801	−0.08109
4	0.43281	−0.07255	−0.85059	1.44121	0.88985	0.75601
5	3.24358	0.8698	1.16641	−0.64174	−0.73097	−0.14877
6	−1.66004	−1.72698	−0.79048	−0.89615	−1.37954	−1.04859
7	3.35238	0.83796	0.48826	1.3759	−0.01863	−0.06741
8	1.81641	0.53427	0.48421	0.22278	0.21089	−0.38724
9	1.14231	−0.30432	−0.34391	0.89259	−0.0458	−0.35593
10	0.45085	−0.11288	−0.60991	0.99177	−0.4897	−0.46577
11	1.91397	0.13138	1.04747	−2.69003	−0.84199	−1.05445
12	1.78333	−0.53285	0.088131	−0.66616	−0.61328	−0.70889
13	1.38143	−0.48333	0.40239	3.02269	0.37746	0.08943
14	−0.82611	−0.7409	−0.79017	2.77024	−0.10919	0.67636
15	1.84442	0.43143	0.073226	2.3632	0.92789	0.88452
16	−1.10043	−0.66814	0.21208	1.88741	1.13694	0.033692
17	0.35454	0.27869	−0.28152	2.28274	0.60812	1.53032
18	1.40024	−0.34504	−0.47339	2.92647	0.042355	0.7366
19	−0.72434	−1.06744	−0.80752	0.42547	−0.35146	−0.34488
20	3.10226	0.17217	0.65241	−0.18506	−0.27782	−0.50361
			r = 0.729			r = 0.742

S/N	Y5	X5	W5	Y6	X6	W6
1	1.44587	1.19835	1.05035	3.76318	1.73313	0.87803
2	2.26596	0.61061	0.038371	−0.19288	−0.98109	−1.02342
3	−5.01228	−1.92376	−2.57276	−1.46449	−1.11048	−0.62862
4	0.30884	−0.24457	−0.86203	3.76293	2.49907	2.48191
5	0.055006	−0.04312	0.15561	2.10668	1.12347	0.64515
6	2.35692	1.22919	1.19883	2.66584	1.20165	1.86077
7	−1.10259	−0.67015	−0.84007	−0.9324	−0.96854	−1.2697
8	−1.11423	−0.87584	−0.30117	5.84296	0.98543	1.24444
9	−0.68929	−0.43176	−1.15634	−3.34169	−1.97007	−1.93834
10	0.70543	−0.51399	−0.48003	3.2876	1.04222	0.55226
11	3.72866	0.60545	1.61636	0.52895	−0.14204	−0.05199
12	1.33587	−0.73989	−0.16524	3.34549	1.44636	1.25108
13	3.74813	−0.1634	0.87186	0.047596	−0.19262	−1.58574
14	4.81353	1.86272	1.75211	−0.87085	−1.54184	−1.22638
15	−0.767	−0.11142	−1.17672	4.64743	1.80949	1.72129
16	1.43639	0.052157	−0.51635	−2.00234	−0.64876	−1.14894
17	2.51367	0.80357	0.17416	−1.26549	−1.49884	−1.13443
18	3.09059	1.20465	0.62043	0.40964	0.039981	0.015115
19	−0.35757	−0.16193	−0.1297	0.90993	0.61347	0.6578
20	1.27697	−1.52352	0.2058	2.12474	0.97295	0.82414
			r = 0.751			r = 0.933

S/N	Y7	X7	W7	Y8	X8	W8
1	−3.50878	−2.18359	−1.55106	0.076254	−0.23418	0.048921
2	2.4156	0.8717	1.39553	3.3074	1.44723	1.09089
3	3.06562	1.48603	1.55243	1.1868	0.20706	−0.6677
4	−2.92895	−2.38346	−2.21631	0.023394	0.000726	−0.59954
5	2.3153	0.42631	0.70587	1.91768	0.86289	0.39131
6	−0.19984	0.064572	−0.62676	2.32583	0.31563	−0.34671
7	−0.36474	−0.78689	−0.43041	−3.02063	−1.41113	−1.38761
8	0.76456	−0.35896	−0.50852	−1.37897	−0.53391	−0.49237
9	2.01999	0.21003	0.065936	1.81162	−0.24907	−0.80945
10	1.38975	0.026969	1.30834	1.28033	0.047409	−0.93872
11	1.81495	0.73097	−0.075	4.38403	0.75039	1.05031
12	3.48246	2.02637	0.28098	0.39661	−0.55226	0.5132
13	3.61022	0.73517	0.97411	−2.91438	−0.8609	−1.21659
14	3.66682	0.71715	0.48678	0.34524	0.28251	−0.43841
15	1.05873	0.21997	−0.23808	0.90967	0.42648	0.19382
16	1.9437	1.10314	0.52012	2.72093	−0.04672	0.14791
17	0.3991	−0.76497	−0.37913	−0.59405	−0.8968	−0.84802
18	0.051408	−0.57242	−0.4514	2.33572	1.28621	0.88596
19	2.44624	0.45133	1.39705	4.51668	0.6872	0.090591
20	−2.64525	−2.07589	−2.31814	1.53907	1.12448	0.88106
			r = 0.833			r = 0.796

S/N	Y9	X9	W9	Y10	X10	W10
1	2.69193	1.09502	0.14693	−2.73936	−1.0867	−1.36837
2	−1.06365	−0.73137	−1.14807	−1.27555	−0.30111	−0.67624
3	2.93547	0.89406	1.44463	0.31707	−1.22302	−0.073
4	−2.15273	−0.72439	−0.76475	0.49865	−0.12398	0.7807
5	−1.66121	−1.1876	−0.5036	1.47445	0.066794	0.65546
6	−2.41449	−1.34946	−1.60326	1.01931	−0.09158	−0.38849
7	1.70228	0.74786	1.02823	0.15263	−0.28873	0.63814
8	1.97525	0.82349	0.8294	1.24492	−0.20136	−0.39735
9	2.43194	0.77227	1.27333	2.48844	0.49592	−0.13759
10	1.48651	−0.3768	−0.08572	−0.84409	−0.4656	−1.25628
11	−1.70744	−1.08339	−1.75188	1.33773	−0.15735	0.062484
12	0.39137	0.71765	0.21859	−0.4288	−0.97345	−0.82371
13	4.86109	1.03945	0.65675	1.17385	0.32869	−0.16841
14	−0.85694	−1.94513	−1.52449	3.77985	1.67905	1.41282
15	4.1642	1.36677	1.61711	−1.60179	−0.6312	−1.67819
16	−1.174	−1.07873	−1.05705	1.94537	0.48315	0.069188
17	−0.53425	−0.44102	−0.60068	1.17771	−0.40438	1.08911
18	−0.20197	−1.24348	−0.41439	0.64786	−0.35464	0.07857
19	−0.94765	−1.2982	0.17942	1.70651	1.10694	0.21645
20	4.41907	1.44245	1.92737	6.26712	2.07987	2.24085
			r=0.868			r=0.697

Note: $\bar{r} = 0.7908$.

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.