Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 10-2024
DOI: 10.5281/zenodo.14196326

AN OVERVIEW ON COMPARATIVE METHODOLOGY OF CLASSICAL OLS AND TWO-STAGE TECHNIQUES IN REGRESSION ANALYSIS MODEL

ELSAYIR, H.A.

Department of Mathematics, Al-Qunfudah University College, Umm Al- Qura University, Mecca, Saudi Arabia. Email: hamusa@uqu.edu.sa, Habibsayiroi@Yahoo.com

Abstract

This article provides a methodological overview for the classical Ordinary Least Square (OLS) and Instrumental variable (or Two-stages) technique by discussing the conditions that satisfy the use of the method. The Ordinary Least Squares (OLS) estimator in the classical linear regression model is considered the Best Linear Unbiased Estimator (BLUE) if the model assumptions hold. When these assumptions are met, OLS provides linear, unbiased estimates with the smallest variance compared to other estimators. Although alternative methods exist, especially for specific issues like multicollinearity or heteroscedasticity, OLS remains the most efficient estimator, particularly in large samples. Even when assumptions are slightly violated, OLS generally performs well, maintaining its reliability due to the central limit theorem.

Keywords: Instrumental Variable; Lagged Variables; Multicollinearity *OLS Estimator.*

1.0. INTRODUCTION

Regression analysis is a method for studying the relationship between a dependent variable and one or more independent variables. The interest is in estimating a regression function that will be useful for forecasting the dependent variable or for testing econometric or economic hypotheses about that relationship through statistical inference. The simplest linear regression model is one in which a single independent variable is used to predict the dependent variable.

Before discussing the two-stage least squares method for a linear regression model with endogenous explanatory variables, it is helpful first to consider the assumptions of the general classical linear model, because all the conclusions derived for the general case will cover the setting where some of the independent variables are potentially endogenous. Moreover, to discuss the properties of the two-stage least squares estimator, we need to discuss some of the properties of the ordinary least squares estimators first.

One of the crucial assumptions of applying the Ordinary Least Squares (OLS) technique in regression analysis is the absence of endogeneity among the explanatory variables. Violation of the endogeneity assumption might lead to serious biases in estimations; thus, the choice of an appropriate method of estimation is of high importance. In the presence of endogeneity problems, instrumental variable regressions (IVRs) play a significant role. Two-Stage Least Squares (2SLS) and Generalized Method of Moments (GMM) estimation techniques are commonly used in IVR settings. This article provides a comparison between 2SLS and OLS techniques, both theoretically. It has been noted that the use of the standard 2SLS technique, rather than OLS, without removing the endogeneity among the variables or using the wrong instruments might severely bias the estimations and lead to

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 10-2024
DOI: 10.5281/zenodo.14196326

irrelevant results. Especially in the OLS case, we might frequently observe insignificant estimations, regardless of the true relationships among the variables. The use of the 2SLS technique might help avoiding that problem and provide more accurate results.

2.0. LITERAURE REVIEW

The comparative analysis of Two Stage Least Squares (2SLS) and Ordinary Least Squares (OLS) in regression models has garnered considerable attention in econometric literature. The foundational work by Lleras-Muney and Dhrymes (Lleras-Muney & J. Dhrymes, 2002) provides critical insights into the relative efficiency of these estimators, particularly in the context of grouped versus ungrouped data. The results presented by Lleras-Muney and Dhrymes (Lleras-Muney & J. Dhrymes, 2002) also suggest that employing individual data in the first stage of the 2SLS process can yield greater efficiency, thereby advocating for a more data-informed approach in econometric modeling.

The exploration of regression analysis techniques, particularly the comparison between two-stage methods and classical Ordinary Least Squares (OLS) estimators, has garnered significant attention in recent years. A foundational contribution to this discourse is presented in "Combined Estimators as alternative to Ordinary Least Square Estimator" by (Ayindea, 2013). This article elucidates the conditions under which the OLS estimator is deemed the Best Linear Unbiased Estimator (BLUE), emphasizing that its efficacy hinges on adherence to the underlying assumptions of the regression model.

(Ayindea, 2013) critically examines the limitations of OLS, particularly in scenarios where the assumptions of independence among regressors and error terms are violated, leading to issues such as multicollinearity and autocorrelation. The research introduces a novel approach by integrating Feasible Generalized Least Squares (FGLS) estimators with Principal Component (PC) Analysis. Through empirical analysis, the study reveals that while the OLS estimator maintains its status as the most efficient estimator, the proposed combined estimators demonstrate competitive performance, especially as sample sizes increase. The insights drawn from (Ayindea, 2013)'s work serve as a critical lens through which to view the ongoing debate surrounding the effectiveness of various estimation techniques in regression analysis. The article not only reinforces the foundational principles of OLS but also highlights the necessity for researchers to consider alternative methodologies in the face of potential violations of regression assumptions.

As the literature progresses, it is essential to consider how the insights from Kuchibhotla et al. (K. Kuchibhotla et al., 2018) inform the understanding of both classical OLS and two-stage regression techniques. The implications of model misspecification raised in this study serve as a crucial backdrop for analyzing the efficacy and application of these methodologies in multiple regression analysis models.

The critical evaluation of OLS provided by the authors sets the stage for a deeper investigation into the comparative advantages and limitations of two-stage regression techniques, as the field continues to evolve.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 10-2024
DOI: 10.5281/zenodo.14196326

Further insights about this topic and related issues could be found in later articles such as (Shin, & Kim, M. (2021), Keane, & Neal. (2021), Young, A. (2022), Fakhreddin, F. (2023) , Keane, M. & Neal, T. (2023) and Elsayir,H.A(2024)).

3.0. THE MATHEMATICAL MODEL

Consider the following linear regression model:

$$Y_i = \beta_{0+}\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \mu_i \ldots [1]$$

where:

- Y_i is the dependent variable (outcome) for observation i
- x_i, are the independent variables (predictors) for observation i.
- $\beta_0, \beta_1, \ldots, \beta_k$ are the parameters (coefficients) to be estimated.
- μ_i is the error term or residual for observation i
- $i = 1,2, \ldots, ni = 1,2, \ldots, n$ refers to the i -th observation.

  In matrix form, this can be written as:

$$Y = X\beta + u \ldots [2]$$

where:

- Y is an $n \times 1$ vector of the dependent variable.
- X is an $n \times (k+1)$ matrix of the independent variables (including a column of ones for the intercept).
- β is a $(k+1) \times 1$ vector of parameters (including the intercept term β_0.
- u is an $n \times 1$ vector of the residuals.

The objective of OLS is to minimize the sum of squared residuals:

Minimize:

$$S(\beta) = \sum_{I=1}^{n} u_i{}^2 = \sum_{I=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{I=1}^{n} (Y_i - \sum_{j=0}^{k} \beta_j X_{JI})^2 \ldots [3]$$

In matrix notation, this becomes:

$$S(\beta) = (Y - X\beta)^T (Y - X\beta) \ldots [4]$$

To minimize the sum of squared residuals, we take the derivative of $S(\beta)$ with respect to β and set it equal to zero:

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^T (Y - X\beta) = 0 \ldots [5]$$

After simplifying this gives the solution to the normal equations and obtained by solving for β:

$$= X^T X\beta - X^T Y \ldots [6]$$

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 10-2024
DOI: 10.5281/zenodo.14196326

$$\hat{\beta} = (X^T X)^{-1} XY \ldots [7]$$

This equation provides the OLS estimator $\hat{\beta}$, which is the vector of estimated regression coefficients.$\hat{\beta}$ is the vector of estimated coefficients that minimizes the sum of squared differences between the observed values Y and the predicted values $\hat{Y} = X\hat{\beta}\{Y\} = \hat{X}\{\hat{\beta}\}\hat{Y} = X\hat{\beta}$ to reach to the formula for calculating the best-fitting line (or hyperplane) in the context of OLS regression.

The predicted values of Y denoted by $\hat{Y}$ can be obtained by:

$$\hat{Y} = X\hat{\beta} \ldots [8]$$

The residuals u, which are the differences between the observed and predicted values, are calculated as:

$$\hat{u} = Y - \hat{Y}$$

$$= Y - X\hat{\beta} \ldots [9]$$

In summary, the OLS method estimates the coefficients by solving the normal equations to minimize the sum of squared residuals. The OLS solution gives the best linear unbiased estimates under the assumption that the error terms are homoscedastic, uncorrelated, and have zero mean.

A two-stage regression model, also known as a Two-Stage Least Squares (2SLS) model, is typically used when dealing with endogeneity in a regression model. Endogeneity can occur due to omitted variables, measurement errors, or simultaneity, where one or more independent variables are correlated with the error term. The 2SLS method solves this by using instrumental variables that are correlated with the endogenous explanatory variables but not with the error term. Here are the steps and the equations for a two-stage regression model:

Stage 1: Instrumental Variable (IV) Regression

In the first stage, you regress the endogenous variable(s) on the instrumental variable(s) and other exogenous variables to estimate the predicted (fitted) values of the endogenous variables. These predicted values will be uncorrelated with the error term and can then be used in the second stage.

Let the original model be:

$$Y = \beta_{0+} + \beta_1 X_{1+} + \beta_2 X_2 + u \ldots [10]$$

where:

- Y is the dependent variable.
- X_1 is an endogenous explanatory variable.
- X_2 is an exogenous explanatory variable.
- u is the error term.
- $\beta_0, \beta_1, \beta_2$ are parameters to be estimated.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 10-2024
DOI: 10.5281/zenodo.14196326

In the first stage, we replace the endogenous variable X_1 with an instrumental variable $Z_{1,}$, which is correlated with X_1 but uncorrelated with the error term. We run the following regression:

$$X_1 = \pi_0 + \pi_1 Z_1 + \pi_2 X_2 + vX_1 \ldots [11]$$

$$= \pi_0 + \pi_1 Z_{1,} + \pi_2 X_2 + vX_1 \ldots [12]$$

$$= \pi_0 + \pi_1 Z_{1,} + \pi_2 X_2 + vX_1 + v \ldots [13]$$

where:

- $Z_{1,}$ is the instrumental variable for $X_1 Z_{1,}$ is the instrumental variable for X_1
- v is the error term.
- π_0 , π_1 , π_2 are parameters to be estimated.

From this regression, we obtain the predicted value of X_1, denoted as $\widehat{X_1}$

Stage 2: Main Regression Using Fitted Values

In the second stage, we substitute the endogenous variable X_1 with its predicted value $\widehat{X_1}$ from the first stage. The second-stage regression equation is:

$$Y = \beta_{0+} + \beta_1 \widehat{X_1} + \beta_2 X_2 + \varepsilon Y \ldots [14]$$

where:

- $\widehat{X_1}$ is the predicted value from the first stage.
- ε is the new error term (which should no longer be correlated with $\widehat{X_1}$.
- **Summary of the Two Stages:**

1. **Stage 1 (Instrumental Variable Regression):**

   $X_1 = \pi_0 + \pi_1 Z_1 + \pi_2 X_2 + vX_1 \ldots [11]$

   $X_1 = \pi_0 + \pi_1 Z_1 + \pi_2 X_2 + v \ldots [13]$

   Obtain predicted values $\widehat{X_1}$.

2. **Stage 2 (Main Regression with Predicted Values):**

   $Y = \beta_{0+} + \beta_1 \widehat{X_1} + \beta_2 X_2 + \varepsilon Y \ldots [14]$

   $= \beta_{0+} + \beta_1 \widehat{X_1} + \beta_2 X_2 + \varepsilon \ldots [15]$

Comparisons conducted between the results with 2SLS and the OLS results show that OLS effects are usually smaller than 2SLS effects by varying degrees. The support for the assumptions of 2SLS and the results and conclusions about the models are dependent upon the research method and a number of factors. In order to implement 2SLS and to make the estimates of the model identification, research would require a minimum dataset with large sample sizes depending upon the nature of the variables.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 10-2024
DOI: 10.5281/zenodo.14196326

Example:

A summary of the classical regression model coefficient containing OLS compared to two stages least squares and significance test is presented in table (1). The value of R-square exceeds 99 % for all methods, and Probability value (as in Table (1)) of regression table is highly significant. In the equation, the R^2 is the coefficient of determination where DW is the Durbin-Watson statistic.The high R^2 (0.999) suggests that dependent variable is almost totally determined by the defined explanatory variable i.e., the variation in this almost fully explained by the existing explanatory variables.

Table (1): The Two stage least squares model results

Model	No. of Obs.	R-squared	S.e of Regression	Sum squared Residuals	D.W.	Prob. value	F-statistic
Model I (OLS)	24	0.99999	1419.94	40324599	2.008839	0.00	6525954
Model II (OLS)	34	0.99999	1159.19	40312099	1.988390	0.00	9932836
Model III (2SLS)	42	0.99376	45.02	40537.7	-	0.00	1061.93

Source: Author's processing using SPSS

4.0. DISCUSSION

The 2SLS method addresses the endogeneity problem by using instrumental variables to create exogenous fitted values for the endogenous variable, allowing for unbiased and consistent estimates of the regression coefficients. Two-Stage Least Squares (2SLS) regression is a statistical method used in structural equation analysis and is an extension of the Ordinary Least Squares (OLS) approach. It is particularly useful when the error terms of the dependent variable are correlated with the independent variables or when feedback loops exist in the model. In structural equation modeling, the path coefficients are typically estimated using the maximum likelihood method.

This issue often arises when a regression model includes numerous independent variables, leading to multicollinearity. Although multicollinearity is not the only violation of OLS assumptions, severe multicollinearity can invalidate the assumption that the matrix of independent variables (X) has full rank, making OLS estimation impossible.

In such cases, when the matrix X cannot be inverted, the model produces an infinite number of possible least squares solutions. 2SLS offers a solution by addressing these challenges and providing more reliable estimates.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN: 1671-5497
E-Publication: Online Open Access
Vol: 43 Issue: 10-2024
DOI: 10.5281/zenodo.14196326

5.0. CONCLUSION

To help researchers apply and discriminate between both ordinary least squares and two-stage least squares techniques, it is necessary to compare the estimated coefficients and their ordinary least squares and two-stage least squares counterparts in the same data and model structure. Many references on econometric techniques estimate the coefficients of known policy variables in each model using ordinary least squares and those of instrumental variables in cases when 2SLS is required. However, only a few studies have determined the factors that force a researcher to use either of the two techniques, given certain criteria. The future direction of this article might include but is not limited to considering the complex association and usage of the instrumental variables with other types of research questions and discussions that can be made with the 2SLS.In addition to that, it is necessary to consider the causality-based econometric methods, modeling causal relationships together with other unknown or unobservable variables in the case of indirect effects. Future studies may be enhanced with the generalization of discussion types as well as further simulation and sensitivity analysis of the results in the assumptions.

References:

1) Ayindea, K. 2013. Combined Estimators as alternative to Ordinary Least Square Estimator. *International Journal of Sciences: Basic and Applied Research (IJSBAR). [PDF]*

2) Elsayir, H.A.2024. An Econometric Analysis for a Dynamic Foreign Trade Model Using Box-Jenkins Methodology.*Kexue Tongbao/Chinese Science Bulletin. ISSN: 0023-074X Volume 69, Issue 02, February 2024. https://www.kexuetongbao-csb.com.*

3) Fakhreddin, F. 2023. Addressing endogeneity in survey research: Application of two-stage least squares (2SLS) regression analysis. *Researching and Analysing Business. [HTML]*

4) Keane, M. P. & Neal, T. 2021. 2SLS using weak instruments: Implications for estimating the Frisch labor supply elasticity. *unsw.edu.au*

5) Keane, M. & Neal, T. 2023. Instrument strength in IV estimation and inference: A guide to theory and practice. *Journal of Econometrics. unsw.edu.au*

6) K. Kuchibhotla, A., D. Brown, L., & Buja, A. 2018. Model-free Study of Ordinary Least Squares Linear Regression. [PDF].

7) 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. *Mathematics. mdpi.com*

8) Young, A. 2022. Consistency without inference: Instrumental variables in practical application. *European Economic Review. sciencedirect.com*.