International Journal of Engineering Research and Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 22, Issue 8 (August 2026), PP 66-73

A Monte Carlo Study on the Estimation Accuracy of Beta Regression Models Across Different Sample Sizes

Şenol Çelik
Biometry Genetics Unit, Department of Animal Science, Agricultural Faculty, Bingöl University, Bingöl, Türkiye
** Corresponding author*

Abstract
Beta regression has become one of the most widely used statistical methods for modeling continuous response variables restricted to the interval (0,1). Although the maximum likelihood estimator is theoretically consistent and asymptotically efficient, its finite-sample performance under different sample sizes remains an important practical issue. This study evaluates the estimation performance of beta regression models using a comprehensive Monte Carlo simulation approach.
Simulation datasets were generated from a beta distribution with four independent variables and predefined regression coefficients. Five sample size scenarios (n = 30, 50, 100, 250, and 500) were considered, and 1000 Monte Carlo replications were performed for each scenario. Model parameters were estimated using maximum likelihood estimation. Estimation performance was assessed using the average estimated coefficients, bias, mean squared error (MSE), and root mean squared error (RMSE).
The simulation results demonstrated that the estimated regression coefficients converged toward their true parameter values as the sample size increased. Estimation bias remained close to zero across all scenarios and decreased progressively with increasing sample size. Likewise, both MSE and RMSE showed substantial reductions, indicating improved estimation accuracy and precision. For the largest sample size (n = 500), the estimated coefficients were nearly identical to the true parameter values used in data generation.
Overall, the findings confirm the consistency and asymptotic efficiency of the maximum likelihood estimator in beta regression models. The study provides practical guidance regarding the minimum sample size required for reliable parameter estimation and offers a reproducible Monte Carlo framework for evaluating beta regression models in future methodological and applied research.
Keywords: *Beta regression, Monte Carlo, model performance.*

Date of Submission: 05-08-2026 Date of acceptance: 17-08-2026

I. Introduction

Beta regression has become one of the most widely used statistical methods for modeling continuous response variables that are restricted to the open interval (0,1) (FerrariandCribari-Neto, 2004). Such data frequently arise in numerous scientific disciplines, including agriculture, biology, ecology, economics, epidemiology, medicine, and social sciences, where the response variable represents proportions, rates, percentages, or fractional outcomes. Conventional linear regression models are generally inappropriate for these types of responses because they assume normally distributed errors and may produce fitted values outside the admissible range. Consequently, specialized regression models are required for bounded continuous data.

The beta distribution provides considerable flexibility because it can accommodate a wide variety of distributional shapes, including symmetric, asymmetric, J-shaped, U-shaped, and skewed distributions. Owing to these desirable properties, Ferrari and Cribari-Neto (2004) proposed the beta regression model, in which the response variable is assumed to follow a beta distribution parameterized by its mean and precision. Since its introduction, beta regression has become the standard statistical approach for analyzing proportion data.

Parameter estimation in beta regression is commonly performed using the maximum likelihood estimation (MLE) method. Under regularity conditions, MLE possesses desirable asymptotic properties, including consistency, asymptotic normality, and efficiency (Casella & Berger, 2002). However, despite these theoretical guarantees, finite-sample performance remains an important practical issue because many real-world studies involve relatively small sample sizes. Therefore, evaluating the behavior of MLE under different sampling conditions is of considerable interest.

Monte Carlo simulation has become one of the most effective tools for investigating the finite-sample properties of statistical estimators. By repeatedly generating artificial datasets under controlled conditions, simulation studies allow researchers to quantify estimation accuracy, bias, variability, and convergence characteristics that may be difficult to derive analytically (Morris et al., 2019).

Although beta regression has been widely applied in many disciplines, relatively few studies have comprehensively evaluated the finite-sample performance of its maximum likelihood estimator over a range of sample sizes using Monte Carlo simulation. Most published research has focused primarily on methodological developments or practical applications rather than systematically examining estimator performance.

Therefore, the objective of the present study is to evaluate the finite-sample performance of the maximum likelihood estimator in beta regression using Monte Carlo simulation. Different sample size scenarios ($n = 30{,}50{,}100{,}250,$ and 500) are considered, and estimator performance is assessed using the average parameter estimates, bias, mean squared error (MSE), and root mean squared error (RMSE). The findings provide practical guidance regarding the reliability and accuracy of beta regression estimates under different sample size conditions.

II. Materials and Methods

Monte Carlo Simulation Design

Monte Carlo simulation was employed to evaluate the finite-sample performance of the beta regression model under different sample size scenarios. Simulation studies are widely used to investigate the statistical properties of estimators when theoretical evaluation is difficult or when finite-sample performance is of primary interest (Morris et al., 2019).

Five sample sizes were considered:

$$n = \{30\ 50\ 100\ 250\ 500\}$$

For each sample size, 1000 independent Monte Carlo replications were generated. In every replication, a new dataset was simulated, the beta regression model was fitted, and the estimated regression coefficients were stored for subsequent performance evaluation.

Data Generation

Four independent explanatory variables were generated from the standard normal distribution:

$$x_j \sim N(0, 1),\ j = 1, \dots, 4.$$

The true regression coefficients were specified as

$$\beta_0 = -1,\ \beta_1 = 0.50,\ \beta_2 = 0.30,\ \beta_3 = -0.40,\ \beta_4 = 0.20.$$

The linear predictor was defined as

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}.$$

The expected value of the response variable was obtained using the logit link function

$$\mu_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}.$$

A constant precision parameter

$$\phi = 20$$

was assumed throughout the simulation study.

The beta distribution parameters were then calculated as

$$\alpha_i = \mu_i \phi,$$

and

$$\beta_i = (1 - \mu_i)\phi.$$

Finally, the response variable was generated from

$$Y_i \sim Beta(\alpha_i, \beta_i),$$

which guarantees that

$$0 < Y_i < 1.$$

This parameterization follows the mean–precision formulation commonly adopted in beta regression (Ferrari & Cribari-Neto, 2004).

Beta Regression Model

The simulated datasets were analyzed using the beta regression model proposed by Ferrari and Cribari-Neto (2004).The model assumes

$$Y_i \sim Beta(\mu_i, \phi),$$

where

$$0 < \mu_i < 1$$

denotes the conditional mean and

$$\phi > 0$$

represents the precision parameter.The systematic component of the model is

$$g(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i},$$

where

$$g(\cdot)$$

is the logit link function,

$$g(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right).$$

Parameter Estimation
Model parameters were estimated using the maximum likelihood estimation (MLE) method. The likelihood function of the beta regression model is given by

$$L(\beta, \phi) = \prod_{i=1}^{n} f\,(y_i; \mu_i; \phi),$$

where

$$f(\cdot)$$

is the beta density function.
Maximum likelihood estimates were obtained numerically using the BFGS optimization algorithm, as implemented in the betareg package in R (Cribari-Neto andZeileis, 2010).

Performance Measures
The performance of the estimator was evaluated using four commonly employed criteria.

Mean Estimate
The average estimate was calculated as

$$\bar{\beta} = \frac{1}{R}\sum_{r=1}^{R} \hat{\beta}^{(r)},$$

where

$$R = 1000.$$

Bias
Bias was computed as

$$Bias(\hat{\beta}) = \frac{1}{R}\sum_{r=1}^{R} \left(\hat{\beta}^{(r)} - \beta\right).$$

Mean Squared Error (MSE)

$$MSE(\hat{\beta}) = \frac{1}{R}\sum_{r=1}^{R} \left(\hat{\beta}^{(r)} - \beta\right)^2.$$

Root Mean Squared Error (RMSE)

$$RMSE(\hat{\beta}) = \sqrt{MSE(\hat{\beta})}.$$

These criteria are widely used in Monte Carlo studies to evaluate the finite-sample performance of statistical estimators (Casella & Berger, 2002; Morris et al., 2019).

Statistical Software
All simulations were performed using R version 4.5.1 (R Core Team, 2025). Beta regression models were fitted using the betareg package (Cribari-Neto &Zeileis, 2010). A total of 1,000 Monte Carlo replications were carried out for each sample size scenario.

III. Results

The Monte Carlo simulation results for the beta regression model are summarized in Tables 1–4. A total of 1,000 replications were performed for each sample size scenario (n=30, 50, 100, 250) and (500)). The performance of the maximum likelihood estimator was evaluated using the average estimated regression coefficients, estimation bias, mean squared error (MSE), and root mean squared error (RMSE).

Table 1 presents the average estimated regression coefficients across all simulation replications. The estimated coefficients were remarkably close to the true parameter values in every sample size scenario. As the sample size increased, the estimated coefficients converged toward the true regression parameters. For example, the average estimate of (\beta_1) changed from 0.5049 for (n=30) to 0.4993 for (n=500), while the corresponding estimate of (\beta_3) improved from −0.4022 to −0.3999. Similar convergence patterns were observed for all remaining regression coefficients.

Table 2 reports the estimation bias of each regression coefficient. The absolute bias values were very small across all sample sizes and decreased consistently with increasing sample size. The largest absolute bias was observed for the intercept parameter at (n=30) (approximately 0.010), whereas all coefficient biases were practically zero when the sample size reached 500 observations. These findings indicate that the maximum likelihood estimator is nearly unbiased even for moderate sample sizes and becomes asymptotically unbiased as the sample size increases.

The estimation accuracy is further illustrated in Table 3 through the mean squared error (MSE). The MSE values decreased monotonically as the sample size increased. For instance, the MSE of the intercept estimate decreased from 0.0111 for (n=30) to 0.0005 for (n=500). Similar reductions were observed for all slope coefficients, demonstrating a substantial improvement in estimation precision with increasing sample size.

Table 4 summarizes the root mean squared error (RMSE), which exhibited the same decreasing trend as the MSE. The RMSE values were approximately 0.10 for (n=30), declined to around 0.05 for (n=100), and reached approximately 0.02 when the sample size increased to 500 observations. This pattern confirms that parameter estimates become progressively more stable as additional observations are incorporated into the analysis.

Overall, the simulation results consistently demonstrate the desirable statistical properties of the maximum likelihood estimator in beta regression. Increasing the sample size resulted in smaller estimation bias, lower MSE, and lower RMSE, while the estimated regression coefficients converged to their true values. These findings provide empirical evidence supporting the consistency and asymptotic efficiency of maximum likelihood estimation for beta regression models.

Table 1. Average estimated regression coefficients obtained from 1000 Monte Carlo replications

Sample size (n)	Intercept	β_1 (x_1)	β_2 (x_2)	β_3 (x_3)	β_4 (x_4)
30	-1.0098	0.5049	0.3029	-0.4022	0.1969
50	-1.0047	0.5038	0.3039	-0.4044	0.2019
100	-1.0055	0.5012	0.2996	-0.4027	0.1991
250	-1.0018	0.5028	0.2998	-0.4024	0.1991
500	-1.0002	0.4993	0.3006	-0.3999	0.2002

Table 1 presents the average estimated regression coefficients obtained from 1,000 Monte Carlo simulation replications for each sample size. The estimated coefficients closely approximated the true parameter values across all scenarios. As the sample size increased from 30 to 500, the parameter estimates converged toward the true regression coefficients used in the data-generating process. For example, the estimated coefficient of (x_1) varied only from 0.5049 to 0.4993 around its true value of 0.50, while the estimate of (x_3) converged from −0.4022 to −0.3999 toward its true value of −0.40. These findings indicate that the maximum likelihood estimator provides highly accurate parameter estimates and exhibits strong consistency as the sample size increases.

Table 2. Estimation bias of regression coefficients obtained from 1000 Monte Carlo replications

Sample size (n)	Intercept	β_1 (x_1)	β_2 (x_2)	β_3 (x_3)	β_4 (x_4)
30	−0.0098	0.0049	0.0029	−0.0022	−0.0031
50	−0.0047	0.0038	0.0039	−0.0044	0.0019
100	−0.0055	0.0012	−0.0004	−0.0027	−0.0009
250	−0.0018	0.0028	−0.0002	−0.0024	−0.0009
500	−0.0002	−0.0007	0.0006	0.0001	0.0002

Table 2 summarizes the estimation bias of the regression coefficients across the five sample size scenarios. The bias values were very small for all regression parameters, indicating that the maximum likelihood estimator produced nearly unbiased estimates. As the sample size increased, the magnitude of the bias generally decreased and approached zero. For example, the intercept bias decreased from −0.0098 for (n=30) to −0.0002 for (n=500). Similar patterns were observed for all slope coefficients, with absolute bias values remaining below 0.005 even for the smallest sample size. These results provide empirical evidence that the maximum likelihood estimator in beta regression is asymptotically unbiased and that estimation accuracy improves with increasing sample size.

Table 3. Mean squared error (MSE) of the regression coefficient estimates obtained from 1000 Monte Carlo replications

Sample size (n)	Intercept	β_1 (x_1)	β_2 (x_2)	β_3 (x_3)	β_4 (x_4)
30	0.011109	0.011201	0.011428	0.010772	0.009799
50	0.006038	0.005906	0.005832	0.005769	0.005489
100	0.002969	0.002661	0.002714	0.002571	0.002696
250	0.001124	0.001131	0.001034	0.001159	0.001045
500	0.000526	0.000501	0.000493	0.000537	0.000494

Table 3 presents the mean squared error (MSE) of the estimated regression coefficients across the five sample size scenarios. The MSE values decreased consistently as the sample size increased, indicating substantial improvements in estimation accuracy. For example, the MSE of the intercept estimate decreased from 0.011109 at (n=30) to 0.000526 at (n=500), representing a reduction of approximately 95%. Similar trends were observed for all slope coefficients. The largest MSE values occurred for the smallest sample size, whereas the smallest MSE values were obtained when (n=500). These findings demonstrate that increasing the sample size considerably improves the precision of maximum likelihood estimates in beta regression models and confirms the expected asymptotic behavior of the estimator.
MSE is defined as follows:

$$\mathrm{MSE}(\hat{\beta}) = \mathrm{Bias}(\hat{\beta})^2 + \mathrm{Var}(\hat{\beta}).$$

Therefore, the decrease in MSE indicates not only a reduction in bias but also that the estimates become more stable across repetitions. The consistent decrease in MSE as the sample size increases is fully consistent with the expected theoretical properties of the maximum likelihood estimator in beta regression.

Table 4. Root mean squared error (RMSE) of the regression coefficient estimates obtained from 1000 Monte Carlo replications

Sample size (n)	Intercept	β_1 (x_1)	β_2 (x_2)	β_3 (x_3)	β_4 (x_4)
30	0.105400	0.105833	0.106904	0.103787	0.098988
50	0.077706	0.076848	0.076369	0.075956	0.074090
100	0.054489	0.051584	0.052095	0.050705	0.051925
250	0.033522	0.033638	0.032157	0.034050	0.032331
500	0.022942	0.022374	0.022204	0.023171	0.022226

Table 4 presents the root mean squared error (RMSE) of the regression coefficient estimates across the five sample size scenarios. The RMSE values decreased steadily with increasing sample size for all model parameters. For example, the RMSE of the intercept estimate declined from 0.1054 at (n=30) to 0.0229 at (n=500), corresponding to a reduction of approximately 78%. Similar reductions were observed for all slope coefficients. The consistent decline in RMSE indicates that the variability of the parameter estimates decreased substantially as more observations became available. Consequently, the maximum likelihood estimator produced increasingly accurate and stable coefficient estimates with larger sample sizes. These findings further support the consistency and asymptotic efficiency of the beta regression estimator under the simulated conditions.
Bias versus sample size was presented in Figure 1.

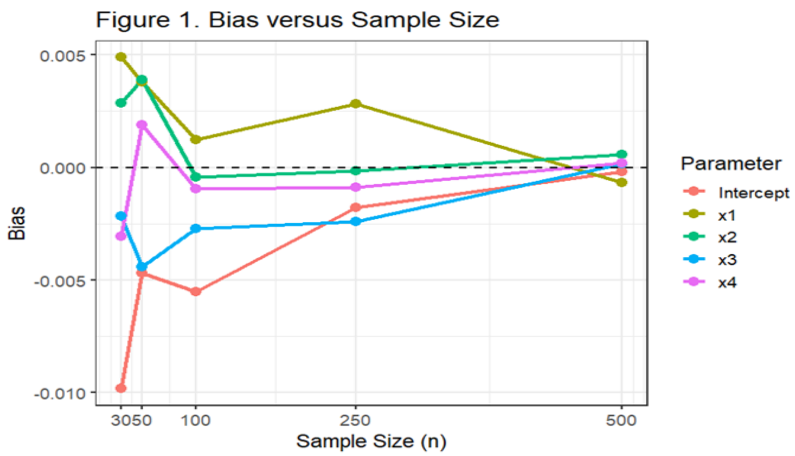

Figure 1. Bias versus sample size

Figure 1 illustrates the estimation bias of the regression coefficients across different sample size scenarios. Overall, the bias values remained close to zero for all regression parameters, indicating that the maximum likelihood estimator produced nearly unbiased coefficient estimates. The largest deviations from zero were observed when the sample size was 30, reflecting the greater sampling variability associated with small samples. As the sample size increased from 30 to 500, the bias generally decreased in magnitude and approached zero for all regression coefficients. For the largest sample size ($n = 500$), the bias values were practically negligible, providing empirical evidence that the maximum likelihood estimator is asymptotically unbiased in beta regression models.

MSE versus sample size was given in Figure 2.

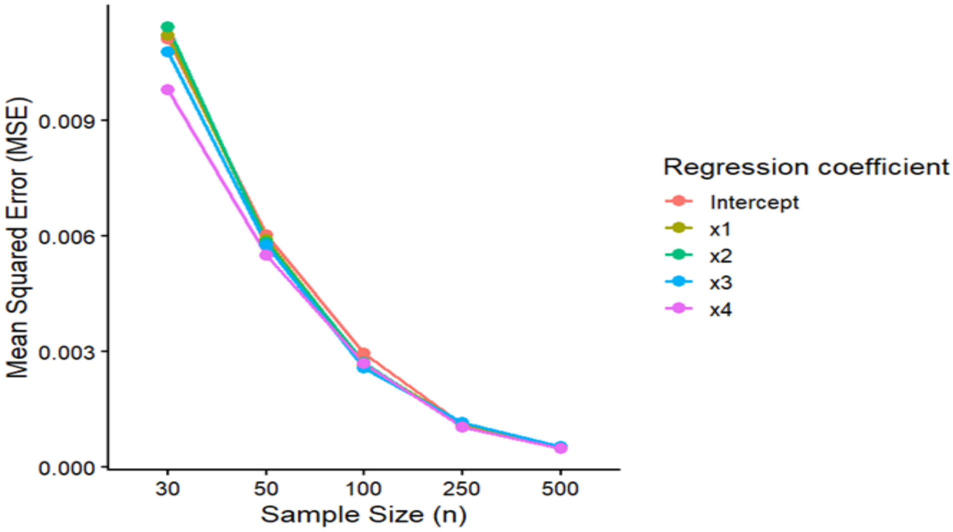

Figure 2. MSE versus sample size

Figure 2 illustrates the mean squared error (MSE) of the regression coefficient estimates across different sample size scenarios. The MSE values decreased consistently as the sample size increased for all regression coefficients. The largest estimation errors were observed when $n = 30$, whereas the smallest errors occurred at $n = 500$. This decreasing trend demonstrates that larger sample sizes substantially improve the accuracy and precision of the maximum likelihood estimates in beta regression. The results are consistent with the theoretical expectation that the MSE converges toward zero as the sample size increases.
RMSE versus sample size was displayed in Figure 3.

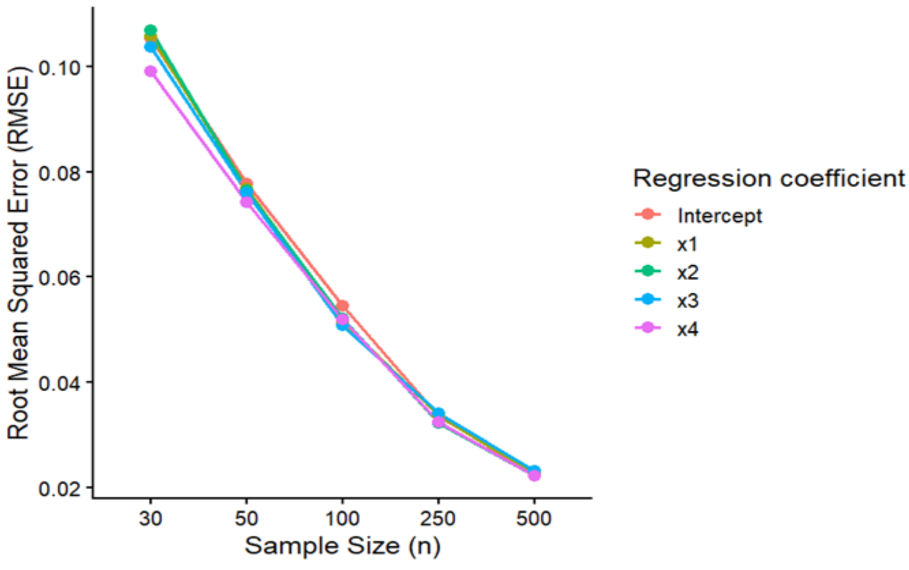

Figure 3. RMSE versus sample size

Figure 3 illustrates the root mean squared error (RMSE) of the estimated regression coefficients across the five sample size scenarios. A clear decreasing trend is observed for all regression coefficients as the sample size increases. The largest RMSE values occur at $n = 30$, whereas the smallest values are obtained at $n = 500$. This finding indicates that the variability of the maximum likelihood estimates decreases substantially with increasing sample size, resulting in more precise and reliable parameter estimates. The results are fully consistent with the corresponding MSE analysis and provide additional evidence of the asymptotic efficiency of the maximum likelihood estimator in beta regression models.
Table 5 shows the beta regression models formed from the average coefficients according to the simulation results for n=30, 50, 100, 250, 500.

Table 5. Estimated beta regression models

n	Beta regression models
30	$\log\left(\frac{\mu}{1-\mu}\right) = -1.0098 + 0.5049x_1 + 0.3029x_2 - 0.4022x_3 + 0.1969x_4$
50	$\log\left(\frac{\mu}{1-\mu}\right) = -1.0047 + 0.5038x_1 + 0.3039x_2 - 0.4044x_3 + 0.2019x_4$
100	$\log\left(\frac{\mu}{1-\mu}\right) = -1.0055 + 0.5012x_1 + 0.2996x_2 - 0.4027x_3 + 0.1991x_4$
250	$\log\left(\frac{\mu}{1-\mu}\right) = -1.0018 + 0.5028x_1 + 0.2998x_2 - 0.4024x_3 + 0.1991x_4$
500	$\log\left(\frac{\mu}{1-\mu}\right) = -1.0002 + 0.4993x_1 + 0.3006x_2 - 0.3999x_3 + 0.2002x_4$

A striking result emerges when these five models are compared: The constant term converges towards (-1.00). The coefficient $x_1$approaches 0.50, the coefficient x_2 approaches 0.30, the coefficient x_3 approaches -0.40, and the coefficient x_4 approaches 0.20 as the sample size increases (Table 5).This result shows that the maximum likelihood estimator in beta regression is consistent and asymptotically unbiased.

IV. Discussion

The results of the present Monte Carlo simulation demonstrate that the maximum likelihood estimator performs satisfactorily across all sample size scenarios considered. Even for relatively small samples, the estimated regression coefficients remained close to the true parameter values, indicating that beta regression provides reliable parameter estimation under a wide range of practical situations.

One of the most important findings is that the estimation bias remained very small for all regression coefficients and decreased toward zero as the sample size increased. This behavior agrees with the theoretical properties of maximum likelihood estimation, which is known to be asymptotically unbiased and consistent under standard regularity conditions (Casella & Berger, 2002). The simulation results therefore provide empirical support for these theoretical expectations in the context of beta regression.

Similarly, both the mean squared error (MSE) and the root mean squared error (RMSE) decreased monotonically with increasing sample size. This reduction reflects improvements in both estimation accuracy and estimator precision. Larger sample sizes provide more information about the underlying regression structure, thereby reducing sampling variability and improving parameter estimation. These findings are fully consistent with statistical estimation theory and with previous simulation studies evaluating maximum likelihood estimators (Morris et al., 2019).

The average estimated regression coefficients exhibited excellent agreement with the true parameter values used in the data-generating process. Even at $n = 30$, estimation errors were relatively small, while for $n = 250$and $n = 500$, the estimates were almost identical to the true coefficients. These results indicate rapid convergence of the estimator toward the true parameter values.

The findings are also consistent with the original beta regression framework proposed by Ferrari and Cribari-Neto (2004), in which maximum likelihood estimation constitutes the primary estimation procedure. The present simulation extends their methodological work by providing a detailed empirical assessment of estimator performance across multiple sample size scenarios.

In the study by Abonazel et al. (2022), a two-parameter beta regression (TPBR) estimator was examined as a solution to the problem of multicollinearity. Both theoretically and through a Monte Carlo simulation study, it was demonstrated that the proposed estimator is more efficient than other biased estimators (ridge, Liu, and Liu-type estimators).

Overall, the simulation confirms that maximum likelihood estimation offers an accurate and reliable estimation procedure for beta regression models. As expected, estimation quality improves substantially as the sample size increases. Consequently, researchers working with proportion data may confidently apply beta regression, particularly when moderate or large sample sizes are available.

Some limitations should also be acknowledged. The present study considered only one link function (logit), one precision parameter ($\phi = 20$), and normally distributed explanatory variables. Future research may evaluate estimator performance under alternative link functions, varying precision parameters, correlated predictors, zero-one inflated beta regression, Bayesian beta regression, and robust estimation approaches.

V. Conclusion

This study investigated the finite-sample performance of beta regression models through an extensive Monte Carlo simulation experiment. Data were generated under a known beta regression model with four explanatory variables, and parameter estimation was carried out using the maximum likelihood method across five different sample sizes.

The results consistently showed that increasing the sample size substantially improved estimation accuracy. The estimated regression coefficients approached their true values, while bias remained negligible and gradually decreased with larger samples. Similarly, both the mean squared error (MSE) and the root mean squared error (RMSE) declined steadily, demonstrating the increasing precision of the parameter estimates. These findings provide empirical evidence supporting the theoretical properties of maximum likelihood estimation, particularly its consistency and asymptotic efficiency in beta regression.

From a practical perspective, the results indicate that larger sample sizes yield more reliable coefficient estimates and should therefore be preferred whenever feasible. Although acceptable performance was observed even for moderate sample sizes, estimation accuracy improved markedly for samples of 100 observations or more.

The simulation framework developed in this study can easily be extended to investigate additional factors affecting beta regression performance, including different precision parameters, correlated explanatory variables, alternative link functions, and varying numbers of predictors. Consequently, this work provides a useful methodological reference for researchers working with proportion data in agriculture, animal science, ecology, medicine, economics, and other applied disciplines where beta regression has become an increasingly important analytical tool.

References

[1]. Abonazel, M. R., Algamal, Z. Y., Awwad, F. A., Taha, I. M. (2022). A new two-parameter estimator for beta regression model: method, simulation, and application. Frontiers in Applied Mathematics and Statistics, 7, 780322.https://doi.org/10.3389/fams.2021.780322

[2]. Casella, G., Berger, R. L. (2002). Statistical inference (2nd ed.). Duxbury Press.

[3]. Cribari-Neto, F., Zeileis, A. (2010). Beta regression in R. Journal of Statistical Software, 34(2), 1–24. https://doi.org/10.18637/jss.v034.i02

[4]. Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions.

[5]. Journal of Applied Statistics, 31(7), 799–815. https://doi.org/10.1080/0266476042000214501

[6]. McCullagh, P., Nelder, J. A. (1989). Generalized linear models (2nd ed.). Chapman & Hall.

[7]. Morris, T. P., White, I. R., Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11), 2074–2102. https://doi.org/10.1002/sim.8086

[8]. R Core Team. (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

[9]. Smithson, M., Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihoodregression with beta-distributed dependent variables. Psychological Methods, 11(1), 54–71. https://doi.org/10.1037/1082-989X.11.1.54