DOI: https://doi.org/10.57233/ijsgs.v11i3.939 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOLXXX WEBSITE: https://fugus-ijsgs.com.ng

INTERNATIONAL JOURNAL OF SCIENCE FOR GLOBAL USTAINABILITY
(A PUBLICATION OF FACULTY OF SCIENCE, FEDERAL UNIVERSITY GUSAU, NIGERIA)

Comparison of Methods for Estimating Correctly Specified and Misspecified Linear Regression Models

Gulumbe S.U, Onwuka G. I, Danchadi N. S.
Department of Statistics, Usmanu Danfodiyo University, Sokoto Nigeria
Department of Mathematics and Statistics Kebbi State University of Science and Technology, Alero
Department of Mathematics and Statistics Umaru Ali ShinkafiPolytechnic,Sokoto Nigeria

Corresponding author's Email:nurasambo7@gmail.com +2348133057408

Received on: August, 2025 **Revised and Accepted on:** September, 2025 **Published on:** October, 2025

ABSTRACT

If a probability model of observed data is misspecified, then the interpretation of its Parameter estimates May be invalid, leading to incomplete or incorrect conclusions. The comparison of Weighted Maximum Likelihood (WLE) and Conditional Maximum Likelihood Estimates (CMLE) on their predictive performance when applied to correctly specified and misspecified linear regression was extensively studied. To carry out this analysis, we utilized simulated data that effectively mimicked various scenarios, allowing us to compare the performance of these estimators under small and large sample sizes. Specifically, we focused on the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as our primary metrics for evaluating the efficiency of the estimated models, as these statistics provide a clear understanding of the average prediction error associated with each method. The results of our study revealed a significant trend: the WLE consistently outperformed the CMLE across all tested conditions. Notably, it exhibited the least MSE for linear regression models characterized by one dependent and one independent variable, irrespective of whether smaller sample sizes or larger ones were utilized for the comparison. Moreover, when we extended our analysis to multiple linear regression scenarios, the superiority of the WLE remained evident, reinforcing its position as a more efficient estimator than its counterpart. Furthermore, we took a step further by applying both methodologies to real-life data sets. The findings demonstrated that the advantages of the WLE were not constrained to theoretical or simulated environments but also persisted in practical applications, further validating its efficiency and robustness.

Keywords: Weighted Maximum Likelihood (WML), Conditional Maximum Likelihood (CML), Mean Square Error (MSE), Root Mean Square Error (RMSE)

1.0 INTRODUCTION

1.1 Background of the Study

Models are only ever an approximation of reality and, thus, correct specification is unlikely. Hence, for a diverse collection of summary statistics, it is unlikely that the assumed model can exactly match all the summaries calculated from the observed data; with this problem likely exacerbated in early phases of model exploration, design, and formulation. When the summaries simulated under the assumed model can not match the observed summaries for any value of the unknown model parameters, we say that the model is misspecified (David *et al*.,2025), when the assumed model is misspecified, neither approach is ideal. Meanwhile, model misspecification is ubiquitous in practice because the true data-generating process is unknown and complicated, and the assumed model often simplifies the reality. Therefore, evaluating estimation uncertainty under model misspecification is a pressing task (Liu et al., 2023).

Misspecification 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, choosing a wrong functional form, errors of measurement etc (Ijomah *et al.*, 2020). If a researcher's probability model of the observed data is not correctly specified, then the interpretation of its parameter estimates may not be valid, leading to incomplete or incorrect conclusions. Thus, whether a model is correctly specified must be considered when analyzing and interpreting data. This issue is critically important in econometrics as well as more general scientific inquiry. For example, in health economics, estimates of the impact of clinical treatments, care systems and health policy interventions on health outcomes are dependent on the underlying assumption that the model to be tested is correctly specified. Further, model misspecification testing is essential for statistical analysis of randomized control trials, and observational studies (Golden *et al.*, 2019) Misspecified models are common. In prediction problems, simple, misspecified models may be used instead of complex models with many parameters in order to avoid over-fitting. In big data problems, true models may be computationally

DOI: https://doi.org/10.57233/ijsgs.v11i3.939 ISSNp: 2488-9229; ISSNe: 3027-1118
IJSGS FUGUSAU VOLXXX WEBSITE: https://fugus-ijsgs.com.ng

intractable, leading to model simplifications which induce some level of misspecification. In many scientific domains there exist sets of well-established models with fast computer implementations. A practitioner with a particular data set may have to choose between using one of these models (even when none are exactly appropriate) and devising, testing and implementing a new model. Pressed for time, the practitioner may use an existing misspecified model(Long, 2017). In this research we will focus mainly on comparing the methods of estimating misspecified models namely weighted likelihood estimation (WLE) and conditional maximum likelihood estimation (CMLE).

We shall consider the classical linear regression model where all the independent variables are nonstochastic and the error terms are homoscedastic and serially independent. Let the two regression equations, of which only one is the truth, be:

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \eta_t \quad 1$$

$$y_t = \alpha_1 x_{1t} + \alpha_2 x_{2t} + \cdots + \alpha_k x_{ki} + \alpha_{k+1} x_{k+1t} + \epsilon_t \quad 2$$

The standard is $t = 1,2, \ldots\ldots, T$

When equation (1) is the truth, equation (2) is a misspecified model because of the presence of the irrelevant variable x_{k+1}. When equation (2) is the truth, equation (1) is a misspecified model because of the left out variable x_{k+1}.

2.1 MATERIALS AND METHOD

2.2 Material

This research obtained simulated multivariate normally distributed data using R software version 4.4.0 with sample sizes 1000, 3000 and 5000 it has a mean vector ($\mu_y = 0.649$, $\mu_{X_1} = 0.017$, $\mu_{X_2} = 0.003$) and variance covariance matrix

	y	X1	X2
y	1.00	0.13	-0.12
X1	0.13	1.00	-0.01
X2	-0.12	-0.01	1.00

And also an irrelevant (I) variable was simulated with the same sample sizes which followed gamma distribution with shape = 0.5, rate = 1 and scale = 1

The real life data used was sourced from Central Bank of Nigeria (CBN) statistical bulleting and website http://statistics.cbn.gov.ng/cbn-onlinestats/FeedBack.aspx, the microeconomics variables used for this research is Total Generated Revenue (TGR) which include Oil revenue (OR) and Non-Oil revenue (NOR)

2.3 Methods

2.3.1 Weighted Likelihood Method

Let us consider a model

$$Y_i = \beta_1 X_{i1} + \cdots \beta_p X_{ip} + \varepsilon_i \quad 3$$

where random noise variables $\varepsilon_i, \cdots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$.

Therefore, the weighted likelihood function is given by

$$\prod_{i=1}^{n} f(\beta; X, Y_i) = \left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y - \beta_1 X_{i1} - \cdots - \beta_P X_{ip})^2\right)\right)^{\lambda_i} \quad 4$$

The weighted likelihood estimate is given by

$$\hat{\beta} = \sum_{i=1}^{n} \lambda_i \beta_i \quad 5$$

The weights are selected by the cross-validation procedure proposed by Wang (2001). The cross-validated weights for equal sample sizes can be calculated as follows

$$\lambda_e^{opt} = (1, 0, 0, \ldots, 0) - e_n^2 \left(A_e^{-1}\hat{\Sigma}_1 - \frac{\mathbf{1}^t A_e^{-1}\hat{\Sigma}_1}{\mathbf{1}^t A_e^{-1}\mathbf{1}} A_e^{-1}\mathbf{1}\right) \quad 6$$

Where

$e_n = {}^{n}/_{n-1}$, $\qquad A_e = {}^{en}/_{n-1}\hat{\Sigma} + \left(e_n^2(n-2) + {}^{e_n}/_{n-1}\right)\bar{X}\bar{X}^t$

2.3.2 Conditional Maximum Likelihood Method

Consider an arbitrary density or probability mass function $f(y; \beta, \gamma)$ where β is a vector of nuisance parameters. If there is a statistic $t(y)$ sufficient for β then the nuisance parameters can be eliminated from the likelihood by conditioning on it. We have in mind cases in which f is, for fixed γ, an exponential family so that t is complete sufficient and of the same dimension as β. Let (t, a) be a one-to-one transformation of y, let $J_1 = \delta t^T/\delta y$ and $J_2 = \delta a^T/\delta y$ and let f_t be the density or probability mass function of t. The parameter of interest γ can be estimated by maximizing the conditional log-likelihood

$$l_{a/t}(y; \gamma) = \log(y; \beta, \gamma) - \log f_{\bar{\beta}}(t; \beta, \gamma) + \frac{1}{2}\log|X^T X| \quad 7$$

Consider the linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ where Y is $n \times 1$ vector of responses, X is $n \times p$ design matrix of full column rank and $\varepsilon \sim N(O, \Omega)$ is a random vector. The covariance matrix Q is a function of a q-dimensional parameter γ and is assumed positive definite for γ in a neighbourhood of the true value. For any fixed value of y, the statistic $t = AX^T\Omega^{-1}y$, where A is any non-singular $p \times p$matrix function of γ, is complete sufficient for β. Given the above form for t, the conditional

DOI: https://doi.org/10.57233/ijsgs.v11i3.939 ISSNp: 2488-9229; ISSNe: 3027-1118
IJSGS FUGUSAU VOLXXX WEBSITE: https://fugus-ijsgs.com.ng

likelihood is well defined and satisfies the properties of a likelihood function In the conditional model, the signals are assumed to be deterministic unknown parameters, denoted by $S = [s(1), \cdots, s(N)]$.The corresponding ML estimator is the minimizer of the normalized negative log-likelihood (Mestre, X & Vallet, P. 2020).

2.3.3 Mean Square Error (MSE) of an Estimator

Suppose that we would like to estimate the value of an unobserved random variable X given that we have observed Y$=y$. In general, the estimate $\hat{x}$ is a function of y:

$$\hat{X} = g(Y) \quad 8$$

The **mean squared error (MSE)** of this estimator is defined as

$$E[(X-\hat{X})^2] = E[(X-g(y))^2] \quad 9$$

2.4 Root Mean Square Error (RMSE) of an Estimator

Suppose that we would like to estimate the value of an unobserved random variable X given that we have observed Y$=y$. In general, the estimate $\hat{x}$ is a function of y:

$$\hat{X} = g(Y) \quad 10$$

The **rootmean squared error (RMSE)** of this estimator is defined as

$$\sqrt{E[(X-\hat{X})^2]} = \sqrt{E[(X-g(y))^2]} \quad 11$$

For an estimated parameter to be more efficient than the other estimated parameter it has to have least values of mean square error (MSE) and root mean square error (RMSE).

3.1 RESULTS AND DISCUSIONS

Table (1): Results of MSE and RMSE for Correctly Specified and Misspecified Linear Models at n=1000

Specification	Estimator	β_1	β_I	MSE(β_1)	RMSE(β_1)	MSE(β_I)	RMSE(β_I)
Correctly specified model	CMLE	0.522	------	0.072	0.268	------	------
	WML	0.115	------	0.014	0.118	------	------
Misspecified model	CMLE	0.605	0.126	0.076	0.276	0.072	0.268
	WML	0.094	------	0.015	0.122	------	------

Table 1 presents the MSE and RMSE for correctly specified linear regression parameter $\beta_1$and misspecified linear regression parameters and parameter of irrelevant variable ($\beta_1$and β_I) when n = 1000. It indicates that, the weighted maximum likelihood estimate (WML) has least values of MSE and RMSE, it's therefore more efficient than CMLE for both correctly specified and misspecified models.

Table (2): Results of MSE and RMSE for Correctly Specified and Misspecified Linear Models at n=3000

Specification	Estimator	β_1	β_I	MSE(β_1)	RMSE(β_1)	MSE(β_I)	RMSE(β_I)
Correctly specified model	CMLE	0.523	-----	0.044	0.209	-----	-----
	WML	0.141	-----	0.010	0.100	-----	-----
Misspecified model	CMLE	0.523	0.0003	0.046	0.214	0.049	0.221
	WML	0.121	-----	0.010	0.100	-----	-----

Table 2 Presents the MSE and RMSE for correctly specified linear regression parameter $\beta_1$and misspecified linear regression parameters and parameter of irrelevant variable ($\beta_1$and β_I) when n = 3000. It indicate that, the Weighted maximum likelihood (WML) has least values of MSE and RMSE, it's therefore more efficient than CMLE for both correctly specified and misspecified models.

Table (3): Results of MSE and RMSE for Correctly Specified and Misspecified Linear Models at n=5000

Specification	Estimator	β_1	β_I	MSE(β_1)	RMSE(β_1)	MSE(β_I)	RMSE(β_I)
Correctly specified model	CMLE	0.498	-----	0.033	0.182	-----	-----
	WML	0.133	-----	0.007	0.084	-----	-----
Misspecified model	CMLE	0.498	0.004	0.033	0.182	0.032	0.198
	WML	0.592	-----	0.007	0.084	-----	-----

Table 3 Presents the MSE and RMSE for correctly specified linear regression parameter $\beta_1$and misspecified linear regression parameters and parameter of irrelevant variable ($\beta_1$and β_I) when n = 5000. It

DOI: https://doi.org/10.57233/ijsgs.v11i3.939 ISSNp: 2488-9229; ISSNe: 3027-1118
IJSGS FUGUSAU VOLXXX WEBSITE: https://fugus-ijsgs.com.ng

indicate that, the Weighted maximum likelihood (WML) has least values of MSE and RMSE, it's therefore more efficient than CMLE for both correctly specified and misspecified models.

Table (4): Results of MSE and RMSE for Correctly Specified ($y = x_1 + x_2$) and Misspecified ($y = x_1 + x_2 + x_I$) Linear Models at $n = 1000$

Specification	Estimator	β_1	β_2	β_I	MSE (β_1)	RMSE (β_1)	MSE (β_2)	RMSE (β_2)	MSE (β_I)	RMSE (β_I)
Correctly specified model	CMLE	0.624	-0.514	-----	0.079	0.281	0.079	0.281	-----	-----
	WML	0.029	-----	-----	0.030	0.173	-----	-----	-----	-----
Misspecified model	CMLE	0.625	-0.514	0.034	0.079	0.281	0.079	0.281	0.075	0.274
	WML	0.044	-----	-----	0.032	0.179	-----	-----	-----	-----

Table 4 Presents the MSE and RMSE for correctly specified linear regression parameters β_1,$\beta_2$and misspecified linear regression parameters and parameter of irrelevant variable (β_1,$\beta_2$and β_I) when n = 1000. It indicate that, the Weighted maximum likelihood (WML) has least values of MSE and RMSE, it's therefore more efficient than CMLE for both correctly specified and misspecified models.

Table (5) Results of MSE and RMSE for Correctly Specified ($y = x_1 + x_2$) and Misspecified ($y = x_1 + x_2 + x_I$) Linear Models at $n = 3000$

Specification	Estimator	β_1	β_2	β_I	MSE (β_1)	RMSE (β_1)	MSE (β_2)	RMSE (β_2)	MSE (β_I)	RMSE (β_I)
Correctly specified model	CMLE	0.527	-0.514	-----	0.046	0.214	0.047	0.217	-----	-----
	WML	0.045	-----	-----	0.018	0.134	-----	-----	-----	-----
Misspecified model	CMLE	0.527	-0.515	0.013	0.046	0.214	0.047	0.217	0.043	0.207
	WML	0.029	-----	-----	0.018	0.134	-----	-----	-----	-----

Table 5 Presents the MSE and RMSE for correctly specified linear regression parameters β_1,$\beta_2$and misspecified linear regression parameters and parameter of irrelevant variable (β_1,$\beta_2$and β_I) when n = 3000. It indicate that, the Weighted maximum likelihood (WML) has least values of MSE and RMSE, it's therefore more efficient than CMLE for both correctly specified and misspecified models.

Table (6) : Results of MSE and RMSE for Correctly Specified ($y = x_1 + x_2$) and Misspecified ($y = x_1 + x_2 + x_I$) Linear Models at $n = 5000$

Specification	Estimator	β_1	β_2	β_I	MSE (β_1)	RMSE (β_1)	MSE (β_2)	RMSE (β_2)	MSE (β_I)	RMSE (β_I)
Correctly specified model	CMLE	0.511	-0.499	-----	0.035	0.187	0.036	0.189	-----	-----
	WML	0.116	-----	-----	0.014	0.118	-----	-----	-----	-----
Misspecified model	CMLE	0.511	-0.500	-0.019	0.035	0.187	0.036	0.189	0.033	0.182
	WML	0.060	-----	-----	0.014	0.118	-----	-----	-----	-----

Table 6 Presents the MSE and RMSE for correctly specified linear regression parameters β_1,$\beta_2$and misspecified linear regression parameters and parameter of irrelevant variable (β_1,$\beta_2$and β_I) when n = 5000. It indicate that, the Weighted maximum likelihood (WML) has least values of MSE and RMSE, it's therefore more efficient than CMLE for both correctly specified and misspecified models.

Table (7) : Results of MSE and RMSE for Correctly Specified ($y = x_1 + x_2$) and Misspecified ($y = x_1 + x_2 + x_I$) Linear Models of Oil and Non-oil Revenue in Nigeria

Specification	Estimator	β_{OR}	β_{NOR}	β_I	MSE (β_{OR})	RMSE (β_{OR})	MSE (β_{NOR})	RMSE (β_{NOR})	MSE (β_I)	RMSE (β_I)
Correctly specified model	CMLE	4.632	2.947	-----	0.599	0.774	0.329	0.574	-----	-----
	WML	0.676	-----	-----	0.017	0.130	-----	-----	-----	-----
Misspecified model	CMLE	4.632	2.947	0.469	0.599	0.774	0.329	0.574	0.090	0.300
	WML	0.468	-----	-----	0.024	0.155	-----	-----	-----	-----

Table 7 Presents the MSE and RMSE for correctly specified linear regression parameters β_{OR},$\beta_{NOR}$and misspecified linear regression parameters and parameter of irrelevant variable (β_{OR},$\beta_{NOR}$and β_{I}), it indicate that, the weighted likelihood estimate (WLE) has the least estimated values of MSE and RMSE and therefore it's more efficient than CMLE for both correctly specified and misspecified models.

4.0 CONCLUSION

The results shows that for the simulated data weighted likelihood estimate (WLE) is efficient than the conditional maximum likelihood estimate (CMLE) when applied to correctly specified and misspecified linear regression model with one dependent and one independent variable and also the WLE is efficient than the CMLE when applied to correctly specified and misspecified linear regression model with two independent and one dependent variable simply because it has the least value of MSE and RMSE, and also for the real life data WLE is more efficient than CMLE.

4.1 Contribution to Knowledge

The study has contributed to knowledge in the following directions:-

i. When estimating the parameters of correctly specified linear regression model weighted likelihood estimation (WLE) gives out better results over conditional maximum likelihood (CMLE).

ii. When estimating the parameters of misspecified linear regression model, weighted likelihood estimation (WLE) also gives out better results over conditional maximum likelihood (CMLE).

iii. When estimating the parameters of misspecified and correctly specified linear regression model using real life data, the weighted likelihood estimation (WLE) gives better result over conditional maximum likelihood (CMLE)

4.2 Further Work

Future research could carried out similar work by employing another method of estimating misspecified linear regression model such as calibrated weighted likelihood estimate and also incorporate with clustered correlated data to compare the estimators.

REFERENCES

David T. Frazier, David J. Nott & Christopher Drovandi (2025) Synthetic Likelihood in Misspecified Models, Journal of the American Statistical Association, 120:550, 884-895, DOI: 10.1080/01621459.2024.2370594

Golden, R. M., Henley, S. S., White, H., & Kashner, T. M. (2019). Consequences of model misspecification for maximum likelihood estimation with missing data. Econometrics, 7(3), 37..

Hausman, J. A. (1978). Specification Tests in Econometrics. Econometrica, 46(6), 1251–1271. https://doi.org/10.2307/1913827

Hu, F., & Zidek, J. V. (2002). The Weighted Likelihood. Canadian Journal of Statistics, 30(3),347–371. https://doi.org/10.2307/3316141

Ijomah Maxwell Azubuike & Nduka Ethelbert Chinaka (2020). Choice Model between Omission of Relevant Variable and Inclusion of Irrelevant Variable in a Multicollinear Regression Model. International Journal of Scientific and Innovative Mathematical Research (IJSIMR), 8(2), pp. 1-9. http://dx.doi.org/ 10.20431/2347 -3142.0802001

Jacinto, G., Filipe, P. A. & Braumann, C. A. (2022). Weighted maximum likelihood estimation forindividual growth models. Optimization. 71 (11), 3295-3311

Liu R., Qin Y. & Li Y (2023). Assessing Estimation Uncertainty Under Model Misspecification (arXiv:2312.1139v1[stat.ME]18)

Long, J. P. (2017). A note on parameter estimation for misspecified regression models with heteroskedastic errors. Electronic Journal of Statistics, 11(1), 1464-1490. https://doi.org/10.1214/17-EJS1255

Mestre, X & Vallet, P. (2020). On the Resolution Probability of Conditional and Unconditional Maximum Likelihood DoA Estimation. IEEE Transactions on signal processing, 68, 4656-4671.

Yuan, R., Jiang, B., & Tian, Y. (2023). A study of the equivalence of inference results in the contexts of true and misspecified multivariate general linear models. AIMS Mathematics, 8(9), 21001+.https://link.gale.com/apps/doc/A767148676/AONE?u=anon~209813d3&sid=sitemap&xid=fd9e5662