# Logistic Regression for Rare Events Data

Quinlin Neuhaus
STAT 6342

May 3, 2026

## 1 Introduction

Rare events data is any binary data which contains a very high(or very low) proportion of ones. Existing literature codes the rare class as ones by convention. Data like this arises in credit and fraud, rare disease presence, manufacturing component failures, and the study of geopolitical conflicts. There is no specific established threshold for what prevalence of ones constitutes a rare event, with results also depending on the given sample size. While traditional logistic regression is asymptotically valid, rare events data pose unique challenges for parameter estimation.

## 2 MLE Logistic Regression

### 2.1 Logistic Regression Overview

Let $Y_i \in \{0, 1\}$ denote a binary response and $X_i \in \mathbb{R}^d$ denote some continuous covariates with

$$Y_i \sim \text{Bernoulli}(p_i), \quad \log\left(\frac{p_i}{1-p_i}\right) = X_i^\top \beta.$$

The log-likelihood is

$$\ell(\beta) = \sum_{i=1}^{n} \left[ y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right],$$

with score function

$$U(\beta) = \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} x_i (y_i - p_i).$$

Maximum likelihood estimation proceeds by solving $U(\beta) = 0$ when a finite solution exists.

Under standard regularity conditions,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, I(\beta)^{-1}\right),$$

where the Fisher information is

$$I(\beta) = X^\top W X, \quad W = \mathrm{diag}(p_i(1 - p_i)).$$

The asymptotic normality of the MLE alone is an important result in the pursuit of parameter estimation, but does not imply parameter estimates are unbiased in finite samples, rare events, or both.

## 2.2 Finite-Sample Bias and Rare Events

Let $\hat{\beta}$ denote the MLE, defined implicitly by the score equation

$$U(\hat{\beta}) = 0.$$

Construct a Taylor expansion of the score function around the true parameter $\beta$,

$$0 = U(\hat{\beta}) = U(\beta) + U^{(1)}(\beta)(\hat{\beta} - \beta) + \frac{1}{2}U^{(2)}(\beta)[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top] + \cdots,$$

where

$$U^{(1)}(\beta) = \frac{\partial U}{\partial \beta^\top}, \quad U^{(2)}(\beta) = \frac{\partial^2 U}{\partial \beta \partial \beta^\top}.$$

Rearrange the expansion,

$$\hat{\beta} - \beta = -U^{(1)}(\beta)^{-1}U(\beta) - \frac{1}{2}U^{(1)}(\beta)^{-1}U^{(2)}(\beta)[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top] + \cdots.$$

Take expectations on both sides,

$$\mathbb{E}[\hat{\beta} - \beta] = -\mathbb{E}[U^{(1)}(\beta)^{-1}U(\beta)] - \frac{1}{2}\mathbb{E}[U^{(1)}(\beta)^{-1}U^{(2)}(\beta)[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top]] + \cdots.$$

The first term vanishes,

$$\mathbb{E}\big[U^{(1)}(\beta)^{-1}U(\beta)\big] \approx I(\beta)^{-1}\mathbb{E}\big[U(\beta)\big] \approx 0.$$

The second term is similar,

$$-\frac{1}{2}\mathbb{E}[U^{(1)}(\beta)^{-1}U^{(2)}(\beta)[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^{\top}]] \approx -\frac{1}{2}I(\beta)^{-1}\mathbb{E}[U^{(2)}(\beta)[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^{\top}]].$$

We let,

$$b(\beta) = -\frac{1}{2}\mathbb{E}[U^{(2)}(\beta)[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^{\top}]].$$

This yields the first-order bias expansion,

$$\mathbb{E}[\hat{\beta}-\beta] = I(\beta)^{-1}b(\beta) + O(n^{-2}),$$

where $b(\beta)$ depends on third derivatives of the log-likelihood.

From the Taylor expansion of the score function, the MLE's bias is proportional to the inverse of the Fisher information.

## 2.3 Why Bias is Amplified in Rare Events

In rare events data, where $p_i \ll 1$ for most observations, the Fisher information degenerates,

$$W_i = p_i(1-p_i) \approx 0, \quad \Longrightarrow \quad I(\beta) = X^{\top}WX \approx 0.$$

Consequently, $I(\beta)^{-1}$ becomes large, inflating the bias of the MLE as shown in the previous section. From an information perspective, the information grows with the size of the ones class, not the total sample size, meaning rare events data has little information to draw inference from. From the asymptotic normality of the MLE, the variance also explodes because its variance is the inverse Fisher information as well.

## 2.4 Resulting Bias Patterns

The biased estimates of model parameters typically result in downward biased intercepts and slopes biased away from 0. As information grows, either because the proportion of ones increases, or the total sample size increases, the bias decreases because the estimator is consistent. It is important to note that these biases are not the result of model misspecification or numerical problems, but inherent to the finite sample size.

## 3 Separation and Non-Existence of the MLE

Another issue with traditional MLE for logistic regression for rare events data comes from numerical methods. Binary data is said to be linearly separable if it can perfectly divided by some hyperplane, that is,

$$X_i^\top \beta > 0 \quad \text{for all } Y_i = 1, \text{ and } X_i^\top \beta < 0 \quad \text{for all } Y_i = 0.$$

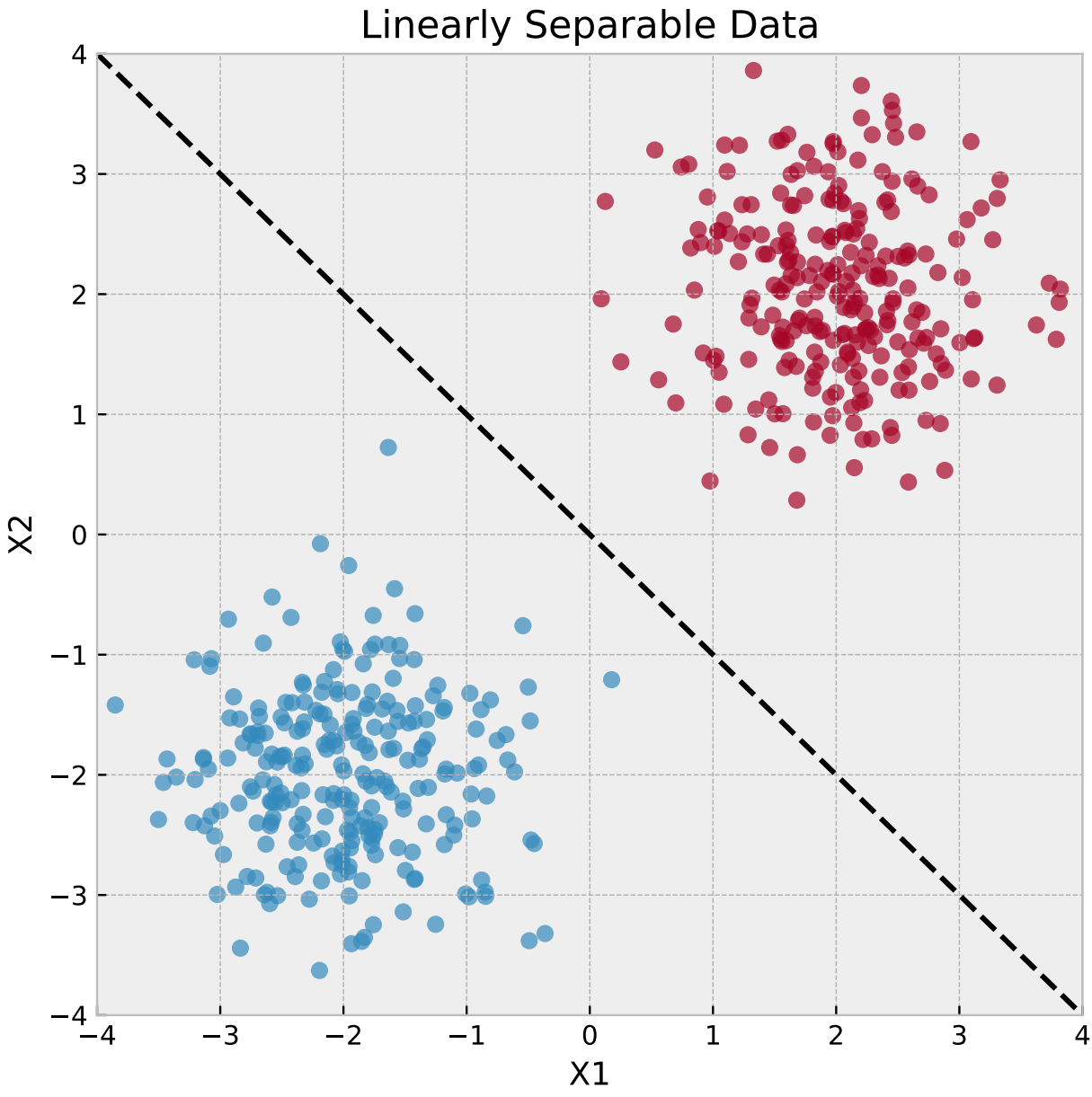


Figure 1: Linear Separability in $\mathbb{R}^2$

When this occurs, the likelihood can grow without bound when separated, and the estimated parameters also diverge,

$$\|\hat{\beta}\| \to \infty.$$

Thus, no finite maximizer exists and

$$p_i \to 1 \text{ or } p_i \to 0 \implies$$

$$p_i(1-p_i) \to 0 \implies$$

$$W_i = p_i(1-p_i) \to 0 \implies$$

$$I(\beta) = X^\top W X \to 0.$$

The most popular numerical methods for logistic regression MLE, Newton-Raphson and IRLS require $I(\beta)^{-1}$, so as the Fisher information becomes nearly singular these methods fail. Methods like gradient descent that do not require inverting the Fisher information also fail because there is no maximum on the likelihood surface to converge to.

In rare events data, separability becomes a greater possibility because of the small number of ones. This can become exacerbated by high dimensional settings, where the "blessing of dimensionality" can become a numeric curse.

# 4 Firth Bias-Reduced Logistic Regression

## 4.1 Firth's Penalty

Firth(1993) introduces the following bias reduced estimator by penalizing the log-likelihood,

$$\ell^*(\beta) = \ell(\beta) + \frac{1}{2}\log|I(\beta)|,$$

where $I(\beta) = X^\top W X$ is the Fisher information matrix and $\ell(\beta)$ is the typical log-likelihood.

The corresponding penalized score function is

$$U^*(\beta) = \frac{\partial \ell^*}{\partial \beta} = U(\beta) + A(\beta),$$

where

$$A(\beta) = \frac{1}{2}\frac{\partial}{\partial \beta}\log|I(\beta)| = I(\beta)b(\beta).$$

When Firth's penalty is applied to MLE for logistic regression, the score has the form,

$$U^*(\beta) = \sum_{i=1}^{n} x_i \left(y_i - p_i + (\tfrac{1}{2} - p_i)h_i\right),$$

where

$$h_i = H_{ii}, \quad H = W^{1/2}X(X^\top WX)^{-1}X^\top W^{1/2}.$$

Maximum likelihood estimation is replaced by solving $U^*(\tilde{\beta}) = 0$.

## 4.2 Bias Expansion of the Penalized Estimator

Let $\tilde{\beta}$ denote the Firth estimator. Expanding the modified score around the true parameter $\beta$ yields

$$0 = U^*(\tilde{\beta}) = U^*(\beta) + U^{*(1)}(\beta)(\tilde{\beta} - \beta) + \frac{1}{2}U^{*(2)}(\beta)\big[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^\top\big] + \cdots.$$

Rearranging the expansion,

$$\tilde{\beta} - \beta = -U^{*(1)}(\beta)^{-1}U^*(\beta) - \frac{1}{2}U^{*(1)}(\beta)^{-1}U^{*(2)}(\beta)\big[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^\top\big] + \cdots.$$

Take expectation on both sides,

$$\mathbb{E}[\tilde{\beta} - \beta] = -\mathbb{E}[U^{*(1)}(\beta)^{-1}U^*(\beta)] - \frac{1}{2}\mathbb{E}\Big[U^{*(1)}(\beta)^{-1}U^{*(2)}(\beta)\big[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^\top\big]\Big] + \cdots.$$

It is important to note that Firth's penalty is not just an ad hoc fix, but rather specifically constructed to eliminate the first bias term of the MLE. The derivative of the penalty term has the following formula,

$$A(\beta) = I(\beta)b(\beta),$$

where $b(\beta)$ is the first-order bias term from the standard expansion.

Therefore,

$$\mathbb{E}[U^*(\beta)] = \mathbb{E}[U(\beta)] + A(\beta) = 0 + I(\beta)b(\beta).$$

Using the same linearization of expectation argument as for the MLE,

$$\mathbb{E}[U^{*(1)}(\beta)^{-1}U^*(\beta)] \approx -I(\beta)^{-1}I(\beta)b(\beta) = -b(\beta).$$

Substituting into the expectation expansion gives

$$\mathbb{E}[\tilde{\beta} - \beta] = b(\beta) - b(\beta) + O(n^{-2}) = O(n^{-2}).$$

Firth's correction does not yield unbiased estimators in finite sample sizes, but eliminates the first order term and makes the estimates more accurate. For rare events data, Firth's estimates still exhibit similar behavior to MLE, better performance for larger sample sizes and more balanced classes.

### 4.3 Firth's Penalty on Numerical Methods

The secondary improvement that Firth's method provides over traditional MLE comes from its effect on the numerical methods used to estimate parameters. Recall Firth's penalized log-likelihood,

$$\ell^*(\beta) = \ell(\beta) + \frac{1}{2}\log|I(\beta)|,$$

Previously, it was noted that under linear separability,

$$I(\beta) = X^\top W X \to 0,$$

clearly,

$$I(\beta) \to 0 \implies \log|I(\beta)| \to -\infty.$$

Since our methods maximize $\ell^*$, there exists some trade off between maximizing the log-likelihood and driving $p_i$ to 0 or 1.

Firth's method guarantees that even under separation, there exists a finite maximizer of the penalized log-likelihood. This keeps the Fisher information non singular, so that numerical methods that invert $I(\beta)$ hold.

## 5 Simulation Study

The following simulation was conducted to demonstrate finite sample bias and compare the performance of traditional MLE logistic regression against Firth's bias-reduced logistic regression.

## 5.1 Design

Points $(X, Y)$ were generated according to

$$X_i \sim N(0,1), \quad \text{logit}(p_i) = \beta_0 + \beta_1 X_i, \quad Y_i \sim \text{Bernoulli}(p_i).$$

With the following parameter values:

- $\beta_1 = 1$,
- $\beta_0 \in \{-6, -5, -4, -3, -2, -1, 0\}$,
- $n \in \{30, 100, 300, 1000, 3000, 10000\}$.

Model parameters were estimated with MLE and Firth's method on 1,000 Monte Carlo simulations of each pair of $\beta_0$ and $n$, except for pairs that would generate samples of all zeroes with high probability. Bias of $\beta_0$ and $\beta_1$ were calculated by subtracting the corresponding population values and plotted against different intercept values. IRLS was chosen as the numerical method to maximize the log-likelihoods, and all the simulations were ran with Python using NumPy, PyTorch, and Matplotlib. Because of the size of the simulation, the code was ran on Google Colab GPU.

## 5.2 Results

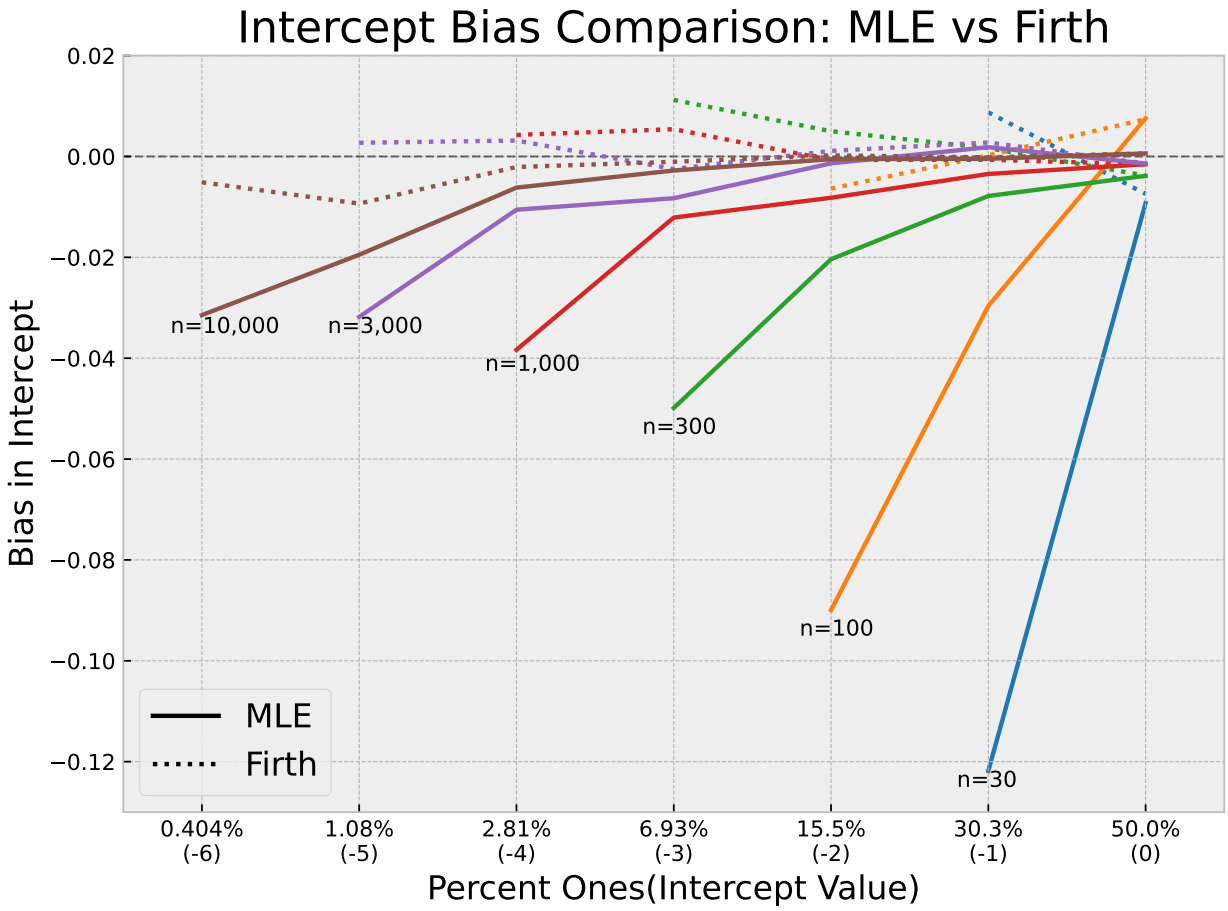


Figure 2: Intercept Bias

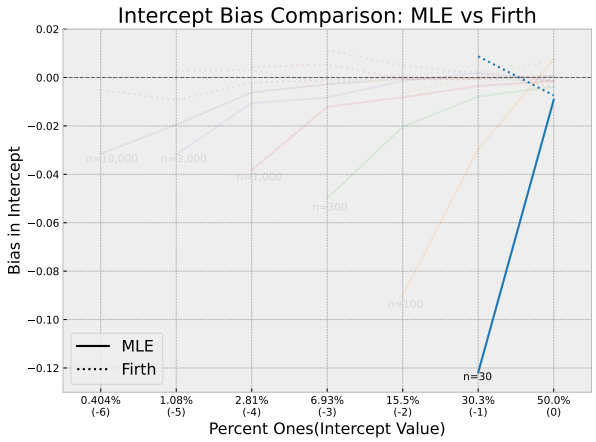


Figure 3: Intercept Bias(n=30)

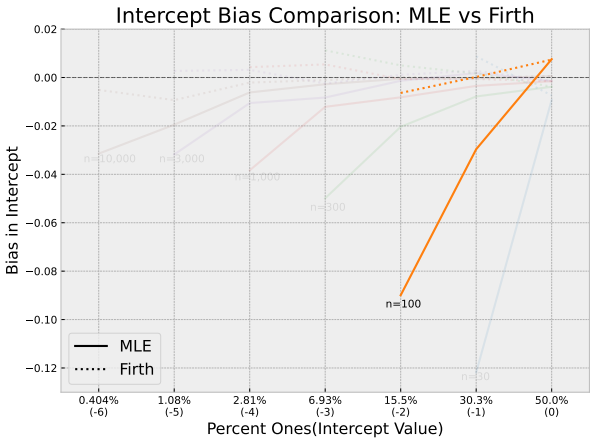


Figure 6: Intercept Bias(n=100)

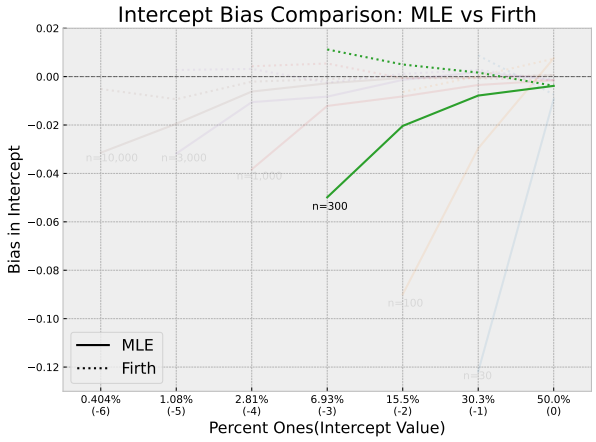


Figure 4: Intercept Bias(n=300)

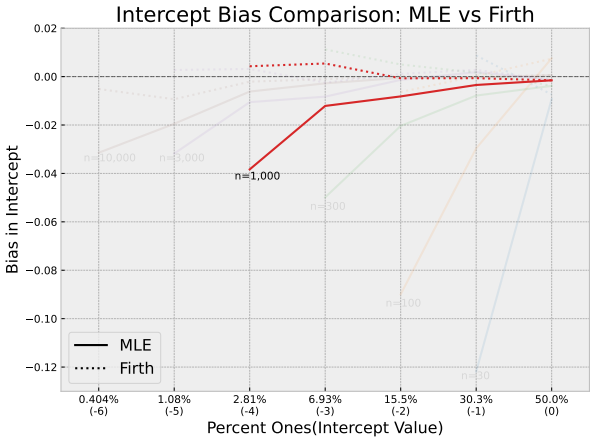


Figure 7: Intercept Bias(n=1000)

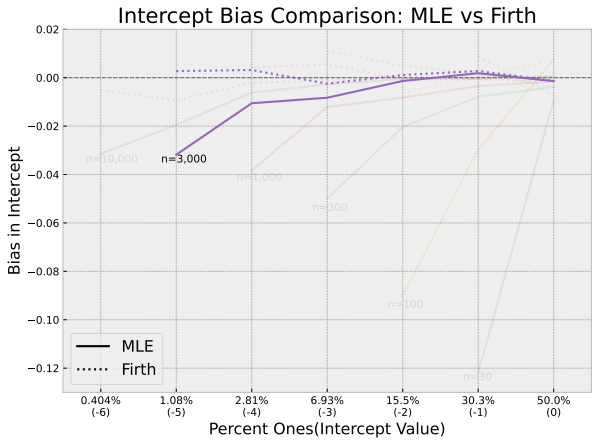


Figure 5: Intercept Bias(n=3000)

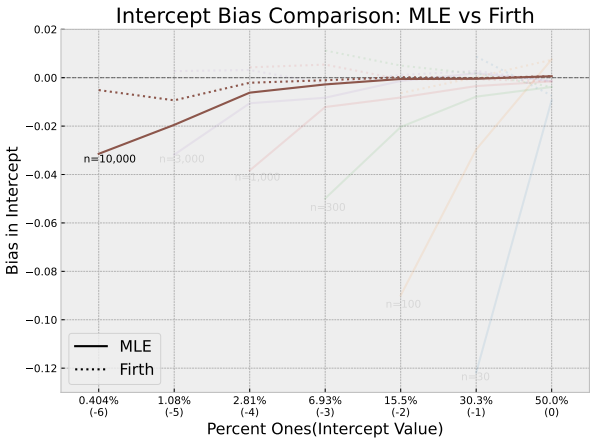


Figure 8: Intercept Bias(n=10000)

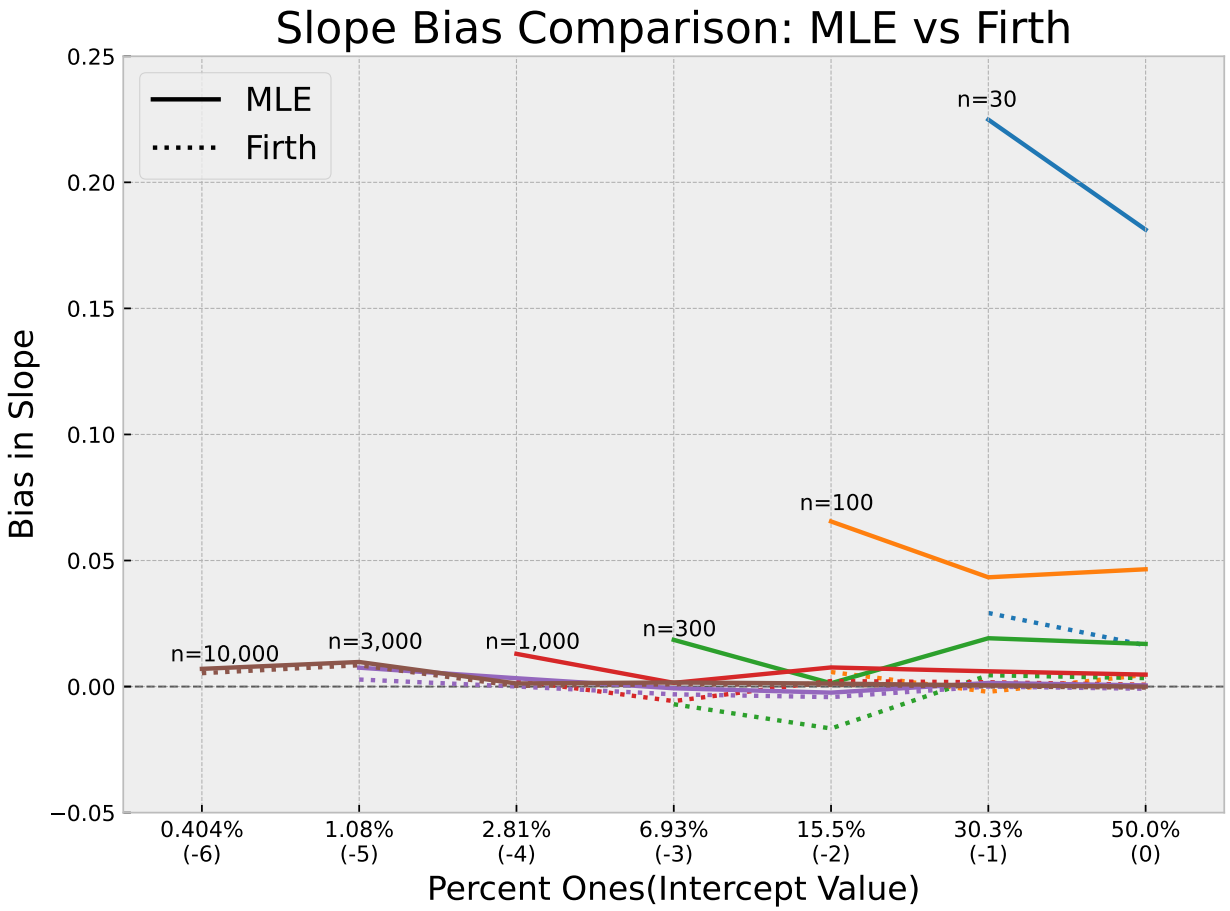


Figure 9: Slope Bias

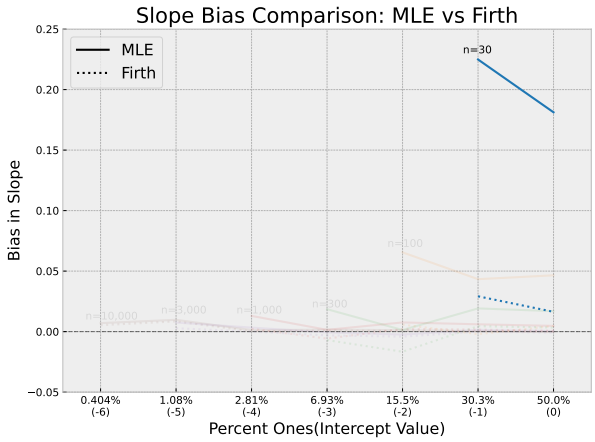


Figure 10: Slope Bias(n=30)

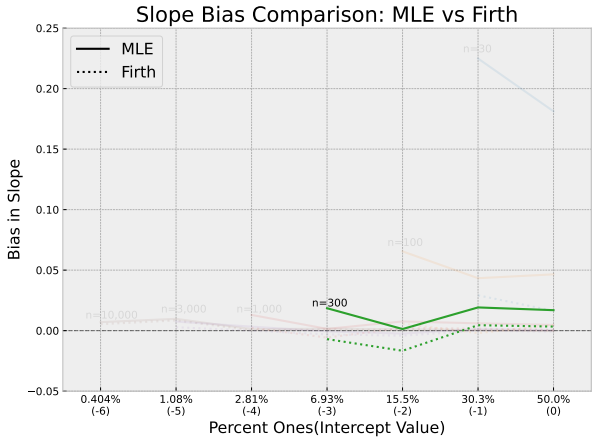


Figure 11: Slope Bias(n=300)

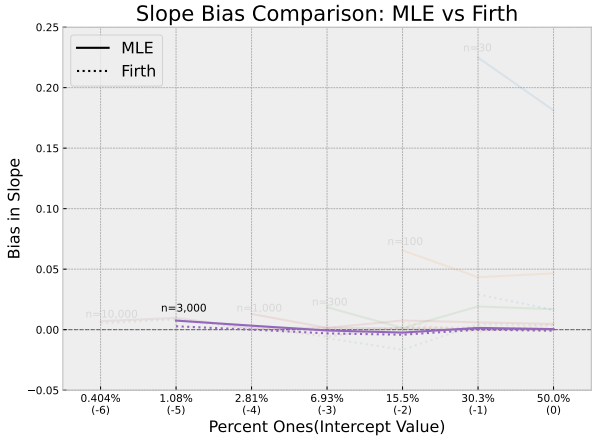


Figure 12: Slope Bias(n=3000)

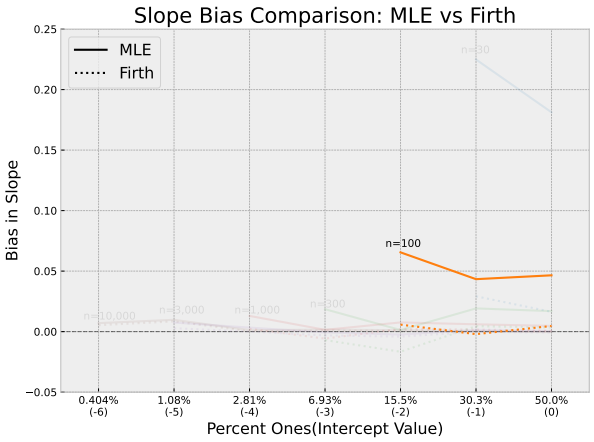


Figure 13: Slope Bias(n=100)

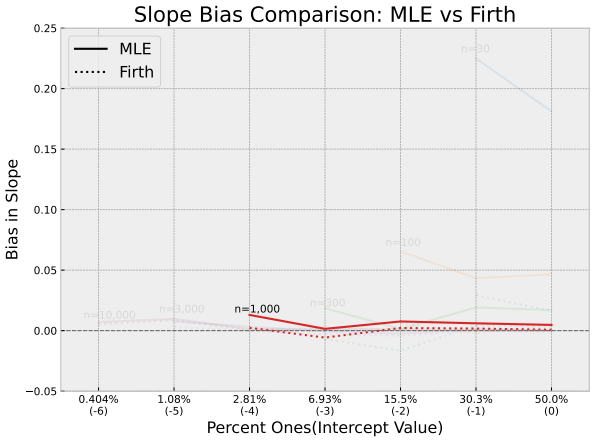


Figure 14: Slope Bias(n=1000)

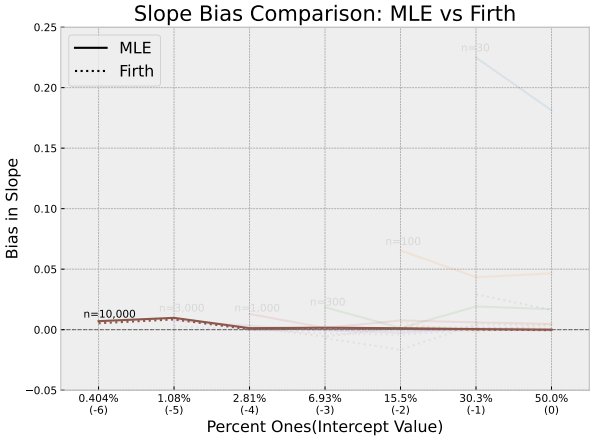


Figure 15: Slope Bias(n=10000)

The following patterns can be observed from the simulation that align with theoretical results:

- MLEs have worse bias than Firth's method for both slope and intercept
- MLE underestimates the intercept and overestimates the slope
- Increasing the sample size improves estimation across both slope and intercept
- Bias diminishes as the classes become closer to balanced

# 6 Real Data Application

The previous simulation demonstrates the significant difference between the MLE and Firth estimates in small sample sizes. The subsequent real world data analysis demonstrate the contrapositive. NFL play-by-play data was collected for the last 27 NFL season(1999-2025). 27,721 total field goal attempts were recorded, with blocked field goals coded as ones, and makes and misses coded as zeroes. Across the sample, 2% of attempts recorded a block. The only covariate includes was the distance in yards on the attempt. MLE and Firth logistic regression were run on the data, which produced the following model parameter estimates:

| Model | $\hat{\beta}_0$ | $\hat{\beta}_1$ |
|---|---|---|
| MLE | -5.44311 | 0.0406545 |
| Firth | -5.44021 | 0.0406247 |

As expected, the estimates are nearly identical, showing that when the problems with rare event regimes, separability, low event rate, and small sample size, are absent these methods are asymptotically equivalent.

# 7 Conclusion

Logistic regression is a flexible method for predicting probabilities and classifying binary tabular data. Its MLE model parameters are asymptotically consistent, but not unbiased, which becomes problematic in rare events data. Combined with separation issues, these concerns can make traditional MLE a dubious choice in rare event settings.

Firth's bias-reduced logistic regression constructs a penalty which eliminates first order bias and tackles numerical instability. Differences in these methods become apparent in rare events data, where Firth's method becomes superior.

# References


[1] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38.

[2] Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16), 2409–2419.

[3] King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2), 137–163.

[4] Puhr, R., et al. (2017). Firth's logistic regression with rare events. *Statistics in Medicine*, 36(14), 2302–2317.

[5] Wang, H. (2020). Logistic regression for massive data with rare events. *PMLR*, 119, 9829–9836.