Bias, Calibration and Stability in Logistic Regression Models: A Comparative Simulation Study of MLE, Firth and Ridge Methods

Dr. Apaka Rangita[1], Ngetich Festus[2]

[1]Department of Statistics and Actuarial Science, Maseno University, Kenya

DOI: https://doi.org/10.51584/IJRIAS.2025.100700075

Received: 02 July 2025; Accepted: 10 July 2025; Published: 11 August 2025

ABSTRACT

Logistic regression is widely used to model binary outcomes but the traditional maximum likelihood estimation performs poorly in small samples, rare events and in cases of correlated predictors. This simulation study compared the MLE, the Firth's bias-reduced and the ridge-penalized logistic regression across diverse sample sizes (n = 20, 100, 1000), event rates (5%, 20%, 50%) and predictor correlations (ρ = 0.1, 0.5, 0.8). The key performance metrics were estimation bias, calibration slope and bootstrap-based coefficient variability. Results show that MLE suffers extreme bias and instability in small samples with rare events but recovers well at n = 1000 thus achieving nearly perfect calibration with slope approximately 1. The Firth's method mitigates bias and complete-separation issues in small samples though it also introduces severe calibration distortion (slopes >50). Ridge regression on the other hand provide the most stable coefficient estimates from the bootstrap SDs significantly lower than those of MLE but shows inconsistent calibration especially under sparse conditions. Overall, the Firth is recommended for inference in sparse data, the ridge for prediction in high-dimensional and multicollinear settings and the MLE for large and well-powered datasets. This study demonstrates the significance of aligning estimation methods with those of data characteristics to ensure accuracy and robustness of logistic regression modeling.

Key Words: Logistic Regression, Maximum Likelihood Estimation, Firth Regression, Ridge Regression and Model Evaluation.

INTRODUCTION

Logistic regression is a foundational statistical technique for modeling binary outcomes across disciplines such as medicine to predict disease presence or absence, social sciences to model survey response behavior such as a member churning or renewing and machine learning for spam detection. Logistic regression appeal stems from its interpretability, ease of implementation and the ability to yield insights towards the relationships between predictors or independent variables and binary outcomes (Hosmer et al., 2013).

However, the logistic regression based on the traditional maximum likelihood estimation (MLE) faces notable limitations under challenging data conditions. In small-sample clinical studies with rare outcomes such as rare side effects of a treatment, MLE estimates can be greatly biased or even non-existent due to separation. Also with high-dimensional omics data or survey datasets with multicollinearity the MLE may produce unstable estimates with high variance. In highly imbalanced datasets such as fraud detection the MLE can overestimate coefficients thereby leading to miscalibrated probabilities.

These issues can be mitigated through alternative estimation approaches such as Firth's Logistic model. Firth's bias-reduced logistic regression model adjusts the score equations with a penalization derived from the Jeffreys prior to offer finite and stable estimates even in the presence of complete or quasi-complete separation (Firth, 1993; Heinze & Schemper, 2002). This is particularly useful in rare event modeling such as the genetic association studies and uncommon surgical complications. Firth's model key strength lies in production of unbiased coefficients and valid inference for small samples (Kosmidis & Firth, 2009). However, this method may yield less optimal classification performance in cases of high-dimensional spaces when compared to shrinkage-based approaches (Hastie et al., 2009).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

Ridge logistic regression, on the other hand seems to address these problems of overfitting and multicollinearity by penalizing the size of regression coefficients through the use of a L_2 norm (Cessie & Houwelingen, 1992). This approach is advantageous in applications like credit scoring or electronic health record modeling with many predictors are present are available in the datasets. Ridge offers improved prediction stability and reduced variance in data with collinear or high-dimensional covariates. Nevertheless, the interpretability of its coefficients is reduced due to the shrinkage and selection of tuning parameter (λ) requires computationally intensive cross-validation (Sokolova et al., 2006).

In contrast, the MLE is most reliable in large-sample contexts with balanced event rates and with its asymptotic properties thereby ensuring efficiency and unbiasedness (Agresti, 2013). As an example the national survey data or large insurance databases often provide the ideal conditions under which MLE performs optimally (Harrell, 2015). However, in smaller studies and when data separation occurs such in rare disease modeling, the MLE can fail with severe consequences by producing infinite estimates (Heinze & Schemper, 2002). The increasing demand for robust, interpretable and well-calibrated models in high-stakes applications like clinical risk prediction, policy evaluation and business analytics necessitates the evaluation of these methods systematically (Pavlou et al., 2016). This simulation study conducts a comparative analysis of MLE, Firth and ridge logistic regression across three key perspectives: bias in parameter estimates, calibration of predicted probabilities and stability across samples. By examining these diverse scenarios, the study aim to inform evidence-based method selection tailored to specific data and research contexts.

Problem Statement

Logistic regression models estimated via maximum likelihood occasionally suffer from instability, bias and convergence failures in data-constrained and structurally complex scenarios like rare event modeling, small-sample studies and/or multicollinear predictor spaces. These challenges stem from the theoretical limitations of MLE under finite-sample conditions and the sensitivity to separation and imbalance. As a result, practitioners and researchers face compromised inference, poor calibration and unreliable predictive performance. This study intervenes by systematically comparing the MLE, Firth and Ridge logistic regression methods through simulation across the key performance criteria of bias, calibration and stability hence offering empirical guidance for selecting robust estimation techniques suited to real-world data conditions.

General Objective

The general objective of this study is to systematically compare the performance of maximum likelihood estimation, Firth's bias-reduced logistic regression and ridge-penalized logistic regression based bias, calibration and stability under varying data conditions using a simulation-based framework.

Specific Objectives

To compare bias in coefficient estimates across MLE, Firth and ridge methods under varying sample sizes and data structures.

To evaluate the calibration of predicted probabilities using calibration slope metrics under each method.

To assess the stability of estimates and predictions using bootstrap variability across methods.

To identify optimal use conditions for each method and provide practical guidelines for method selection.

Theoretical Background and Model Formulations

Maximum Likelihood Estimation (MLE)

The MLE is the standard approach for estimating logistic regression parameters. Let $y_i \in \{0, \quad 1\}$ be the binary response and a vector of predictors X_i, then logistic regression model that specifies the log-odds of the probability, $\pi_i = P(y_i = 1|X_i)$ is given by equation (1);

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = X_i{}^T\beta \ldots \text{Equation [1]}$$

Considering that the study considers a single trial, the Bernoulli distribution is adhered to. The likelihood function for a sample of size n is given by equation (2);

$$L(\beta) = \prod_{i=1}^{n}(\pi_i)^{y_i}\,(1-\pi_1)^{1-y_i} \ldots \text{Equation [2]}$$

With the continuous variables considered then the corresponding log-likelihood function is given in equation (3);

$$l(\beta) = \sum_{i=1}^{n}[y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)] \ldots \text{Equation [3]}$$

The MLE estimates $\hat{\beta}$ are obtained by solving the score equations and equating them to zero as shown in equation (4);

$$\frac{dl(\beta)}{d\beta} = X^T(y-\pi) = 0 \ldots \text{Equation [4]}$$

where is X is the design matrix and π is the vector of predicted probabilities.

The MLE possesses strong asymptotic properties with respect to consistency, efficiency and normality (Hosmer, Lemeshow & Sturdivant, 2013). However with respect to small-sample contexts and in presence of separation and/or multicollinearity, the MLE can yield biased and unstable estimates and in some extreme cases there is failure in convergence (Albert & Anderson, 1984).

Firth's Logistic Regression

Firth's logistic regression modifies the conventional likelihood function that aid reduce the small-sample bias and to provide finite estimates even for data that shows presence of complete of uasi-complete separation. The method adjusts the score function by incorporating a penalty derived from the Jeffreys' invariant prior; a non-informative prior distribution that is invariant under the reparameterization and proportional to the square root of the determinant of Fisher's information matrix (Jeffreys, 1946; Kass & Wasserman, 1996). This penalization corresponds to a correction term that involves the observed Fisher's information matrix (Firth, 1993).

The penalized likelihood function is given by equations (5).

$$l^*(\beta) = l(\beta) + \frac{1}{2}\log|I(\beta)| \ldots \text{Equation [5]}$$

with $l(\beta)$ been the regular log-likelihood while $I(\beta)$ as the Fisher information matrix. The added term is a penalty that shrinks extreme estimates and eliminate the infinite predictor values.

The modified score equations for Firth's estimator is given by equation (6).

$$U^*(\beta) = U(\beta) + A(\beta) = 0 \ldots \text{Equation [6]}$$

where $A(\beta) = -I(\beta) * b(\beta)$ and $b(\beta)$ is the first order bias for MLE. This results in an implicit correction in the standard score function by directly countering the bias term ensuring that the adjusted score function has an expected value of zero up to order $0(n^{-1})$. Computationally, this is implemented through modifying the Fisher scoring algorithm in equation (7) to incorporate bias adjustment at each level of iteration involving the hat matrix and the leverage values derived from the design matrix and Fisher information (Kosmidis & Firth, 2009).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

$$\beta^{(t+1)} = \beta^t + [I(\beta^t)]^{-1} U^*(\beta^t) \ldots \text{Equation [7]}$$

With the modified score function given in equation (8).

$$U^*(\beta) = X^T(y - \pi) + X^T(h - 0.5) \ldots \text{Equation [8]}$$

With $W = \text{diag}(\pi_i(1 - \pi_i))$ and the vector of leverages diagonal elements of the hat matrix, $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$. The term $X^T(h - 0.5)$ represents bias adjustment derived from the trace of the influence of each observation thereby aligning the score with Jeffreys' prior penalty (Kosmidis & Firth, 2009).

The Firth's method therefore yields estimates with reduced first-order bias and performs well under small sample sizes, imbalanced outcome distributions and complete or quasi-complete separation (Heinze & Schemper, 2002). The method also retains interpretability similar to MLE through enhanced numerical stability. However, it is computationally more intensive due to the inclusion of the bias correction term and lacks a closed-form solution therefore requires iterative estimation procedures.

Ridge Logistic Regression

Ridge logistic regression introduces an penalty L_2 to the likelihood function to prevent overfitting and also improve the stability in presence of multicollinearity among predictors. he penalized likelihood function is given by equation (9).

$$l_{ridge}(\beta) = l(\beta) - \frac{\lambda}{2}\beta^T\beta \ldots Equation\ [9]$$

where $\lambda \geq 0$ is the regularization parameter that controls the amount of shrinkage applied to the regression coefficients.

The penalized score equations take the form presented in equation (10).

$$U_{ridge}(\beta) = X^T(y - \pi) - \lambda\beta = 0 \ldots Equation\ [10]$$

which do not generally have closed-form solutions hence requiring iterative optimization. The Newton-Raphson update for the penalized objective is as in equation (11).

$$\beta^{(t+1)} = \beta^{(t)} + \left[I\left(\beta^{(t)}\right) + \lambda I_p\right]^{-1}\left[X^T(y - \pi) - \lambda\beta^{(t)}\right] \ldots Equation\ [11]$$

Where $I(\beta) = X^TWX$ is the Fisher information matrix and I_p is the identity matrix.

Theoretically the ridge regression balances between bias and variance to minimize the mean squared error of the model's estimates. Through introduction of the penalty, ridge regression reduces coefficient variance thus enhancing the generalization in predictive settings. In Bayesian interpretation the ridge regression corresponds to the Gaussian prior $\beta \sim N(0,\ \tau^2 I)$, where $\lambda = 1/\tau^2$ offers principled shrinkage mechanism (Hoerl & Kennard, 1970).

Despite its advantages in reduction of overfitting and management of collinearity the ridge regression compromises interpretability due to the shrinkage and thereby necessitates cross-validation and/or information-theoretic criteria for optimal λ selection (Hastie et al., 2009).

Simulation Design and Evaluation Framework

Data Generating Process (DGP)

The data generating DGP assumes a true underlying logistic regression model presented in equation (12).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

$$logit(\pi_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 Xi_2 + \beta_p X_{i3} + \beta_p X_{i4} \dots Equation\ [12]$$

Where $\pi_i = P(y_i = 1|X_i)$ is the probability of success and $X_{ij} \sim N(0,1)$ for $j = 0,1,2,3, \dots, p$. The predictors correlations was introduced via the multivariate normal distributions based on a Toeplitz correlation matrix Σ as shown in equation [13].

$$\Sigma_{ij} = \rho^{|i-j|} \dots Equation\ [13]$$

Where $\rho \epsilon \{0.1, 0.5, 0.8\}$ corresponds to low, moderate and high levels of correlation between model's predictors.

The binary outcome y_i was then generated from a Bernoulli distribution based on the linear predictor highlighted in equation [14].

$$y_i \sim Bernoulli(\pi_i) \dots Equation\ [14]$$

To ensure meaningful effects, the true regression coefficients were selected to reflect moderate associations. There were four predictor variables in the study; $X_1, X_2, X_3\ and\ X_4$ with the corresponding logit co-efficients $\beta_1 = 0.5, \beta_2 = -0.5, \beta_3 = 1.1$ & $\beta_4 = -1.1$. The intercept β_0 was adjusted to achieve specific outcome event rates (5%, 20%, 50%) through marginal logit transformation of the average predicted probability. The calibration ensured comparison of estimation methods under the conditions of rare, moderate and balanced events.

The simulations were conducted across a comprehensive grid of design factors of sample sizes, predictor correlations and the event rates;

Sample sizes: $n = 20, 100, 1000$

Predictor correlations: $\rho = 0.1, 0.5, 0.8$

Event rates: $5\%, 20\%, 50\%$

These combinations yielded 27 distinct data-generating scenarios (3 × 3 × 3). In each scenario data was simulated 1000 replications for ensured robust estimation of performance metrics.

Evaluation Metrics

The performance of the three logistic regression estimation method: MLE, Firth and Ridge was evaluated across the three essential statistical dimensions: bias, calibration and stability.

Bias of Coefficient Estimates

The bias refers to the average deviation of estimated regression coefficients from their true values across simulation replications. For a given coefficient β_j, the bias is computed as in equation [15].

$$Bias(\widehat{\beta_j}) = \frac{1}{R}\sum_{i=1}^{R}\left(\widehat{\beta_j}^{(r)} - \beta_j\right) \dots Equation\ [15]$$

Where: $\widehat{\beta_j}^{(r)}$ is the estimate of β_j in the r^{th} simulation, β_j is the true value of the parameter and R is the number of simulation replications (R=1000 in this study).

Low or near-zero bias shows an unbiased estimator while large positive or negative values indicate systematic over-estimation or under-estimation.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

Calibration of Predicted Probabilities

Calibration assesses the extent to which the predicted probabilities and observed outcomes are in agreement. A well-calibrated model yields predicted risks that directly match the actual observed event proportions. The calibration slope, γ, is obtained by fitting a logistic regression of observed outcomes Y_{ij} on the log-odds of the predicted probabilities $\hat{\pi}_i$ as in equation [16].

$$\text{logit}(Y_{ij} = 1) = \alpha + \gamma * \text{logit}(\hat{\pi}_i) \ldots \text{Equation [16]}$$

Where: A slope γ=1 indicates perfect calibration, slope γ<1 suggests overfitting or predictions are too extreme and a slope γ>1 indicates excessive shrinkage or underfitting.

Stability of Coefficient Estimates (Bootstrap SD)

Stability reflects to the sensitivity of model estimates to sample variation. This is quantified by the bootstrap standard deviation (SD) of estimated coefficients.

For a coefficient β_j, the bootstrap-based standard deviation is defined in equation [17].

$$SD_{boot}(\hat{\beta}_J) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_J^{*(b)} - \hat{\beta}_J^{*}\right)^2} \ldots \text{Equation [17]}$$

Where: B is the number of bootstrap replicates (B=100 for this), $\hat{\beta}_J^{*(b)}$ is the estimate from the b^{th} bootstrap sample and $\hat{\beta}_J^{*}$ is the mean of the bootstrap estimates.

The Smaller the SDs the more stable or less variable the estimates are.

Simulation Analysis and findings

Bias Comparison of Logistic Regression Methods

Table 1: Bias Comparison of Logistic Methods

		X1	X2	X3	X4			X1	X2	X3	X4			X1	X2	X3	X4
n=20,	**MLE**	-2.33E+11	-1.54E+12	9.84E+11	2.82E+10	**n=100,**	**MLE**	0.55	-0.65	0.32	-0.38	**n=1000,**	**MLE**	0.01	-0.01	0.01	-0.01
p=0.05	**Firth**	-0.24	0.27	-0.34	0.35	**p=0.05**	**Firth**	0.00	-0.01	-0.01	0.03	**p=0.05,**	**Firth**	0.00	-0.01	0.00	0.00
, r=0.1	**Ridge**	-0.24	0.28	-0.42	0.41	**, r=0.1**	**Ridge**	-0.19	0.20	-0.29	0.29	**r=0.1**	**Ridge**	-0.04	0.04	-0.06	0.06
n=20,	**MLE**	-4.08E+11	8.29E+11	-1.14E+11	1.09E+12	**n=100,**	**MLE**	1.69	-2.04	2.91	-2.90	**n=1000,**	**MLE**	0.01	-0.01	0.01	-0.01
p=0.05	**Firth**	-0.28	0.27	-0.38	0.36	**p=0.05**	**Firth**	-0.01	0.02	-0.01	0.01	**p=0.05,**	**Firth**	0.00	0.00	0.00	0.00
, r=0.5	**Ridge**	-0.33	0.40	-0.50	0.40	**, r=0.5**	**Ridge**	-0.26	0.32	-0.43	0.40	**r=0.5**	**Ridge**	-0.05	0.08	-0.09	0.07
n=20,	**MLE**	5.40	4.77	23.19	-35.94	**n=100,**	**MLE**	25.22	-9.87	10.93	-15.84	**n=1000,**	**MLE**	0.00	0.00	0.00	-0.01
p=0.05	**Firth**	-0.30	0.26	-0.35	0.36	**p=0.05**	**Firth**	0.00	-0.03	0.03	-0.01	**p=0.05,**	**Firth**	0.00	0.01	0.00	0.00
, r=0.8	**Ridge**	-0.37	0.48	-0.65	0.51	**, r=0.8**	**Ridge**	-0.33	0.39	-0.54	0.47	**r=0.8**	**Ridge**	-0.16	0.24	-0.30	0.23
n=20,	**MLE**	8.70	-31.05	37.29	-23.96	**n=100,**	**MLE**	0.04	-0.04	0.07	-0.07	**n=1000,**	**MLE**	0.01	-0.01	0.00	-0.01
p=0.2,	**Firth**	-0.01	0.05	-0.05	0.05	**p=0.2,**	**Firth**	-0.01	0.00	0.00	0.00	**p=0.2,**	**Firth**	0.00	0.00	0.00	0.00
r=0.1	**Ridge**	-0.26	0.29	-0.39	0.37	**r=0.1**	**Ridge**	-0.12	0.13	-0.17	0.16	**r=0.1**	**Ridge**	-0.04	0.05	-0.07	0.06
n=20,	**MLE**	22.08	-19.01	25.62	-19.93	**n=100,**	**MLE**	0.06	-0.05	0.06	-0.06	**n=1000,**	**MLE**	0.00	0.00	0.00	0.00
p=0.2,	**Firth**	0.00	0.05	0.00	-0.02	**p=0.2,**	**Firth**	0.01	0.00	0.00	0.01	**p=0.2,**	**Firth**	0.00	0.00	0.00	0.01
r=0.5	**Ridge**	-0.28	0.38	-0.48	0.42	**r=0.5**	**Ridge**	-0.15	0.21	-0.28	0.23	**r=0.5**	**Ridge**	-0.05	0.07	-0.09	0.07
n=20,	**MLE**	34.45	-39.04	126.11	-45.71	**n=100,**	**MLE**	0.05	-0.05	0.03	-0.04	**n=1000,**	**MLE**	0.00	-0.01	0.01	0.00
p=0.2,	**Firth**	0.01	0.04	-0.11	0.06	**p=0.2,**	**Firth**	0.01	0.00	-0.03	0.02	**p=0.2,**	**Firth**	0.00	0.00	0.00	0.01
r=0.8	**Ridge**	-0.36	0.44	-0.60	0.52	**r=0.8**	**Ridge**	-0.25	0.34	-0.47	0.38	**r=0.8**	**Ridge**	-0.06	0.11	-0.14	0.09
n=20,	**MLE**	8.04	-13.66	16.89	-22.51	**n=100,**	**MLE**	0.05	-0.05	0.06	-0.07	**n=1000,**	**MLE**	0.01	0.00	0.00	0.00
p=0.5,	**Firth**	0.00	0.01	0.04	-0.07	**p=0.5,**	**Firth**	0.01	-0.01	0.00	-0.01	**p=0.5,**	**Firth**	0.00	0.00	0.00	0.00
r=0.1	**Ridge**	-0.26	0.26	-0.35	0.34	**r=0.1**	**Ridge**	-0.09	0.10	-0.14	0.12	**r=0.1**	**Ridge**	-0.04	0.05	-0.07	0.07
n=20,	**MLE**	1.84	18.82	10.85	-39.62	**n=100,**	**MLE**	0.05	-0.05	0.06	-0.06	**n=1000,**	**MLE**	0.00	0.00	0.00	-0.01
p=0.5,	**Firth**	0.03	-0.02	-0.02	-0.03	**p=0.5,**	**Firth**	0.01	-0.01	0.01	-0.01	**p=0.5,**	**Firth**	0.00	0.00	0.00	-0.01
r=0.5	**Ridge**	-0.30	0.36	-0.49	0.42	**r=0.5**	**Ridge**	-0.10	0.15	-0.20	0.16	**r=0.5**	**Ridge**	-0.05	0.07	-0.09	0.06
n=20,	**MLE**	5.44	0.71	3.56	-9.72	**n=100,**	**MLE**	0.05	-0.02	0.03	-0.04	**n=1000,**	**MLE**	0.00	-0.01	0.01	-0.01
p=0.5,	**Firth**	0.10	-0.06	0.03	-0.08	**p=0.5,**	**Firth**	0.01	0.01	-0.02	0.01	**p=0.5,**	**Firth**	0.00	-0.01	0.00	0.00
r=0.8	**Ridge**	-0.32	0.42	-0.57	0.47	**r=0.8**	**Ridge**	-0.21	0.30	-0.39	0.32	**r=0.8**	**Ridge**	-0.06	0.10	-0.13	0.08

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

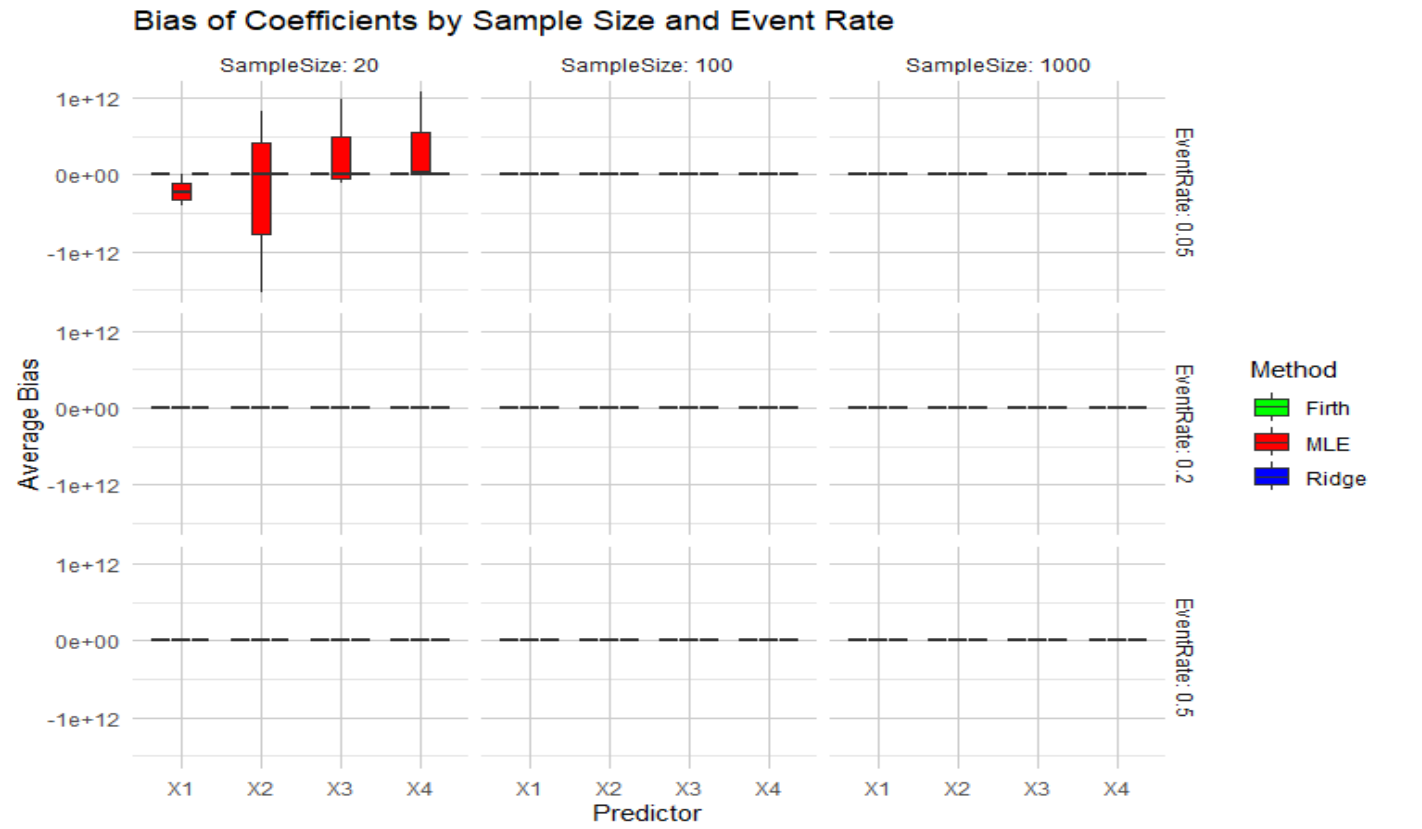


Figure 1: Bias of Coefficients by Sample Size and Event Rate

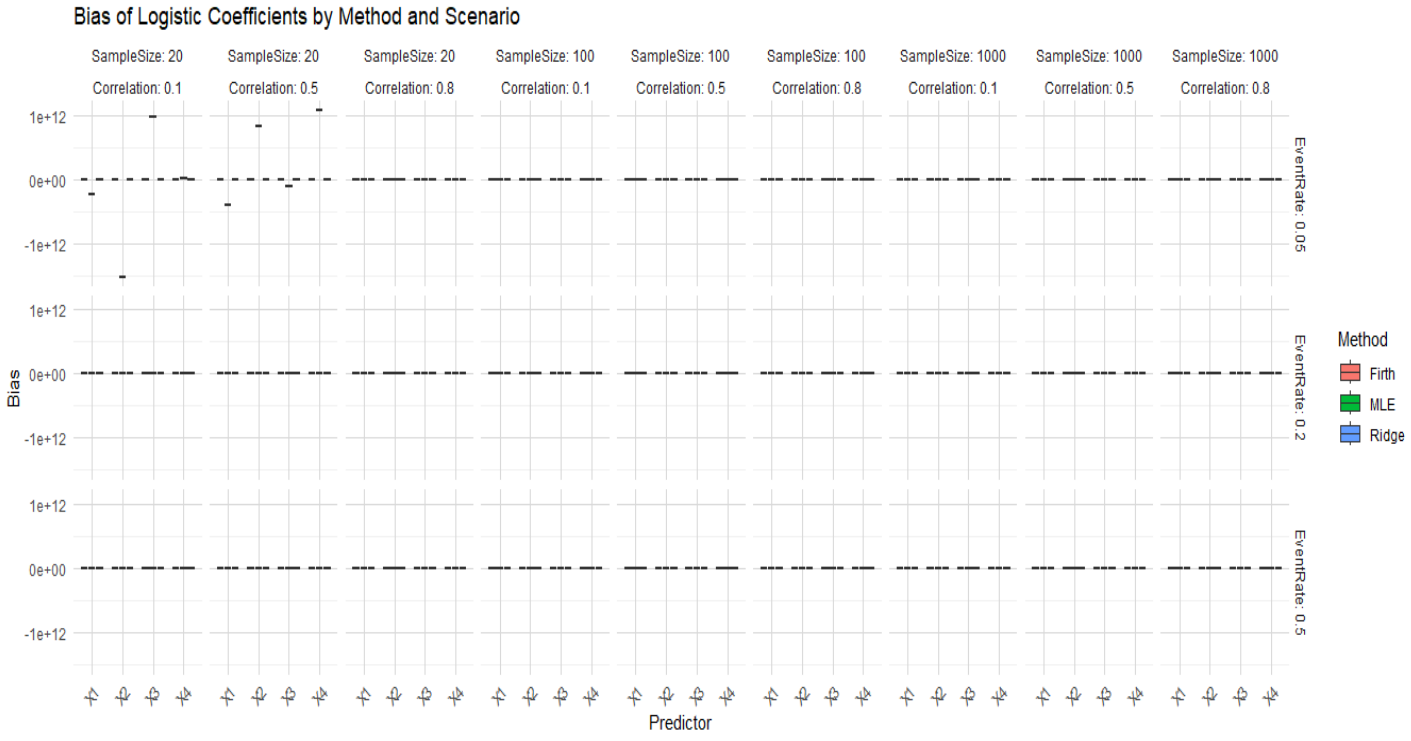


Figure 2: Bias of Logistic Coefficients by Method and Scenario

This simulation study evaluated the bias in logistic regression coefficients estimated by the MLE, the Firth's penalized likelihood and the Ridge regression across varying sample sizes, event rates and predictor correlation structures. The findings were summarized in Table 1 and visualized using Figures 1 and 2, revealing significant differences in performance across the methods more so under challenging conditions like small samples (n=20) and rare events (p=0.05).

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

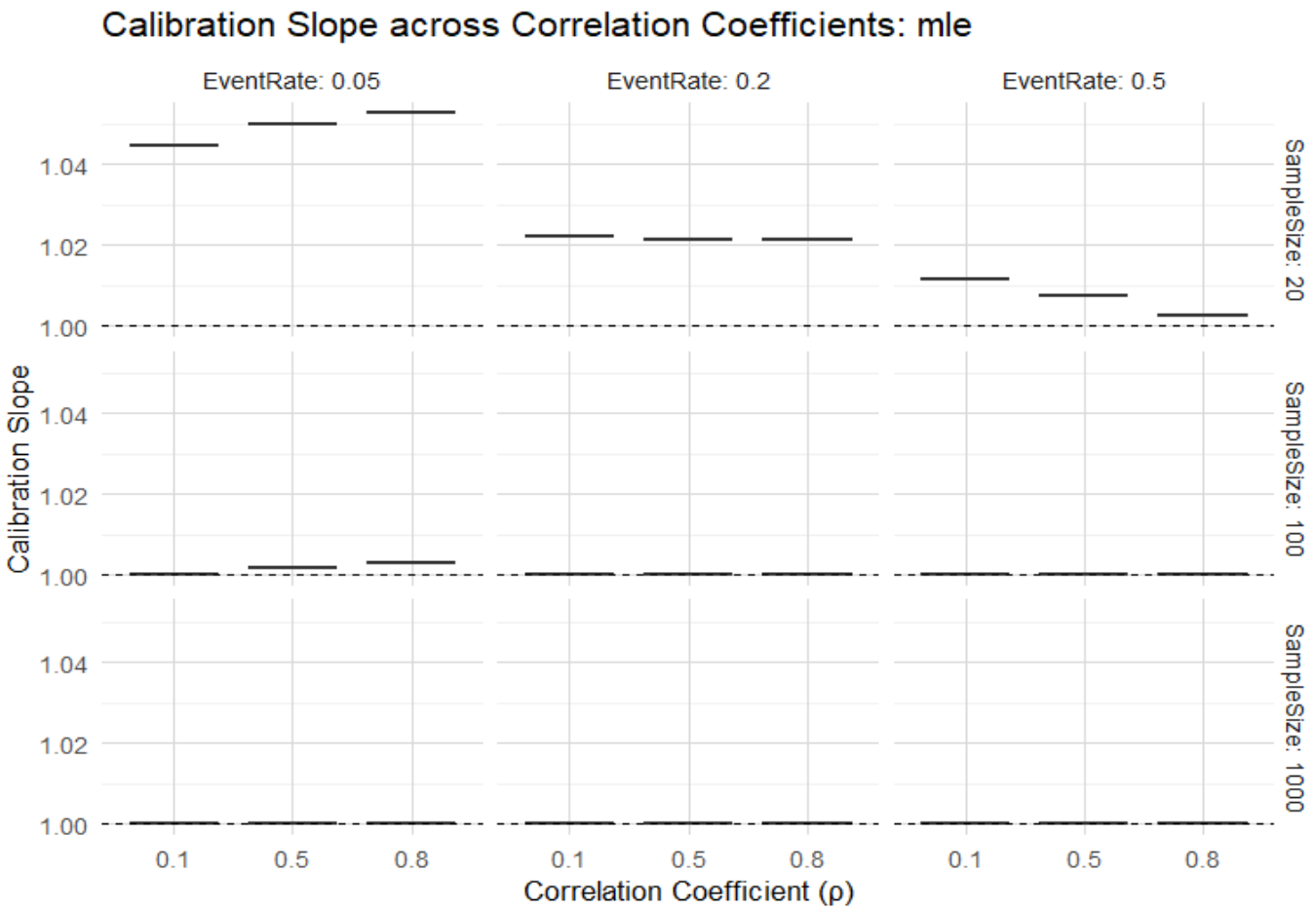
MLE showed severe instability in small samples with bias and its average values reaching magnitudes as high 10^{12} highlighting complete separation and estimation breakdown. In contrast, the Firth's regression showed remarkable resilience under similar conditions averaging biases typically within ±0.5. The ridge regression similarly performed well demonstrating moderate and consistent bias reduction from its L2 regularization. As sample sizes increased to n = 100 the bias of MLE decreased substantially with some instability persisting under low event rates while that of the Firth and Ridge remained stable. At large sample size; n = 1000, it was demonstrated that all three methods yielded bias near zero thereby aligning them with their theoretical asymptotic properties. These findings suggest that with small-sample and/or low-event-rate scenarios the penalized methods, especially the Firth give more reliable estimation while in well-powered settings all methods perform comparably.

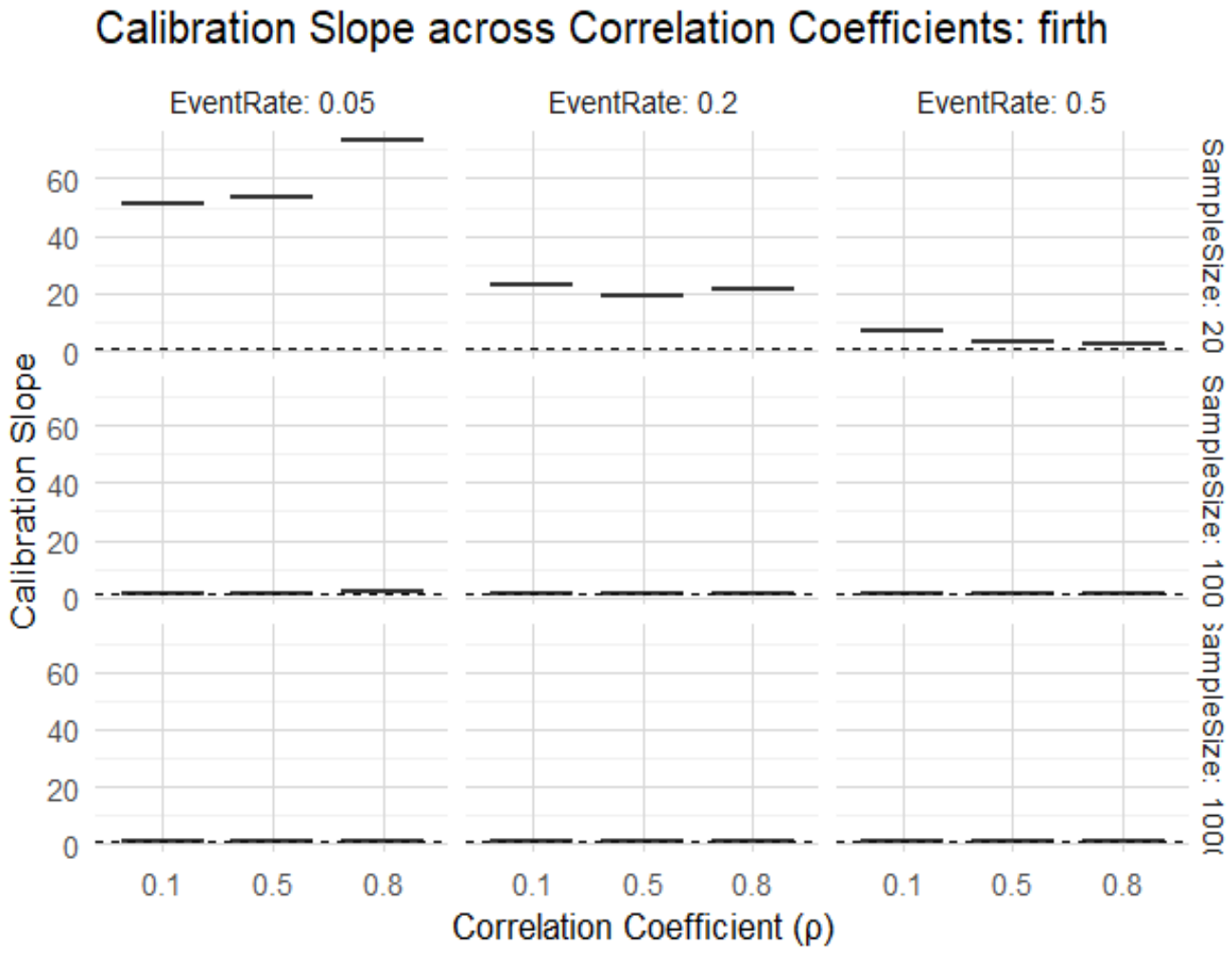
Evaluation of Calibration Performance Using Calibration Slope Across Estimation Methods

Table 2: Calibration Performance Metrics

	n=20, p=0.05, r=0.1	n=20, p=0.05, r=0.5	n=20, p=0.05, r=0.8	n=20, p=0.2, r=0.1	n=20, p=0.2, r=0.5	n=20, p=0.2, r=0.8	n=20, p=0.5, r=0.1	n=20, p=0.5, r=0.5	n=20, p=0.5, r=0.8
MLE	1.0445	1.0498	1.0528	1.0220	1.0214	1.0215	1.0115	1.0073	1.0028
Firth	51.2390	53.5977	73.0605	23.2722	18.8788	21.3450	6.6044	3.3233	2.7222
Ridge	226.244	234.2875	119.6356	958.236	174.3931	98.2613	91.7362	114.8836	74.0821
	n=100, p=0.05, r=0.1	n=100, p=0.05, r=0.5	n=100, p=0.05, r=0.8	n=100, p=0.2, r=0.1	n=100, p=0.2, r=0.5	n=100, p=0.2, r=0.8	n=100, p=0.5, r=0.1	n=100, p=0.5, r=0.5	n=100, p=0.5, r=0.8
MLE	1.0003	1.0019	1.0030	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Firth	1.1913	1.5152	1.7422	1.0912	1.0884	1.0856	1.0776	1.0735	1.0713
Ridge	36.0741	84.9469	59.0560	4.6380	23.3928	25.7325	1.3754	6.6045	14.7098
	n=1000, p=0.05, r=0.1	n=1000, p=0.05, r=0.5	n=1000, p=0.05, r=0.8	n=1000, p=0.2, r=0.1	n=1000, p=0.2, r=0.5	n=1000, p=0.2, r=0.8	n=1000, p=0.5, r=0.1	n=1000, p=0.5, r=0.5	n=1000, p=0.5, r=0.8
MLE	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Firth	1.0127	1.0116	1.0101	1.0083	1.0080	1.0077	1.0073	1.0071	1.0070
Ridge	1.1050	1.3419	30.6990	1.1101	1.1153	3.2216	1.1129	1.1207	2.0147

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025





ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

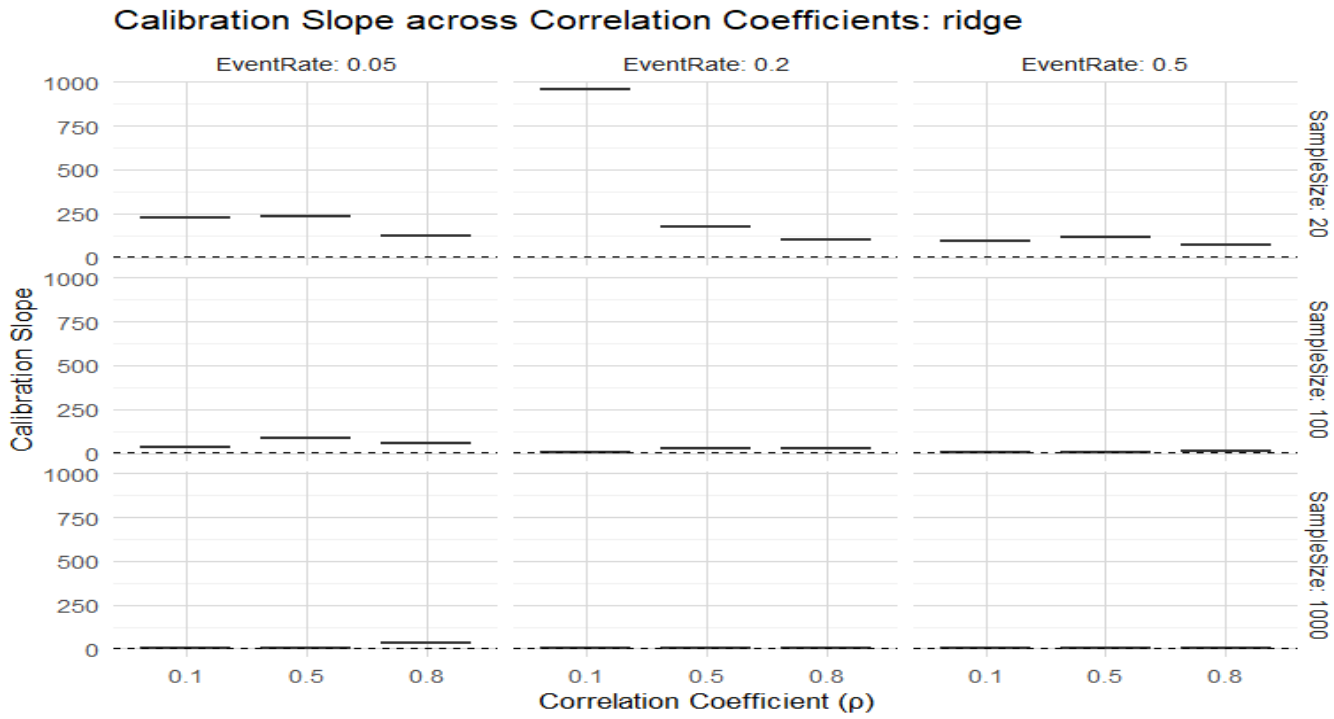


Figure 3: Calibration Slope across Correlation Coefficients

Presentations in Table 2 and Figures 3 shows the calibration slope results from the simulation study by assessing how well the predicted probabilities align with the observed outcomes across three estimation methods: MLE, Firth's penalized likelihood and ridge regression.

The MLE demonstrated strong calibration performance across most of the conditions more so in moderate to large sample sizes. In sample sizes of n=100 and n=1000, the calibration slopes were consistently close to 1.0 thereby giving a reflection of well-calibrated predictions. Although, under the most challenging scenarios of small samples (n=20) with very low event rates (p=0.05) MLE showed slight deterioration with the calibration slopes marginally exceeding 1.04. The findings thereby reaffirm the asymptotic properties of MLE, suggesting its suitability majorly when adequate sample sizes are available.

The Firth's penalized likelihood method even though known for its ability in bias reduction in small samples showed substantial overestimation in the calibration slope under low-information conditions. In scenarios with n=20 and p=0.05, the slopes ranged between 51 and 73 demonstrating extreme overfitting and miscalibration. Although Firth's method showed improvement with larger sample sizes the slope values remained above 1.0 even at n=1000 especially when the predictor correlations were high. These results highlight that while Firth regression mitigates separation and improve coefficient stability it can compromise calibration particularly under sparse data conditions.

The ridge regression demonstrated the most unstable calibration behavior across the methods. In small samples (n=20) with low event rates (p=0.05) the calibration slopes exceeded 900 in certain conditions thus highlighted significant overfitting. Similarly, in moderate samples (n=100), the ridge calibration was still inconsistent and often poor. Even though, the performance improved at n=1000 the slopes remained above the ideal levels of around 1 more so with increase in predictor correlation. The findings thus imply that although that ridge regression provides regularization, it requires careful tuning in achieving reliable calibration under varying data complexities.

Assessment of Stability of Estimates and Predictions Using Bootstrap Variability Across Methods

Table 3: Co-efficients Standard Deviations

Coefficient	MLE	Firth	Ridge
Beta1	0.3114	0.2852	0.2343
Beta2	0.4232	0.3871	0.3043
Beta3	0.4390	0.3966	0.3477
Beta4	0.4339	0.3887	0.3264

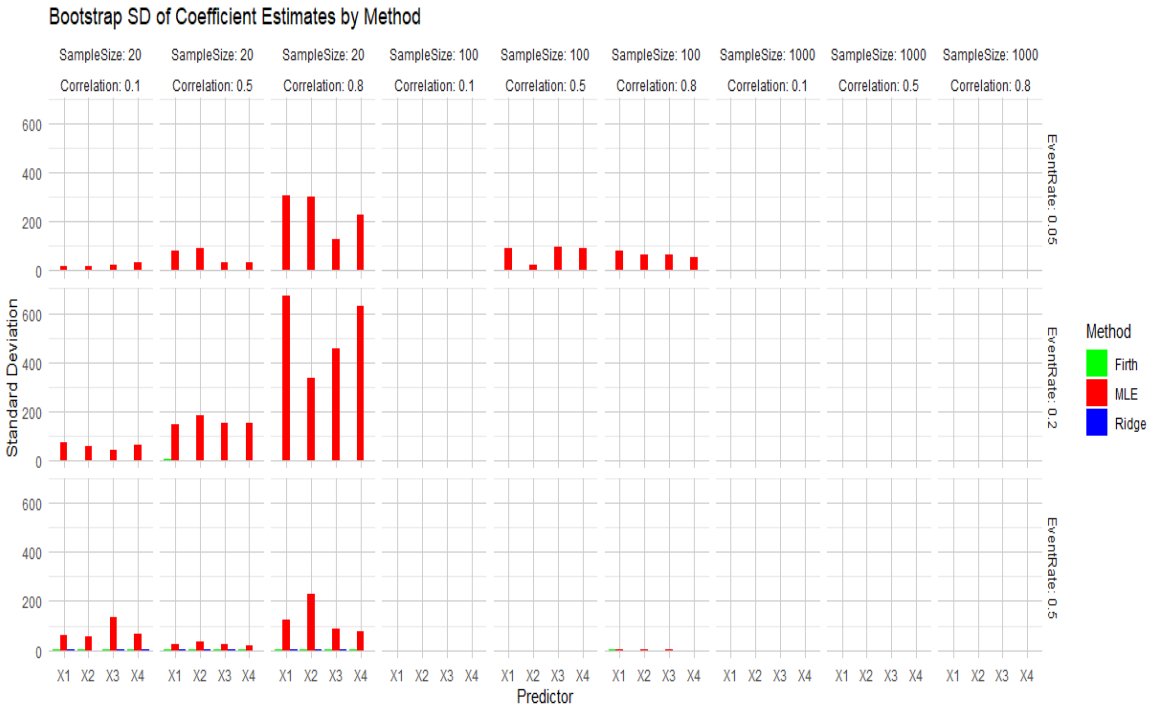


Figure 4: Bootstrap SD of Coefficient Estimates by Method

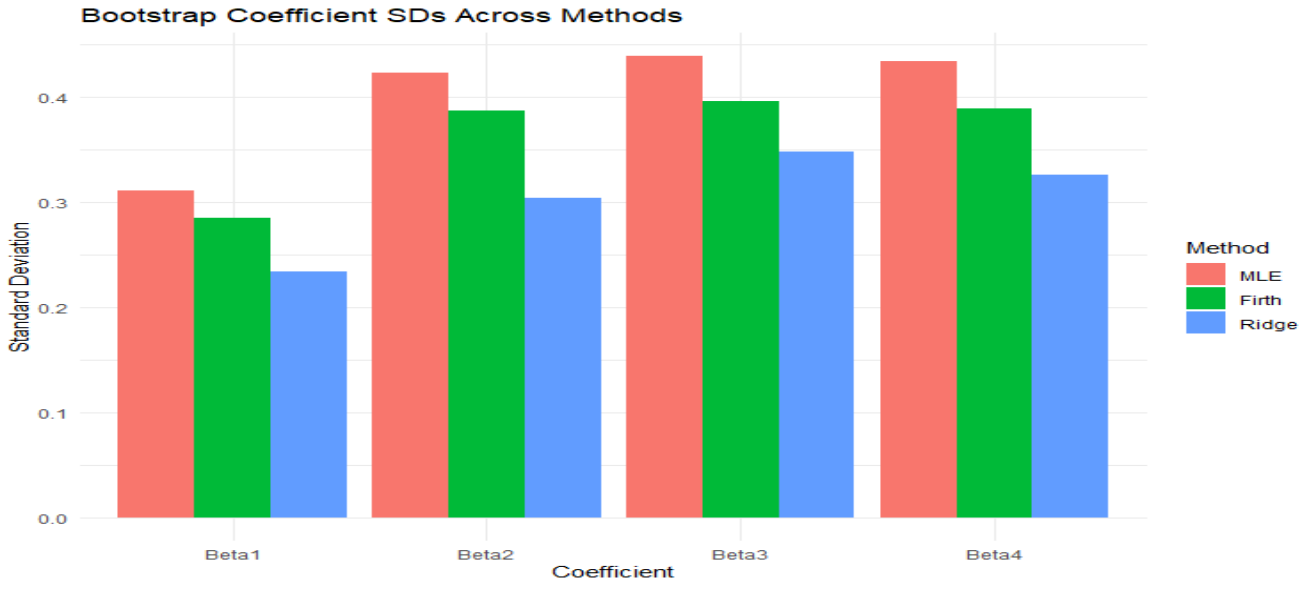


Figure 5: Bootstrap Coefficient SDs Across Methods

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

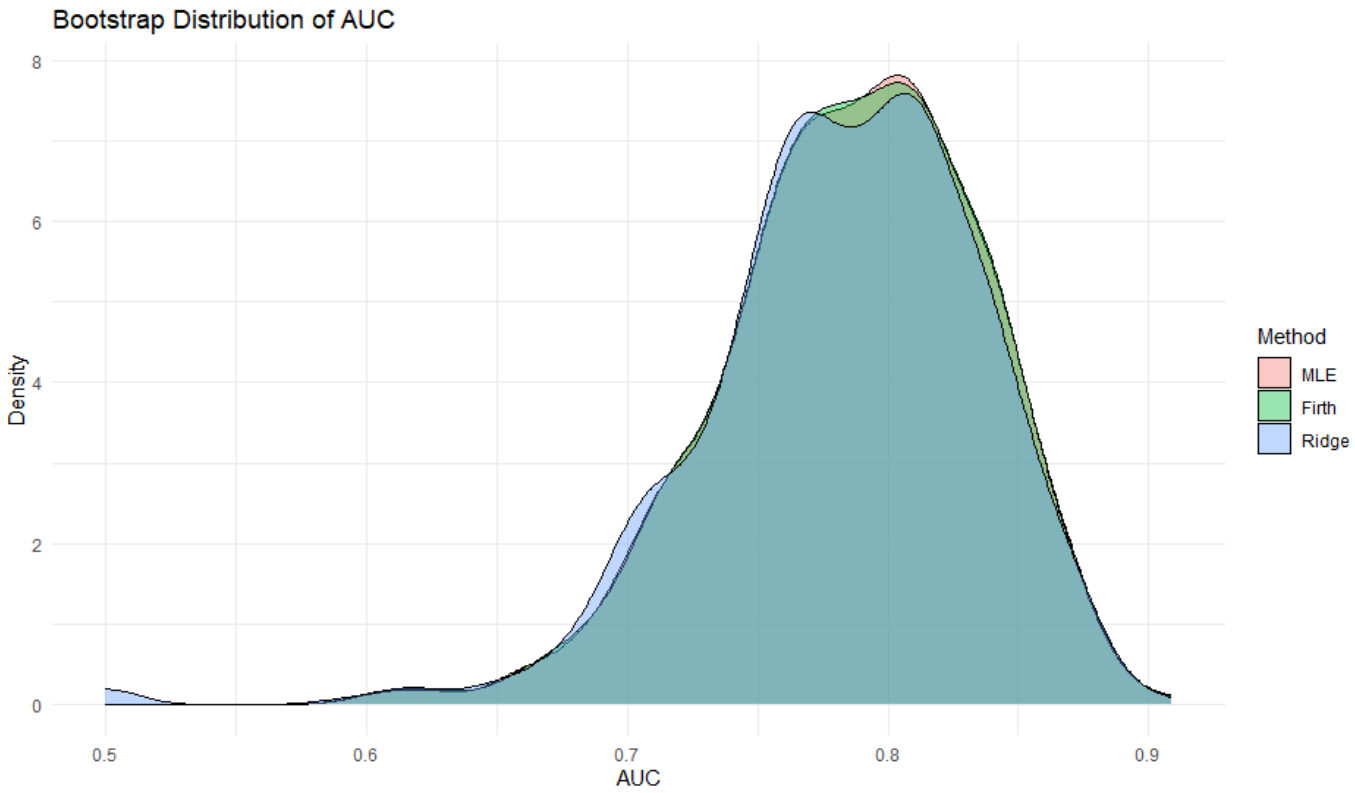


Figure 6: Bootstrap Distribution of AUC

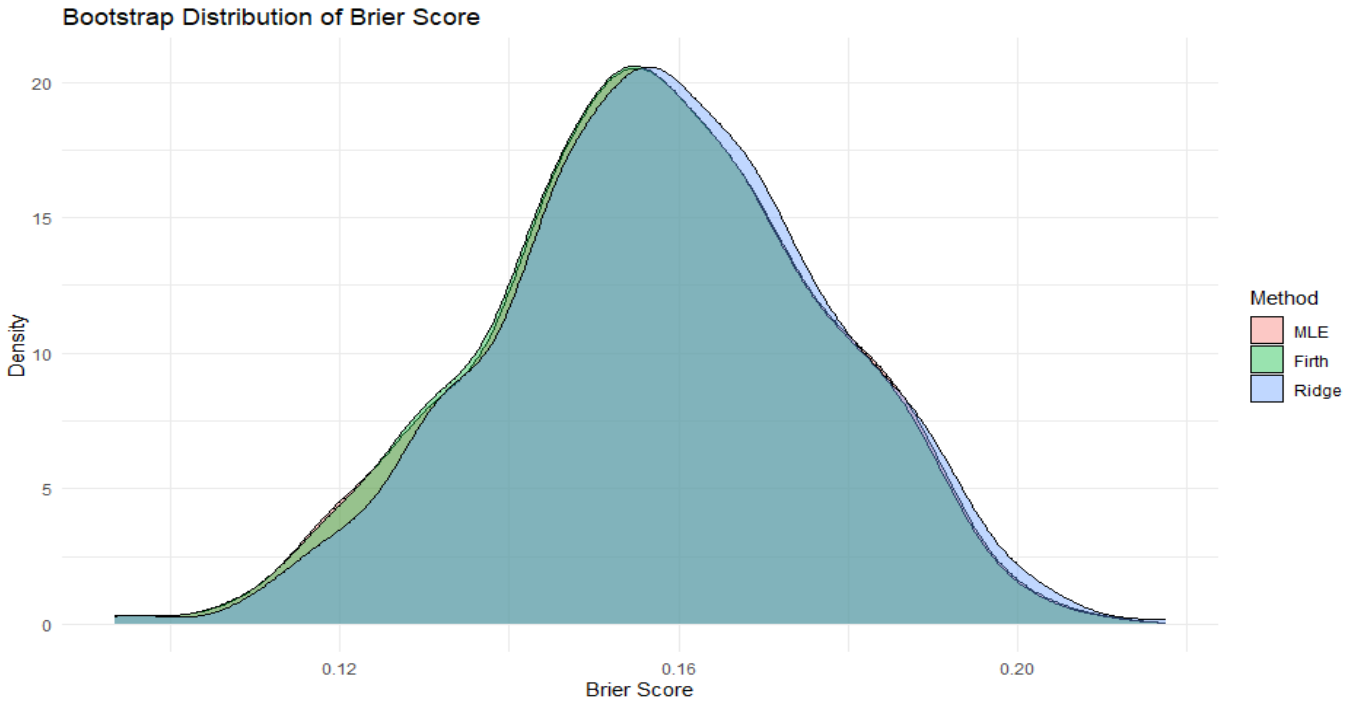


Figure 7: Bootstrap Distribution of Brier Score

Objective 3 is addressed by assessing the stability of coefficient estimates and predictions using bootstrap variability. The bootstrap procedure with 100 replicates per scenario was employed for each of the three estimation methods: MLE, Firth's penalized likelihood and Ridge regression. Stability was evaluated using the standard deviation (SD) of the regression coefficients across the bootstrap samples and the variability in predictive metrics of Area Under the Curve (AUC) and Brier score.

The results, summarized in Table 3, provide a comprehensive overview of the bootstrap standard deviations for all regression coefficients (X1 to X4) under combinations; sample size (n = 20, 100, 1000), event rate (5%, 20%,

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

50%) and predictor correlation (ρ = 0.1, 0.5, 0.8). Figure 4 visualizes the coefficient SDs by estimation method that clearly shows that the Ridge regression yields the most stable estimates especially in small-sample scenarios. In contrast, the MLE shows extreme variability in cases of small samples with the coefficient SDs sometimes exceeding 300. The Firth regression performs better than the MLE but does not consistently match the Ridge in terms of its stability.

The inspection of how bootstrap variability responds to changes in design factors was presented in Figure 5. This stratifies the coefficient SDs by method, sample size, event rate and predictors correlation. This visual evidence reaffirms that all the methods show improved stability with increasing sample sizes, with the Ridge regression highlighting the lowest SDs consistently across diverse settings. By n = 1000 all the regression methods largely converge thereby confirming that large samples lessened the differences in estimation variability.

In terms of prediction performance, the bootstrap distributions for AUC and Brier scores are shown in Figure 6 and Figure 7 respectively shows that despite the large disparities in coefficient SDs and calibration metrics these figures reveal that the AUC and Brier score distributions remain relatively stable across the methods. Despite the ridge regression producing slightly narrower distributions under challenging conditions, the differences in the measures of central tendency among the three methods are consistently minimal. This highlights a key insight that shows that while estimation instability and calibration issues may arise, the discrimination and overall predictive accuracy as measured by AUC and Brier score are comparatively robust to the decision of estimation method.

These findings demonstrate that the ridge regression offers superior coefficient stability, particularly in small samples and high-correlation contexts. However, the MLE's performance improves dramatically with increased sample size and its predictive metrics remain solid when the coefficient estimates are unstable. The Firth regression shows moderate variability though its use may be limited by pertinacious calibration challenges. The use of bootstrap methods thereby offers an essential insight towards the estimator reliability, thus guiding method selection for both simulation and applied research contexts.

DISCUSSION OF FINDINGS

The simulation results for Objective 1 reveal a clear pattern in bias estimation across different estimation methods and scenarios. The MLE produced highly unstable and severely biased coefficient estimates in small samples with rare events (n=20, p=0.05) with some bias values reaching astronomical magnitudes due to variable separation. In contrast, the Firth's penalized likelihood demonstrated robustness by maintaining bias values close to zero even under the most challenging conditions. Similarly, the ridge regression offers improved stability over ML by exhibiting moderate and consistently bounded bias. As sample sizes increased, especially at n=100 and n=1000, all the three methods converged toward negligible bias thus aligning it with the asymptotic properties (Agresti, 2013).

Under Objective 2, the calibration performance was evaluated using Calibration Slope demonstrated significant differences across the methods. The MLE showed a near-perfect calibration in moderate to large samples with slopes very close to the ideal of 1.0 and intercepts near to zero, thus reaffirms its strong asymptotic behavior in well-powered datasets (Harrell, 2015). However, with small samples the Firth and Ridge regression exhibited extreme slope inflation like slope > 900 for Ridge at n=20), thus indicating cases of severe overfitting and/or shrinkage distortions. Notably while Firth corrects for bias it also introduces calibration bias under sparse data conditions (Heinze & Schemper, 2002) and despite the Ridge having stable coefficients, it frequently under-corrects or over-corrects the probabilities Cessie & van Houwelingen, 1992). These findings reveal that calibration performance, especially slope behavior remains a key limitation for penalized methods in low-information contexts.

For Objective 3, the bootstrap analysis of coefficient standard deviations provided clearer insights into the stability of estimation and prediction. In cases of small samples, the MLE and Firth displayed substantial coefficient variability with some SDs exceeding 600 in some settings whereas the ridge regression consistently yielded the lowest SDs thus demonstrating superior stability (Sokolova et al., 2006). Despite differences in

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume X Issue VII July 2025

coefficient variability and calibration, AUC and Brier score distributions across methods were relatively stable. This aligns with findings by Pavlou et al. (2016), suggesting that discrimination metrics like AUC may remain robust even when coefficient estimates are unstable. Ridge offers the best stability in extreme conditions while the MLE is preferable when calibration accuracy is essential and with adequacy in sample size.

CONCLUSION

Maximum Likelihood Estimation (MLE) is best suited for large sample sizes with balanced event rates. The MLE delivers accurate, well-calibrated and unbiased estimates when sufficient data are available. However, it becomes unreliable in small samples and when events are rare due to bias and convergence issues.

Firth Regression is the most appropriate for small samples and when datasets are of rare events. The Firth logistic regression effectively addresses bias and separation problems, offering stable coefficient estimates. Nonetheless, the Forth may lead to poor calibration especially when data are sparse.

Ridge Regression is found to be ideal for scenarios with high predictor correlation and/or many covariates. It provides the most stable estimates and handles multicollinearity well. Though despite its stability, it occasionally suffers from calibration issues in low-information contexts.

RECOMMENDATIONS

The selection of logistic regression estimation methods should be based on study design especially the sample size, event rate and predictor correlation. In small samples (n=20) with low event rates (≤5%), MLE is highly unstable thus producing extreme bias with unreliable estimates. In such contexts, the Firth's method is recommended based on its bias reduction capabilities while Ridge regression offers superior stability of predictor coefficients and predictive consistency making it suitable for prediction-oriented models even compromised calibration.

For moderate sample all the methods improve in model's reliability but caution is needed under the low event rates. The Ridge remains as the most stable method whereas the Firth ensures finite estimates in borderline separation settings. The MLE becomes competitive especially with recalibration. For datasets with large samples (n=1000), the MLE is an optimal choice as it consistently delivers minimal bias, excellent calibration and stable performance metrics without the need for penalization.

Practitioners should evaluate method choice not only based on model convergence but also through calibration slope, and the bootstrap variability. Where stability is crucial or data are sparse, penalized methods are valuable. However, for well-powered studies, MLE remains the gold standard. To support robust inference, researchers should routinely report calibration and variability diagnostics alongside standard performance metrics such as AUC and Brier scores.

REFERENCES

1. Agresti, A. (2013). Categorical data analysis (3rd ed.). Wiley. https://doi.org/10.1111/biom.12128
2. Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71(1), 1–10. https://doi.org/10.1093/biomet/71.1.1
3. Cessie, S., & van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. Applied Statistics, 41(1), 191–201. https://doi.org/10.2307/2347628
4. Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika, 80(1), 27–38. https://doi.org/10.1093/biomet/80.1.27
5. Harrell, F. E. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-19425-7
6. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

7. Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. Statistics in Medicine, 21(16), 2409–2419. https://doi.org/10.1002/sim.1047
8. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634
9. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387
10. Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 186(1007), 453–461. https://doi.org/10.1098/rspa.1946.0056
11. Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. Journal of the American Statistical Association, 91(435), 1343–1370. https://doi.org/10.1080/01621459.1996.10477003
12. Kosmidis, I., & Firth, D. (2009). Bias reduction in exponential family nonlinear models. Biometrika, 96(4), 793–804. https://doi.org/10.1093/biomet/asp055
13. Pavlou, M., Ambler, G., Seaman, S., De Iorio, M., & Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Statistics in Medicine, 35(7), 1159–1177. https://doi.org/10.1002/sim.6782
14. Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. AI 2006: Advances in Artificial Intelligence, 4304, 1015–1021. https://doi.org/10.1007/11941439114