International Journal of Modern Mathematical Sciences, 2014, 11(1): 49-57

Modern Scientific Press

International Journal of Modern Mathematical Sciences **ISSN:2166-286X**

Journal homepage:www.ModernScientificPress.com/Journals/ijmms.aspx *Florida, USA*

Article

Effect of Omitted Variable due to Misspecification Error in Regression Analysis

Babatunde O.S[1], Ikughur A.J[2], Ogunmola A.O[2], Oguntunde P. E[3]

[1]*Mathematics, Statistics and Computer Science Department, Kwararafa University Wukari, P.M.B 1019, Wukari, Taraba State, Nigeria*

[2]*Mathematics and Statistics Department, Federal University Wukari, P.M.B 1020, Wukari, Taraba State, Nigeria*

[3]*Department of Mathematics, Covenant University, Ota, Ogun State, Nigeria*

* Author to whom correspondence should be addressed; Email:olamee2323@gmail.com; atsua2002@hotmail.com; adeniyiogunmola@gmail.com

Article history: Received 19 February 2014, Received in revised form 19 June 2014, Accepted 23 July 2014, Published 28 July 2014.

Abstract: The practical problem is not why specification errors are made but how to detect them. There are number of tests for specification error in detecting the errors of omitted variables from a regression analysis. Using the observations on the dependent variables generated from Microsoft Excel according to the specification labeled true, a bootstrap simulation approach was used for the data generated for each of the models at different sample sizes 20, 30, 50, and 80 respectively each with 100 replications. Using bootstrapping experiment and some properties which estimators should possess if their estimates are to be accepted as good and satisfactory estimates of the parameters, namely, the bias, variance, mean square error, and root mean square error. The models investigated in the bootstrapping experiment consist of the problem of omitted variables. For the models considered, the experiment reveals that the estimated $\beta's$ changes the effect of omitted variable as the coefficient varies in the different models. The effect of omitted variable becomes unstable which produces a bias and inconsistent

Keywords: Specification error, Omitted variables, Bootstrapping, Inconsistent estimator, Estimators.

2000 Mathematical Subject Classification: 62J05

1. Introduction

Misspecification error are the errors associated with the specification of the model, which can take many forms such as omission of relevant variable, inclusion of unnecessary variables, errors of measurement etc.

When an irrelevant variable is included in the model, the presence of such variable gives rise to error. This specification error in the model does not affect the properties of OLS estimators, however, the estimators will generally be inefficient. For example,

$True: y_t = \beta_1 + \beta_2 x_{2t} + u_t$

$Null: y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$

where the true model is specifying a correct model while the null model is specifying the model with irrelevant variable been included.

When a relevant variable in the model is excluded, the specification error will affect the properties of OLS estimator, in the presence of such error, OLS estimators will be bias.

$True: y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t$

$Null: y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$

where the true model is specifying a correct model while the null model is specifying the model when a relevant variable been omitted.

An economic investigation begins with the specification of the econometric model underlying the phenomenon of interest. Some important questions that arise in the specification of models include what variables should be included in the model, what are the probabilistic assumptions made about the y_t(y_t =dependent variable), x_t (x_t = independent variable and u_t(u_t = random error term).

According to Kelvin A. Clarke (2006), when a model is misspecified due to omitted variable, there is always the fear of omitted variable bias. He said a key underlying assumption is that the danger posed by omitted variable can be ameliorated by the inclusion of control variables. Also small amount of nonlinearity in control variables can also have a deleterious effect on the models considered (Achen 2005, Welch 1975).

In a classic regression equation, the estimated $\hat{\beta}$ is little affected by omitted variables provided these are orthogonal to the remaining regressors (J.S. Cramer 2005). He also said, the estimates are still inconsistent and unbiased, and the only inconvenience is an increase of the residual variance and hence of the estimated standard deviation of $\hat{\beta}$. There are two conditions that must hold true for omitted-variable bias to exist in linear regression: the omitted variable must be a determinant of the dependent variable (i.e., its true regression coefficient is not zero); and the omitted variable must be correlated

with one or more of the included independent variables (i.e. the covariance of the omitted variable and the independent variable, cov(x, y), is not equal to zero). Any model that fulfills the classical linear regression model assumptions provide the best, linear and unbiased estimator (BLUE), with respect to OLS. The relevant assumption of the classical linear regression model is that the error term is uncorrelated with the regressors. The presence of omitted variable bias violates this assumption which causes estimators to be bias and inconsistent.

Bootstrapping involves resampling the data with replacement many times in order to generate an empherical estimate of the entire sampling distribution of a statistic (Efron 1979, Efron and Tibshirani 1993).

Inoue and Shintani 2003 reveals that bootstrap provides asymptotic refinements for the generalized method of moment estimator of over-identified linear models when autocorrelation structures of moment functions are unknown.

2. Methods

Consider the standard linear regression model given as

$$Y = X\beta + U$$

where: Y is an n x 1 vector of dependent variables

X is an n x k matrix of regressors

β is a k x 1 vector of parameter

U is an n x 1 vector of disturbance and is normally distributed with covariance matrix proportional to the identity matrix.

A three model of the form

Model	Specification	Problem
i.	True: $y_t = 1.0 - 0.4x_{3t} + x_{4t} + 0.1x_{2t} + u_t$	
	Null: $y_t = \beta_0 + \beta_1 x_{4t} + \beta_2 x_{3t} u_t$	Omitted Variable
ii.	True: $y_t = 1.0 - 0.4x_{3t} + x_{4t} + x_{2t} + u_t$	
	Null: $y_t = \beta_0 + \beta_1 x_{4t} + \beta_2 x_{3t} u_t$	Omitted Variable
iii.	True: $y_t = 1.0 - 0.4x_{3t} + x_{4t} + 2.0x_{2t} + u_t$	
	Null: $y_t = \beta_0 + \beta_1 x_{4t} + \beta_2 x_{3t} u_t$	Omitted Variable

The true model is the model that has been specified correctly without any specification error and the null model is the model that contained the problem of omitted variable i.e x_{2t} is been omitted for all the three models. Observations on the dependent variables are generated according to one of the specification labeled true.

2.1. Criteria for Evaluating the Performance of the Estimators

In this study, the following criteria were used for comparing evaluation of the performance of our estimators;

- Average or mean of estimators in comparison with the true parameter, let $\hat{\beta}$ be the estimates of the parameter β obtained in the i[th] bootstrap replication, we compute

$\hat{\beta} = \frac{\sum_{i=1}^{r} \beta}{r}$, where r = number of replications

- Bias$(\hat{\beta}) = \hat{\beta} - \beta$
- Variance$(\hat{\beta}) = \frac{1}{r}\sum_{i=1}^{r}(\hat{\beta} - \beta)^2$
- MSE$(\hat{\beta}) = \frac{1}{n}E(\hat{\beta} - \beta)^2$
- RMSE$(\hat{\beta}) = \sqrt{}$ MSE$(\hat{\beta})$

2.2. Generation of Data

For the bootstrap experiment, the study consider the specification labeled true model from the above model i.e. $y_t = 1.0 - 0.4x_{3t} + x_{4t} + 0.1x_{2t} + u_t$, assigned a numerical values to all the parameter $(\beta_0 = 1, \beta_2 = 0.1, \beta_3 = -0.4, \beta_4 = 1)$ in the model, the variance σ^2 is also assigned a numerical value on the basis of assumed σ^2, and then the disturbance term U is generated. The U generated was standardized. A random sample of size (n) of X was then selected from a pool of random numbers and numerical values of $y_t = 1.0 - 0.4x_{3t} + x_{4t} + 0.1x_{2t} + u_t$ was computed for each sample size using Microsoft Excel software. The $x's$ and $y's$ generated were then copied from Microsoft Excel into STATA and then bootstrapped and replicated 100 times using a STATA command, each replication produces a bootstrap sample which give distinct values of y which leads to

different estimate of β's for each bootstrap sample regression of y on fixed x. The procedure above is then repeated for different sample sizes and was also performed on each of the three models.

3. Results and Discussion

These are the main results of the paper.

Table 3.1: Comparison of the models with the sample sizes

	β_0	β_1	β_2
Actual	1	-0.4	1
Model1 n=20	1.4070	-0.6916	0.5371
n=30	1.4941	-0.2814	0.0941
n=50	0.8183	0.0963	1.0107
n=80	1.0960	0.0096	0.5986
Model 2 n=20	1.7338	-0.5067	0.5632
n=30	2.0942	-0.2664	-0.2676
n=50	1.3126	0.0616	0.9942
n=80	1.4563	0.2702	0.5369
Model 3n=20	1.9565	-0.2539	0.7176
n=30	1.3712	-0.3814	0.2647
n=50	1.8628	0.1619	0.8303
n=80	1.9576	0.4788	0.4062

Table 3.2: Comparison of coefficients of the models with the same sample size

	β_0	β_1	β_2
Actual	1	-0.4	1
Model 1 n=20	1.4070	-0.6916	0.5371
Model 2 n=20	1.7338	-0.5067	0.5632
Model 3 n=20	1.9565	-0.2539	0.7176
Model 1 n=30	1.4941	-0.2814	0.0941
Model 2 n=30	2.0492	-0.2664	-0.2647
Model 3 n=30	1.3712	-0.3814	0.2647
Model 1 n=50	0.8183	0.0963	1.0107
Model 2 n=50	1.3126	0.0616	0.9942
Model 3 n=50	1.8628	0.1619	0.8303
Model 1 n=80	1.0960	0.0096	0.5986
Model 2 n=80	1.4563	0.2702	0.5369
Model 3 n=80	1.9576	0.4788	0.4062

From table 3.1, the actual $\beta's$ are obtained from the $\hat{\beta}$'s are seriously affected by the omitted variable x_2 in the different models as the coefficient varies because of the unstable nature of the $\hat{\beta}$'s .

From model considered in this study, the $\hat{\beta}'s$ exhibited a damped oscillation in nature apart from β_0 in model 3 where there is an upward trend after n= 30, β_1 rose consistently in model 2 and β_2 oscillated with no negative value.

From table 3.2, when the sample size is 20, the β_0 exhibit a steady upwards positive bias at which model 1 is best,β_1 increase steadily with model 2 as best and for β_2,it rose steadily with model 3 as best. When the sample size is 30, β_0 produces a break at model 2 at which model 1 is best, in β_1, there is a downward trend with model 3 as best, and at β_2, it oscillated with a negative value at model 2. At n=50, β_0 there is a steady increase at which model 1 is best, β_1 fluctuates as all values been positive and atβ_2, the three models performed well with 1 and 2 giving the least bias. At n=80, β_0 increased steadily in which model 1 is best, β_1 produces a steady increase with all values been positive and at β_2, the values dropped consistently as we move from model 1 to model 3. The three models have nearly similar bias.

Table 3.3: Bias table based on 3.1

	β_0	β_1	β_2
Actual	1	-0.4	1
Model1 n=20	0.4070	-0.2916	-0.4629
n=30	0.4941	0.1186	-0.9059
n=50	-0.1817	0.4963	0.0107
n=80	0.0960	0.4096	-0.4014
Model 2 n=20	0.7338	-0.1067	-0.4368
n=30	1.0492	0.1336	-1.2676
n=50	0.3126	0.4616	-0.0058
n=80	0.4563	0.6702	-0.4631
Model 3 n=20	0.9565	0.1461	-0.2824
n=30	0.3712	0.0186	-0.7353
n=50	0.8628	0.5619	-0.1697
n=80	0.9576	0.8788	-0.5938

From table 3.3, most of the biases in the entire β_0 are positive. There are high positive biases in estimate of β_1 as the sample size increases and at β_2, bias are negative in most cases. The β's has been affected by the omitted variable in the different models as the coefficients of the explanatory variables vary. The effects have made the estimates to fluctuate and are inconsistence, but after removing the estimated bias from the estimated β's it becomes stable with the actual. It is obvious that the omitted variable has a serious effect on the true models which has made the model not to be stable.

Table 3.4: Mean square error

	β_0	β_1	β_2
Actual	1	-0.4	1
Model1 n=20	0.0083	0.0045	0.0105
n=30	0.0080	0.0470	0.0275
n=50	0.0006	0.0050	0.0000
n=80	0.0001	0.0021	0.0020
Model 2 n=20	0.0270	0.0005	0.0009
n=30	0.0367	0.0007	0.0537
n=50	0.0020	0.0042	0.0001
n=80	0.0026	0.0056	0.0026
Model 3 n=20	0.0455	0.0010	0.0040
n=30	0.0047	0.0000	0.0180
n=50	0.0148	0.0064	0.0006
n=80	0.0115	0.0096	0.0044

Table 3.5: Ranking of the MSE

	β_0	β_1	β_2	Mean
Actual	1	-0.4	1	
Model1 n=20	4	2	3	3.0
n=30	3	4	4	3.7
n=50	2	3	1	2
n=80	1	1	2	1.3
Model 2 n=20	3	1	3	2.3
n=30	4	2	4	3.3
n=50	1	3	1	1.7
n=80	2	4	2	2.7
Model 3 n=20	4	2	2	2.7
n=30	1	1	4	2.0
n=50	3	3	1	2.3
n=80	2	4	3	3.0

From table 3.4 and 4.5, after the ranking of the MSE, in model1, the higher the sample the better the model except at n=30. At n=80 model is best because of the minimum mean given. In model 2 and 3, the result exhibited a damped oscillation with n=50 and 30 respectively given the best performance.

When the root mean square error is been ranked, at n=20, model 2 produced the best result when the coefficient of the omitted variable is 1. At n=30, model 3 tends to be the best when the coefficient of the omitted variable is 2, when n=50, model 1 produced the best result and at n=80, model 1 produced the best result when the coefficient of the omitted variable is 0.1.

Table 3.6: Root Mean Square Error

	β_0	β_1	β_2
Actual	1	-0.4	1
Model1 n=20	0.0911	0.0671	0.1025
n=30	0.0894	0.2168	0.1652
n=50	0.0245	0.0707	0.0000
n=80	0.0100	0.0458	0.0447
Model 2 n=20	0.1643	0.0224	0.0975
n=30	0.1916	0.0265	0.2317
n=50	0.0447	0.0648	0.0100
n=80	0.0510	0.0748	0.0510
Model 3 n=20	0.2133	0.0316	0.0632
n=30	0.0686	0.0000	0.1342
n=50	0.1217	0.0800	0.0245
n=80	0.1072	0.0980	0.0663

It can be concluded at large that model 1 for each of the sample for the omitted variable in the RMSE has the best result. The MSE and RMSE do not have a stable effect when compared based on sample sizes which may be due to the change in values of the coefficient of the omitted variable. But when compared based on the RMSE, model 1 produced the best result.

4. Conclusions

In this study, we have examined the performances of the estimators i.e. (the bias, variance, mean square error, and root mean square error) in estimating effect of omitted variable in misspecification error in regression analysis. Based on the estimators considered, the estimate has been seriously affected by the omitted variable. As the sample size increases and the coefficient changes the effect becomes unstable. It produces an unreliable and less precise estimates i.e. bias and inconsistence estimates and the change in the coefficient of the omitted variable from 0.1 through 1 and 2 has seriously affected the instability of the results obtained.

Further study is being carried out to increase the number of models considered, bootstrap replication, sample size to see if the effect of the omitted variable will be more noticeable.

References

[1] Achen, Christopher H., Why Lagged Dependent Variables Can Suppress the Explanatory Power of Other Independent Variables. *Annual Meetings of the Political Methodology Section of the American Political Science Association, UCLA*, July 20-22, **2000**.

[2] Anya McGuirk& ArisSpanos. The Linear Regression Model with Autocorrelated Errors: Just Say No to Error Autocorrelation. *Annual Meeting of the American Agricultural Economics Association*, July 28-31, **2002**.

[3] Babatunde O. S, Ikughur A. J, Ogunmola A.O, Oguntunde P. E. On the Effect of Autocorrelation in Regression Model due to Specification Error. International Journal of Modern Mathematical Sciences, 2014, 10(3): 239-246

[4] Beck, Nathaniel. Comparing Dynamic Specifications: The Case of Presidential Approval. *Political Analysis* 3(**1991**):51-87

[5] Clarke Kelvin A. Non-Parametric Model Discrimination in International Relation, *Journal of Conflict resolution*, 47(1) (**2003**): 75-80.

[6] Effron, B. and Tibshirani R.J. An Introduction to the Bootstrap, *Chapman & Hall/CRC*. New York, **1993**.

[7] KayodeAyinde, Emmanuel O. Apata, Oluwayemisi O. Alaba. Estimators of Linear Regression Model and Prediction under Some Assumption Violation.*Open Journal of Statistics*, 2(**2012**): 534-546. http://dx.doi.org/10.4236/ojs.2012.25069

[8] Hesterberg, T., Simulation and Bootstrapping for Teaching Statistics.*American Statistical Association: Proceedings of the Section on Statistical Education*, (**1998**): 44-52.

[9] KayodeAyinde and B.A Oyejola. A Comparative Study of the Performance of the OLS and Some GLS Estimators When Stochastic Regressors are Correlated with the Error Terms.*Research Journal of Applied Sciences*, 2(**2007**): 215-220.

[10] Thursby, J. G. Alternative Specification Error Tests: A Comparative Study. *Journal of the American Statistical Association*. 74 (365)(**1979**): 222-225. DOI:10.1080/01621459.1979.10481641

[11] Thursby, J. G and Schmidt, P., Some Properties of Tests for Specification Error in a Linear Regression Model.*Journal of the American Statistical Association.*72 (**1977**): 359.

[12] Vaduva, I., Bootstrap Method in Density Estimation, *St.Cerc.Mat., Tom*, 46(3)(**1994**): 397-406

[13] Yatchew and Griliches, Effects of misspecification, including omitted variables in the estimation of probit models. *The Review of economics and Statistics*, 67(**1985**): 134-137