SigmaMu 4(1) (2026)

Journal of Mathematics, Statictics and Data Science

https://journals.balaipublikasi.id

On instrumental variable regression method for estimating econometric model perturbed with endogenous variable

M.K. Garba[1*], A.A. Osobase[1], S.B. Akanni[2*]

[1] M.K., Garba Department of Statistics, University of Ilorin, Ilorin, Nigeria.
[1] A.A. Osobase, Department of Statistics, University of Ilorin, Ilorin, Nigeria.
[2*]S.B., Akanni, Department of Mathematical Sciences, Crescent University, Abeokuta, Ogun State, Nigeria.

Received:
Revised:
Accepted:

DOI:
https://doi.org/10.56566/sigmamu.v4i1.621

Abstract: Regression techniques are essential tools utilized to formulate, describe and evaluate econometric models. These techniques rely on some assumptions which, if one or more are violated the naive approach of estimating econometric models will be characterized with one problem or the other. Most often in real life situations, one or more of these assumptions cannot go unfulfilled while modelling econometric data. This study therefore, focuses on the consequences of violation of the assumption that error terms are linearly independent of explanatory variables in classical linear econometric model. For the Ordinary Least Squares (OLS) estimator to be sufficient, the expected value of the error term given the explanatory variable should be zero. And for OLS estimator to be consistent, the covariance between the error term and any of the explanatory variables should be zero. Endogeneity is one of the major challenges of econometric analyses. The effect of endogeneity is bias in estimates and therefore inducing the likelihood of committing the Types I and II errors more rapidly. To examine the behaviours of OLS estimators in the presence of endogeneity and compare its performances with Two-Stage Least Squares (2-SLS) as an alternative method of estimation, data were simulated in the environment of R statistical package in which endogeneity problem was infused into the data. It was discovered that relative to OLS, 2-SLS is consistent and less biased when modelling econometric data that are perturbed with endogeneity problem. Although, the 2-SLS might not be more efficient than the OLS under certain condition, but when there is problem of endogeneity in the model, the choice between OLS and 2-SLS depends on whether the Analyst is willing to trade-off efficiency for biasedness or vice versa in finite sample and asymptotically.

KewWords: Endogeneity, Two-Stage Least Squares, Instrumental variable regression, Classical Multiple Linear Regression Model and Simultaneity.

*Corresponding Author Email: garba.mk@unilorin.edu.ng

Introduction

Explanatory variables of econometric models are usually categorized into endogenous and exogenous variables. The endogenous variables are those variables that are generated by a statistical model whose values are explained by the relationships between functions within the model in which their values are determined by the current workings of the model. Exogenous variables, on the other hand, are the variables whose values are determined outside the model. More formally, an exogenous variable is one that is assumed to be statistically independent of all stochastic disturbance terms of the model while endogenous variables are not statistically independent of those terms.

One important assumption of the classical multiple linear regression (CMLR) model is that the expected value of the error term given the explanatory variable is zero in order for the estimator to be sufficient. Also, the covariance between the error terms and any of the explanatory variables should be zero for the estimator to be consistent. Mathematically, the assumptions are stated as $\mathrm{E}\{e_i/x_i\} = 0$ for sufficiency and that $\mathrm{Cov}\{e_i/x_i\} = 0$ for consistency (Wooldridge, 2009).

The presence of endogeneity in econometric model is a violation of one of the assumptions of CMLR in that the explanatory variables are not independent of the error terms. Endogeneity is one of the major challenges of econometric analyses in management and social sciences.

The effect of endogeneity is bias in estimates and hence causing an investigator to reject a hypothesis that is true (type I error) as well as failing to reject a hypothesis that is false (type II error) more rapidly. The OLS estimates are biased, inconsistent and none of the usual hypothesis testing tools is valid (Hills *et al*., 2012).

Consider a true population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad (1)$$

And assume that the effect of x_2 is omitted from the equation (1) leaving $\widehat{\beta_1}$ to be estimated from

$$y = \beta_0 + \beta_1 x_1 + u \quad (2)$$

The summary of bias in $\widehat{\beta_1}$ is such as given as equation (3)

$$\text{Bias} = \beta_2 \times cor(x_1, x_2) \quad (3)$$

Table 1 depicts the direction of bias with respect to the correlation between variables as well as the magnitude of the effect of the omitted explanatory variable of the response as illustrated by (Woodridge, 2009).

Table 1: Direction of bias with respect to correlation

	$\boldsymbol{cor(x_1, x_2) > 0}$	$\boldsymbol{cor(x_1, x_2) < 0}$
$\boldsymbol{\beta_2 > 0}$	Positive Bias	Negative Bias
$\boldsymbol{\beta_2 < 0}$	Negative Bias	Positive Bias

There are three most popular cases where endogeneity bias might be a problem:

Case I: Correlated missing regressors (omitted variables) - If we omit an important variable z from the regression, which has an effect on y, and this variable is correlated with the explanatory variable *x*, i.e. $cov(z, x) \neq 0$. In this case the effect of the omitted variable z is attributed to the explanatory variable x (Hills *et al*., 2012).

Case II: Measurement Errors - If an approximation *t* of the explanatory variable *x* is used such that $t = x + v$, *v* being the measurement error. The true model is estimated as seen below

$y = \beta_0 + \beta_1 t + \varepsilon$

Where $\varepsilon = -\alpha_1 v + \epsilon$

And $cov(\varepsilon, t) = cov(-\alpha_1 v + \epsilon, x + v) = -\alpha_1 var(v)$

Case III: Simultaneity/ Reverse Causality - A case where y also causes x. Since the error term is contained in y, it is also contained in x.

Therefore, the main focus of this work is endogeneity caused by an Omitted Variable.

Method

In multiple linear regression, the regression coefficient β_1 is interpreted as the change in the dependent variable (y), given a unit increase in the independent variable (x_1), while keeping all other independent variables constant. This however is not usually the case, as in observational data it is not uncommon to find that two or more independent variables are correlated. It is also not uncommon that a particular independent variable that is crucial to explaining the response (y) is numerically unobservable and therefore omitted. For example, job seekers' scores in an aptitude test can be modelled using their individual levels of education and their intellectual abilities as shown in equation (4)

$$Test\ Score = \beta_0 + Education\beta_1 + IntAbility\beta_2 + e \quad (4)$$

A person's intellectual ability can be described in terms of grades and scores, how long it takes them to finish an exercise, how quickly they recall, etc. But it cannot be numerically observed in isolation. That being noted, the regression equation omits intellectual ability and equation (4) becomes

$$Test\ Score = \beta_0 + Education\beta_1 + u \quad (5)$$

It is understood from the basics that the error term accounts for any other entity outside the already stated independent variables that could also affect the values of the dependent variable (y). In this case, we can assume that the new error term consists of the effect of the omitted variable and the initial error term. That is, $* = IntAbility\beta_2 + e$.

The OLS estimate of β_1 from (5) above $\widehat{\beta_1} \rightarrow \beta_1 + \beta_2 \frac{cov(\epsilon,x)}{var(x)}$ might overstate the effect of an individual's level of education on their test scores.

2.1 Detection of Endogeneity

Endogeneity in a regression equation can be detected in several ways. The following are some of them

1. **Measures of Association:** One simple and straightforward way is to find the relationship between the error term in question and all of the independent variables involved in the equation. This could be in form of covariance or correlation coefficient. i.e $cov(\epsilon, x_i)$ or $cor(\epsilon, x_i)$.
   Endogeneity is said to exist where the degree of correlation between the error term and at least one of the explanatory variables is non-negligible. Although, it might not necessarily be zero absolutely.
2. **Hausman Specification Test:** If all criteria are met, we would expect that the estimates produced when OLS is used and those produced using another method of estimation that is supposed to be more tolerant in cases of endogeneity would be the same (Reichstein, 2015).
   Steps
   The Hausman specification test was carried out using the *hausman.systemfit* code under the library (systemfit) in R to check for the consistency of the OLS estimator against that of the 2-SLS estimator. Comparing both to see if there is a significant difference between the coefficients estimated using the OLS estimator and those estimated using the 2-SLS estimator.
   The hypotheses of the Hausman test were stated as

   $$\boldsymbol{H_0}: 2SLS\ is\ more\ consistent \quad \text{versus} \quad \boldsymbol{H_1}: not\ H_0$$

   Test Statistic is

   $$\boldsymbol{H} = (\beta^{OLS} - \beta^{2SLS})' [var(\beta^{OLS}) - var(\beta^{2SLS})](\beta^{OLS} - \beta^{2SLS}) \quad (6)$$

   Steps in carrying out the Hausman test for endogeneity are

   i. Run the two regressions (OLS and 2SLS) and save the outcomes.
   ii. Use the Hausman test to check if the coefficients are different and which is more consistent.

   The decision rule is to reject the null hypothesis if the p-value is less than the set level of significance. Otherwise, do not reject the null hypothesis.

This is where the Two-Stage Least Squares method of estimation comes in, otherwise known as the Instrumental Variable (IV) estimator. If there are endogeneity effects, there would be a significant difference between the estimates of 2SLS and those of OLS. Using the idea that IV estimation will always be asymptotically unbiased whereas OLS will only be unbiased if $Cov(x,u) = 0$. Visual comparison may give a hint but is too weak. Therefore, the Hausman test is required to be carried out.

3. **Durbin-Wu-Hausman Test:** This is a way to individually check if a suspected independent variable suffers from endogeneity.

The hypotheses for Durbin-Wu-Hausman test are stated as

$\boldsymbol{H_0}: x_i\ is\ exogenous$ versus $\boldsymbol{H_1}: x_i\ is\ endogenous$

And the steps for carrying out the Durbin-Wu-Hausman test are as follows:

i. Estimate the reduced form regression against the endogenous variable
ii. Extract the residuals.
iii. Run the initial equation but this time, including the residuals as additional explanatory variables.
iv. Test that the residual is significantly different from zero using any standard test.

The decision rule is to reject the null hypothesis of no endogeneity if the p-value is less that the set level of significance. Otherwise, the null hypothesis will not be rejected (see Woodridge, 2009 and Hills *et al*., 2012).

2.2 Instrumental Variable Regression

The theory of instrumental variables was first introduced by Wright (1928). Overtime, experts like Woodridge (2009), Hills *et al*. (2012), Reichstein (2015) among others have come up with corrections, additions and interpreted the concept in many ways. While Schroeder *et al*. (1986) and Hills *et al*. (2012) opined that instruments as variables should be correlated with the endogenous variable but not with the error term, Reichstein (2015) argued that a good instrument should be correlated with the endogenous variable and not with the dependent variable. These two submissions established the fact that the error term mirrors the dependent variable. Meanwhile, Mogstad *et al*. (2021) explored how 2SLS can be interpreted causally when multiple instrumental variables are used, especially under treatment effect heterogeneity.

Since we cannot correctly estimate any of the $\beta_r's$ in equation (7) when the error term is correlated with any of the regressors x_r.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdot \quad \cdot \quad \cdot + \beta_r x_{ri} + e_i \qquad (7)$$

The estimate of β_r consists of the True β_r and an omitted variable bias. That is, the effect of x_r is overestimated.

Instrumental variables are designed to control for the pollution in the endogenous variable. They are selected such that they are correlated with the problem (endogenous) variable but uncorrelated with the omitted variable which in turn makes it uncorrelated with the error term, since the correlation between the endogenous variable and the error term is caused by the omitted variable being correlated with the endogenous variable.

Instrumental variables do not have a direct impact on y, they are correlated but only indirectly through the endogenous variable which is not an explanatory variable.

More formally, an instrument z for a specific variable of x satisfies the following conditions in order for the IV estimator to be unbiased and consistent

a. Cov (x, z) ≠ 0, which means the instrument variable is correlated with the affected variable.
b. Cov (z, u) = 0, implying that the instrument variable is uncorrelated with the residual.

It is possible to have multiple instrumental variables. Though, it is sufficient enough to use one instrument but more efficient to use all. Instrumental variable regression is carried out using the two stage least squares method of estimation.

Steps in Carrying out Two Stage Least Squares

The steps to be followed in carrying out 2SLS are demonstrated below

Given the model in equation (8) where x_{1i} is endogenous, x_{ji} are the exogenous variables and z_{ni} is an instrumental variable for x_{1i},

$$y = \beta_0 + \beta_1 x_{1i} + \beta_{ji} x_{ji} + \epsilon_i \quad (8)$$

The 2-SLS model becomes

$$y = \beta_0 + \beta_1 x_{1i} + \beta_j x_{ji} + \epsilon_i \quad (9)$$

$$x_{1i} = \alpha_0 + \alpha_n z_{ni} + \alpha_j x_{ji} + v_i \quad (10)$$

It is necessary to note that all the other independent variables should be added in the regression equation against the endogenous variable as it is observed in equations (9) and (10) (see Wooldridge, 2012).

Since $cov(z_{ni}, \epsilon_i) = 0$ then it follows that $cov(\alpha_0 + \alpha_n z_{ni}, \epsilon_i) = 0$

Basically, the 2-SLS method of estimation follows same procedures as OLS

Stage I: After the initial regression equation has been specified, the regression against the endogenous variable is carried out and the estimated value of x_{1i} (i.e $\widehat{x_{1i}}$) is obtained.

Stage II: The estimated value $\widehat{x_{1i}}$ is then used in the initial regression against y instead of the observed value x_{1i}.

What the first stage does is to clean the endogenous variable of the pollution of the error term so that the estimated variable $\widehat{x_{1i}}$ does not suffer from endogeneity.

The estimate for the instrumental variable method of estimation is derived as follows:

Given the model $y = \beta_0 + \beta_1 x + \epsilon$, multiply through by an instrument z, we have

$$\mathbf{z}y = \mathbf{z}\beta_0 + \beta_1 \mathbf{z}x + \mathbf{z}\epsilon \quad (11)$$

which applies that

$$cov(\mathbf{z}, y) = cov(\mathbf{z}\beta_0 + \mathbf{z}\beta_1 x + \mathbf{z}\epsilon) \quad (12)$$

$$= cov(\mathbf{z}\beta_0) + cov(\mathbf{z}\beta_1, x) + cov(\mathbf{z}, \epsilon) \quad (13)$$

Recall that the covariance of a constant equals zero. Also, from the assumptions stated earlier for the instrumental variable, it can be concluded that $cov(z, \epsilon) = 0$.

Hence, $cov(z, y) = \beta_1 cov(\mathbf{z}, x)$

$$\beta_1^{IV} = \frac{cov(\mathbf{z},y)}{cov(\mathbf{z},x)} \quad (14)$$

It can be recalled that

$$\beta_1^{OLS} = \frac{cov(y,x)}{var(x)}) \quad (15)$$

Equation (14) can be represented as

$$\beta_1^{IV} = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})} \quad (16)$$

Bound *et al*. (1995) highlighted the significance of strength of relationship between the endogenous variable *x* and the instrument *z*. They submitted that weak instruments (instruments that are weakly correlated with the endogenous explanatory variable) can lead to large biases in IV estimates and yield misleading inference, even when instruments are valid in theory. Therefore, it is imperative to pay attention to the degree of relationship between the endogenous variable *x* and the instrument *z* in order to avoid results worse than those of OLS.

2.3 Simulation Scheme

The regression equation model used for simulation is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e \quad (17)$$

In order to satisfy the instrumental variable constraints, the variables x_1, x_4 *and* z (the instrument) were jointly generated from a multivariate normal distribution in the R environment. The variables x_2 & x_3 were also generated with specified means and standard deviations. The parameters β_0, β_1, β_2, β_3, β_4 were fixed at 43, 32, 6, 21 and 98 respectively. The relationship between the response and all the independent variables are expressed as equation (18)

$$y = 43 + 32x_1 + 6x_2 + 21x_3 + 98x_4 + e \quad (18)$$

It should be noted that the choice of specified values for the parameters was arbitrary. Also, the conclusions made are not based on the specified values themselves but on the behaviour of the estimated values in relation to the fixed values of the parameters.

The intent is to ensure that the error term is correlated with at least one of the explanatory variables, we randomly assumed that the variable x_4 is unobservable and is therefore omitted from the equation. Hence, x_4 is assumed to be contained in the error term and the regression equation reduced to

$$y = 43 + 32x_1 + 6x_2 + 21x_3 + v \quad (19)$$

where $v = (\beta_4 * x_4) + e$

This indicates that the effect of variable x_4 on y is masked in the error term. And since the omitted variable x_4 is correlated with x_1, then error term v is also correlated with x_1 thereby infusing endogeneity in x_1 which can be stated mathematically as **corr (x_1, v) ≠ 0**.

2.4 Assessment Criteria for the Performances of the Estimators

The assessments of the performances of the estimators considered in this work were based on their absolute biases and mean squared errors as expressed in equations (20) and (21).

2.4.1 Absolute Bias

The absolute bias (AB) criterion measures the deviation of the estimate from the fixed or true value of the parameter. The smaller the absolute bias of an estimator the better.

$$\text{AB}(\hat{\beta}_k) = |\hat{\beta}_k - \beta_k| \quad (20)$$

2.4.2 Mean Squared Error

The mean squared error (MSE) of an estimator measures the efficiency of the estimator, and it is computed using

$$MSE(\hat{\beta}_k) = E(\hat{\beta}_k - \beta_k)^2 \quad (21)$$

The smaller the mean squared error of an estimator the better that estimator is.

The mean squared error of an estimator $\hat{\theta}$ can be expressed mathematically as

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E\{\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\}^2$$
$$= E\{\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\}^2$$
$$= E\{[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]\}^2$$
$$= E\{[\hat{\theta} - E(\hat{\theta})]^2 + 2[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta] + [E(\hat{\theta}) - \theta]^2\}$$
$$= E[\hat{\theta} - E(\hat{\theta})]^2 + 2[0] + [E(\hat{\theta}) - \theta]^2$$
$$= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2$$

That is, $MSE(\hat{\theta})$ = Variance of $\hat{\theta}$ + square bias of $\hat{\theta}$.

It should be noted that if the bias of $\hat{\theta}$ is zero, then the MSE ($\hat{\theta}$) = Var ($\hat{\theta}$).

2.5 Confidence Intervals for the Estimated Parameters

The estimates of β_i obtained using both OLS and 2SLS are not fixed. If another sample of n subjects is taken from the same population, it is more probable to obtain different estimates for β_i. Hence, it is imperative to build a level of confidence around our parameter estimates. So, a 100(1 − α)% confidence limits for β_i given by (22) was used. It will give us an idea of how close our estimates are to the true values of population parameters.

$$\hat{\beta}_i \pm t_{\frac{\alpha}{2},n-p} SE(\hat{\beta}_i) \quad (22)$$

for i = 0, 1, 2, 3, . . ., p, where p is number of parameters in the model. Note that n-p is the degrees of freedom for residual on the ANOVA table.

Result and Discussion

Here, results of several analyses are presented and discussions are made accordingly.

3.1 Diagnosing endogeneity

After getting the residual values from the 1st stage model of the 2-SLS estimator, the values were added in the initial model as an extra independent variable and checked for significance. The results presented in Table 2 were obtained

Table 2: Summary Statistics for the Model after including Residuals

	Estimate	Standard Error	t-value	P-value
Intercept	555.1	3.177e-03	174688	< 0.0001
x_1	31.82	6.375e-04	49910	< 0.0001
x_2	-1.913	7.965e-04	-2401	< 0.0001
x_3	23.01	2.230e-04	103188	< 0.0001
Residual	5.501	3.255e-06	1690106	< 0.0001

The decision rule is that the null hypothesis is not rejected if the p-value associated with the residual term is greater than the level of significance implying that the residual is not significant. Otherwise, the null hypothesis rejected. In this case, the p-value for the residual term is statistically significant in predicting the response. This leads to rejection of the null hypothesis and conclude that x_1 is indeed endogenous.

3.2 Hausman Specification Test

The test statistic reported in R after the Hausman specification test was carried out to check for the most consistent estimator is

$$\textit{Hausman} = -50367,\ df = 4,\ p\text{-}value = 1$$

Here, the decision rule is not to reject the null hypothesis if the p-value is greater that the level of significance. Otherwise, reject the null hypothesis. With respect to the hypotheses stated earlier, it is safe to conclude that 2SLS is more consistent since the null hypothesis was not rejected at 5% level of significance.

Table 3 has absolute biases and mean squared errors for OLS and 2SLS together with their respective parameter estimates across various sample sizes. Aside from the intercept $(\widehat{\boldsymbol{\beta}}_0)$, $\widehat{\boldsymbol{\beta}}_1$ for OLS has higher deviations from the true value of the parameter. This is unconnected to the fact that the endogeneity was infused into the model through variable x_1. This is technically known as omitted variable bias. Globally, it can be deduced that 2SLS is preferable to OLS using absolute bias yardstick.
Also from Table 3, it be observed that 2SLS has smaller MSE consistently over the various sample sizes considered. Hence, 2SLS is said to be more efficient than OLS in estimating the econometric model perturbed with endogeneity.

Table 3: Parameter Estimates with associated Absolute Biases & Mean Squared Errors

Sample Size	Estimator	β_0 = 43	AB for β_0	β_1 = 32	AB for β_1	β_2 = 6	AB for β_2	β_3 = 21	AB for β_3	MSE
30	OLS	332.366	289.366	100.889	68.889	4.206	1.794	20.445	0.555	3132.25
	2SLS	*38.904*	*4.096*	*32.5*	*0.5*	*5.929*	*0.071*	*20.972*	*0.028*	*0.572*
60	OLS	326.908	283.908	100.565	68.565	4.393	1.607	21.493	0.493	1466.12
	2SLS	*38.891*	*4.109*	*32.499*	*0.499*	*5.983*	*0.017*	*20.955*	*0.045*	*0.287*
80	OLS	323.93	280.93	100.615	68.615	7.303	1.303	21.319	0.319	1072.43
	2SLS	*38.891*	*4.109*	*32.497*	*0.497*	*5.978*	*0.022*	*20.956*	*0.044*	*0.28*
120	OLS	327.973	284.973	100.587	68.587	4.962	1.038	21.278	0.278	726.108
	2SLS	*38.889*	*4.111*	*32.488*	*0.488*	*5.933*	*0.067*	*20.976*	*0.024*	*0.143*
250	OLS	327.833	284.833	100.593	68.593	4.975	1.025	21.223	0.223	345.641
	2SLS	*38.883*	*4.117*	*32.485*	*0.485*	*5.936*	*0.064*	*20.966*	*0.034*	*0.069*
350	OLS	328.417	285.417	100.586	68.586	5.113	0.887	21.201	0.201	247.393
	2SLS	*38.889*	*4.111*	*32.472*	*0.472*	*5.934*	*0.066*	*20.971*	*0.029*	*0.049*
500	OLS	327.826	284.826	100.614	68.614	5.294	0.706	20.822	0.178	172.245
	2SLS	*38.882*	*4.118*	*32.464*	*0.464*	*5.937*	*0.063*	*20.974*	*0.026*	*0.034*
1000	OLS	327.77	284.77	100.601	68.601	5.313	0.687	20.896	0.104	85.934
	2SLS	*38.873*	*4.127*	*32.397*	*0.397*	*5.94*	*0.06*	*20.977*	*0.023*	*0.017*

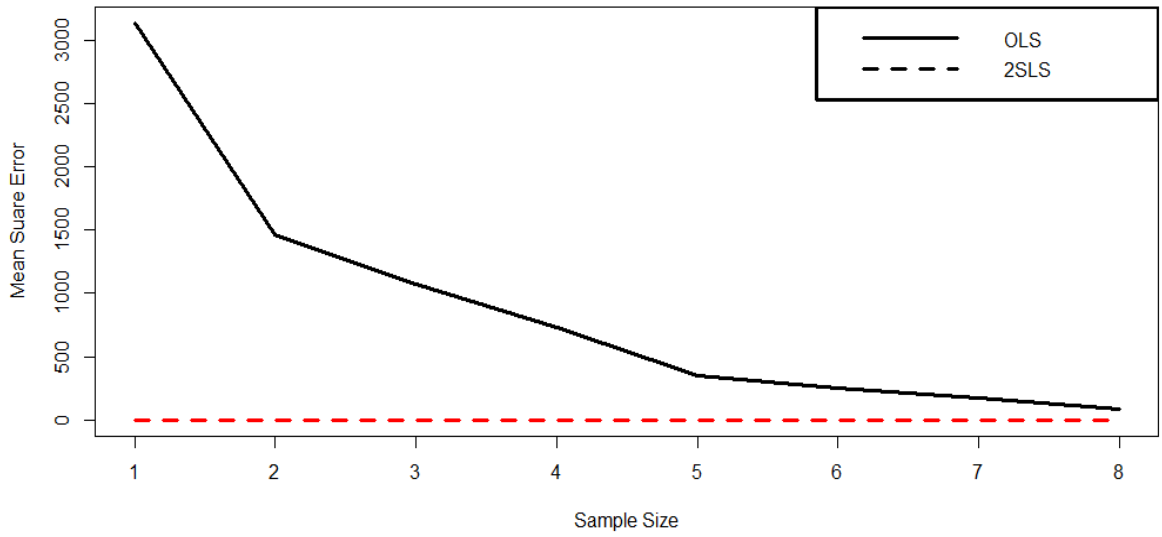

Figure 1: Graph Showing the Estimated Mean Square Error at Various Sample Sizes

The graph in Figure 1 has the mean squared errors for both OLS and 2SLS at various sample sizes. It corroborates the consistency of 2SLS for estimating econometric model with endogeneity. Thus, 2SLS is superior to OLS for modelling non-exogeneity econometric model.

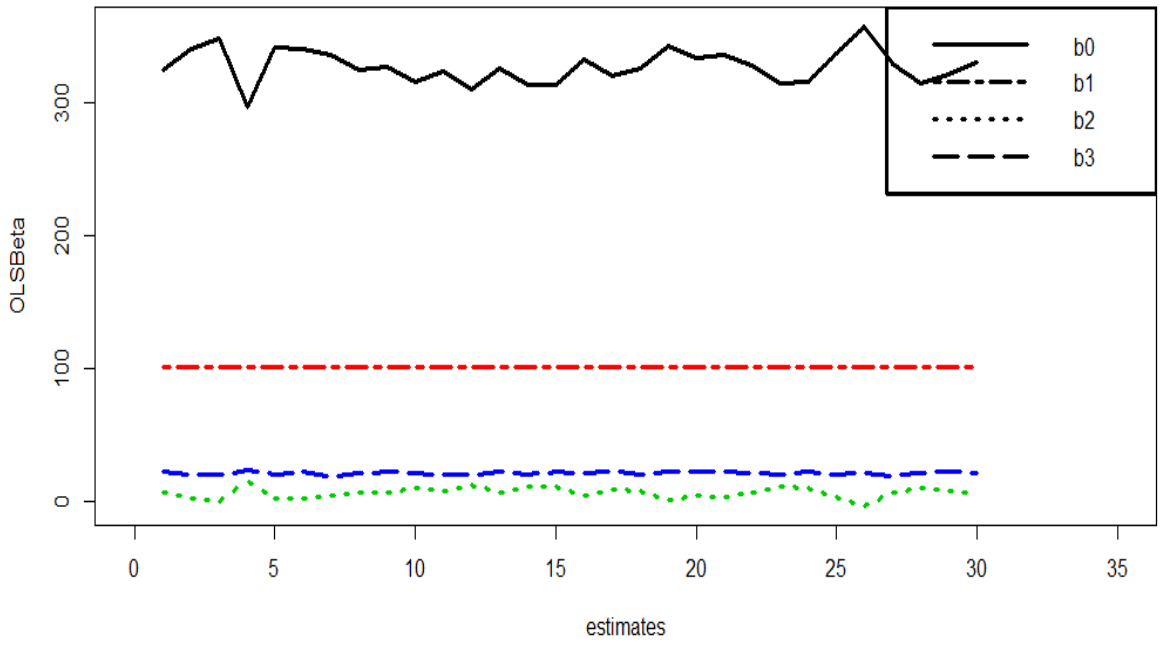

Figure 2: Graph of OLS Parameter Estimates over 30 Iterations at n = 1000

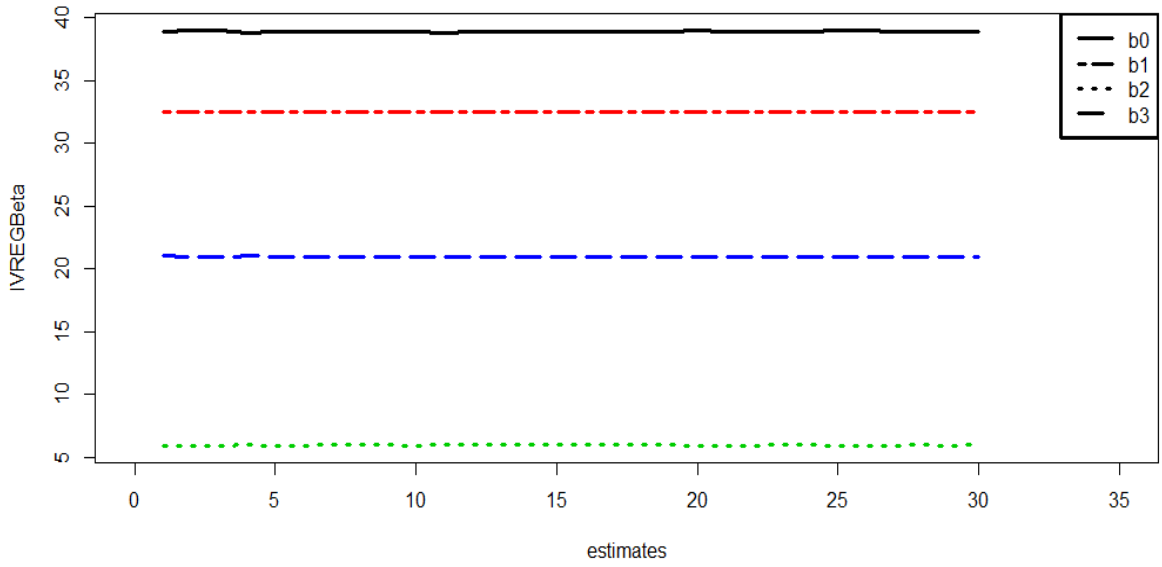

Figure 3: Graph of 2SLS Parameter Estimates over 30 Iterations at n = 1000

Figures 2 and 3 show the behaviours of the estimated parameters of OLS and 2SLS respectively. Apparently, parameters of 2SLS behaved relatively stable compared to those of OLS. The deductions from the graphs underscore the preference of 2SLS over OLS.

Table 4: Confidence Interval of Estimates at n = 100

True Parameter	OLS		2SLS	
Value	**LCL**	**UCL**	**LCL**	**UCL**
$b_0 = 43$	243.2673	412.3758	38.4363	39.3419
$b_1 = 32$	86.3717	114.7435	32.4997	32.5003
$b_2 = 6$	-17.9485	29.6536	5.8282	6.0390
$b_3 = 21$	13.84045	28.1137	20.9432	21.0064

Table 5: Confidence Interval of Estimates at n = 1000

True Parameter	OLS		2SLS	
Values	**LCL**	**UCL**	**LCL**	**UCL**
$b_0 = 43$	301.9745	353.7519	38.7458	39.0234
$b_1 = 32$	96.2448	104.9549	32.4998	32.5001
$b_2 = 6$	-1.4357	13.0925	5.9026	5.9669
$b_3 = 21$	18.7083	23.0611	20.9663	20.9855

The estimates of the constructed confidence intervals for both OLS and 2SLS are presented in Tables 4 and 5 for small and large samples respectively. It is apparent that the intervals for 2SLS estimated parameters are narrower than those of OLS. It implies that 2SLS is more precise than OLS in both small and large samples.

Conclusion

The inconsistent and highly biased estimates resulting from the OLS estimator cannot be overlooked. The MSE reduces as sample size increases and it becomes more stable as seen in Figure 1 but when viewed closely, it is observed that the effect of the endogenous variable is being overestimated and even when it seems to be stable, the bias in ever present as it is the major effect of endogeneity. Moreover in a real life situation, getting more samples is not always possible.

Meanwhile, the parameter estimates from the **2SLS** estimator were quite close to the true values even at a low sample size and the mean squared error is quite low in comparison to those of **OLS**. The confidence intervals of OLS are wide and unreliable in finite sample and asymptotically.

Generally speaking, the OLS estimator loses some of its finest properties when underlying assumptions are violated. Particularly, in the presence of endogeneity, the 2- stage least square/Instrumental variable technique of estimation produces estimates that are more consistent and stable with less bias than those of the ordinary least squares. It also yields predictors that are applicable in the theoretical sense making 2-SLS a less biased estimator than OLS estimator in the presence of endogeneity.

This study demonstrated that in the presence of endogeneity, the ordinary least squares (OLS) estimator is biased and not stable. Two-stage least squares (2SLS) estimator outperformed OLS when endogeneity exists in the dataset. The results from this study is in agreement with Greene (2003), Andren (2007) and Hills *et al*. (2012). It was discovered that the values of the estimated parameters using 2SLS are closer to the true parameter values than those obtained using the ordinary least squares estimator.

Acknowledgements

We acknowledge the contributions of God, the lead author and the co-authors. No fund was received for this work.

References

Andren, T. (2007). Econometrics. Ventus Publishing (Bookboon).

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association, 90*(430), 443–450.

Chatterjee, S., & Hadi, A. S. (2006). Regression analysis by example (4th ed.). John Wiley & Sons.

Greene, W. H. (2003). Econometric Analysis (5th ed.). Prentice Hall.

Gujarati, D. N. (2004). Basic Econometrics (4th ed.). McGraw-Hill.

Hills, R. C., Griffiths, W. E., & Lim, G. C. (2012). Principles of econometrics. John Wiley & Sons.

Lili, I., & Kosta, A. (2024). Applying an Ordinary Least Squares (OLS) Regression Model on Processed Air Quality and Environment Data. British Journal of Environmental Sciences, 12(2), 49–58

Mogstad, M., Torgovitsky, A., & Walters, C. R. (2021). The Causal Interpretation of Two-Stage Least Squares With Multiple Instrumental Variables. American Economic Review, 111(11), 3663–3698.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). Applied Regression Analysis: A Research Tool (2nd ed.). Springer-Verlag.

Reichstein, T. (2015). Econometrics, Endogeneity. Department of Innovation & Organization Economics, Copenhagen Business School.

Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). Understanding Regression Analysis: An introductory Guide (Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-057). Sage.

Shin, K., You, S., & Kim, M. (2021). A Comparison of Two-Stage Least Squares (TSLS) and Ordinary Least Squares (OLS) in Estimating the Structural Relationship between After-School Exercise and Academic Performance. Mathematics, 9(23), 3105.

Wooldridge, J. M. (2012). Introductory Econometrics: A Modern Approach (4th ed.). Cengage Learning.

Wright, P. G. (1928). The Tariff on Animal and Vegetable Oils. Macmillan,

Zhao, A., Ding, P., & Li, F. (2025*). Interacted Two-Stage Least Squares with Treatment Effect Heterogeneity*. arXiv:2502.00251v2