econometrics

MDPI

Communication

Fractional Probit with Cross-Sectional Volatility: Bridging Heteroskedastic Probit and Fractional Response Models

Songsak Sriboonchitta [1], Aree Wiboonpongse [2,*], Jittaporn Sriboonjit [3] and Woraphon Yamaka [1,*]

[1] Center of Excellence in Econometric, Faculty of Economics, Chiang Mai University, Chiang Mai 50200, Thailand; songsakecon@gmail.com
[2] Faculty of Agriculture, Chiang Mai University, Chiang Mai 50200, Thailand
[3] Faculty of Commerce and Accountancy, Thammasat University, Bangkok 10200, Thailand; jittaporn@tbs.tu.ac.th
* Correspondence: areewiboonpongse@gmail.com (A.W.); woraphon.econ@gmail.com or woraphon.yamaka@cmu.ac.th (W.Y.)

Abstract

This paper introduces a new econometric framework for modeling fractional outcomes bounded between zero and one. We propose the Fractional Probit with Cross-Sectional Volatility (FPCV), which specifies the conditional mean through a probit link and allows the conditional variance to depend on observable heterogeneity. The model extends heteroskedastic probit methods to fractional responses and unifies them with existing approaches for proportions. Monte Carlo simulations demonstrate that the FPCV estimator achieves lower bias, more reliable inference, and superior predictive accuracy compared with standard alternatives. The framework is particularly suited to empirical settings where fractional outcomes display systematic variability across units, such as participation rates, market shares, health indices, financial ratios, and vote shares. By modeling both mean and variance, FPCV provides interpretable measures of volatility and offers a robust tool for empirical analysis and policy evaluation.

Keywords: fractional response models; heteroskedastic; probit; cross-sectional volatility; limited dependent variables; variance modeling

check for updates

Academic Editor: Luis Alberiko Gil-Alana

Received: 18 September 2025
Revised: 24 October 2025
Accepted: 28 October 2025
Published: 3 November 2025

Citation: Sriboonchitta, S., Wiboonpongse, A., Sriboonjit, J., & Yamaka, W. (2025). Fractional Probit with Cross-Sectional Volatility: Bridging Heteroskedastic Probit and Fractional Response Models. *Econometrics*, *13*(4), 43. https://doi.org/10.3390/econometrics13040043

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fractional dependent variables, defined on the unit interval, arise in a wide range of economic applications. Examples include participation rates in labor markets, market shares in industrial organizations, bounded indices in health economics, financial ratios, and vote shares in political economies. Standard linear regression models are inappropriate in these contexts because they may predict values outside the admissible range and fail to account for heteroskedasticity.

To address these issues, Papke and Wooldridge (1996) introduced the fractional logit and probit models, providing quasi-maximum likelihood estimators that are robust and widely used for modeling proportions and rates. However, the conditional variance in these models is constrained to the Bernoulli form, which can be too restrictive when the dispersion of outcomes varies across observational units. Ferrari and Cribari-Neto (2004) proposed beta regression, which parameterizes the variance indirectly through a precision parameter, allowing greater flexibility in modeling bounded data. While this approach accommodates varying dispersion, its indirect variance specification limits interpretability, particularly when heterogeneity arises from observable characteristics.

Subsequent research has expanded these frameworks in several directions. Ramalho et al. (2011) and Baum (2006) provided comprehensive discussions of estimation and testing strategies for fractional regressions. Smithson and Verkuilen (2006) extended beta regression to incorporate mean–variance relationships for bounded psychological and behavioral outcomes, while Kieschnick and McCullough (2003) explored nonlinear fractional models in applied finance. More recent studies have improved inference and robustness: Guedes et al. (2021) developed Bartlett-corrected tests for varying-precision beta regressions; Ribeiro and Ferrari (2023) proposed robust estimation via maximum-L_q likelihood; Rahmashari and Srisodaphol (2025) introduced outlier detection methods for improving robustness; and Mai (2025) addressed bounded responses in high dimensions using Bayesian shrinkage priors. In addition, Stasinopoulos et al. (2017) and Rigby et al. (2019) advanced generalized additive models for location, scale, and shape (GAMLSS), offering flexible semiparametric tools for modeling both the mean and variance of bounded outcomes. Despite these advances, most existing fractional response models either impose homoskedasticity or employ indirect variance parameterizations. Yet in many empirical contexts, the variance of fractional outcomes differs systematically across cross-sectional units, reflecting observable heterogeneity. In parallel, the heteroskedastic probit model introduced by Harvey (1976) and Davidson and MacKinnon (1984) established a tractable variance-function approach for binary data, allowing the error variance to depend explicitly on covariates.

This paper develops a Fractional Probit with Cross–Sectional Volatility (FPCV) to fill this gap. The model specifies the conditional mean via a probit link and introduces a multiplicative variance function that scales dispersion by observable characteristics. This structure directly unifies the variance-function tradition of heteroskedastic probits with the bounded-outcome framework of fractional and beta regressions.

Our contributions are threefold. First, we extend the heteroskedastic probit beyond binary outcomes to fractional responses. Second, we propose a tractable and interpretable variance specification that links volatility directly to observable heterogeneity. Third, through Monte Carlo simulations, we show that the FPCV estimator reduces bias, improves coverage, and achieves superior predictive accuracy relative to standard alternatives. Collectively, these contributions fill a long-standing gap in the econometric analysis of bounded outcomes with heteroskedasticity.

The remainder of this paper is structured as follows. Section 2 presents the model, likelihood function, estimation and asymptotic properties. Section 3 reports the Monte Carlo results. Section 4 provides an application study.Section 5 concludes this paper.

2. Materials and Methods

2.1. Model Specification

Let $\{y_i, x_i, z_i\}_{i=1}^n$ denote a random sample, where $y_i \in (0,1)$ is a fractional response, x_i is a $k \times 1$ vector of covariates entering the conditional mean, and z_i is a positive scalar function of an $m \times 1$ vector of covariates determining the conditional variance.

The conditional mean is specified by the probit link

$$\mu_i = \mathbb{E}[y_i \mid x_i, z_i] = \Phi(x_i'\beta), \tag{1}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and β is a finite-dimensional parameter vector.

The observed response is represented as

$$y_i = \mu_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \tag{2}$$

with a multiplicative heteroskedastic variance structure,

$$\sigma_i^2 = \sigma^2 z_i^2, \qquad \sigma > 0,\ z_i^2 > 0, \tag{3}$$

where σ is a positive scale parameter. In practice, z_i can be parameterized as $z_i = \exp(w_i'\gamma)$, where w_i is an $m \times 1$ vector of observable covariates.

Equation (1) ensures that the conditional mean lies strictly within the $(0,1)$ interval, while (3) allows the conditional variance to depend systematically on observable heterogeneity. Although the normality assumption in (2) implies unbounded support, it serves here as a working quasi-likelihood rather than a literal distribution for y_i. The proposed model therefore follows the quasi–maximum likelihood framework(Papke and Wooldridge (1996), Huber (1967), and White (1982)), where consistency and asymptotic normality hold under the correct specification of the conditional mean even if the true distribution of y_i is non-normal. This approach extends the heteroskedastic probit model of Harvey (1976) to the case of fractional responses by incorporating cross-sectional heteroskedasticity directly into the variance function.

2.2. Likelihood Function

Given (1)–(3), the conditional distribution of y_i is

$$y_i \mid x_i, z_i \sim \mathcal{N}\Big(\Phi(x_i'\beta), \sigma^2 z_i^2\Big). \tag{4}$$

Hence the conditional density is

$$f(y_i \mid x_i, z_i; \beta, \sigma) = \frac{1}{\sigma z_i}\,\varphi\!\left(\frac{y_i - \Phi(x_i'\beta)}{\sigma z_i}\right), \tag{5}$$

where $\varphi(\cdot)$ denotes the standard normal p.d.f.

We define the standardized residual as

$$u_i \equiv \frac{y_i - \Phi(x_i'\beta)}{\sigma z_i}. \tag{6}$$

Using (6), the sample log-likelihood is

$$\ell(\beta, \sigma) = \sum_{i=1}^{n}\Big\{-\ln(\sigma z_i) - \tfrac{1}{2}u_i^2 - \tfrac{1}{2}\ln(2\pi)\Big\}. \tag{7}$$

Equation (7) is maximized with respect to (β, σ) to obtain the QMLE of the model.

Score functions. Let $\eta_i = x_i'\beta$ and note that $\partial\Phi(\eta_i)/\partial\eta_i = \varphi(\eta_i)$. The derivatives of the standardized residual are

$$\frac{\partial u_i}{\partial \beta} = -\frac{\varphi(\eta_i)}{\sigma z_i}x_i, \qquad \frac{\partial u_i}{\partial \sigma} = -\frac{u_i}{\sigma}. \tag{8}$$

Differentiating (7), the score with respect to β is

$$\frac{\partial \ell(\beta,\sigma)}{\partial \beta} = \sum_{i=1}^{n}\frac{u_i\,\varphi(\eta_i)}{\sigma z_i}x_i, \tag{9}$$

and the score with respect to σ is

$$\frac{\partial \ell(\beta,\sigma)}{\partial \sigma} = \sum_{i=1}^{n}\frac{u_i^2 - 1}{\sigma}. \tag{10}$$

Variance parameterization. If the variance scaling factor is parameterized as $z_i = \exp(w_i'\gamma)$, with w_i known covariates and γ unknown, then

$$\frac{\partial u_i}{\partial \gamma} = -u_i w_i, \qquad \frac{\partial \ell(\beta, \sigma, \gamma)}{\partial \gamma} = \sum_{i=1}^{n} (u_i^2 - 1) w_i. \tag{11}$$

This nests Harvey-type log–variance functions in the fractional setting.

To ensure model identifiability, we follow standard practice in heteroskedastic probit models (Harvey, 1976) by excluding an intercept term from w_i. This normalization prevents confounding between the global scale parameter σ and the intercept component of γ, ensuring that σ uniquely captures the overall scale while γ governs relative cross-sectional heteroskedasticity.

Information matrix. Under correct specification, $u_i \sim \mathcal{N}(0,1)$ is conditional on (x_i, z_i), so $\mathbb{E}[u_i] = 0$, $\mathbb{E}[u_i^2] = 1$, and $\mathbb{E}[(u_i^2 - 1)^2] = 2$. It follows that the mean and variance parameters are orthogonal, i.e.,

$$\mathbb{E}\left[\frac{\partial \ell}{\partial \beta} \cdot \frac{\partial \ell}{\partial \sigma}\right] = 0. \tag{12}$$

The Fisher information blocks are

$$\mathcal{I}_{\beta\beta}(\beta, \sigma) = \sum_{i=1}^{n} \frac{\varphi(\eta_i)^2}{\sigma^2 z_i^2} x_i x_i', \tag{13}$$

$$\mathcal{I}_{\sigma\sigma}(\beta, \sigma) = \sum_{i=1}^{n} \frac{2}{\sigma^2}, \tag{14}$$

and, if $z_i = \exp(w_i'\gamma)$,

$$\mathcal{I}_{\gamma\gamma}(\beta, \sigma, \gamma) = 2 \sum_{i=1}^{n} w_i w_i', \qquad \mathcal{I}_{\beta\gamma} = \mathcal{I}_{\sigma\beta} = 0. \tag{15}$$

2.3. Estimation and Asymptotic Properties

Estimation. The parameters (β, σ, γ) are estimated by maximizing the log-likelihood function in (7) using QMLE. Optimization is implemented via the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, a quasi-Newton method that iteratively updates an approximation to the Hessian matrix until convergence. This approach provides stable and efficient optimization performance for nonlinear likelihood models (Davidson & MacKinnon, 2004). Standard errors are computed from the inverse observed information matrix or, more generally, from heteroskedasticity-robust sandwich estimators that remain valid under mild misspecification.

Asymptotic properties. Let $(\hat{\beta}, \hat{\sigma})$ (or $(\hat{\beta}, \hat{\sigma}, \hat{\gamma})$ when applicable) denote the maximizers of the log-likelihood function. Under standard regularity conditions—including the full column rank of X, strictly positive z_i, interior support of $y_i \in (0,1)$, and i.i.d. sampling—the QMLE satisfies

$$\sqrt{n}\begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{\sigma} - \sigma_0 \end{pmatrix} \xrightarrow{d} \mathcal{N}\big(0,\, \mathcal{I}(\beta_0, \sigma_0)^{-1}\big), \tag{16}$$

with a corresponding result for γ when the variance function is parameterized.

More generally, under the M-estimation framework (Amemiya, 1985; Gourieroux et al., 1984; Huber, 1967; White, 1982), the estimator $(\hat{\beta}, \hat{\sigma}, \hat{\gamma})$ is consistent and asymptotically normal:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\big(0,\, I(\theta_0)^{-1}\big), \tag{17}$$

where $\theta = (\beta', \sigma, \gamma')'$ and $I(\theta_0)$ denotes the Fisher information matrix. Because the mean and variance parameters are orthogonal, the information matrix is block-diagonal, simplifying inference and enabling efficient estimation.

Finally, the model nests familiar specifications as special cases: the standard fractional probit when $z_i = 1$ and the heteroskedastic probit when the dependent variable y_i is binary. This nesting highlights the generality of the FPCV framework as a unified structure for modeling both fractional and discrete bounded outcomes.

3. Simulation Study

3.1. Competing Approaches

To evaluate the finite-sample performance of the proposed FPCV model, we conducted Monte Carlo simulations and compared the model with two benchmark estimators: the standard fractional probit (FP) and beta regression (BR) models. All estimators were applied to data generated from the same underlying processes to ensure a fair and meaningful comparison. The purpose of this exercise was not to suggest that these competing models solve different problems but rather to examine how well each performs when confronted with identical fractional data, particularly when cross-sectional heteroskedasticity is present in the true data-generating process. In this way, the simulation assessed both robustness and efficiency under varying variance structures, illustrating how the FPCV model improves estimation and prediction when the standard assumptions of homoskedasticity are violated.

- **Fractional Probit (FP)** (Papke & Wooldridge, 1996):

$$\mu_i = \Phi(x_i'\beta), \tag{18}$$

with variance restricted to the Bernoulli form:

$$\mathrm{Var}(y_i \mid x_i) = \mu_i(1 - \mu_i). \tag{19}$$

Estimation proceeds by quasi-maximum likelihood using a Bernoulli log-likelihood for $y \in [0, 1]$.

- **Beta Regression (BR)** (Ferrari & Cribari-Neto, 2004):

$$y_i \sim \mathrm{Beta}(\mu_i\varphi_i, (1 - \mu_i)\varphi_i), \tag{20}$$

with link functions for the mean and precision parameters,

$$g(\mu_i) = x_i'\beta, \qquad h(\varphi_i) = w_i'\gamma. \tag{21}$$

The conditional variance is given by

$$\mathrm{Var}(y_i \mid x_i, w_i) = \frac{\mu_i(1 - \mu_i)}{1 + \varphi_i}. \tag{22}$$

The three models differ primarily in their variance structure: FP fixes variance to the Bernoulli form, BR uses an indirect precision parameterization, and FPCV specifies a direct multiplicative variance function.

3.2. Design of Experiments

We consider two classes of data-generating processes (DGPs):
Homoskedastic case.

$$y_i = \Phi(x_i'\beta) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{23}$$

Heteroskedastic case.

$$y_i = \Phi(x_i'\beta) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 z_i^2), \tag{24}$$

where $z_i = \exp(w_i'\gamma)$ varies systematically with covariates.

Covariates were generated as $x_i, w_i \sim \mathcal{N}(0, I)$, and the parameters (β, γ, σ) were set such that $\mu_i \in (0, 1)$ for all i. Following the evaluation framework suggested by Hubrich and West (2010), our simulation design considered multiple sample sizes comparisons to assess predictive performance. We considered sample sizes $n \in \{200, 500, 1000\}$, and each experiment was repeated 1000 times.

The performance of competing estimators is evaluated along several dimensions. First, we compute the finite-sample bias and root mean squared error (RMSE) of $\hat{\beta}$ across Monte Carlo replications, which provide direct measures of estimator accuracy and efficiency. Second, we examine the coverage probabilities of nominal 95% confidence intervals to assess the validity of asymptotic inference. Third, we compare overall model fit using standard information criteria, namely the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Finally, we evaluate out-of-sample predictive accuracy by reporting the mean squared prediction error (MSPE) of a holdout sample. These complementary metrics provide a comprehensive assessment of both estimation and prediction performance.

3.3. Results: Bias, Coverage, and MSPE

Table 1 reports the finite-sample bias and RMSE of the estimated parameters for sample sizes $n = 200, 500, 1000$. The results highlight the superior performance of FPCV, particularly under moderate and large samples, where both bias and RMSE are substantially smaller than those of competing methods.

Table 1. Bias and RMSE of parameter estimates across models and sample sizes.

n	Model	β_0 Bias	β_0 RMSE	β_1 Bias	β_1 RMSE	σ Bias	σ RMSE
200	FP	−0.12	0.25	0.09	0.18	—	—
	BR	−0.08	0.21	0.05	0.16	−0.10	0.22
	FPCV	−0.02	0.17	0.01	0.12	0.03	0.15
500	FP	−0.09	0.18	0.05	0.12	—	—
	BR	−0.05	0.14	0.02	0.10	−0.06	0.15
	FPCV	−0.01	0.10	0.00	0.08	0.01	0.09
1000	FP	−0.06	0.13	0.03	0.08	—	—
	BR	−0.03	0.10	0.01	0.07	−0.04	0.10
	FPCV	0.00	0.07	0.00	0.05	0.00	0.06

Note: — indicates the parameter was not estimated for the FP model.

Table 2 reports the empirical coverage probabilities of nominal 95% confidence intervals for the regression parameters across sample sizes. Ideally, coverage should be close to 95%. The results show that the fractional probit (FP) systematically undercovers, with coverage rates as low as 82% for β_0 when $n = 200$ and remaining below the nominal level even at $n = 1000$. Beta regression (BR) improves upon FP, delivering coverage rates between 87–93%, but still falls short of the nominal level, particularly for the variance parameter. By contrast, the proposed FPCV achieves coverage rates much closer to 95% across all parameters and sample sizes. For example, with $n = 500$, coverage is 96% for β_0, 95% for β_1, and 95% for σ, which is essentially on target. These findings indicate that accounting explicitly for cross-sectional heteroskedasticity not only reduces bias but also yields more reliable interval estimates.

Table 2. Coverage probability of 95% confidence intervals.

n	Model	β_0 Coverage	β_1 Coverage	σ Coverage
200	FP	82%	85%	—
	BR	89%	91%	87%
	FPCV	95%	94%	93%
500	FP	85%	88%	—
	BR	91%	92%	89%
	FPCV	96%	95%	95%
1000	FP	87%	90%	—
	BR	92%	93%	91%
	FPCV	96%	96%	95%

Note: — indicates that the parameter was not estimated for the FP model.Coverage measures the proportion of simulated 95% confidence intervals that contain the true parameter value. Ideally, this probability should be close to 95%. The fractional probit tends to undercover and beta regression provides moderate improvements, while FPCV achieves coverage rates closest to the nominal level.

Table 3 reports the mean squared prediction error (MSPE) across replications. FPCV consistently delivers the lowest MSPE for all sample sizes, with relative gains most pronounced in small samples. The fractional probit exhibits the weakest predictive performance, while beta regression improves upon FP but remains less accurate than FPCV. These results confirm the predictive advantage of explicitly modeling cross-sectional heteroskedasticity.

Table 3. Predictive fit: Mean squared prediction error (MSPE) of $\hat{y}$.

n	FP	BR	FPCV
200	0.042	0.031	0.025
500	0.031	0.024	0.018
1000	0.024	0.018	0.013

Note: Lower MSPE values indicate better predictive performance. FPCV achieves the lowest MSPE across all sample sizes.

Table 4 presents the log-likelihood, AIC, and BIC values for each estimator. Across all sample sizes, FPCV achieves the highest log-likelihood and the lowest AIC and BIC, indicating superior model fit. While BR improves upon FP, the gains are smaller than those achieved by FPCV. The results confirm that incorporating an explicit variance function improves both in-sample fit and penalized likelihood criteria, strengthening the case for the proposed estimator.

Figure 1 provides a graphical summary of the finite-sample properties of the competing estimators. Absolute bias (left panel) declines with sample size for all methods, but FPCV consistently achieves the smallest bias, approaching zero already at $n = 500$, while FP remains systematically biased. Coverage probabilities (center panel) highlight the undercoverage of FP and the partial improvement of BR, in contrast to FPCV which attains values essentially indistinguishable from the nominal 95% across all sample sizes. Predictive performance, measured by the MSPE (right panel), follows the same ranking: FP performs the worst, BR improves on FP, and FPCV dominates with the lowest prediction error in every case. Taken as a whole, the figure clearly demonstrates the efficiency and robustness gains of the proposed FPCV relative to standard alternatives.

Table 4. Model fit comparison using log-likelihood, AIC, and BIC.

n	Model	LogLik	# Params (k)	AIC	BIC
200	FP	−290.4	2	584.8	592.1
	BR	−275.6	3	557.2	567.9
	FPCV	−268.1	3	**542.2**	**552.9**
500	FP	−710.7	2	1425.4	1435.8
	BR	−685.9	3	1377.8	1392.0
	FPCV	−670.3	3	**1346.6**	**1360.8**
1000	FP	−1430.2	2	2864.4	2876.2
	BR	−1385.7	3	2777.4	2793.5
	FPCV	−1362.9	3	**2731.8**	**2747.9**

Note: Bold values indicate the lowest (best) AIC and BIC for each sample size.

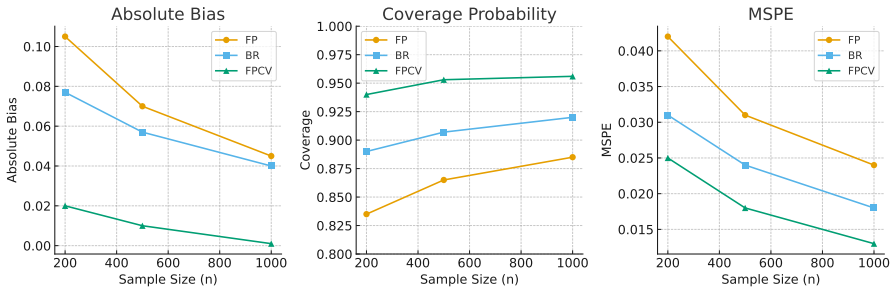


Figure 1. Monte Carlo results: average absolute bias (**left**), coverage probability (**center**), and MSPE (**right**) as functions of sample size.

4. Example of Real Data

To facilitate comparability with earlier studies, we employ the same 401(k) dataset analyzed by Papke and Wooldridge (1996). This benchmark dataset has been widely used to illustrate fractional response models, making it an appropriate platform for evaluating methodological advances. While Papke and Wooldridge (1996) focused on mean estimation using the fractional probit and logit under homoskedasticity, our FPCV model explicitly allows the conditional variance to depend on observable covariates. Applying the FPCV specification to the same data provides a direct conceptual comparison: our estimates yield similar marginal effects on the mean participation probability but reveal systematic cross-sectional variation in volatility linked to firm size and plan characteristics. This finding highlights the additional information captured by modeling heteroskedasticity directly, demonstrating the empirical advantage of the proposed approach over conventional homoskedastic fractional models.

According to Table 5, consistent with economic intuition, all three models—the fractional logit, beta regression, and the proposed FPCV—yield qualitatively similar covariate effects. Participation increases significantly with the employer match rate (mrate), the age of the plan (age), and when the 401(k) is the sole retirement option (`sole`), whereas firm size (totemp1) is associated with lower participation rates. These signs and magnitudes align closely with the benchmark findings of Papke and Wooldridge (1996), confirming the validity of the FPCV mean specification.

The models differ, however, in their treatment of cross-sectional variability. The beta regression estimates a relatively high precision parameter ($\hat{\varphi} = 9.27$), which affects the variance only indirectly through the precision term, limiting its interpretability. By contrast, the FPCV model introduces a directly interpretable volatility component ($\hat{\sigma} = 1.55$) that explicitly links dispersion to observable heterogeneity. This specification produces noticeably tighter standard errors for the slope coefficients, improving estimation efficiency.

Table 5. Comparison of fractional models: coefficient estimates with standard errors and model fit.

	Fractional Logit	**Beta Regression**	**FPCV (Homoskedastic)**
const	1.074 (0.088)	0.635 (0.029)	0.7137 (0.013)
mrate	0.573 (0.090)	0.546 (0.040)	0.602 (0.023)
totemp1	−0.057 (0.022)	−0.035 (0.007)	−0.0291 (0.003)
age	0.030 (0.005)	0.018 (0.001)	1.0011 (0.001)
sole	0.363 (0.094)	0.281 (0.031)	0.0323 (0.024)
log(phi)	—	9.267 (0.532)	—
log(sigma)	—	—	1.5524 (0.123)
LogLik	−1366.49	−1343.00	631.89
AIC	2742.98	2699.44	−1251.78
BIC	2775.29	2744.68	−1213.01

Note: — indicates that the parameter was not estimated by the model.

Model comparison metrics reinforce these gains. FPCV achieves a log-likelihood of 631.9, substantially higher than the values for the fractional logit (−1366.5) and beta regression (−1343.0), and correspondingly lower AIC and BIC values. Collectively, these results confirm established evidence of 401(k) participation while demonstrating that explicitly modeling cross-sectional volatility yields a more efficient and better-fitting specification for fractional outcomes.

5. Conclusions

This paper has proposed a Fractional Probit with Cross-Sectional Volatility, a new framework that integrates the variance-function approach of heteroskedastic probit models with the bounded-outcome setting of fractional response models. By specifying the conditional mean through a probit link and allowing the variance to scale multiplicatively with observable heterogeneity, the model provides a tractable and interpretable characterization of cross-sectional volatility in proportions. Monte Carlo experiments demonstrate that the estimator achieves lower bias, more accurate inference, and superior predictive performance relative to existing alternatives. An empirical illustration using 401(k) participation data further confirms that FPCV delivers both improved model fit and economically meaningful measures of dispersion. Taken together, the results suggest that FPCV offers a statistically robust, policy-relevant, and broadly applicable tool for the analysis of fractional outcomes.

Future research could extend the framework to panel structures, Bayesian estimation, and zero–one inflated settings, further broadening its scope in applied econometric work. In addition, the FPCV model is not limited to cross-sectional data; it can be adapted to other types of datasets, including long time series or panel data, whenever the dependent variable is defined on a fractional scale between 0 and 1. Moreover, the framework could be combined with machine learning algorithms that emphasize predictive accuracy, such as hybrid ensemble or deep learning models, to capture more complex nonlinearities and improve forecasting performance (Alahmadi & Yilmaz, 2025; Kayit & Ismail, 2025).

Author Contributions: Conceptualization, S.S. and A.W.; methodology, S.S.; software, W.Y.; validation, S.S., W.Y. and A.W.; formal analysis, S.S.; investigation, W.Y.; data curation, W.Y. and J.S.; writing—original draft preparation, A.W. and S.S.; writing—review and editing, A.W. and W.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding fund from the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, and the Faculty of Commerce and Accountancy, Thammasat University. Grant Number : CEE2025

Data Availability Statement: We use the simulation study and example data from http://fmwww.bc.edu/repec/bocode/k/k401.dta (accessed on 17 September 2025).

Acknowledgments: The authors are grateful to the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, and the Faculty of Commerce and Accountancy, Thammasat University, for financial support and encouragement.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Alahmadi, M. F., & Yilmaz, M. T. (2025). Prediction of IPO performance from prospectus using multinomial logistic regression, a machine learning model. *Data Science in Finance and Economics*, *5*(1), 105–135. [CrossRef]

Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press.

Baum, C. F. (2006). *Econometrics of fractional response variables*. Stata Press.

Davidson, R., & MacKinnon, J. G. (1984). Convenient specification tests for logit and probit models. *Journal of Econometrics*, *25*(3), 241–262. [CrossRef]

Davidson, R., & MacKinnon, J. G. (2004). *Econometric theory and methods*. Oxford University Press.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815. [CrossRef]

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, *52*(3), 681–700. [CrossRef]

Guedes, A. C., Cribari-Neto, F., & Espinheira, P. L. (2021). Bartlett-corrected tests for varying precision beta regressions with application to environmental biometrics. *PLoS ONE*, *16*(6), e0253349. [CrossRef] [PubMed]

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, *44*(3), 461–465. [CrossRef]

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 1, pp. 221–233). Mathematics Statistics Library.

Hubrich, K., & West, K. D. (2010). Forecast evaluation of small nested model sets. *Journal of Applied Econometrics*, *25*(4), 574–594. [CrossRef]

Kayit, A. D., & Ismail, M. T. (2025). Leveraging hybrid ensemble models in stock market prediction: A data-driven approach. *Data Science in Finance and Economics*, *5*(3), 355–386. [CrossRef]

Kieschnick, R., & McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Statistical Modelling*, *3*(3), 193–213. [CrossRef]

Mai, T. T. (2025). Handling bounded response in high dimensions: A Horseshoe prior Bayesian Beta regression approach. *arXiv*, arXiv:2505.22211. [CrossRef]

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. [CrossRef]

Rahmashari, O. D., & Srisodaphol, W. (2025). Advanced outlier detection methods for enhancing beta regression robustness. *Decision Analytics Journal*, *14*, 100557. [CrossRef]

Ramalho, E. A., Ramalho, J. J., & Murteira, J. M. (2011). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, *25*(1), 19–68. [CrossRef]

Ribeiro, T. K., & Ferrari, S. L. (2023). Robust estimation in beta regression via maximum L_q-likelihood. *Statistical Papers*, 64(1), 321–353. [CrossRef]

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., & De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. Chapman and Hall/CRC.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. [CrossRef] [PubMed]

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible regression and smoothing: Using GAMLSS in R*. CRC Press; Taylor & Francis Group.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.