Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/120223

Estimated Path Analysis Parameters Using Weighted Least Square to Overcome Heteroskedasticity at Various Sample Sizes

A. D. Adyatama[*], *Solimun, S. A. Efendi, Nurjannah, and M. B. T. Mitakda*

Department of Statistic, Brawijaya University, Malang 65141, Indonesia

Received: 22 May 2022, Revised: 22 Jul. 2022, Accepted: 30 Jul. 2022
Published online: 1 May 2023.

Abstract: This study aims to determine a better parameter estimation method between the OLS and WLS methods to overcome the problem of heteroscedasticity in path analysis and to find out the comparison of standard and adjusted errors between the two methods at various sample sizes R^2. Path analysis is a complex regression analysis with a direct or indirect causal relationship between several variables. The data used in this research is simulation data. The path analysis model formed consists of two correlated exogenous variables, one endogenous variable and one intermediate variable with the relationship between variables used limited to linear form. The results of this study indicate that the WLS parameter estimation method is better than the OLS method in estimating the path analysis parameters that have heteroscedasticity problems. The parameter estimator between the two methods has no significant difference, but the standard error of the WLS method is smaller than that of the OLS. In line with this, the p-value of the significance of the WLS method parameters was almost entirely significant for the five relationships at various levels of heteroscedasticity. it can also be concluded that the larger the sample size, the smaller the standard error for both the OLS and WLS methods. The model's goodness from the adjusted value of the WLS method is higher than the adjusted value of the OLS method.

Keywords: Heteroscedasticity, Path Analysis, Sample Size, Simulation Study, Weighted Least Square.

1 Introduction

According to [1], there is a more complex regression analysis where it is possible to have a relationship between several variables, namely path analysis. Parameter estimation method can be done by Ordinary Least Square (OLS) method. The homogeneity of the variance of errors is an assumption of the OLS estimation. If these assumptions are not met, the estimator obtained remains unbiased and consistent, but not efficient, which means that the estimator does not have the smallest variance among other unbiased estimators for both small and large samples [2]. In addition, if the variance of the model estimator does not meet the assumption of homoscedasticity, then the inference and prediction of the sample coefficients will be wrong. Therefore, violations of this assumption must be addressed.

Handling heteroscedasticity can be done using two approaches. If known, heteroscedasticity is handled using the Weighted Least Square (WLS) method, where the estimators produced are BLUE (Best Linear Unbiased Estimator) [3]. The second approach is to use White's Robust Standard Error. The second method is used if the cause of heteroscedasticity is unknown. In dealing with a problem, it is best to first know the cause of the problem.

Simulation studies are often applied in several studies with the reason to make it easier to get data according to the desired conditions. In this study, simulations were carried out on three sample sizes, namely small (25 observations), medium (50 observations) and large (100 observations) [4]. In addition, the level of heteroscedasticity of error in each data always varies according to the existing pattern. There are three conditions that will be applied in this study, namely low MAPD values (between 0.01-0.33), medium (between 0.34-0.67) and high (between 0.68-1.00).

This study aims to determine the better parameter estimation method between the OLS and WLS methods to overcome the heteroscedasticity problem in path analysis. On the other hand, different from previous studies, this study also wanted to

[*]Corresponding author e-mail: anggadwiee@student.ub.ac.id

know the comparison of standard errors and between OLS and WLS methods at various sample sizes.R^2_{adj}

2 Literature review

2.1. Path Analysis

Path analysis is a complex regression analysis that allows direct or indirect causal relationships between several variables with standardized data.

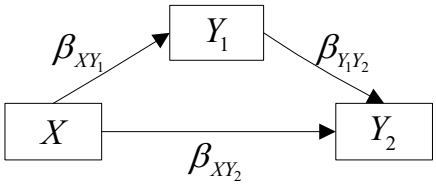


Fig. 1. Path diagram of two relationships

Standardization can be done by standard normal transformation which has a mean of zero and variance of one such as equation (1) [5].

$$Z_{X_i} = \frac{X_i - \bar{X}}{S} \text{ and } Z_{Y_i} = \frac{Y_i - \bar{Y}}{S_Y} \quad (1)$$

Based on Figure 1. It can be formed a path analysis model that has been standardized as in equation (2)

$$Z_{Y1i} = \beta_{Z_X Z_{Y_1}} Z_{X_i} + \varepsilon_{Z_{Y1i}}$$
$$Z_{Y2i} = \beta_{Z_X Z_{Y_2}} Z_{X_i} + \beta_{Z_{Y_1} Z_{Y_2}} Z_{Y1i} + \varepsilon_{Y2i} \quad (2)$$

In addition, equation (2) can be written in the form of equation (3)

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \text{M} \\ Y_{1n} \\ Y_{21} \\ Y_{22} \\ \text{M} \\ Y_{2n} \end{bmatrix} = \begin{bmatrix} 1 & X_1 & 0 & 0 & 0 \\ 1 & X_2 & 0 & 0 & 0 \\ \text{M} & \text{M} & \text{M} & \text{M} & \text{M} \\ 1 & X_n & 0 & 0 & 0 \\ 0 & 0 & 1 & X_1 & Y_{11} \\ 0 & 0 & 1 & X_2 & Y_{12} \\ \text{M} & \text{M} & \text{M} & \text{M} & \text{M} \\ 0 & 0 & 1 & X_n & Y_{1n} \end{bmatrix} \begin{bmatrix} \beta_{01} \\ \beta_{XY_1} \\ \beta_{02} \\ \beta_{XY_2} \\ \beta_{Y_1Y_2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \text{M} \\ \varepsilon_{1n} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \text{M} \\ \varepsilon_{2n} \end{bmatrix} \quad (3)$$

Another form of equation (3) can be written as equation (4).

$$Z_{Y} = \mathbf{Z_X}\beta + \varepsilon \quad (4)$$

2.2. Ordinary Least Square (OLS) Method

According to [1] there is a more complex regression analysis where it is possible to have a relationship between several variables, namely path analysis. Because path analysis is an extension of regression analysis, parameter estimation can be done using the Ordinary Least Square (OLS) method [6]. This method is used if the model is linear in parameters and is carried out by minimizing the number of squares of errors [7].

$$\hat{\beta}_{ols} = (X^T X)^{-1} X^T Y \quad (5)$$

2.3. Path Analysis Assumptions

The following are path analysis assumptions [8].

- In the path analysis model, the relationship between variables is linear and additive.

- Recursive model, namely the causal flow system in one direction.
- Minimal endogenous variables in the interval measuring scale.
- Variables are measured without error.
- The analyzed model is properly specified based on relevant theories and concepts.

Estimation of path analysis coefficients using the OLS method. The assumption of homoscedasticity is one of the assumptions that must be met when using this method [3]; [9].

2.4. Heteroscedasticity Level

In general, Mean Absolute Percentage Error (MAPE) is a relative accuracy measure used to determine the percentage deviation of forecasting results [10]. [11] said that MAPE can also be called the Mean Absolute Percentage Deviation (MAPD). MAPD is used as a reference to calculate the level of heterogeneity of various deviations in each data and has a value between 0 and 1 [12]. The calculation of MAPD can be seen in equation (6).

$$MAPD = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\sigma_i^2 - \bar{\sigma}^2}{\bar{\sigma}^2}\right| \qquad (6)$$

where:

σ_i^2 = variance of the predicted i-th data error with the square of the i-th data remainder

$\bar{\sigma}^2$ = error range

2.5. Weighted Least Square (WLS) Method

The WLS method is one of the parameter estimation methods to deal with heteroscedasticity problems when the cause of heteroscedasticity is known [3].

$$\hat{\beta} = (X^T \sum^{-1} X)^{-1} X^T \sum^{-1} Y \qquad (7)$$

2.6. Simulation

Simulation is a technique of compiling a model from a real situation and then conducting experiments on the model [13]. Simulation studies are often applied in several studies with the reason of making it easier to get data according to the desired conditions.

3 Research methodology

3.1 Research data

The concept of path analysis used in this study can be seen in Figure 2.

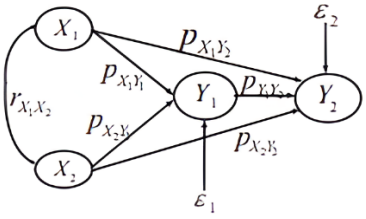


Fig. 2. Research path diagram

Based on Figure 2. it can be formed a path analysis model that has been standardized as in equation (8).

$$Z_{Y_1} = \beta_{X_1Y_1} Z_{X_1} + \beta_{X_2Y_1} Z_{X_2} + \varepsilon_1$$
$$Z_{Y_1} = \beta_{X_1Y_2} Z_{X_1} + \beta_{X_2Y_2} Z_{X_2} + \beta_{Y_1Y_2} Z_{Y_2} + \varepsilon_2 \quad (8)$$

This research uses software R 3.5.1 in generating data and analyzing data. The following is how to generate simulation data

in this study.

1. Generate values with a normal spread of multivariate 0, so that correlations between and can be formed. R is a diagonal matrix, where the diagonal element is indicated in the equation (9).

$$\mathbf{R} = \left(\begin{array}{cccc|cccc} 1 & 0 & \cdots & 0 & r_{X_1X_2} & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & r_{X_1X_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & r_{X_1X_2} \\ \hline r_{X_1X_2} & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & r_{X_1X_2} & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{X_1X_2} & 0 & 0 & \cdots & 1 \end{array}\right) \quad (9)$$

Then, the results of generation are grouped into five categories of values. Values in accordance with the Likert scale are often used using a scale of 5 and further standardized data

Table 3.1 Categories of data rise

Value X_i	Value X_i^*	Information
1	$X_i^* \leq -1{,}5$	Strongly Disagree
2	$-1{,}5 < X_i^* \leq -0{,}5$	Disagree
3	$-0{,}5 < X_i^* \leq 0{,}5$	Neutral
4	$0{,}5 < X_i^* \leq 1{,}5$	Agree
5	$X_i^* > 1{,}5$	Strongly Agree

2. The generated data has three conditions of heteroskedasticity level based on MAPD values, namely low (MAPD between 0.01-0.33), medium (MAPD between 0.34-0.67) and high (MAPD between 0.68-1.00).
3. Determining the sample size includes small (n=25), medium (n=50) and large (n=100) samples. In addition, it determines the Error Variance (EV) where the error variance is 1.
4. Generate error (ε) which follows a multivariate normal distribution with the median value) $E(\varepsilon) = 0$ and the variance matrix$Var\ (\varepsilon) = \Sigma\ .\Sigma$ is a diagonal matrix, where the diagonal element is generated with a uniform distribution $diagsig_i \sim U(c_i, d_i)\ diagsig_i\ \Sigma, c_1 d_i$ where the diagonal element is the diagonal element is the lower limit and is the upper limit formulated in equation (10) and equation (11) [14]
$c_i = \sigma^2(1 - MAPD_i) = EV(1 - MAPD_i)$ (10)
$d_i = \sigma^2(1 - MAPD_i) = EV(1 - MAPD_i)$ (11)
5. Calculate endogenous variables according to equation (8).

3.2 steps

a. Determine the initial value of the path coefficient (equal to 0.5) which describes the level of closeness of the relationship or commonly known as correlation.
b. Estimation using the OLS method before handling heteroscedasticity.
c. Estimation with WLS is used to deal with the problem of heteroscedasticity in each condition of the MAPD value. The weights used are equal to the reason for the variance of the error in which the assumption of homoscedasticity is not met. In order for the results obtained are accurate results it takes repetition. Each replication is assumed to be a sample, sample 50 is a sample that is considered large enough so that 50 repetitions were selected.$\frac{1}{e^2}e^2$
d. Steps a to c were repeated according to the three sample size conditions.
e. Comparing the average results of the path analysis parameter estimator, the standard error of the path analysis parameter estimator, the p-value of the significance test of the path analysis parameter estimator and the goodness of the model () between the OLS and WLS methods with paired t-test in each condition the MAPD value and three

conditions sample size.R^2_{adj}

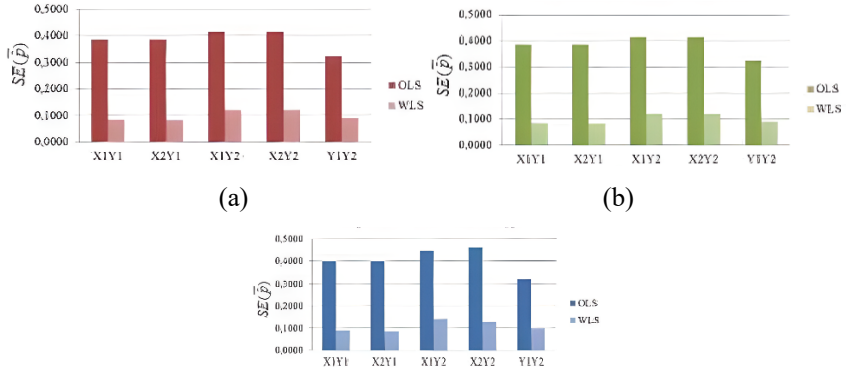
4 Results and Discussion

4.1. Comparison of Path Analysis Using OLS Method and WLS Method

4.2.1 At Small Sample Size (n=25)

Simulations were carried out with 50 repetitions at each level of heteroscedasticity. Pairwise t-test was performed on these results to compare the OLS and WLS methods. The results obtained are that both methods have the same average path analysis estimator. Meanwhile, the average standard error of parameter estimators and the p-value of parameter significance between the OLS and WLS methods are different for various heteroscedasticity conditions and for the five existing relationships.

In general, the expected model is a model with a smaller standard error or error. Based on Figures 3 (a) to (c), it shows that the standard error resulting from the WLS method is smaller than the OLS method in all relationships and various levels of heteroscedasticity.



(a) (b)

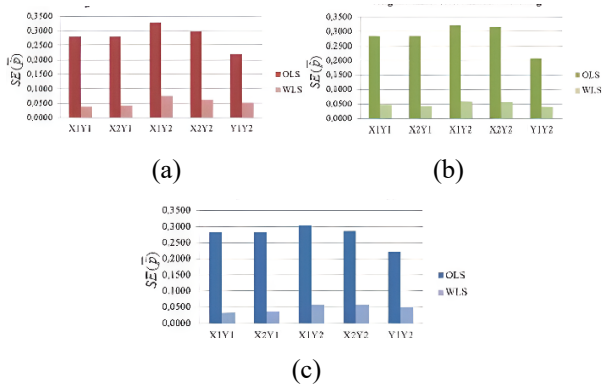
(c)

Fig. 3. Histogram mean standard error path analysis on small sample size with (a) low heteroscedasticity level, (b) medium and (c) high

Regarding the existing results, it is evident that the estimation of path analysis parameters using the WLS method is better than the OLS method at various levels of heteroscedasticity.

4.2.2 At Medium Sample Size (n=50)

As with the small sample size, the mean parameter estimates between the two methods were not significantly different. However, the average standard error of parameter estimators and the p-value of parameter significance between the OLS and WLS methods are different both for various heteroscedasticity conditions and for the five existing relationships.



(a) (b)

(c)

Fig. 4. Histogram mean standard error path analysis on medium sample size with (a) low heteroscedasticity level, (b) medium and (c) high

Based on Figures 4 (a) to (c) it shows that the standard error resulting from the WLS method is smaller than the OLS method in all relationships and various levels of heteroscedasticity. Regarding the existing results, it is evident that the estimation of path analysis parameters using the WLS method is better than the OLS method at various levels of heteroscedasticity.

4.2.3 On Large Sample Sizes (n=100)

The same as in the small and medium sample sizes, the mean parameter estimates between the two methods were not significantly different. However, the average standard error of parameter estimators and the p-value of parameter significance between the OLS and WLS methods are different both for various heteroscedasticity conditions and for the five existing relationships.

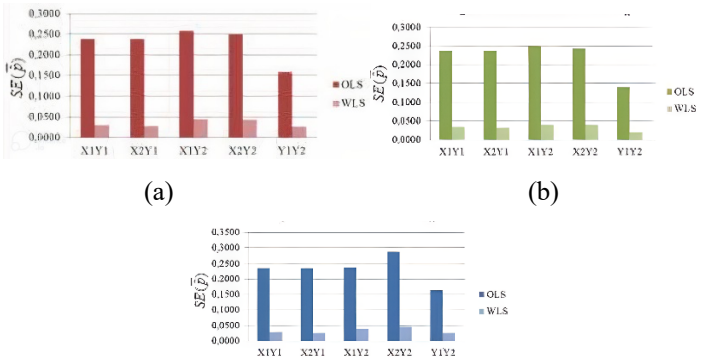


Fig. 5. Histogram mean standard error path analysis on large sample size with (a) low heteroscedasticity level, (b) medium and (c) high

Based on Figures 5 (a) to (c) it shows that the standard error resulting from the WLS method is smaller than the OLS method in all relationships and various levels of heteroscedasticity. Regarding the existing results, it is evident that the estimation of path analysis parameters using the WLS method is better than the OLS method at various levels of heteroscedasticity. Mathematically, the calculation of the standard error is closely related to the sample size. The size of the standard error is inversely proportional to the sample size. If the sample size is large, the resulting variance is small, and vice versa

4.2. Comparison of the Goodness of the Model (Adjusted R2) in Path Analysis Using the OLS Method and the WLS Method

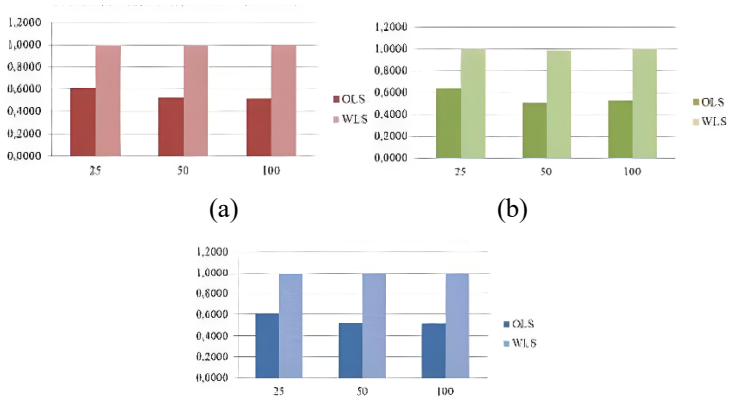


Fig. 6. Histogram average adjusted R2 path analysis model at various sample sizes with (a) low heteroscedasticity levels, (b) medium and (c) high

Based on Figures 6 (a) to (c) the value of the path analysis model using the WLS method is higher than the OLS method. Therefore, it can be said that the WLS method is better used for cases where the assumption of homoscedasticity is not met both at low, medium, and high heteroscedasticity levels. The large sample size has the greatest change compared to the small and medium sample sizes in the three heteroscedasticity conditions. This shows that the larger the sample size, the greater

the adjusted R2 produced with the assumption that R2 and many exogenous variables remain.

5 Conclusion

The conclusions of this study are:

1. The Weighted Least Square (WLS) parameter estimation method is better than the Ordinary Least Square (OLS) method in estimating the path analysis parameters that have heteroscedasticity problems.

2. Comparison of standard errors and between OLS and WLS methods at various sample sizes are:

a. The estimator parameter between the two methods has no significant difference, but the WLS standard error method is smaller than OLS. In line with this, the WLS method p-values were almost entirely significant for the five relationships at various levels of heteroscedasticity. While the p-value of the OLS method parameter significance was almost entirely insignificant for the five relationships at various levels of heteroscedasticity. In addition, it can be said that the larger the sample size, the smaller the standard error in both the OLS and WLS methods.

b. The model's goodness from the adjusted R2 value of the WLS method is higher than the adjusted R^2 value of the OLS method.

Suggestions

Based on the conclusions of the study, suggestions that can be given are:

1. If there is a heteroscedasticity problem in the path analysis, the parameter estimation can use the Weighted Least Square (WLS) method.
2. In future research, it can be compared which heteroscedasticity treatment is better between the WLS method or the OLS method with the White's Robust Standard Error.

References

[1] Streiner, D. L. Finding Our Way: An Introduction to Path Analysis. *The Canadian Journal of Psychiatry*, 50:2., 115-122, (2005).

[2] Sari, A. Q., Sukestiyarno, Y. L. & Agoestanto, A. Batasan Prasyarat Uji Normalitas dan Uji Homogenitas Pada Model Regresi Linier. *Unnes Journal of Mathematics,* 6:2,. 168-177, (2017).

[3] Gujarati, D. N. & Porter, D. C. *Dasar-dasar Ekonometrika*. Jakarta: Salemba, (2015).

[4] Fernandes, A. A. R. Moderating effects orientation and innovation strategy on the effect of uncertainty on the performance of business environment. *International Journal of Law and Management*. (2017).

[5] Li, C. C. *Path Analysis – a Primer*. California: Pacific Grove. (1975).

[6] Fernandes, A. A. R., Budiantara, I. N., Otok, B. W., & Suhartono, B. W. Reproducing Kernel Hilbert space and penalized weighted least square in nonparametric regression. *Applied Mathematical Sciences*, *8*(146), 7289-7300. (2014).

[7] Dillon, W. R. &Goldstein, M. *Multivariate Analysis (Methods and Applications)*. New York: John Wiley & Sons, (1984).

[8] Solimun, A. A. R. F. Moderating effects orientation and innovation strategy on the effect of uncertainty on the performance of business environment. *International Journal of Law and Management*, *59*(6), 1211-1219, (2017).

[9] Fernandes, A. A. R., Solimun, F. U., Aryandani, A., Chairunissa, A., Alifa, A., Krisnawati, E., ... & Rasyidah12, F. L. N. Comparison Of Cluster Validity Index Using Integrated Cluster Analysis With Structural Equation Modelingthe War-Pls Approach. *Journal of Theoretical and Applied Information Technology*, *99*(18), (2021).

[10] Megasari, R. T. & Sungkawa, I. Penerapan Ukuran Ketepatan Nilai Ramalan Data Deret Waktu Dalam Selesi Model Peramalan Volume Penjualan PT Satria Mandiri Citra Mulia. *ComTech* 2:2, 636-645, (2011).

[11] Empat. Hill, A. V. *The Encyclopedia of Operations Management*. New Jersey: Pearson Education, Inc, (2012).

[12] Fernandes, A. A. R., Hutayan, B., Solimun, Arisoesilaningsih, E., Yanti, I., Astuti, A. B., Nurjannah & Amaliana, L. Comparison of Curve Estimation of the Smoothing Spline Nonparametric Function Path Based on PLS and PWLS

In Various Levels of Heteroscedasticity. *IOP Conference Series: Materials Science and Engineering*, (2019).

[13] Siregar, K. *Simulasi dan Pemodelan (Aplikasi Untuk Keteknikan Pertanian).* Yogyakarta: DEEPUBLISH, (2016).

[14] Hidayat, M. F., & Achmad, R. F. A. Estimation of Truncated Spline Function in Non-parametric Path Analysis Based on Weighted Least Square (WLS). In *IOP Conference Series: Materials Science and Engineering* ,546, No. 5, p. 052027, (2019).