Modeling Fractional Outcome: An Empirical Practice Illustration

by

Mohsen Fazeli
B.Sc., Shahid Beheshti University, 2021

An Extended Essay Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF ARTS

in the Department of Economics

Dr. Tao Wang, Supervisor (Department of Economics)

Dr. Kenneth G. Stewart, Member (Department of Economics)

Contents

List of Figures

List of Tables

1 Introduction

Fractional variables are percentages, proportions, or rates bounded within the range of [0, 1]. Some examples of fractional variables in economics include interest rates, unemployment, participation rates, health insurance coverage, etc. Researchers place a significant emphasis on modeling these types of variables to gain insights into various economic phenomena; see Maddala (1991), Cox (1995), Kieschhnick & McCullough (2003), among others. Given various applications of fractional variables in economic research, understanding their properties is essential to conducting reliable economic analysis.

The bounded nature of the fractional variable leads to its nonlinear relationship with the independent variables. In other words, no change in the independent variable can cause the fractional variable to be outside the range [0, 1]; therefore, the relation is forced to be nonlinear, especially when the data points approach the values of 0 or 1 (Wu, Baleanu, & Luo, 2017). Another significant characteristic of fractional variables is heteroskedasticity, where the expectation of the dependent variable conditional on the independent variables is not constant. Instead, the residuals are more likely to exhibit changes in variance at different levels of the independent variable, which can complicate the modeling process and impact the reliability of the estimates (Elsas & Florysiak, 2013). Due to these characteristics, the reliability of using traditional linear regression (LR) to model fractional variables has been questioned, which is represented as

$$\mathbb{E}(y_i \mid X_i) = X_i^T \beta, \tag{1}$$

where y_i is the fractional dependent variable constrained within the unit interval, and X_i represents the independent variables. LR in (1) is commonly used in economic research due to its simple application and interpretation. LR utilizes the Ordinary

Least Squares (OLS) method to minimize the mean squared errors and fit a linear line to the data. Linearity and homoskedasticity are fundamental assumptions of the OLS method. The OLS method assumes that each term in the model is either a constant or a coefficient multiplied by an independent variable and that the error terms have a constant variance regardless of the values of the independent variables. However, OLS assumptions of linearity and homoskedasticity are more likely to be violated if the dependent variable is fractional, which can cause biased estimators. Furthermore, using LR in the case of fractional variables can lead to predictions that fall outside the bounded values. For instance, predicting a negative unemployment rate or a 110 percent market share is impractical. Therefore, while LR is favored for its simplicity, its application to fractional variables requires caution and the adoption of alternative modeling approaches that account for the bounded nature and potential issues.

It is argued that using a log-odds transformation of fractional variables enables the application of a linear model. The log-odds model is defined as

$$\mathbb{E}\left(\log \frac{y_i}{1-y_i} \mid X_i\right) = X_i^T \beta. \tag{2}$$

While the log-odds model (2) is popular because it ensures that predictions fall within the range of [0, 1], it does not entirely address the issue associated with modeling fractional variables. Papke & Wooldridge (1996) demonstrate that the log-odds model is invalid when fractional variables take on values of 0 and 1. This limitation arises because the logarithm of zero or the division by zero is undefined, making the transformation problematic at these boundary values. In addition, Papke & Wooldridge (1996) point out that using log-odds models requires estimating a conditional density function, which is either difficult or tends to produce non-robust results. Therefore, a log-odd model is not feasible in practice besides providing the possibility of using a

linear model for fractional data while maintaining predictions within the unit interval. The Tobit model is another common approach for handling fractional variables, particularly in cases with numerous observations clustered at the boundary values. Ramalho & Ramalho (2011) highlight the Tobit model's application in the variables that their values are only observed up to a specific limit named censored data. However, the rationale for using the Tobit model becomes challenging to justify for fractional variables, which are naturally limited within a bounded interval. Therefore, theoretically, the Tobit model application in fractional variables may not be suitable.

To address the challenges of modeling fractional variables, Papke & Wooldridge (1996) propose a model that assumes the existence of a known function $G(\cdot)$, ensuring all predicted values fall within the unit interval (0 to 1). The model is expressed as

$$\mathbb{E}(y_i \mid X_i) = G(X_i^T \beta), \tag{3}$$

where $G(\cdot)$ is a distribution function, typically a logistic function $G(z) = \frac{\exp(z)}{1+\exp(z)}$, mapping z between 0 and 1. To estimate the parameters in the fractional model (3), Papke & Wooldridge (1996) suggest using the specific quasi-maximum likelihood (QML) method proposed by Gourieroux, Monfort, & Trognon (1984) and McCullagh & Nelder (1989) to estimate the parameters in the model. This estimation involves maximizing the Bernoulli log-likelihood function, as shown below

$$l_i(\beta) = y_i \log[G(X_i^T \beta)] + (1 - y_i) \log[1 - G(X_i^T \beta)]. \tag{4}$$

They demonstrate that QML estimation is consistent and asymptotically normal, regardless of the distribution of y_i conditional on X_i, which provides greater flexibility for this method. Furthermore, based on the sandwich formula (Cameron & Trivedi, 2005) and the nonlinear conditional mean $G(\cdot)$, the asymptotic variance of β can also be

estimated (Papke & Wooldridge, 1996). Fractional logistic regression (FLR) proposed by Papke & Wooldridge (1996) provides a flexible and robust approach to model fractional variables, even in the presence of heteroskedasticity, and ensures all the predicted values remain within the unit interval. FLR has been applied in various studies across different fields. Molowny-Horas, Basnou, & Pino (2017) utilize FLR to model land use and cover dynamics in a Mediterranean landscape. Their findings indicate that fragmentation is influenced not only by geographical and environmental factors but also by the surrounding landscape. Fang & Ma (2012) use FLR to investigate Chinese insurance coverage rates and their relationship with household size, income, expense, and chronic disease. Papke & Wooldridge (2008) use FLR to examine the estimation of spending on math test results for fourth-grade students in Michigan, demonstrating how changes in school funding in 1994 impacted students' academic performance. Martins (2018) compares FLR with LR and Tobit in handling the fractional nature of efficiency scores of banks and concludes that FLR outperforms LR and Tobit in accurately capturing the efficiency of Portuguese banks. Villadsen & Wulff (2021) show the applicability of FLR in diverse research areas by replicating two published papers in strategy and management research using FLR. These studies collectively demonstrate the utility and versatility of FLR in addressing the complexities of fractional variables across different research contexts.

While FLR offers advantages for analyzing fractional variables, it has yet to gain widespread popularity among researchers. LR, log-odds, and Tobit continue to be more commonly used despite FLR's benefits. Villadsen & Wulff (2021) find that only 6 percent of published papers working with fractional variables in top journals utilized FLR. Considering the best-suited modeling approach is essential to any research, as unsuitable models can lead to incorrect estimations, misinterpreting results, and potentially misleading decision-making. Therefore, researchers must carefully consider

the characteristics of their data, the assumptions of different models, and the research objectives before deciding on a modeling technique. While FLR may offer advantages in particular contexts, the choice of model should always be guided by a thorough understanding of the data, methodology, and research goals to ensure accurate and meaningful results.

This study contributes to understanding the best practices for modeling fractional variables. In Section 2, the paper briefly discusses the theoretical background of FLR, such as model specification, assumptions, estimation procedure, and asymptotic properties, while comparing it to the LR. Then, in Section 3, the paper conducts Monte Carlo simulations, where both FLR and LR are applied to simulated datasets with different characteristics, providing model performance evaluation and a guideline for model selection in the case of fractional variables. In Section 4, the paper validates FLR applicability across various fields by replicating findings from three published papers in the *Journal of Applied Econometrics* using FLR. In Section 5, the paper conducts an empirical analysis examining the nonlinear relationship between poverty rate and educational level. Finally, the paper concludes the findings in Section 6. Overall, this study enhances our understanding of modeling fractional variables and demonstrates the versatility of FLR in addressing complex economic relationships. Findings of this research can serve as a valuable resource for researchers, practitioners, and policymakers seeking to leverage FLR for more accurate and insightful analyses in economics and related fields.

2 Methodology

This section outlines FLR's theoretical background, including the model specification, assumptions, estimation procedures, and asymptotic results, to highlight its advantages

over LR in handling fractional dependent variables.

2.1 Model Specification

FLR is designed for fractional outcomes. The expected value of the dependent variable y_i given the explanatory variables X_i is expressed as

$$\mathbb{E}(y_i \mid X_i) = G(X_i^T \beta), \tag{5}$$

where $G(\cdot)$ is the logistic function defined as

$$G(z) = \frac{\exp(z)}{1 + \exp(z)}, \tag{6}$$

and β represents the vector of parameters to be estimated. The logistic function (6) addresses the bounded nature of the fractional dependent variable, preventing predictions from falling outside the [0, 1] range.

2.2 Assumptions

Several assumptions must be met to utilize FLR. The dependent variable y_i must be bounded between 0 and 1, and also the logistic function $G(X_i^T \beta)$ must correctly specify the relationship between the dependent and independent variables. The observations must also be independently and identically distributed (i.i.d), and there should be no perfect multicollinearity among the independent variables X_i. Finally, the error term should follow a logistic distribution to be able to use FLR.

2.3 Estimation Procedure

The parameters in FLR are estimated using the QML method, which maximizes the Bernoulli log-likelihood function, defined as

$$l_i(\beta) = y_i \log[G(X_i^T \beta)] + (1 - y_i) \log[1 - G(X_i^T \beta)]. \tag{7}$$

This function in (7) represents the likelihood of observing the data given the parameters β. This maximization occurs through an iterative optimization process called the Newton-Raphson method, which can be defined as

$$\beta^{(k+1)} = \beta^{(k)} - H^{-1}(\beta^{(k)}) U(\beta^{(k)}), \tag{8}$$

where k is the iteration indicator, $U(\beta^{(k)})$ is the gradient of the log-likelihood function, providing the direction and magnitude of the adjustments. $H^{-1}(\beta^{(k)})$ is the second derivative of the log-likelihood function indicates the curvature of the log-likelihood function. This method refines the parameter in each iteration until the changes in the parameter values tend to zero.

2.4 Asymptotic Properties

Asymptotic properties of QML, including consistency and asymptotic normality, enable accurate inferences about parameters as the sample size increases. Consistency implies that as the sample size n increases, the QML estimator $\hat{\beta}$ converges to the true parameter value β. The asymptotic normality means that as the sample size becomes large enough, the distribution of the QML estimator $\hat{\beta}$ approximates a normal

distribution, which can be expressed as

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, V_\beta), \tag{9}$$

where $\xrightarrow{d}$ denotes convergence in distribution, and $\mathcal{N}(0, V_\beta)$ represents a normal distribution with mean 0 and variance-covariance matrix V_β. This property (9) is useful as it enables us to construct confidence intervals and test the hypothesis. The variance-covariance matrix V_β determines the variability of the estimator and can be obtained using a sandwich formula specified as

$$\begin{aligned} V_\beta = & \left[\frac{1}{n}\sum_{i=1}^{n} G'(X_i^T\beta)X_iX_i^T\right]^{-1} \\ & \left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - G(X_i^T\beta)}{G(X_i^T\beta)[1-G(X_i^T\beta)]}\right)^2 X_iX_i^T\right] \\ & \left[\frac{1}{n}\sum_{i=1}^{n} G'(X_i^T\beta)X_iX_i^T\right]^{-1}, \end{aligned} \tag{10}$$

where $G'(\cdot)$ is the first derivative of $G(cdot)$.

2.5 FLR vs LR

LR uses the OLS method to estimate the results, which is defined as

$$y_i = X_i^T\beta + \epsilon_i. \tag{11}$$

The parameters β are estimated by minimizing the sum of squared residuals, resulting in

$$\hat{\beta} = \left(X^TX\right)^{-1}X^Ty. \tag{12}$$

The variance of the OLS estimator in (12) is given by

$$\text{Var}(\hat{\beta}) = \sigma^2(X^T X)^{-1}, \tag{13}$$

where σ^2 is the variance of the error term. While LR is appreciated for its simple application and interpretation, it faces challenges when applied to fractional dependent variables. Specifically, LR does not constrain the predicted values within the [0, 1] interval and assumes a linear relationship and homoskedastic errors, which are often violated in the case of fractional data. On the other hand, FLR effectively addresses these limitations. As discussed previously, FLR ensures that all predicted values remain within the unit interval, captures non-linear relationships, and provides consistent estimates even in the presence of heteroskedasticity. Table 1 summarizes the critical differences between FLR and LR in terms of Model Specification (1), Assumptions (2), Estimation Procedure (3), and Asymptotic Properties (4).

While LR remains a powerful and widely used technique, FLR is a more appropriate method from a theoretical perspective for fractional outcomes. The following section, Monte Carlo Simulations, will further illustrate these two regression techniques' practical applications and comparative performance, highlighting scenarios where FLR demonstrates clear advantages.

Table 1: Comparison between FLR and LR

#	**FLR**	**LR**
(1)	$\mathbb{E}(y_i \mid X_i) = G(X_i^T\beta)$	$y_i = X_i^T\beta + \epsilon_i$
(2)	Dependent variable bounded in [0, 1] Logistic function specifies relationship i.i.d observations No perfect multicollinearity Logistic distribution of errors	Linearity Independence of errors Homoskedasticity Normality of errors
(3)	QML Bernoulli log-likelihood function Iterative optimization	OLS $\min_\beta \sum_{i=1}^n (y_i - X_i^T\beta)^2$
(4)	Consistency Asymptotic normality Variance-covariance matrix estimation	Consistency Asymptotic normality Variance: $\sigma^2(X^TX)^{-1}$

3 Monte Carlo Simulations

3.1 One Independent Variable

Monte Carlo simulations compare FLR's and LR's performance in modeling fractional outcomes. Each simulation involves 500 iterations, and the average of estimated coefficients (Coef), Mean Squared Error (MSE), and Standard Error of Estimates (SEE) are reported. The simulation starts by drawing 1000 random values from a normal distribution to construct an independent variable (X_i). To generate y_i within $[0, 1]$,

the logistic model is used, which is defined by

$$y_i = \frac{1}{1 + \exp\left(-(\epsilon_i + X_i\beta)\right)}, \tag{14}$$

where $\beta = 1$ and ϵ_i is a random noise drawn from a normal distribution $\mathcal{N}(0, 0.5)$.[1] To create datasets with diverse characteristics, the mean and variance of the normal distribution from which X_i is drawn are systematically varied. The mean is varied through values -5, 0, and 5; the variance is varied through values 0.1, 1, and 5. This range of means and variances ensures that a wide spectrum of possible data scenarios are captured. The finite sample performance of both estimators is investigated by varying the sample size through 100, 500, 1000, and 5000 to understand asymptotic results for both FLR and LR. Main scenarios are characterized by low, medium, and high variation (defined based on the value of σ_x^2). Each of these main scenarios is further subdivided into three subgroups based on the value of μ_X (μ_X equals -5, 0, and 5). Each subgroup represents cases where the fractional variable y_i is mostly scattered around 0, in the middle, and 1.

3.1.1 Low Variation

Figure 1 illustrates both model predictions in case of low variation in the data, where the variance of X_i is set at 0.1. Given the known relationship in the simulation between X_i and y_i, low variation in X_i also leads to relatively low variation in y_i. This results in data being concentrated around a specific point. To examine different subgroups under low variation scenarios, the mean of the normal distribution from which the 1000 X_i values are drawn is varied. This allows the creation of cases where the fractional variable y_i is scattered close to 0 (top plot), around 0.5 (middle plot), and close to 1

[1] After examining various distributions for ϵ_i, the normal distribution $\mathcal{N}(0, 0.5)$ is found to provide the best exercises for the model.

(bottom plot). Each plot in Figure 1 displays the dependent variable y_i on the vertical axis and the independent variable X_i on the horizontal axis. The blue dots represent the simulated data, while the green plus markers (+) and red dots indicate the predictions from the LR and FLR, respectively. The vertical axis in each plot is adjusted to different ranges to zoom into the specific section, allowing for a more straightforward presentation of the predictions for both models. LR and FLR fit the model similarly in all three sub-plots, and their predictions are mostly identical. This observation suggests that when the variation in X_i is minimal, both models can accurately capture the relationship between X_i and y_i. This is reasonable since the scenarios depicted in Figure 1 provide a controlled environment where the data points are closely clustered, thereby minimizing the inherent complexity in the relationship between the independent and dependent variables.

Table 2 provides the results of the Monte Carlo simulations, comparing the performance of FLR and LR in case of low variation in data. The table summarizes MSE, SEE, and Coef for both models across different subgroups with varying means of X_i. For $\mu_X = -5$, the LR model has an MSE of 0.9850 and a SEE of 0.0013, while the FLR model shows a significantly lower MSE of 0.0298 and a SEE of 0.1676. The average coefficient estimates differ, with FLR at 0.9953 and LR at 0.0075. For $\mu_X = 0$, the MSE for the LR model is 0.5817, and the SEE is 0.0378, while FLR has a much lower MSE of 0.0255 and a SEE of 0.1523. The average coefficient estimates are 0.9553 for FLR and 0.2382 for LR. For $\mu_X = 5$, LR has an MSE of 0.9850 and a SEE of 0.0013, while FLR again shows a lower MSE of 0.0281 and a SEE of 0.1676. The average coefficient estimates are 1.0010 for FLR and 0.0075 for LR. Overall, Table 2 indicates that for low variation scenario, FLR consistently outperforms LR by providing better model fit with lower MSE and more accurate coefficient estimates. This accuracy in coefficient estimation is critical for correctly understanding the relationship between

the dependent and independent variables.

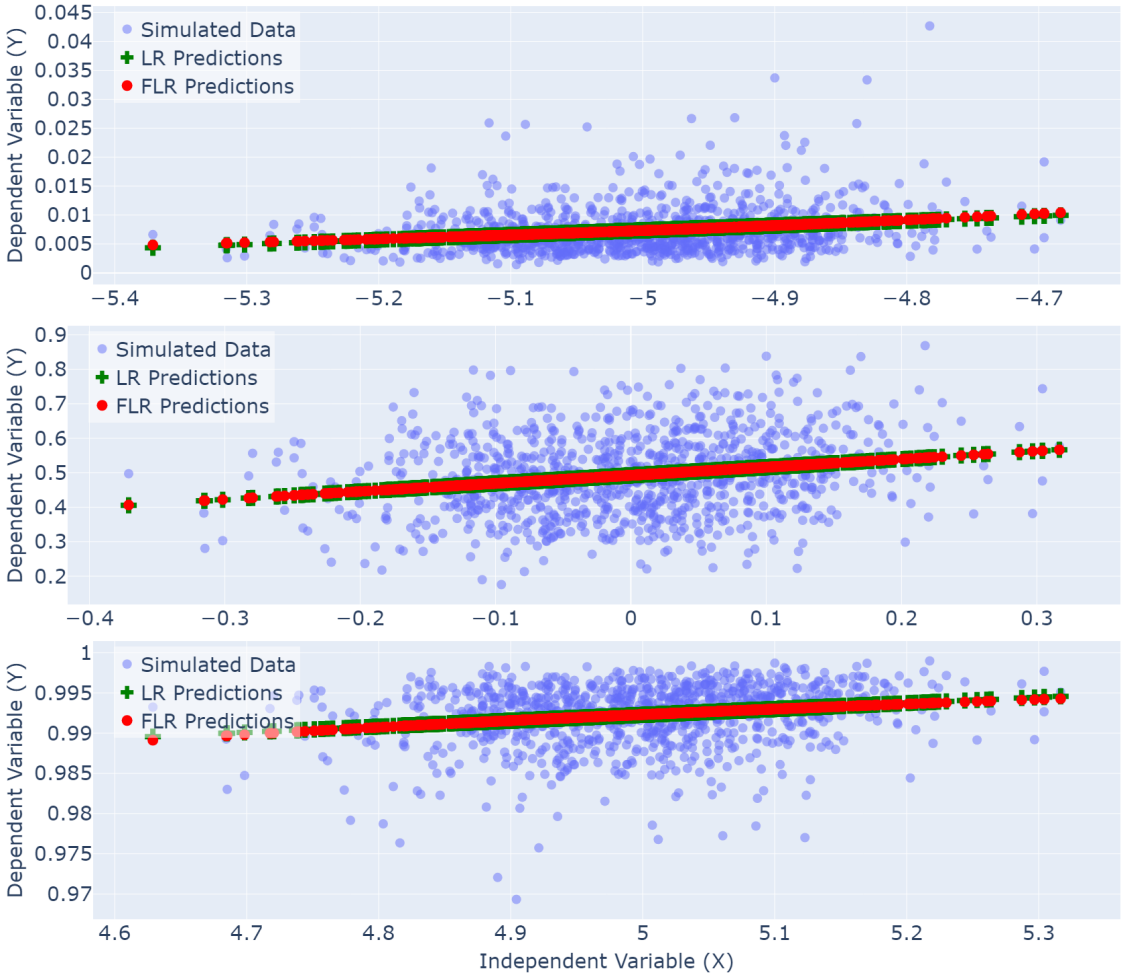


Figure 1: Model Predictions in Low Variation Case (3.1)

Table 2: Monte Carlo Simulation Results for Low Variation Case (3.1)

			LR			FLR		
n	σ_x^2	μ_x	MSE	SEE	Coef	MSE	SEE	Coef
1000	0.1	-5	0.9850	0.0013	0.0075	0.0298	0.1676	0.9953
1000	0.1	0	0.5817	0.0378	0.2382	0.0255	0.1523	0.9553
1000	0.1	5	0.9850	0.0013	0.0075	0.0281	0.1676	1.0010

3.1.2 Medium Variation

Figure 2 represents the case where X_i and consequently y_i have medium variation. To construct medium variation in data, the normal distribution that X_i is drawn from is set to have a variance of 1. The top plot shows when the fractional variable y_i is mostly scattered around 0, the middle plot is when y_i is scattered within the unit interval, and the bottom one shows when y_i is mostly scattered around 1. As can be seen, the model fit slightly differs, particularly in cases where the data are around the bounded values. In the top plot, the LR predictions are negative in several cases, while FLR predictions stay within the bounded value of 0. This behavior demonstrates the limitation of LR in handling fractional outcomes as it can produce invalid negative predictions for y_i. In the middle plot, where y_i values are scattered within the unit interval, the gap between the model fits of FLR and LR becomes smaller. However, FLR continues to offer more accurate predictions that respect the bounds of the unit interval. In the bottom plot, where y_i values are mostly around 1, LR predictions exceed the value of 1 and cannot capture the non-linearity that existed in the data. Conversely, FLR predictions show nonlinear behaviour and remain within the bounded value of 1. Overall, the difference in model fitting between FLR and LR becomes more pronounced with greater variation in the data, especially near the boundaries. This increased variability makes it clear that FLR is better suited for modeling fractional outcomes, as it considers non-linearity and naturally constrains the predictions within the valid range of [0, 1]. LR, on the other hand, cannot account for the non-linearity and may produce invalid predictions outside the range, leading to potential inaccuracies and misinterpretations in practical applications.

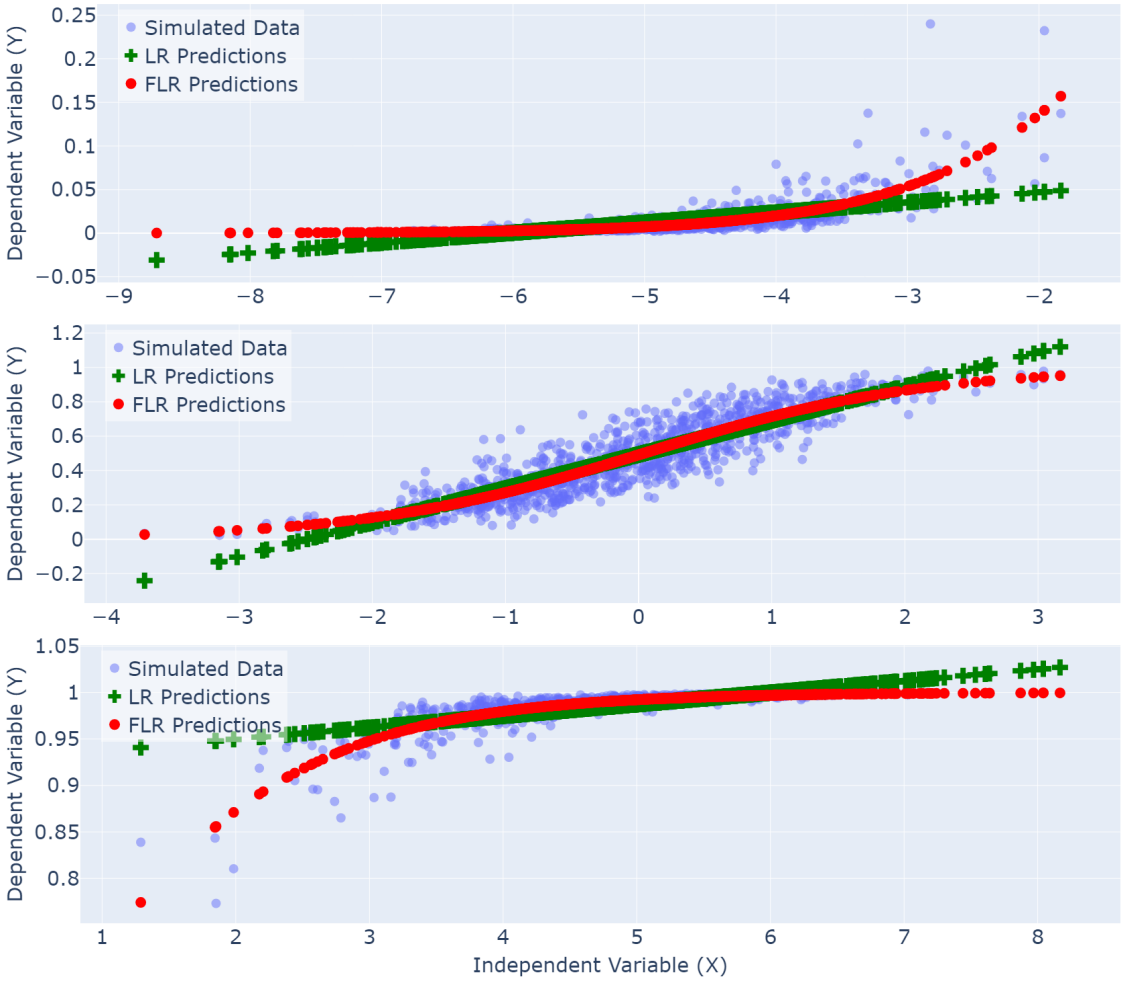


Figure 2: Model Predictions in Medium Variation Case (3.1)

Table 3 presents the results of Monte Carlo simulations in case of medium variation in the data. For $\mu = -5$, the MSE for LR is 0.9769 with a standard error of 0.0009 and an average coefficient estimate of 0.0116, while FLR shows a significantly smaller MSE of 0.0011, a higher SEE of 0.0326, and a closer to accurate average coefficient estimate of 0.9944. These results suggest that FLR offers a better fit and more precise coefficient estimates in cases where the fractional variable y_i is concentrated around 0. When $\mu = 0$, LR and FLR have MSEs of 0.6409 and 0.0027, respectively, with FLR maintaining a slight edge in MSE and a more precise average coefficient estimate

of 0.9507 compared to LR's 0.1995. This indicates that FLR provides a better fit and more reliable coefficient estimation when y_i is scattered within the unit interval. For $\mu = 5$, the MSE for LR is 0.9769 with a SEE of 0.0008, while FLR again shows better performance with an MSE of 0.0010, a SEE of 0.0310, and an average coefficient estimate of 0.9932. These results highlight FLR's consistent ability to deliver a better fit and more accurate coefficient estimates, particularly when y_i is concentrated around 1. Overall, Table 3 suggests that FLR consistently provides better fit and more accurate coefficient estimates than LR, especially when data exhibit medium variation. The improved performance of FLR is particularly evident in scenarios where y_i is close to the bounded values of 0 or 1. This behavior underscores FLR's advantage in maintaining the validity of predictions within the bounded range of fractional outcomes, which becomes more pronounced with medium variation in the data.

Table 3: Monte Carlo Simulation Results for Medium Variation Case (3.1)

			LR			FLR		
n	σ_x^2	μ_x	MSE	SEE	Coef	MSE	SEE	Coef
1000	1	-5	0.9769	0.0009	0.0116	0.0011	0.0326	0.9944
1000	1	0	0.6409	0.0028	0.1995	0.0027	0.0157	0.9507
1000	1	5	0.9769	0.0008	0.0116	0.0010	0.0310	0.9932

3.1.3 High Variation

Figure 3 illustrates both model predictions in case of high variation ($\sigma^2 = 5$) in the data. High variation causes a more significant discrepancy in model fit and predictions between FLR and LR. Across all subgroups, the LR predictions extend beyond the unit interval, whereas the FLR predictions remain within [0, 1]. The LR model produces numerous negative predictions in the top plot, where y_i values are mostly

scattered around 0. This is problematic because negative values are invalid for fractional outcomes. In contrast, the FLR predictions stay within the bounded value of 0, maintaining the validity of the predictions. In the middle plot, where y_i is scattered within the unit interval, the LR predictions exceed the unit interval on both ends. This results in predictions that are not feasible within fractional data. Meanwhile, the FLR model continues to provide more accurate predictions that remain within the valid range. In the bottom plot, where y_i is mostly scattered around 1, numerous LR predictions surpass the value of 1, which is again impossible for fractional outcomes. The FLR model, on the other hand, keeps its predictions within the bounded value of 1. Additionally, the fitted linear line from LR fails to adequately capture the nonlinear relationship between X_i and y_i, as seen in the scattered nature of the linear predictions, which do not align well with the actual data distribution. The FLR model, however, effectively captures this nonlinear relationship, providing a better fit to the data. This ability to model non-linearity is particularly important in cases where the relationship between variables is not strictly linear, as with fractional outcomes.

Table 4 illustrates the results of Monte Carlo simulations in case of high variation in the data. For $\mu = -5$, the LR shows a high MSE of 0.9059, with a SEE of 0.0018 and a Coef of 0.0482. In contrast, the FLR model demonstrates a much lower MSE of 0.0014, a SEE of 0.0107, and a closer to true parameter Coef of 0.9639. Similarly, for $\mu_X = 0$, the LR has an MSE of 0.8556, a SEE of 0.0016, and a Coef of 0.0750. Meanwhile, the FLR model shows better results with an MSE of 0.0015, a SEE of 0.0087, and a Coef of 0.9625. For $\mu_X = 5$, the LR has an MSE of 0.9060, a SEE of 0.0017, and a Coef of 0.0481. The FLR outperforms LR with an MSE of 0.0015, a SEE of 0.0114, and a Coef of 0.9636. Overall, the significantly lower MSE and a closer to the true parameter estimation of FLR underscore the importance of using FLR for fractional outcomes with high variability to ensure reliable and meaningful findings.

LR's inability to account for non-linearity and tendency to produce invalid predictions outside the range becomes more pronounced as data variability increases.

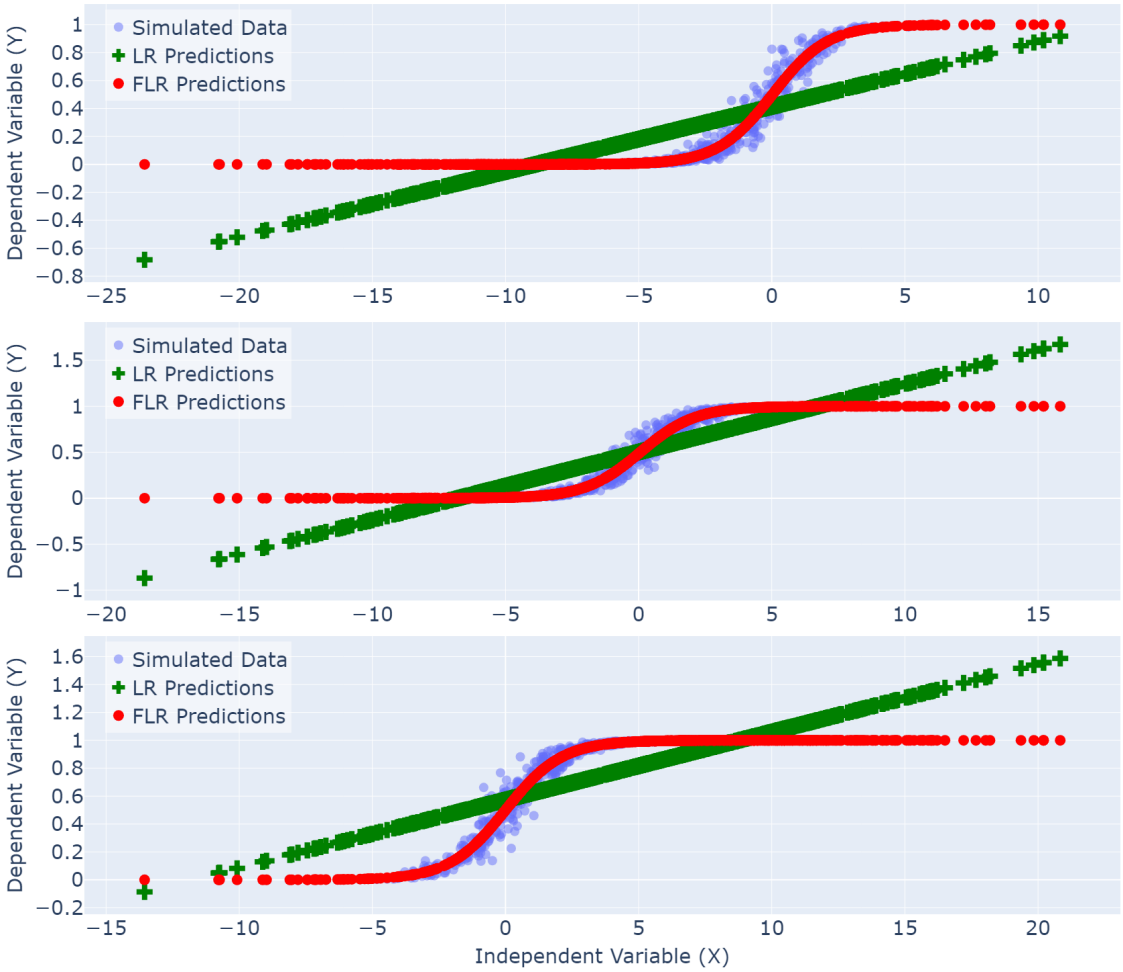


Figure 3: Model Predictions in High Variation Case (3.1)

Table 4: Monte Carlo Simulation Results for High Variation Case (3.1)

			LR			FLR		
n	σ_x^2	μ_x	MSE	SEE	Coef	MSE	SEE	Coef
1000	5	-5	0.9059	0.0018	0.0482	0.0014	0.0107	0.9639
1000	5	0	0.8556	0.0016	0.0750	0.0015	0.0087	0.9625
1000	5	5	0.9060	0.0017	0.0481	0.0015	0.0114	0.9636

3.1.4 Increasing Sample Sizes

Figure 4 shows the performance of FLR and LR models as the sample size increases. The values of $\mu_x = 0$ and $\sigma_x^2 = 2$ are selected to ensure the data cover the entire range of [0, 1]. This allows us to evaluate the models' performance around bounded values and within the unit interval as we increase the sample size. The plots, from top to bottom, represent sample sizes of 100, 500, 1000, and 5000. As the sample size increases, the difference in how well the models fit values around 0 and 1 becomes more noticeable, while the predictions within the unit interval do not show much difference. This demonstrates that FLR fits better than LR when the data are near the bounded values.

Table 5 shows the results of Monte Carlo simulations for increasing sample sizes. Overall, as the sample size increases, the FLR model consistently outperforms the LR model in terms of MSE and the average coefficient estimate. The Coef for the FLR model ranges from 0.9559 to 0.9579, whereas for the LR model, it ranges from 0.1487 to 0.1498, showing the FLR model provides estimates much closer to the true parameter across all sample sizes. The MSE remains consistently high for the LR model, ranging from 0.7229 to 0.7247, whereas the FLR model shows much lower MSE values, ranging from 0.0020 to 0.0029, indicating a better fit for the FLR model in all sample sizes.

It is important to note that increasing the sample size cannot mitigate the problem of biased estimates for LR.

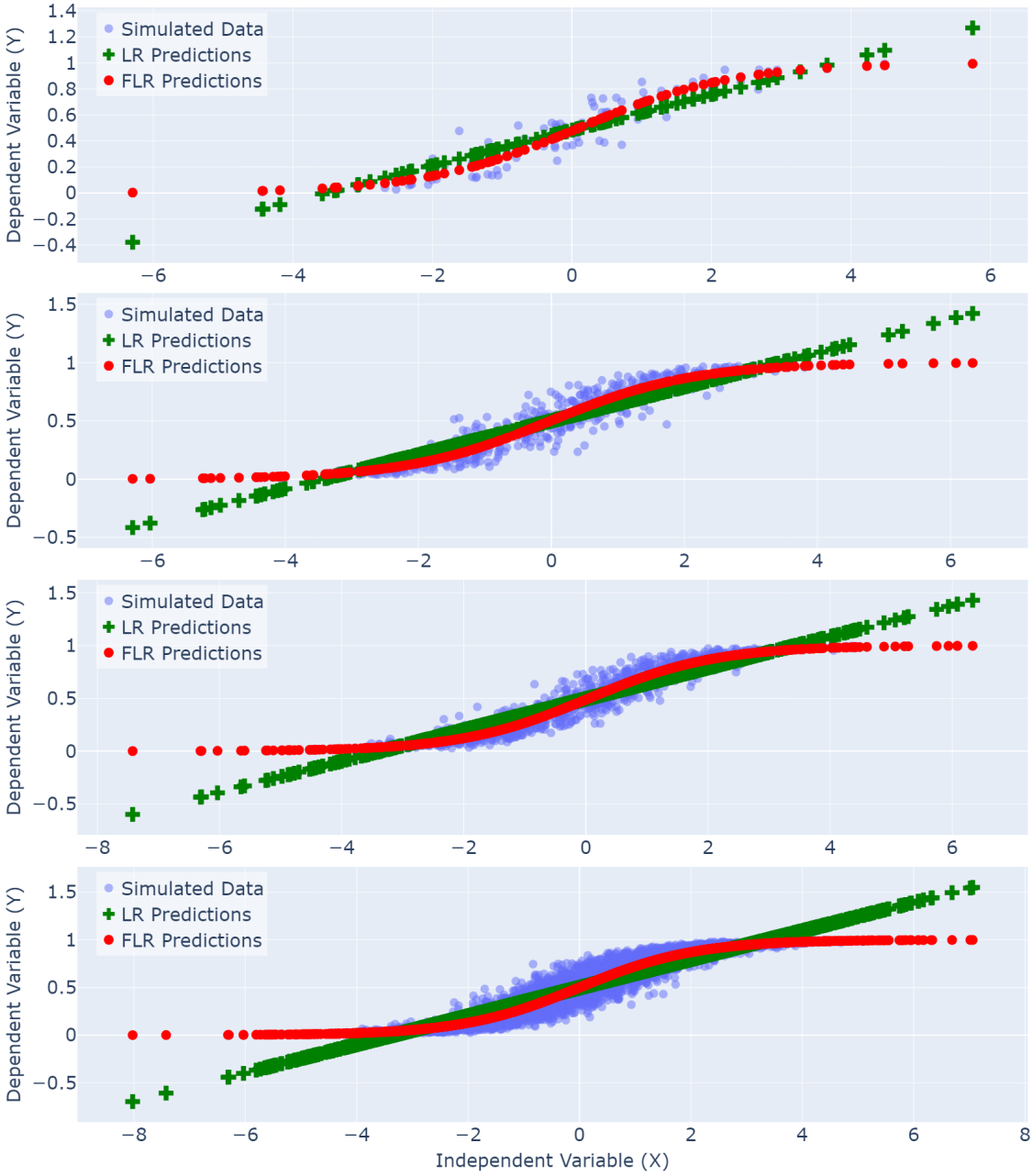


Figure 4: Model Predictions with Increasing Sample Sizes (3.1)

Table 5: Monte Carlo Simulation Results for Increasing Sample Sizes (3.1)

			LR			FLR		
n	σ_x^2	μ_x	MSE	SEE	Coef	MSE	SEE	Coef
100	2	0	0.7229	0.0078	0.1498	0.0029	0.0332	0.9579
500	2	0	0.7244	0.0036	0.1489	0.0022	0.0152	0.9559
1000	2	0	0.7246	0.0024	0.1488	0.0020	0.0104	0.9560
5000	2	0	0.7247	0.0011	0.1487	0.0020	0.0045	0.9559

3.2 Two Independent Variables

In this part of Monte Carlo simulations, two independent variables denoted as X_i and Z_i are considered to explore how FLR and LR handle the added complexity of multiple predictors. The variable X_i is constructed by drawing random values from a normal distribution, while Z_i is drawn from a uniform distribution [-2, 2]. The fractional dependent variable, Y_i, is generated using a logistic model that incorporates both X_i and Z_i. The logistic model for generating Y_i is defined by

$$Y_i = \frac{1}{1 + \exp\left(-(\epsilon_i + X_i + Z_i)\right)}, \tag{15}$$

where ϵ_i is a random noise drawn from a normal distribution $\mathcal{N}(0, 0.5)$. The mean and standard deviation of the normal distribution for X_i are systematically varied. Specifically, the mean varies through values -5, 0, and 5; the standard deviation varies through 0.1, 1, and 5. Three major scenarios are considered: low, medium, and high variation in X based on the value of σ_x^2. Within each major scenario, three minor cases are considered based on the value of μ_x: data primarily concentrated around 0, the middle, and around 1. The finite sample performance of both models is investigated by

varying the sample size through 100, 500, 1000, and 5000 to observe how the estimators perform from small to large datasets.

3.2.1 Low Variation

Figure 5 shows a 3D plot comparing the actual data, LR, and FLR predictions in case of low variation. The axes represent the values of the independent variables X, Z, and the dependent variable Y. The data points are in blue, linear predictions are in red, and FLR predictions are in green. The top plot illustrates the case where the fractional variable y_i is close to 0, the middle plot depicts the case where y_i falls within the unit interval, and the bottom plot represents when y_i is close to 1. According to Figure 5, the FLR predictions better capture the non-linearity in the actual data, mainly when y_i is around bounded values 0 and 1. For instance, in the top plot, the FLR model predicts higher values of Y relative to LR for high values of Z and X, which aligns more closely with the actual data. On the other hand, for low values of Z and X, the linear model's predictions exceed the boundary value, while the FLR model keeps its predictions within the boundary.

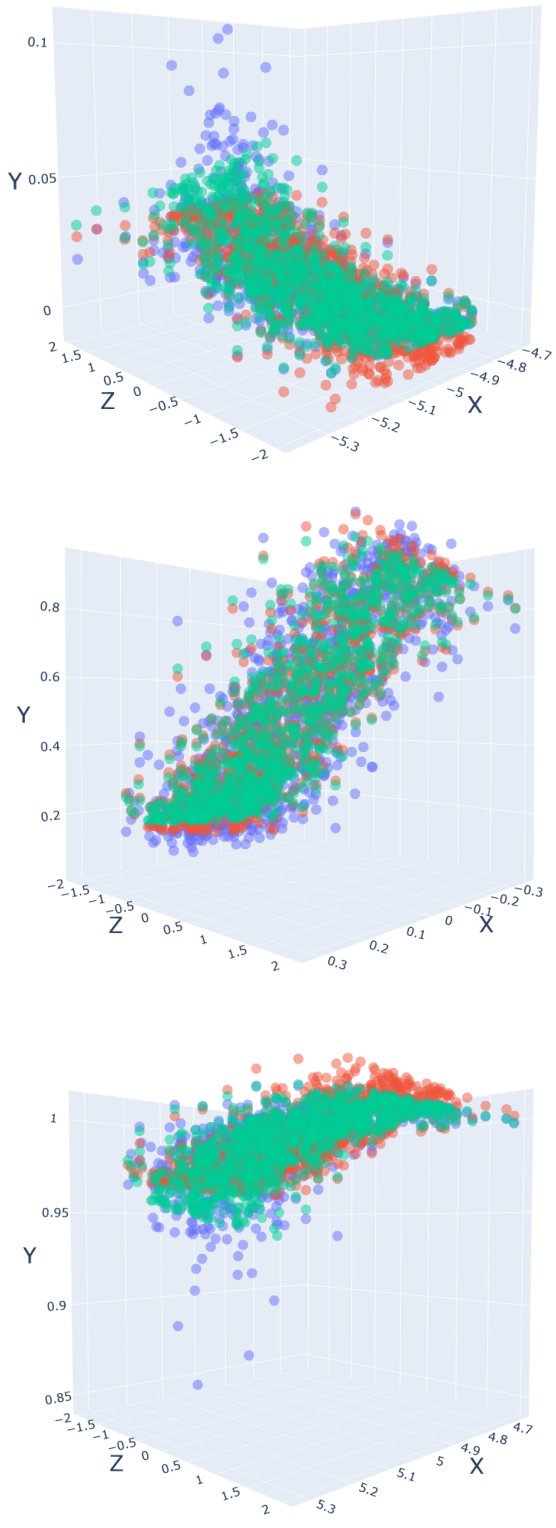


Figure 5: Model Predictions in Low Variation Case (3.2)

Table 6 shows the results of Monte Carlo simulations in case of low variation ($\sigma_x^2 = 0.1$) in the data. It is important to note that the targeted variable X results are only reported in the table. For $\mu = -5$, LR has an MSE of 0.9744, while FLR has a lower MSE of 0.0579, indicating a better model fit by FLR. The SEE for the FLR is 0.2385, which is higher than the LR's SEE of 0.0037. However, the FLR's Coef of 0.9674 is much closer to the true parameter value ($\beta = 1$) than the LR's Coef of 0.0127. For $\mu = 0$, the LR has an MSE of 0.6626, and the FLR performs better with a lower MSE of 0.0248. The FLR's SEE is 0.1521, higher than the LR's 0.0299, but the FLR's Coef of 0.9593 is closer to the true parameter than the LR's 0.1865. For $\mu = 5$, the FLR performs better with an MSE of 0.0589 compared to the LR's 0.9748. The FLR has a higher SEE of 0.2426, while the LR model's SEE is 0.0037. Despite the higher SEE, the FLR's coefficient estimate of 1.0035 is closer to the true parameter value than the LR's estimate of 0.0127. Overall, the results of Monte Carlo simulations in low variation case demonstrate that the FLR consistently outperforms the LR in terms of MSE and coefficient accuracy.

Table 6: Monte Carlo Simulation Results for Low Variation Case (3.2)

			LR			FLR		
n	σ_x^2	μ_x	MSE	SEE	Coef	MSE	SEE	Coef
1000	0.1	-5	0.9744	0.0037	0.0127	0.0579	0.2385	0.9674
1000	0.1	0	0.6626	0.0299	0.1865	0.0248	0.1521	0.9593
1000	0.1	5	0.9748	0.0037	0.0127	0.0589	0.2426	1.0035

3.2.2 Medium Variation

According to Figure 6, in the medium variation case, the FLR continues to outperform LR in capturing the non-linearity of the data. For example, in the top plot, LR cannot

produce predictions for a higher value of y_i while the FLR does. Moreover, the FLR predictions align more closely with the actual data and do not exceed the boundaries, while in all three plots, it can be seen that LR produces predictions outside the range of [0,1]. Table 7 provides the Monte Carlo simulation results in medium variation case ($\sigma_x^2 = 1$). When $\mu = -5$, LR shows an MSE of 0.9637, while FLR achieves a significantly lower MSE of 0.0014, suggesting a better model fit of FLR. The FLR's Coef of 0.9838 is much closer to the true parameter value than the LR model's estimate of 0.0183. The SEE is 0.0339 for FLR and 0.0017 for LR. When $\mu = 0$, the FLR continues outperforming the LR with a lower MSE of 0.0023 than the LR's 0.6913. The Coef of 0.9553 for the FLR is closer to the true parameter value than the LR's 0.1686. The SEE is 0.0168 for FLR and 0.0034 for LR. For $\mu = 5$, the FLR again demonstrates better performance, with a lower MSE of 0.0015 compared to the LR's 0.9641. The FLR's Coef of 0.9825 is much closer to the true parameter value than the LR's estimate of 0.0181. The SEE is 0.0346 for FLR and 0.0018 for LR. Overall, the results of Monte Carlo simulations in medium variation case highlight that the FLR consistently outperforms the LR in terms of MSE and coefficient accuracy.

Table 7: Monte Carlo Simulation Results for Medium Variation Case (3.2)

			LR			FLR		
n	σ_x^2	μ_x	MSE	SEE	Coef	MSE	SEE	Coef
1000	1	-5	0.9637	0.0017	0.0183	0.0014	0.0339	0.9838
1000	1	0	0.6913	0.0034	0.1686	0.0023	0.0168	0.9553
1000	1	5	0.9641	0.0018	0.0181	0.0015	0.0346	0.9825

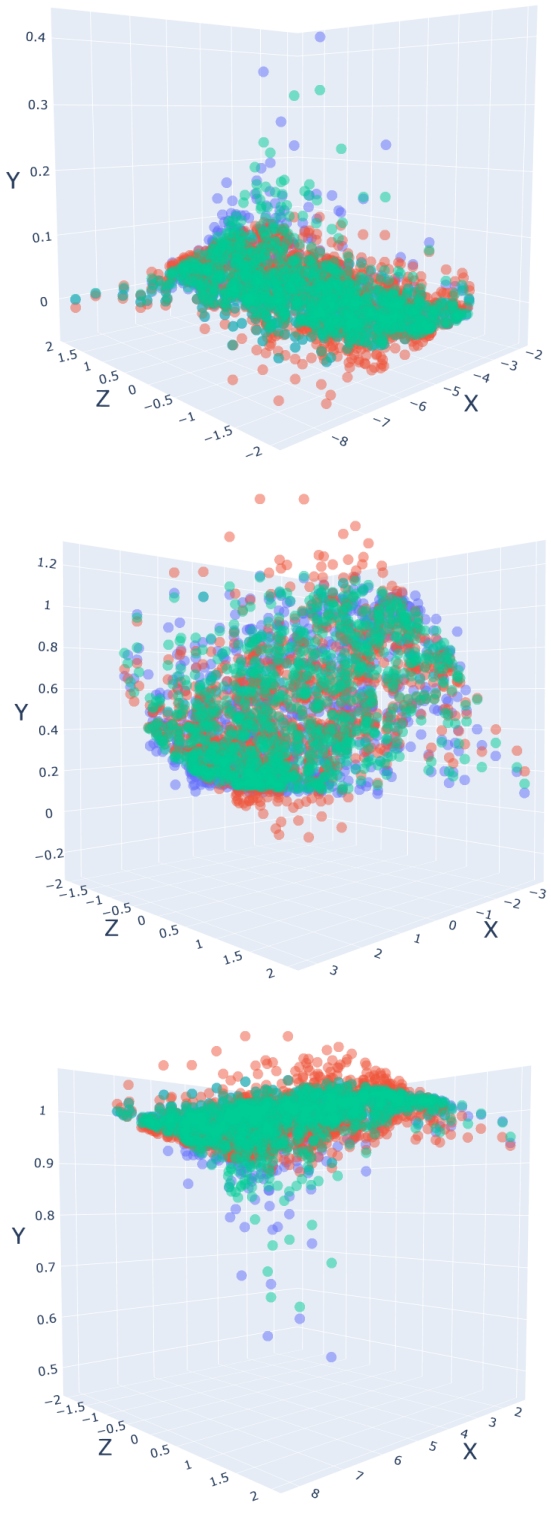


Figure 6: Model Predictions in Medium Variation Case (3.2)

3.2.3 High Variation

According to Figure 7, FLR continues to outperform LR in high variation case. LR struggles to account for the non-linear relationship in the data, leading to numerous predictions that fall outside the range of [0, 1], which can be seen in all three plots. In contrast, the FLR's predictions remain within the boundary. Moreover, it can be seen that in the top plot, LR cannot effectively produce predictions for the high value of y_i while FLR does. Similarly, in the bottom plot, the LR cannot effectively produce values for the low value of y_i while FLR performs better. The Table 8 provides the results of Monte Carlo simulations in high variation case ($\sigma_x^2 = 5$). When $\mu = 5$, the LR has an MSE of 0.9063, while the FLR has a significantly lower MSE of 0.0014, indicating a better model fit by the FLR. The FLR's Coef of 0.9640 is much closer to the true parameter value than the LR's Coef of 0.0480. The SEE is 0.0112 for FLR and 0.0017 for LR. When $\mu = 0$, the FLR continues to outperform the LR with an MSE of 0.0015, compared to the LR's 0.8591. The Coef of 0.9623 for the FLR is closer to the true parameter value than the LR's 0.0731. The SEE is 0/0095 for the FLR and 0.0016 for the LR. Finally, when $\mu = -5$, the FLR demonstrates better performance again with an MSE of 0.0014 compared to the LR's 0.9063, and the FLR's Coef of 0.9640 is much closer to the true parameter value than the LR's 0.0480. The SEE is 0.0095 for the FLR and 0.0016 for the LR. Overall, the results of Monte Carlo simulations in high variation case highlight that the FLR consistently outperforms the LR regarding MSE and coefficient accuracy.

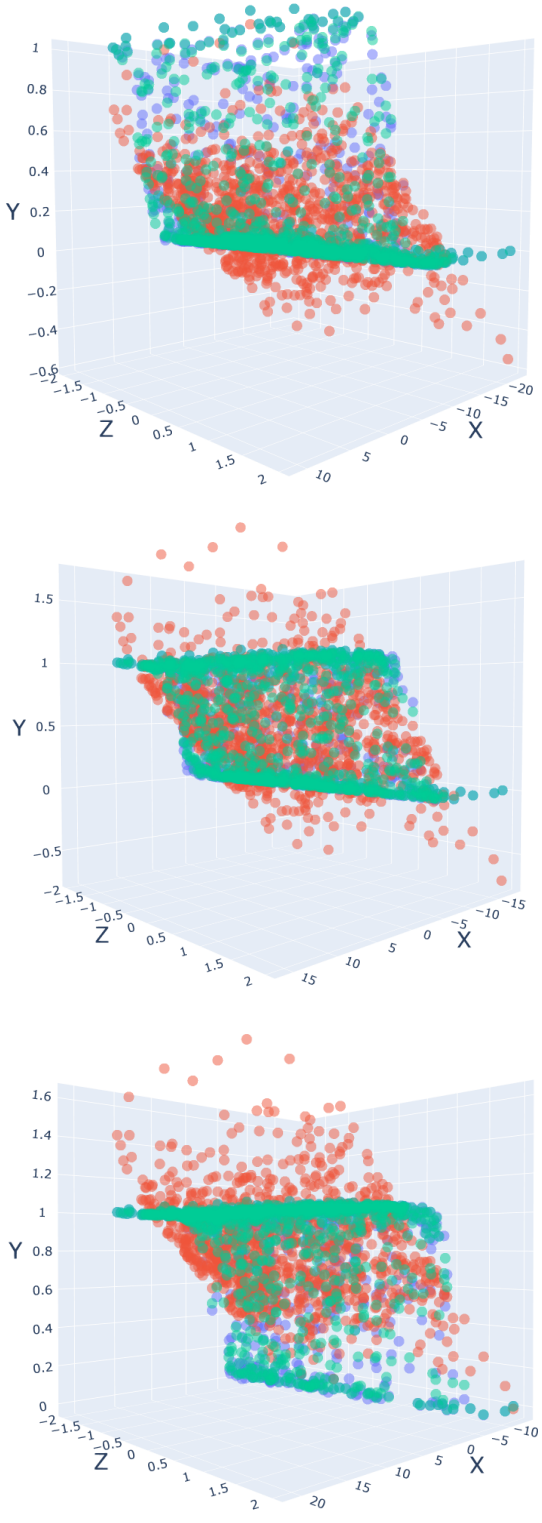


Figure 7: Model Predictions in High Variation Case (3.2)

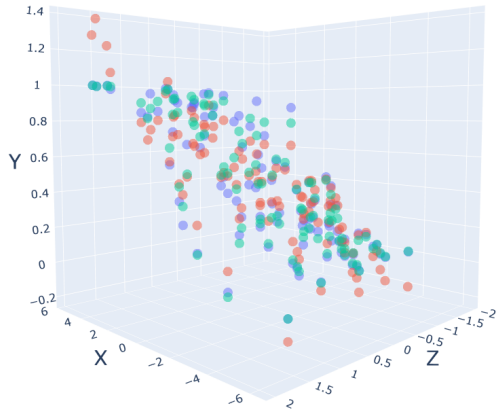
Table 8: Monte Carlo Simulation Results for High Variation Case (3.2)

			LR			FLR		
n	σ_x^2	μ_x	MSE	SEE	Coef	MSE	SEE	Coef
1000	5	-5	0.9063	0.0017	0.0480	0.0014	0.0112	0.9640
1000	5	0	0.8591	0.0016	0.0731	0.0015	0.0095	0.9623
1000	5	5	0.9064	0.0017	0.0479	0.0014	0.0116	0.9642

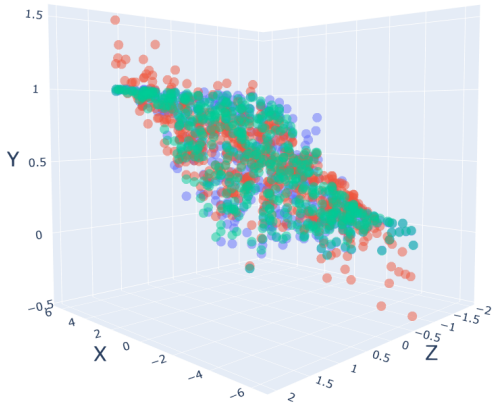
3.2.4 Increasing Sample Sizes

Figure 8 displays the performance of FLR and LR models in terms of prediction as the sample size increases. As shown in the figure, FLR predictions align more closely with the actual data than LR predictions, particularly near the boundaries of 0 and 1. This difference between the two models becomes more considerable by increasing the sample size. As the sample size increases, the non-linear behavior in the actual data becomes more noticeable. This non-linearity is captured by FLR but not by the LR, and the difference in capturing this non-linearity becomes more apparent as the sample size grows. This inability of LR to capture non-linearity ends up in predictions outside boundaries.

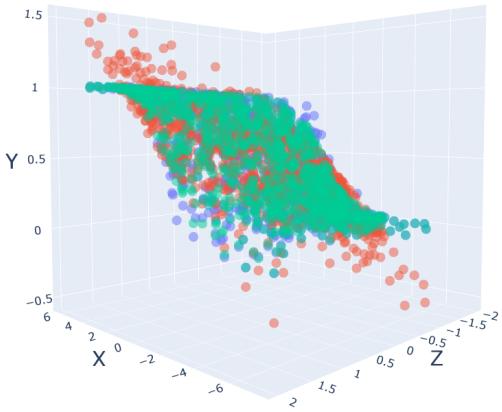
Table 9 presents the results of the Monte Carlo simulations as the sample size increases. We set $\mu_x = 0$ and $\sigma_x^2 = 2$ to ensure the fractional variable is spread across the entire unit interval [0, 1]. As the sample size increases, the SEE for both models decreases, indicating improved precision. For example, when the sample size grows from 100 to 5000, the SEE improves from 0.0083 to 0.0011 for LR and from 0.0353 to 0.0047 for FLR. The average coefficient estimates in both models remain relatively stable across different sample sizes. However, FLR consistently produces estimates closer to the true parameter than LR. For instance, at $n = 100$, the coefficient estimate for FLR



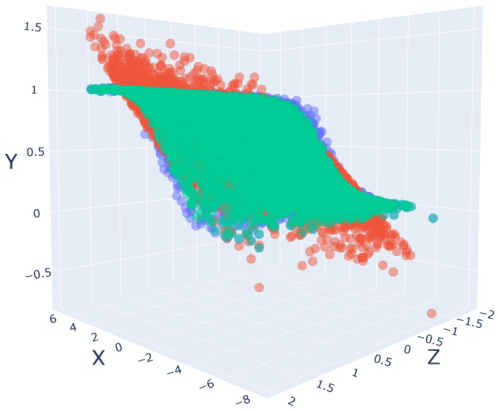
((a)) $N = 100$



((b)) $N = 500$



((c)) $N = 1000$



((d)) $N = 5000$

Figure 8: Model Predictions with Increasing Sample Size (3.2)

is 0.9593, while for LR, it is only 0.1361. FLR also consistently provides a smaller MSE across all sample sizes, indicating a better model fit. As the sample size increases, the MSE for FLR improves, going from 0.0029 to 0.0018, while the MSE for LR worsens slightly, increasing from 0.7464 to 0.7471. This demonstrates that FLR performs better as the sample size grows, while LR does not show the same improvement.

Table 9: Monte Carlo Simulation Results for Increasing Sample Size (3.2)

			LR			FLR		
n	σ^2	μ	MSE	SEE	Coef	MSE	SEE	Coef
100	2	0	0.7464	0.0083	0.1361	0.0029	0.0353	0.9593
500	2	0	0.7467	0.0036	0.1357	0.0021	0.0152	0.9564
1000	2	0	0.7470	0.0027	0.1357	0.0019	0.0106	0.9576
5000	2	0	0.7471	0.0011	0.1356	0.0018	0.0047	0.9576

3.3 Summary of Findings

The LR and FLR provide identical predictions in low variation cases even when the fractional variable is mostly scattered around bounded values zero or one. The difference in predictions of both models is negligible when the data has medium variation and is mainly scattered in the middle, so the number of data points valued at 0 and 1 are minimal. However, when the data with medium variation scatter mostly around bounded values, the difference becomes noticeable, and LR can no longer provide solid predictions. At the same time, FLR continues to have accurate predictions. Regarding high variation in data, it is observed that LR cannot provide accurate predictions even if most data points are scattered in the middle, and only a small proportion of data has values of 0 and 1. However, FLR can produce solid predictions. In addition, FLR

provides estimations close to the true parameter and low MSE in all examined conditions, while LR provides biased estimation and a large MSE. Importantly, increasing the sample size cannot mitigate the problem of a biased estimate in LR. Using LR in case of low variation can be reasonable even if the data points take values around 1 or 0, as it provides a more straightforward interpretation. However, in high or medium variation cases, especially when data is mostly scattered around 0 or 1, the FLR appears to be a better method.

4 Replication Study

In this section, the analysis from three papers published in the *Journal of Applied Econometrics* are replicated, each using fractional outcomes for their research. These papers cover various topics, highlighting the diverse applications of fractional outcomes in economic studies. This replication study applies both LR and FLR. Finally, comparing the results of these two models demonstrates how model selection can significantly impact the findings and interpretations of the analysis and provide a clear understanding of their performance in real-world applications.

4.1 Employee Participation Rates

4.1.1 Review of Original Paper

Papke & Wooldridge (1996) suggest FLR in modeling fractional outcomes. They use FLR to examine the relationship between match rate and employee participation rate in 401(k) pension plans as an empirical analysis. The dataset used in their analysis consists of 4,734 observations and includes variables such as the participation rate (prate), match rate (mrate), total employment (totemp), the average age of the plan (age), and whether the plan is the sole plan offered by the employer (sole). The average

participation rate across the dataset is 0.87, with a standard deviation of 0.17. Notably, over 40 percent of the plans in their research had 100 percent participation. The match rate has an average of 0.75 with a standard deviation of 0.84, a minimum of 0.01, and a maximum of 5. In addition, the total employment across plans has a mean of 4,621. The average plan's age is roughly 13 years, and close to 41 percent of the plans in the dataset are sole plans offered by the employer. Notably, the original paper uses a restricted sample of the data, focusing only on cases where the match rate (mrate) is less than or equal to one (mrate ≤ 1). To ensure consistency, the same restriction and independent variables as the original paper are used in this replication study to explain the participation rate.

Table 10: Summary Statistics of Variables (4.1)

	prate	**mrate**	**totemp**	**age**	**sole**
Count	4734	4734	4734	4734	4734
Mean	0.8696	0.7463	4621.07	13.14	0.4149
Std	0.1668	0.8444	16299.64	9.63	0.4928
Min	0.0232	0.0110	53.00	4.00	0.00
25%	0.7803	0.2701	278.00	7.00	0.00
50%	0.9367	0.4398	628.00	8.00	0.00
75%	1.0000	0.8359	2173.25	17.00	1.00
Max	1.0000	5.0000	44304.00	76.00	1.00

4.1.2 Replication Results

Figure 9 shows the actual data for participation rate and match rate in blue, the LR's predictions in green, and the FLR's predictions in red. The fractional variable *prate* mostly stays around 1, and the FLR keeps predictions within this range, while the

linear regression model predicted participation over 100 percent. The FLR also seems to capture the non-linear pattern, while LR cannot.

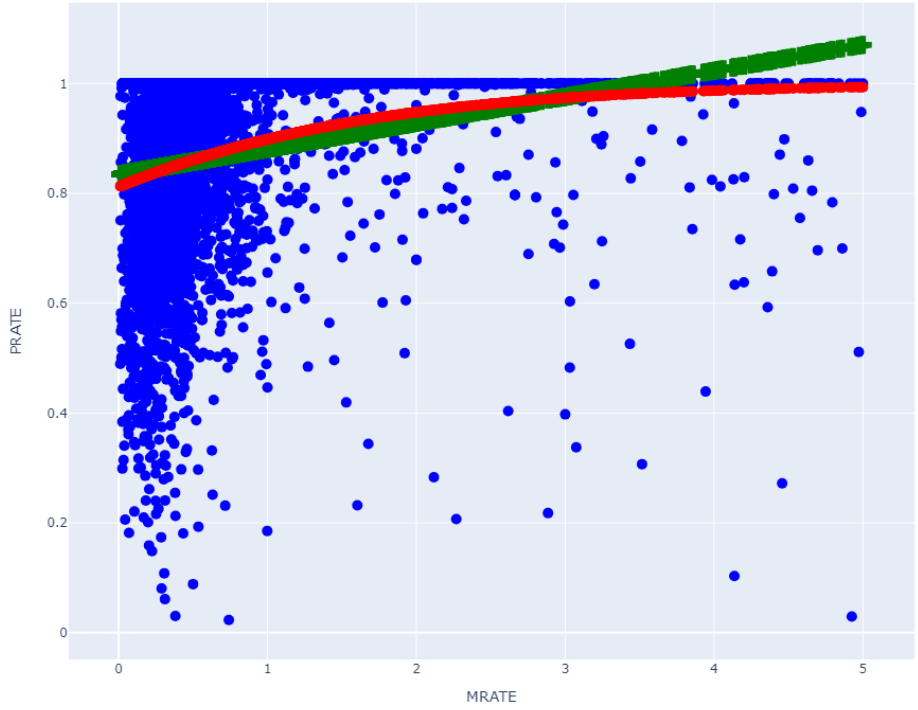


Figure 9: Participation Rate Predictions (4.1)

According to Table 11, the results from the FLR and LR show significant differences in both the magnitude of the coefficients and the associated standard errors. The FLR captures more complex relationships, particularly in variables like mrate and log(totemp). However, these gains come with increased variability, as indicated by the higher standard errors in the FLR estimates. For example, the LR's coefficient is 0.2394 for mrate variable, significantly lower than the 1.2180 estimated by the FLR model, suggesting that LR may underestimate the impact of match rates on participation. The FLR model captures a stronger relationship but with more significant uncertainty, as reflected by its higher standard error (0.780 compared to 0.043 in LR). Additionally, the squared term mrate^2 shows opposite signs in the two models, with LR estimating

a negative effect and FLR estimating a positive one. The variable log(totemp) also exhibits substantial differences between the models, where the LR estimate is -0.1117, while the FLR estimate is much larger in magnitude at -1.0021, suggesting that FLR captures a stronger negative effect of firm size on participation rates. Again, the FLR model shows a higher standard error (0.254 versus 0.014 in LR), which reflects more variability in its estimates. These patterns extend to other variables, such as age, where FLR estimates a stronger positive effect on participation rates than LR. The standard errors in FLR remain consistently larger, indicating that the flexibility of the logistic model introduces additional uncertainty.

The conclusion can be that while the FLR model provides a better fit for capturing complex, non-linear relationships in fractional data, it also comes with increased uncertainty, as shown by the higher standard errors. Although LR provides simpler and lower standard error results, it may need to capture these relationships more effectively. These findings emphasize the importance of choosing the appropriate model for the specific characteristics of the data. Simply adding quadratic terms to a linear model, as seen with the mrate variable, may not be sufficient to capture the actual relationships in the data, and a more flexible functional form, like FLR, may be necessary.

Table 11: Comparison of LR and FLR Results (4.1)

Variable	**LR**	**FLR**
mrate	0.2394	1.2180
	(0.043)	(0.780)
mrate^2	-0.0873	0.1960
	(0.043)	(0.850)
log(totemp)	-0.1117	-1.0021
	(0.014)	(0.254)
log(totemp)^2	0.0057	0.0522
	(0.001)	(0.016)
age	0.0059	0.0503
	(0.001)	(0.020)
age^2	-6.653e-05	-1.80e-05
	(2.33e-05)	(0.0001)
sole	0.0008	0.0006
	(0.010)	(0.107)
Observations	3784	3784

4.2 Investment-to-GDP Ratio

4.2.1 Review of Original Paper

Hacıoğlu Hoke & Kapetanios (2020) explore how the relationship between national savings and investment changes as countries become more open to trade. They utilize

the model expressed as

$$y_{it} = \beta_{01} x_{it} + \beta_{02} x_{it} g(q_{it} : \gamma, c) + e_{it}, \tag{16}$$

where y_{it} is the investment-to-GDP ratio in country i at time t, and x_{it} is the savings rate. The function $g(q_{it} : \gamma, c)$ called interaction term allows the link between savings and investment to change based on how open a country is, defined as

$$g(q_{it} : \gamma, c) = \left[1 + \exp(-\gamma(q_{it} - c))\right]^{-1}, \tag{17}$$

where q_{it} represents the level of openness, and γ and c are given parameters. The dataset of this paper includes 27 OECD countries from 1951 to 2006. Table 12 provides a summary statistics of variables. The saving ratio has an average of 24.36 percent, with a standard deviation of 8.07 percent, a minimum of -1.64 percent, and a maximum of 57.84 percent. The Investment-to-GDP ratio has an average of 24.36 percent, with a standard deviation of 6.01 percent, minimum of 5.97 percent, and maximum of 46.76 percent, meaning the Investment-to-GDP ratio does not take any bounded value. The openness variable has an average of 59.26 and a standard deviation of 39.99, indicating considerable differences in the level of openness across countries. It has a minimum value of 0 and a maximum value of 319.55.

In the original paper, although a nonlinear transition for the variable Openness is considered, the relation between dependent and independent variables is captured by LR. Since the dependent variable is the investment-to-GDP ratio, a fractional variable bounded between zero and one, FLR is deemed more appropriate.

Table 12: Summary Statistics of Variables (4.2)

Statistic	Saving ratio	Investment-to-GDP ratio	Openness
Count	1620	1620	1620
Mean	24.36	24.36	59.26
Std Dev	8.07	6.01	39.99
Min	-1.64	5.97	0.00
25%	19.15	20.27	32.78
50%	23.64	24.19	53.50
75%	29.07	28.03	71.66
Max	57.84	46.76	319.55

4.2.2 Replication Results

Figure 10 shows the actual data for the interaction term and investment-to-GDP ratio in blue, along with the LR and the FLR predictions, which are overlaid in green and red, respectively. In this case, both models produce nearly identical predictions, capturing the same linear trend in the data.

According to Table 13, the results from the LR and FLR show significant differences in both the magnitude of the coefficients and their interpretations. The LR model, with a constant term of 0.0998, suggests that the average investment-to-GDP ratio, when other variables are zero, is approximately 9.98 percent. On the other hand, the FLR model estimates a negative constant of -0.0468, indicating a lower base level for the investment-to-GDP ratio when considering the bounded nature of the data. The LR shows a positive coefficient of 0.0065 for the saving ratio with a very small standard error, indicating a clear positive relationship between savings and investment, aligning with traditional economic theories. However, in the FLR model, the coefficient

for the saving ratio turns negative (-0.0014). It comes with a larger standard error (0.006), suggesting that once the bounded nature of the investment-to-GDP ratio is accounted for, the positive relationship observed in LR may not hold. The relationship could even be negative. The interactive term, which captures the effect of openness on the savings-investment relationship, also shows contrasting results between the models. The LR model's coefficient is slightly negative (-0.0014), indicating that higher openness weakens the savings-investment link. In contrast, the FLR model shows a positive coefficient (0.0154), suggesting that openness could strengthen the relationship between savings and investment. However, the larger standard error in the FLR model indicates more uncertainty in this estimate.

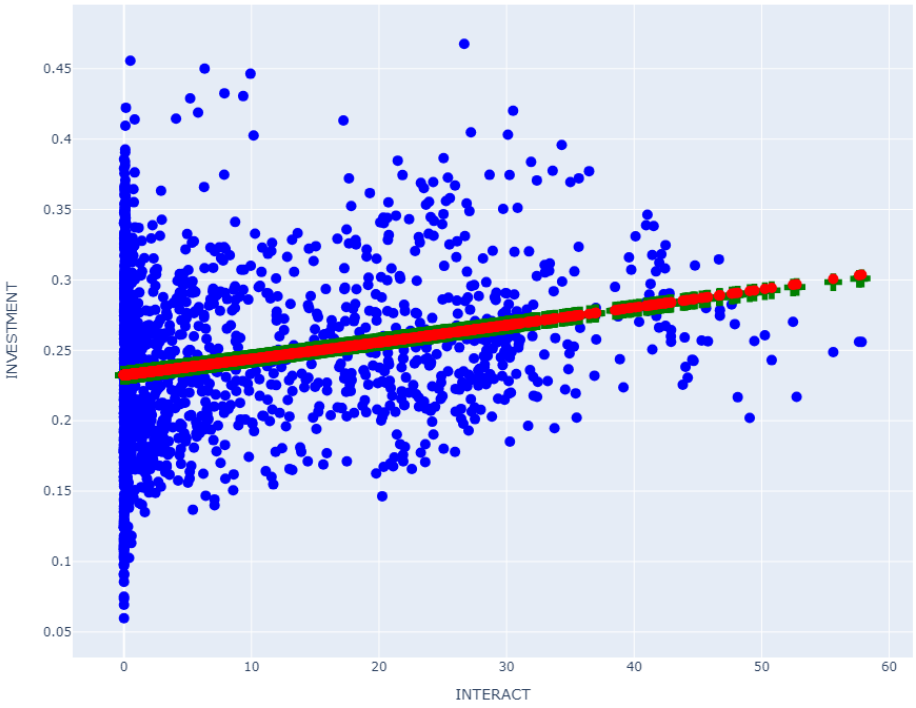


Figure 10: Investment-to-GDP Ratio Predictions (4.2)

Table 13: Comparison of LR and FLR Results (4.2)

Variable	**LR**	**FLR**
Constant	0.0998	-0.0468
	(0.004)	(0.003)
Saving ratio	0.0065	-0.0014
	(0.000)	(0.006)
Interactive term	-0.0014	0.0154
	(0.000)	(0.006)
Observations	1620	1620

In summary, while the LR provides a simpler and more direct interpretation, it may not fully capture the complexities of the data, particularly with respect to the bounded nature of the investment-to-GDP ratio. The FLR, despite of introducing more variability and uncertainty, offers a more detailed understanding of the relationship between savings, investment, and openness. These differences highlight the importance of selecting the appropriate model based on the characteristics of the data.

4.3 Return On Investment

4.3.1 Review of Original Paper

Gallizo, Gargallo, & Salvador (2008) examine how a company's sector and size affect financial ratios. We use the dataset from this paper to replicate the findings, focusing on Return on Investment (ROI) using both FLR and LR. Since ROI is bounded between -1 and 1, we transform it using the formula $(ROI + 1)/2$, allowing us to maintain its non-linear nature while bounding it between zero and one for the FLR model. The data used in this replication are drawn from the AMADEUS database, as in the original

study, and cover a balanced panel of European manufacturing firms from 1994 to 2003. The dataset includes 3,950 observations, with key variables such as ROI, SIZE, a binary variable indicating whether the firm is small/medium (coded as 0) or large (coded as 1), and SECTOR, a categorical variable identifying the sector in which the firm operates, with values 1 (Wood and Paper), 2 (Chemicals and Petroleum Products), and 3 (Minerals and Machinery). Table 14 presents the summary statistics of variables. The ROI variable has a mean of 0.0699 and a standard deviation of 0.0920, indicating a relatively low average return with moderate variability. The SIZE variable is evenly distributed between small/medium and large firms, while SECTOR shows a higher concentration in sectors 2 (Chemicals and Petroleum Products) and 3 (Minerals and Machinery).

Table 14: Summary Statistics of Variables (4.3)

	SIZE	**SECTOR**	**ROI**
Count	3950	3950	3950
Mean	0.5063	2.2785	0.0699
Std	0.5000	0.8073	0.0920
Min	0.00	1.00	-0.5883
25%	0.00	2.00	0.0235
50%	1.00	2.00	0.0666
75%	1.00	3.00	0.1171
Max	1.00	3.00	0.9003

4.3.2 Replication Results

Figure 11 shows the actual data for the ROI and variable sector in blue, along with the predictions from both the LR and FLR, which are overlaid in green and red,

respectively. In this case, both models produce nearly identical predictions.

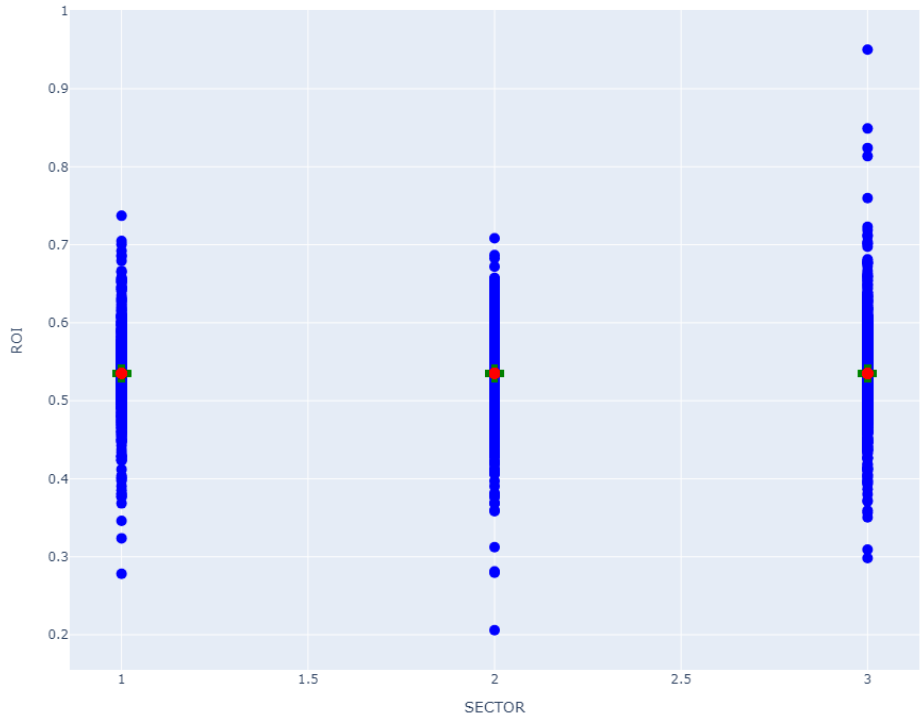


Figure 11: Return On Investment Predictions (4.3)

According to Table 15, the results from the LR and FLR show noticeable differences in both the magnitude of the coefficients and their interpretations. The LR, with a constant term of 0.5299, suggests that the average dependent variable, when other predictors are zero, is approximately 52.99 percent. On the other hand, the FLR estimates a lower constant of 0.1197, indicating a smaller base level when accounting for the characteristics of the data. For the SIZE variable, the LR shows a positive coefficient of 0.0109 with a very small standard error, indicating a significant positive relationship. This is in line with expectations from traditional linear modeling approaches. However, the FLR provides a much larger coefficient of 0.0437, though with a higher standard error (0.064), suggesting that the relationship might be stronger, but also more uncertain, when considering the specific nature of the data. The SECTOR

variable shows a near-zero coefficient in both models, with the LR estimating a coefficient of -0.0002 and the FLR showing a coefficient of -0.0008. The small magnitude and the lack of statistical significance indicate that SECTOR might not have a strong influence on the dependent variable in this context.

Table 15: Comparison of LR and FLR Results (4.3)

Variable	**LR**	**FLR**
Constant	0.5299	0.1197
	(0.002)	(0.100)
SIZE	0.0109	0.0437
	(0.001)	(0.064)
SECTOR	-0.0002	-0.0008
	(0.001)	(0.040)
Observations	3950	3950

In summary, while the LR provides a simpler and more direct interpretation, the FLR offers a different perspective with potentially stronger but more uncertain relationships. These differences underscore the importance of model selection in understanding the complexities of the data, as each model provides unique insights into the underlying relationships between variables.

5 Empirical Application: Impact of Education on Poverty rate

5.1 Background

Education is widely seen as a way to reduce poverty. It gives people the required skills and knowledge to improve their economic situation. Many studies have shown that people with more education are less likely to be poor. Majumder & Biswas (2017) find that in Bangladesh, households lead by people with secondary education or higher have a much lower chance of being poor. Chaudhary et al. (2010) find that although primary and middle school education for household heads have some benefits, they are not as strong as higher education. Zuluaga (2006) finds that education improves incomes and living conditions in Colombia, especially for poorer families. Education does not just raise income; it also significantly improves health and housing conditions. Other studies, such as Qureshi & Arif (2001), also support that education significantly reduces poverty beyond just earning more money but also by impacting health conditions. Appleton (1997) estimates that each additional year of primary schooling reduces the risk of poverty by 2.5 percent, with even greater benefits at the secondary level. Haughton & Khandker (2009), Sackey (2005), and Tilak (2005) emphasize that higher education leads to long-term economic growth and poverty reduction. In this empirical application, we provide more evidence on how education impacts poverty using FLR and compare the results with LR to see how different approaches might lead to different conclusions.

5.2 Data Summary and Methodology

To explore the relationship between education and poverty rates, both FLR and LR are used. In FLR, the dependent variable is the poverty rate for each country, while the independent variables are the level of education and real GDP per capita. The constant in the model is represented by β_0, and β_1 and β_2 represents the coefficients of the independent variables, with ϵ_i capturing the error term. Same variables and notation are also used in LR. The poverty rates come from the World Bank and show the percentage of the population living on less than $2.15 per day. Educational attainment refers to the percentage of the population aged 15-24 that has completed secondary education, and this dataset is sourced from UNESCO's Education Attainment Database. Economic performance is measured by real GDP per capita (adjusted for inflation), which is obtained from the Penn World Table, measuring a country's economic output per person. The equations for the models are as follows

$$\mathbb{E}(poverty_i \mid edu_i, rgdp_i) = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 \cdot edu_i + \beta_2 \cdot rgdp_i)\right)}, \tag{18}$$

$$poverty_i = \beta_0 + \beta_1 \cdot edu_i + \beta_2 \cdot rgdp_i + \epsilon_i. \tag{19}$$

Table 16 provides a summary of the data, which includes 73 countries and covers poverty rates, education levels, and real GDP per capita. On average, the poverty rate is 5.45 percent, with most countries having a rate below 4.6 percent. However, some countries, such as Benin, experience significantly higher rates, reaching 60.80 percent. At the other extreme, countries like Austria have a poverty rate of 0. Education levels also show significant variation. On average, 34.85 percent of the population has completed secondary education, but this number varies widely, with Niger at only 3.53 percent, while Armenia has 65.04 percent completion. Real GDP per capita also varies

considerably, with an average of $25,647.99 and a standard deviation of $20,266.20. For instance, Mozambique has a low GDP per capita of $1,786.91, while wealthier countries like Luxembourg reach $105,998.20.

Table 16: Summary Statistics of Variables

	Poverty Rate (%)	**Education (%)**	**Real GDP per capita (US$)**
Count	73	73	73
Mean	5.45	34.85	25,647.99
Std. Dev.	12.16	13.02	20,266.20
Min	0.00	3.53	1,786.91
25%	0.10	25.26	10,879.27
50%	0.70	34.56	23,051.01
75%	4.60	44.56	36,712.50
Max	60.80	65.04	105,998.20

5.3 Regression Results

Figure 12 compares the predictions of LR and FLR with the actual data. The blue dots represent the actual data, the green dots show the FLR's predictions, and the red dots represent the LR's predictions. As can be seen, the LR struggles to capture the pattern in the data, especially for higher poverty rates, where its predictions fall short. On the other hand, the FLR performs better, especially in predicting higher poverty rates, showing a closer fit to the actual data.

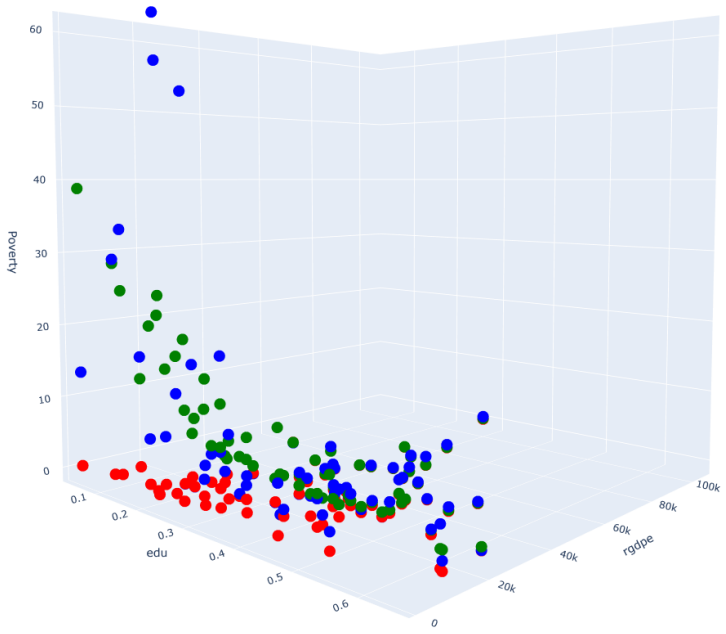


Figure 12: Model Prediction Comparison

Table 17 compares the LR and FLR for analyzing the impact of education and real GDP per capita on the poverty rate. The models estimate the relationship between these variables differently. In the LR, the constant term is estimated at 0.2261, a result that is statistically significant at the 1% significance level. This indicates that even when education and real GDP per capita are held constant, the poverty rate is positively affected by other factors. In contrast, the FLR model provides a constant estimate of -0.0726, with a larger standard error of 1.248. This suggests the FLR model does not provide strong evidence for an intercept when controlling for education and GDP. When examining the impact of education, the LR model estimates the coefficient at -0.3256, with a relatively small standard error of 0.094. This suggests that higher levels of education are associated with decreased poverty rates. The FLR model also finds a negative relationship between education and poverty, but with a much larger coefficient of -3.3321 and an extremely large standard error of 4.640. While the direction

of the effect remains consistent, the FLR results imply a much stronger relationship between education and poverty but with considerable uncertainty due to the high standard error. For the real GDP per capita variable, the OLS model estimates a coefficient of -2.264e-06, which is statistically significant at the 1% level, indicating a very small but significant negative effect of real GDP per capita on poverty. The FLR model, however, estimates the coefficient at -0.0001, with a standard error of 0.000. Though this result suggests a similarly small real GDP per capita effect on poverty, it is not significant, which contrasts with the LR findings.

Table 17: Comparison of LR and FLR Results

Variable	**LR**	**FLR**
Constant	0.2261***	-0.0726
	(0.036)	(1.248)
Education	-0.3256***	-3.3321
	(0.094)	(4.640)
Real GDP per Capita	-2.264e-06***	-0.0001
	(6.05e-07)	(0.000)
Observations	73	73

While both models aim to explain the relationship between poverty, education, and real GDP per capita, the results are quite different. The LR model offers more precise estimates with smaller standard errors, particularly for the education and real GDP variables, suggesting a well-defined linear relationship. However, it may fail to capture more complex non-linear effects. On the other hand, the FLR attempts to account for non-linearities in the data. However, this comes with the cost of much higher variability in the estimates, especially for education, making the results less precise.

6 Conclusion

FLR is a valuable method for analyzing dependent variables bounded between 0 and 1 without adjusting the data, as it estimates the conditional expectation directly. The Monte Carlo simulations in this study support that FLR provides better fit and predictions than LR, especially when there is high variation in the dependent variable or medium variation with data points clustered around bounded values 0 or 1. However, this better model fit for FLR is achieved with the cost of more uncertainty in the estimation relative to LR. The replication study shows the versatility of FLR in economic research and the differences between FLR and LR, which depend on the characteristics of the data. For instance, when analyzing employee participation rates, where many observations were equal to 1, the FLR model performs quite differently than LR. However, in cases like the investment-to-GDP ratio, where values of 0 and 1 are minimal, the difference in predictions and estimates between the two models is much smaller. Finally, this paper's empirical analysis shows how FLR effectively addresses the relationship between poverty rates and education levels. It captures the data's nonlinearity and provides reasonable predictions for high poverty rates, which the LR struggles to explain. While the FLR shows a higher magnitude in the coefficient for education, it also comes with greater uncertainty, whereas the LR offered more precise and statistically significant estimates. In conclusion, FLR is particularly well-suited when the fractional dependent variable significantly varies or when the data contain a medium variation and a substantial number of 0 and 1 values.

7 References

Appleton, S. (1997). Education and Poverty in Uganda. *World Bank Economic Review*, 11 (2), 1-16.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Chaudhary, R. A., Malik, S., & Hassan, A. (2010). Education and Poverty: Evidence from Pakistan. *Journal of Economic Development*, 35 (1), 129-144.

Cox, D. R. (1995). *The Analysis of Binary Data*. Chapman & Hall/CRC.

Elsas, R., & Florysiak, D. (2013). Heteroscedasticity in Financial Data: An Overview. *Financial Econometrics Series*, 7 (4), 57-68.

Fang, H., & Ma, L. (2012). Insurance Coverage in China: A Fractional Logistic Regression Approach. *China Economic Review*, 23 (3), 782-795.

Gallizo, J. L., Gargallo, P., & Salvador, M. (2008). Multivariate Partial Adjustment of Financial Ratios: A Bayesian Hierarchical Approach. *Journal of Applied Econometrics*, 23 (2), 205-227.

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo Maximum Likelihood Methods: Applications to Poisson Models. *Econometrica*, 52 (3), 701-720.

Hacıoğlu Hoke, S., & Kapetanios, G. (2020). Estimating the Relationship between National Savings and Investment for Open Economies. *Journal of Applied Econometrics*, 35 (3), 123-147.

Haughton, J., & Khandker, S. R. (2009). *Handbook on Poverty and Inequality*. World Bank.

Kieschhnick, R., & McCullough, B. D. (2003). Regression Analysis of Proportions in Finance. *Journal of Financial and Quantitative Analysis*, 38 (5), 797-824.

Maddala, G. S. (1991). *Introduction to Econometrics*. Macmillan.

Majumder, R., & Biswas, D. (2017). Education and Poverty: A Household-Level Analysis in Bangladesh. *Journal of Development Studies*, 53 (12), 2057-2073.

Martins, D. (2018). Modeling Efficiency Scores in Portuguese Banks: A Comparison of FLR and Tobit. *Finance Research Letters*, 25, 146-153.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models.* Chapman & Hall/CRC.

Molowny-Horas, R., Basnou, C., & Pino, J. (2017). Land Use and Cover Dynamics in Mediterranean Landscapes: FLR Model Approach. *Ecological Modelling*, 354, 71-82.

Papke, L. E., & Wooldridge, J. M. (1996). Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates. *Journal of Applied Econometrics*, 11 (6), 619-632.

Papke, L. E., & Wooldridge, J. M. (2008). Estimation of Fractional Response Models with an Application to 401(k) Plan Participation. *Journal of Applied Econometrics*, 23 (3), 605-627.

Qureshi, S. K., & Arif, G. M. (2001). Education and Poverty in Pakistan. *Pakistan Development Review*, 40 (4), 781-804.

Ramalho, J. S., & Ramalho, E. A. (2011). Alternative Models for Fractional Regression. *Econometrics Journal*, 14 (2), 231-253.

Sackey, H. A. (2005). Education and Economic Development in Africa. *African Development Review*, 17 (3), 567-593.

Tilak, J. B. G. (2005). Education and its Relation to Economic Growth, Poverty, and Income Distribution. *World Bank Research Observer*, 20 (2), 267-290.

Villadsen, F., & Wulff, T. (2021). Reproducibility in Strategy and Management Research: FLR Applications. *Journal of Strategy and Management*, 14 (1), 12-29.

Wu, G.-C., Baleanu, D., & Luo, A. C. (2017). *Fractional Calculus and Fractional Dynamics*. World Scientific.

Zuluaga, B. (2006). The Impact of Education on Poverty in Colombia. *Journal of International Development*, 18 (3), 479-491.