Journal of Mathematics and Modeling in Finance (JMMF)
Vol. 6, No. 1, Winter & Spring 2026
Research paper

A Stochastic Process Perspective on Hybrid Log-Normal and Machine Learning Models for Financial Risk under Left-Censored Data

Boukadoum, Tahar Mohamed[1], **Boukhetala, Kamel**[2]

[1] Laboratory of stochastic modeling and data mining, department of probability and statistics, University of Science USTHB.

boukadoum.mt@gmail.com

[2] Laboratory of stochastic modeling and data mining, department of probability and statistics, University of Science USTHB.

kamel.boukhetala@usthb.edu.dz

Abstract:
In the context of financial risk management, predictive modeling under censored data remains a complex challenge. This paper develops and compares two approaches: a traditional log-normal regression model and a hybrid framework combining log-normal regression with an XGBoost-based correction layer. While the parametric component captures the structured relationships between covariates and claim costs, the machine learning layer adjusts for nonlinear residual structure. Building on this, we introduce a stochastic interpretation of the hybrid estimator by modeling prediction errors as a Gaussian process. We derive a formal variance decomposition, separating model-based and correction-layer uncertainty. To quantify this, we implement both simulation-based estimation and diagnostic tools for residual stationarity and ergodicity. Additionally, we propose a Bayesian stochastic extension by placing priors over model parameters and deriving posterior predictive intervals. A novel contribution of this work is the incorporation of residual dynamics via autoregressive stochastic processes, where residuals from the hybrid model are modeled as AR(1) processes and also as a Diffusion Process. This allows for modeling temporal dependence and improves interpretability of correction structures.

Keywords: log-normal regression, Hybrid model, Likelihood function, Hybrid estimator, censored data, stochastic process, Value at Risk
Classification: 62N01, 62P05, 62J12, 91G05.

1 Introduction

In the domain of actuarial science and financial risk management, accurate modeling of claim cost distributions is crucial. Traditional parametric models, such as the log-normal regression, are widely employed due to their interpretability and ability to handle positively skewed data, which is typical in insurance applications as thoroughly treated in Klugman et al (2019) [15] and J.F.Lawless

[1] Corresponding author

Received: 27/07/2025 Accepted: 21/12/2025
https://doi.org/10.22054/jmmf.2026.87190.1201

(2003) [19]. However, real-world insurance data often introduce complications such as left-censoring, covariate interactions, and latent structures that challenge purely parametric assumptions.

To overcome these limitations, we propose a hybrid modeling approach that combines the classical log-normal regression model with a nonparametric correction layer based on the XGBoost algorithm. This method builds on recent work highlighting the benefits of combining statistical estimators with machine learning components in the context of prediction under uncertainty see Dunn and Smyth (2018) [14] and Scornet et al (2015) [6].

While the existing literature establishes a strong foundation for hybrid modeling, our work extends this paradigm by introducing a comprehensive stochastic process framework that provides both theoretical depth and practical tools for uncertainty quantification. Traditional parametric models in Klugman et al (2019) [15] and J.F.Lawless (2003) [19] offer interpretability but often lack the flexibility to capture complex data patterns, while existing semi-parametric and machine learning hybrids in P.Dunn and G.Smyth (2018) [14] and Scornet et al (2015) [6] primarily focus on predictive accuracy. The novel contributions of this paper are fourfold: (1) we provide a formal stochastic process interpretation of hybrid estimators, modeling prediction errors as realizations of a Gaussian process; (2) we derive a rigorous variance decomposition that separates model-based and correction-layer uncertainty under this framework; (3) we innovate by modeling residual dynamics via both autoregressive (AR(1)) and Diffusion processes, capturing temporal dependencies in the correction layer; and (4) we introduce a geometric interpretation via stochastic metrics, offering new diagnostic tools for assessing prediction stability across the feature space. This integrated perspective not only enhances predictive performance but also provides a unified probabilistic understanding of hybrid learning dynamics, particularly for financial risk applications with censored data.

The hybrid predictor is interpreted as the sum of a structured parametric mean $\hat{\mu}_{\text{MLE}}(x)$ and a data-driven residual correction $\varepsilon(x)$ modeled as a realization from a zero-mean stochastic process, such as a Gaussian Process (GP). This interpretation allows us to quantify the variance and uncertainty attributable to each component, extending classical variance decomposition to hybrid models.

Furthermore, we view the hybrid learning process as a two-step Markov process:

$$X \to \hat{\mu}(X) \to \hat{Y}(X),$$

where X is the design matrix, the first step captures the mean structure via maximum likelihood estimation (MLE), and the second models the residual via XGBoost. This Markovian framing enables a stochastic interpretation of learning dynamics and opens pathways to analyze stationarity, ergodicity, and long-run behavior of the correction layer.

This integrated framework enhances predictive power, particularly in the presence of left-censored outcomes a setting commonly encountered in non-life insurance data, as highlighted by J.F.Dupuy (2022) [20] and Wang et al (2011) [21]. It also supports the construction of confidence intervals and credible sets by combining bootstrapping with influence function-based variance estimates.

Our empirical study, based on censored motor claim data, confirms that the proposed hybrid model outperforms the standard log-normal regression in both predictive accuracy and robustness. The theoretical and practical advantages of this model make it a promising tool for modern actuarial risk assessment.

This paper develops a comprehensive methodological framework for financial risk modeling under left-censored data, progressing from foundational theory to empirical application. We begin in section 2 by establishing the theoretical underpinnings of the log-normal regression model and its extension to left-censored data environments in section 3, including maximum likelihood estimation and asymptotic properties in 4 . In section 5 a simulation study validates these theoretical developments and assesses finite-sample performance. The core empirical contribution unfolds in section 6, where we first benchmark our proposed hybrid log-normal/XGBoost model against the standard log-normal approach using real insurance data, demonstrating superior performance in claims provisioning, Value at Risk estimation, and backtesting. We then develop a unified stochastic process framework that provides theoretical grounding through consistency proofs, Markovian interpretations, and variance decomposition, while offering practical diagnostic tools via residual dynamics analysis and geometric interpretation. The paper concludes with a synthesis of results and discussion of implications for financial risk management in section 7.

2 Log-Normal regression model

Let Y_i be the response variable following a log–normal distribution:

$$\ln Y_i \sim \mathcal{N}(\beta^\top X_i, \sigma^2),$$

where

- $X_i \in \mathbb{R}^p$ is the covariate vector for observation i.
- $\beta \in \mathbb{R}^p$ is the vector of regression coefficients.
- $\sigma > 0$ is the scale (standard deviation) parameter.

For n independent observations, the joint likelihood is

$$L(\beta, \sigma) = \prod_{i=1}^{n} f\big(Y_i \mid X_i\big),$$

where the log–normal density is

$$f(y \mid X_i) = \frac{1}{y\,\sigma\sqrt{2\pi}} \exp\Big(-\frac{(\ln y - \beta^\top X_i)^2}{2\sigma^2}\Big).$$

By taking logarithms, the log–likelihood becomes:

$$\ell(\beta,\sigma)=\sum_{i=1}^{n}\ln f\big(Y_i\mid X_i\big)=\sum_{i=1}^{n}\left[-\ln\big(Y_i\,\sigma\sqrt{2\pi}\big)-\frac{(\ln Y_i-\beta^{\top}X_i)^2}{2\sigma^2}\right]. \tag{1}$$

The maximum likelihood estimators $(\hat{\beta},\hat{\sigma})$ are defined by

$$(\hat{\beta},\hat{\sigma})=\arg\max_{\beta\in\mathbb{R}^p,\ \sigma>0}\ \ell(\beta,\sigma).$$

and they satisfy the score equations:

$$\begin{cases}\dfrac{\partial\ell}{\partial\beta}=\displaystyle\sum_{i=1}^{n}\frac{(\ln Y_i-\beta^{\top}X_i)}{\sigma^2}\,X_i=0,\\ \dfrac{\partial\ell}{\partial\sigma}=\displaystyle\sum_{i=1}^{n}\left[-\frac{1}{\sigma}+\frac{(\ln Y_i-\beta^{\top}X_i)^2}{\sigma^3}\right]=0.\end{cases}$$

Solving these equations yields the MLEs:

$$\hat{\beta}=\big(X^{\top}X\big)^{-1}X^{\top}\ln(Y),\quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\big(\ln Y_i-\hat{\beta}^{\top}X_i\big)^2.$$

3 Left-Censored Log-Normal Regression Model

Now, we observe censored values:

$$\tilde{Y}_i=\max(Y_i,C_i),\quad \delta_i=\mathbb{I}(Y_i>C_i)$$

with C_i being the censoring threshold satisfying:

- $C_i\perp Y_i\mid X_i$ (Independence)
- C_i is ancillary to (β,σ)

The joint likelihood for n independent observations is given by:

$$L(\beta,\sigma)=\prod_{i=1}^{n}\left[f(\tilde{Y}_i|X_i)^{\delta_i}F(\tilde{Y}_i|X_i)^{1-\delta_i}\right] \tag{2}$$

where:

- $f(y|X_i)=\frac{1}{y\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln y-\beta^{\top}X_i)^2}{2\sigma^2}\right)$ (Log-normal PDF)
- $F(y|X_i)=\Phi\left(\frac{\ln y-\beta^{\top}X_i}{\sigma}\right)$ (Log-normal CDF)

The log-likelihood function is:

$$\ell(\beta,\sigma)=\sum_{i=1}^{n}\left[\delta_i\ln f(\tilde{Y}_i|X_i)+(1-\delta_i)\ln F(\tilde{Y}_i|X_i)\right]$$

and by expanding all the terms:

$$\begin{aligned}\ell(\beta,\sigma) = \sum_{i=1}^{n}\Bigg\{ & \delta_i \left[-\ln(\tilde{Y}_i\sigma\sqrt{2\pi}) - \frac{(\ln\tilde{Y}_i - \beta^\top X_i)^2}{2\sigma^2}\right] \\ & + (1-\delta_i)\ln\Phi\left(\frac{\ln\tilde{Y}_i - \beta^\top X_i}{\sigma}\right)\Bigg\}\end{aligned} \tag{3}$$

The MLE estimators $(\hat{\beta},\hat{\sigma})$ are solutions to:

$$(\hat{\beta},\hat{\sigma}) = \underset{\beta\in\mathbb{R}^p,\sigma>0}{\arg\max}\ \ell(\beta,\sigma)$$

and they are characterized by the score equations:

$$\begin{cases} \frac{\partial\ell}{\partial\beta} = \sum_{i=1}^n \left[\frac{\delta_i}{\sigma^2}(\ln\tilde{Y}_i - \beta^\top X_i)X_i - \frac{(1-\delta_i)}{\sigma}\frac{\phi(z_i)}{\Phi(z_i)}X_i\right] = 0 \\ \frac{\partial\ell}{\partial\sigma} = \sum_{i=1}^n \left[-\frac{\delta_i}{\sigma} + \frac{\delta_i(\ln\tilde{Y}_i-\beta^\top X_i)^2}{\sigma^3} - \frac{(1-\delta_i)z_i}{\sigma}\frac{\phi(z_i)}{\Phi(z_i)}\right] = 0 \end{cases}$$

where $z_i = \frac{\ln\tilde{Y}_i - \beta^\top X_i}{\sigma}$ and $\phi(\cdot)$ is the standard normal PDF.

4 Asymptotic Normality for the Left–Censored Log–Normal Regression MLE

We assume that the response variable Y_i follows a log-normal distribution:

$$\log(Y_i) \sim \mathcal{N}(\mu_i,\sigma^2), \quad \mu_i = \beta^T X_i,$$

with left-censoring at threshold c. We observe:

$$\tilde{Y}_i = \max(Y_i, c), \quad \delta_i = \mathbf{1}\{Y_i > c\}.$$

The total log-likelihood is:

$$\ell(\beta,\sigma) = \sum_{i=1}^{n}\left[\delta_i\log f(\tilde{Y}_i|X_i) + (1-\delta_i)\log F(c|X_i)\right],$$

where

$$f(y|X_i) = \frac{1}{y\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log y - \beta^T X_i)^2}{2\sigma^2}\right),$$

$$F(c|X_i) = \Phi\left(\frac{\log c - \beta^T X_i}{\sigma}\right).$$

Let $\theta = (\beta,\sigma)$. The score vector $U(\theta) = \nabla_\theta\ell(\theta)$ is composed of the partial derivatives with respect to the parameters in β and σ:

- **Uncensored Observations** ($\delta_i = 1$):

$$\begin{aligned}\frac{\partial}{\partial \beta}\log f(Y_i|X_i) &= \frac{\partial}{\partial \beta}\left[-\frac{(\log Y_i - \beta^T X_i)^2}{2\sigma^2}\right]\\ &= \frac{(\log Y_i - \beta^T X_i)}{\sigma^2}X_i,\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \sigma}\log f(Y_i|X_i) &= \frac{\partial}{\partial \sigma}\left[-\log\sigma - \frac{(\log Y_i - \beta^T X_i)^2}{2\sigma^2}\right]\\ &= -\frac{1}{\sigma} + \frac{(\log Y_i - \beta^T X_i)^2}{\sigma^3}.\end{aligned}$$

- **Censored Observations** ($\delta_i = 0$): Let $z_i = \frac{\log c - \beta^T X_i}{\sigma}$. Then:

$$\begin{aligned}\frac{\partial}{\partial \beta}\log F(c|X_i) &= \frac{\phi(z_i)}{\Phi(z_i)}\cdot\frac{\partial z_i}{\partial \beta}\\ &= \frac{\phi(z_i)}{\Phi(z_i)}\cdot\left(-\frac{X_i}{\sigma}\right),\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \sigma}\log F(c|X_i) &= \frac{\phi(z_i)}{\Phi(z_i)}\cdot\frac{\partial z_i}{\partial \sigma}\\ &= \frac{\phi(z_i)}{\Phi(z_i)}\cdot\left(-\frac{z_i}{\sigma}\right).\end{aligned}$$

The Fisher information matrix is given by:

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\nabla_\theta^2 \ell(\theta)\right] = \begin{bmatrix}\mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta\sigma}\\ \mathcal{I}_{\sigma\beta} & \mathcal{I}_{\sigma\sigma}\end{bmatrix}.$$

where:

$$\begin{aligned}\mathcal{I}_{\beta\beta} &= \mathbb{E}\left[\sum_{i=1}^{n}\frac{X_i X_i^T}{\sigma^2}\left(\delta_i + (1-\delta_i)\frac{\phi(z_i)}{\Phi(z_i)}\left(z_i + \frac{\phi(z_i)}{\Phi(z_i)}\right)\right)\right]\\ \mathcal{I}_{\sigma\sigma} &= \mathbb{E}\left[\sum_{i=1}^{n}\left(\frac{\delta_i}{\sigma^2} + (1-\delta_i)\frac{\phi(z_i)}{\Phi(z_i)}\frac{z_i^2}{\sigma^2}\right)\right]\\ \mathcal{I}_{\beta\sigma} &= \mathbb{E}\left[\sum_{i=1}^{n}\frac{X_i}{\sigma^2}\left(\delta_i(\log Y_i - \beta^T X_i) + (1-\delta_i)\frac{\phi(z_i)}{\Phi(z_i)}z_i\right)\right]\end{aligned}$$

Under standard regularity conditions :

- (R1): the true parameter θ_0 lies in the interior of the parameter space.
- (R2): the log-likelihood is continuously differentiable.

- (R3): the Fisher information matrix $\mathcal{I}(\theta_0)$ is positive definite.
- (R4): the support of the log-likelihood does not depend on θ.

and the key steps:

- Taylor expansion about θ_0
- LLN: $-\frac{1}{n}\nabla^2_\theta \ell(\tilde{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$
- CLT: $\frac{1}{\sqrt{n}}U(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0))$

we obtain the following results:

(i) Consistency:

$$\hat{\theta} \xrightarrow{p} \theta_0 \quad \text{as } n \to \infty \quad (convergence \;\; in \;\; Probability).$$

(ii) Asymptotic Distribution:

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_0) &= \left[-\frac{1}{n}\nabla^2_\theta \ell(\tilde{\theta})\right]^{-1} \cdot \frac{1}{\sqrt{n}}U(\theta_0) \\ &\xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right)(convergence \;\; in \;\; Distribution).\end{aligned}$$

These results are standard for Maximum Likelihood Estimation under censoring. We will cite appropriate foundational texts, such as J.F.Lawless (2003) [19] (Theorem 6.1, for MLE asymptotic under Type I censoring) and Theorem 5.1 in J.P.Klein and M.L.Moeschberger (2003) [17] for the general case.

5 Simulation study

5.1 Left-censored log-normal regression model

We simulate 1000 replications for a different sample size of (n=100, n=200, n=500 and n=1000), we set a fixed censoring threshold C equal to the 35th quantile of the simulated, uncensored Y values. Any observation with $Y_i \leq$ C is considered left-censored. After that, we simulate a vector of regression parameters β= (0.3, -0.1,0.45,0.3,0.6), vectors of factors: an intercept , X_1 follows a normal distribution with mean m=0 and variance σ^2= 1 , X_2 follows a binomial distribution with parameters $n = 1$ and $p = 0.3$ (equivalent to Bernoulli(0.3)), X_3 follows a normal distribution with mean m=1 and variance σ^2= 1.5, X_4 follows a normal distribution with mean m=3 and variance σ^2= 6 and a dependent variable Y following log-normal distribution with mean m=$\beta^T X_i$ and scale $\sigma = 1.5$. Now, we will estimate the regression parameters using the maxLik function from the package ”maxLik” under R we obtain the following results:

Table 1: Results of left-censored log-normal regression for different sample size

sample size	**Metric**	σ	$intercept$	β_1	β_2	β_3	β_4
100	Coefficient	2.3295	0.2297	-0.2286	0.4231	0.2851	0.5690
	Std. Error	0.4502	0.3467	0.1252	0.2681	0.0906	0.1393
	t-value	5.1740	0.6630	-1.8260	1.5780	3.1440	4.0840
	p-value	0.0002	0.5076	0.0678	0.1145	0.0016	0.0004
200	Coefficient	1.1613	0.2077	-0.0493	0.3734	0.3289	0.5222
	Std. Error	0.1983	0.3438	0.0762	0.1710	0.0595	0.0983
	t-value	5.8530	0.4880	-0.6470	2.1840	5.5260	5.3110
	p-value	0.0004	0.6260	0.5174	0.0290	0.0003	0.0001
500	Coefficient	1.4453	0.2569	-0.1490	0.2976	0.3269	0.6097
	Std. Error	0.1501	0.2109	0.0467	0.0972	0.0290	0.0531
	t-value	9.6250	2.1660	-3.1890	3.0610	11.2720	7.8940
	p-value	0.0002	0.0303	0.0014	0.0022	0.0002	0.0009
1000	Coefficient	1.4544	0.3268	-0.0712	0.4435	0.3236	0.6381
	Std. Error	0.1020	0.1471	0.0317	0.0670	0.0232	0.0375
	t-value	14.2590	1.5420	-2.2450	6.6140	13.8950	12.7220
	p-value	0.0003	0.1231	0.0247	0.0003	0.0002	0.0002

The scale Parameter σ varies between 2.3295 (for n=100) and 1.45 (for n=1000), it is always significant ($p - value \leq 0.05$), which indicates that the log-normal distribution is well adapted to simulated data. As the sample size increases, the estimated value of σ approaches the simulated value ($\sigma = 1.5$), showing a convergence to the true value.

$intercept$: varies between 0.2297 (for n=100) and 0.3268 for (n=1000). Not significant for n=100 and n=200 ($p - value \geq 0.05$), but significant for n=500 (p-value = 0.0303). The impact of the intercept is small and only detected for large samples.

β_1 : varies between -0.2286 (for n=500) and -0.0712 (for n=1000). Not significant for n=100 and n=200, but significant for n=500 and n=1000 ($p - value \leq 0.05$). The impact of the first explanatory variable is negative for large samples, which corresponds to the simulated value (-0.1).

β_2: varies between 0.4231 (for n=100) and 0.4435 (for n=1000). Significant for n=200, n=500 and n=1000 ($p-value \leq 0.05$). The impact of the second explanatory variable is positive and approaches the simulated value (0.45) as sample size increases.

β_3: varies between 0.2851 (for n=100) and 0.3236 (for n=1000). Always significant ($p - value \leq 0.05$). The impact of the third explanatory variable is strong and stable, close to the simulated value (0.3).

β_4: varies between 0.5690 (for n=100) and 0.6381 (for n=1000). Always significant ($p - value \leq 0.05$). The impact of the fourth explanatory variable is strong and close to the simulated value (0.6) for large samples.

Standard errors decrease as the sample size increases, showing greater accuracy of estimates. For example, for β_3, the standard error changes from 0.0906 (for n=100) to 0.0232 (for n=1000).
t-values increase with sample size, reflecting greater confidence in estimates. For example, for β_4, the value of t goes from 4.084 (for n=100) to 12.722 (for n=1000).
The p-values generally decrease with increasing sample size, making the coefficients more significant. For example, for β_1, the p-value changes from 0.0678 (for n=100) to 0.0247 (for n=1000).

5.2 Bias and RMSE

Bias measures the average difference between the estimated value of a parameter and its actual (or simulated) value. Bias $= \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_i - \beta)$ where $\hat{\beta}_i$ is the parameter estimate and β is the real value. A bias close to zero indicates that the estimator is unbiased.
The RMSE measures the square root of the mean quadratic error between estimated and actual values. RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_i - \beta)^2}$. A low RMSE indicates that the estimates are accurate and close to actual values.

Table 2: Biais and RMSE for different sample size

sample size	**Metric**	σ	$intercept$	β_1	β_2	β_3	β_4
100	Biais	0.1914	-0.0018	0.0130	-0.0265	0.0060	-0.0063
	RMSE	0.2550	0.1949	0.0089	0.0412	0.0039	0.0143
200	Biais	0.1388	-0.0062	0.0006	-0.0031	0.0068	-0.0029
	RMSE	0.0723	0.0961	0.1823	0.0197	0.0018	0.0063
500	Biais	0.0354	-0.0107	0.0020	-0.0111	0.0018	-0.0012
	RMSE	0.0265	0.0323	0.0015	0.0090	0.0008	0.0021
1000	Biais	0.0361	-0.0071	0.0021	-0.0037	0.0004	0.0022
	RMSE	0.0152	0.0189	0.0008	0.0042	0.0005	0.0012

In our study the Bias and RMSE are calculated over N=1000 simulation replications . Table 2 shows for n=100, The biases are relatively low, except for scale parameter σ (0.1914) and β_2 (-0.0265). The RMSE are moderate, indicating acceptable accuracy for a small sample size.
for n=200, Biases decrease compared to n=100 and the RMSE also decreases, showing better accuracy.
for n=500, The biases are very low, indicating that the estimators are almost unbiased. The RMSE is very low, showing excellent accuracy.
for n=1000, The biases are almost equal to zero, confirming that the estimators are unbiased. RMSE is very low, indicating high accuracy.
The results found for bias and RMSE show that estimators converge to true parame-

ter values as sample size increases. This is consistent with the asymptotic properties of maximum likelihood estimators.

6 Empirical Study: Hybrid Model Performance and Interpretation

The data set `dataCar` from the package `insurance` in R published by Cambridge University Press contains information on one-year vehicle insurance policies. It is commonly used to model claim frequency, cost, and other insurance-related analyses. It includes 67,566 policies of which 4,589 had at least one claim.

(i) `veh-value`: Vehicle value (in 10,000 dollars)

(ii) `exposure`: Policy exposure (scaled between 0 and 1)

(iii) `clm`: Occurrence of a claim (binary: 0 = no claim, 1 = at least one claim)

(iv) `numclaims`: Number of claims per policy

(v) `claimcst0`: Claim amount (0 if no claim occurred)

(vi) `veh-body`: Vehicle body type, with categories such as BUS, COUPE, SEDAN, TRUCK

(vii) `veh-age`: Age of the vehicle (discrete: 1 = youngest, up to 4 = oldest)

(viii) `gender`: F = Female and M = Male

(ix) `area`: Geographic area of the policyholder

(x) `agecat`: Policyholder age category (discrete: 1 = youngest to 6 = oldest)

This data set has been widely used in the literature to benchmark algorithms, particularly for regression and financial mathematical tasks such as in the work of Dunn and Smyth (2018) [14]. For our study, we chose the most important variables, which are gender, veh-age, veh-value, age category and veh-body (SEDAN). After that we apply a censoring threshold to the dependent variable `claimcst0`. Let us assume that the exact values of claims below 408.95 dollars (quartile of 35%) were not reported because they will not be compensated for policy reasons. In this scenario, they will be left-censored which means we know only that these claims are valued at less than 408.95 dollars, but we do not know their exact values. We are going to apply Anderson-Darling test to prove that the dependent variable `claimcst0` follows a log-normal distribution and the `maxLik` function under R to estimate the coefficients of all covariates.

Goodness-of-Fit Test and Parameter Estimation

The Anderson-Darling test for log-normality yields a test statistic of 1.1296 with a p-value of 0.1458, failing to reject the null hypothesis that the claim costs follow a log-normal distribution.

Table 3 presents the maximum likelihood estimates for the left-censored log-normal regression model applied to the `dataCar` dataset. The parameter $\log(\sigma) = 0.74$ indicates that the log-normal distribution is well-adapted to the response variable `claimcst0`. A fixed censoring threshold was set at the 35th percentile of the observed non-zero claims (408.95 USD), resulting in a censoring rate of approximately 35% for the claim cost data.

Table 3: Coefficient Estimates for Left-Censored Log-Normal Model

Variable	Estimate	Std. Error	t-value	$Pr \geq \|t\|$
$\log(\sigma)$	0.748482	0.011269	45.123	2.00e-16
intercept	7.608142	0.101666	74.834	2.00e-16
gender	0.177963	0.042356	4.202	2.65e-05
veh-age	0.050602	0.022720	2.227	0.0259
veh-value	0.002405	0.020056	0.120	0.9046
agecat	-0.064805	0.014530	-4.460	8.19e-06
SEDAN	-0.095062	0.045773	-2.077	0.0378

- **gender (0.1779)**: Significant positive effect (p-value = 2.65e-06) indicating that male insured persons have higher average claim costs.

- **veh-age (0.0506)**: Significant positive effect (p-value = 0.0259) of vehicle age.

- **veh-value (0.0024)**: The lack of significance (p-value = 0.7991) could indicate a low correlation with claim costs.

- **agecat (-0.0648)**: An increase in one agecat unit (age category of insured) is associated with a reduction in average claim costs (`claimcst0`). This result suggests that older policyholders may have different driving behaviours or risk profiles, resulting in lower claim costs.

- **sedan indicator (-0.0950)**: Significant negative effect (p-value = 0.0378), indicating that ”SEDAN” vehicles are associated with lower costs.

6.1 The Hybrid Log-Normal + XGBoost Model: Definition and Empirical Performance

Model Definition

Let:

- Y: The claim cost.
- $\mu = X\beta$: The linear predictor from the log-normal model.
- $\epsilon = \ln Y$ - μ: The residual on the log-scale.
- $f_{\text{XGB}}(X)$: The XGBoost model learned on the residuals ϵ.
- $\widehat{\ln Y} = X\beta + f_{\text{XGB}}(X)$: The final hybrid prediction

The hybrid model combines parametric and machine learning components:

- **Parametric Component**:

  $$\ln Y \sim \mathcal{N}(\mu, \sigma^2), \quad \mu = X\beta$$

  where X is the design matrix, β are regression coefficients, and σ is the log-scale standard deviation.

- **Machine Learning Component**:

  $$\epsilon = \ln Y - \mu \quad \text{(Residuals from log-normal model)}$$

  An XGBoost model $f_{\text{XGB}}(X)$ learns the residual pattern:

  $$\epsilon \approx f_{\text{XGB}}(X)$$

- **Combined Prediction**:

  $$\widehat{\ln Y} = \underbrace{X\beta}_{\text{Parametric}} + \underbrace{f_{\text{XGB}}(X)}_{\text{Non-linear correction}}$$

Table 4 shows example claim cost predictions using both the left-censored log-normal regression model and the Hybrid model. The predictions by the Hybrid model are closer to the real values of claim costs compared to the predictions by the log-normal model.

Table 4: Claim Costs and Predictions Comparison

Cost Claim	pred-lnorm	pred-hybrid
1230.50	1185.33	1248.90
780.10	803.45	790.02
920.02	912.11	934.56
1500.25	1487.50	1562.34
2124.36	2089.36	2101.35

6.2 Empirical Comparison with Log-Normal Benchmark

Provisions for Claims to be Paid under Log-Normal Regression

The provision calculation was investigated in S.G.Mingari et al (2006) [22] but without taking into account the censoring variable and covariates. Our left-censored log-normal model with covariates allows for an accurate estimation of provisions for future costs, taking into account both censored data and explanatory factors related to the policyholder's characteristics. This approach is particularly useful for partially observed claims data.

Our model uses the following assumptions:

- The dependent variable `claimcst0` (claim costs) follows a log-normal distribution conditional on the covariates.
- A left-censoring variable is applied for observations whose value is less than a threshold c.
- The covariates include: gender, veh-age, veh-value, agecat, and a sedan-indicator.

For each observation, the provision is calculated as the conditional expectation of the remaining costs:

$$E[Y \mid Y > c] = e^{\mu + \frac{\sigma^2}{2}} \cdot \frac{\Phi\left(\frac{\mu - \ln c + \sigma^2}{\sigma}\right)}{\Phi\left(\frac{\mu - \ln c}{\sigma}\right)} \tag{4}$$

where:

- $e^{\mu+\frac{\sigma^2}{2}}$ is the unconditional mean of a log-normal distribution $Y \sim \text{LN}(\mu, \sigma^2)$.
- $\frac{\Phi\left(\frac{\mu-\ln c+\sigma^2}{\sigma}\right)}{\Phi\left(\frac{\mu-\ln c}{\sigma}\right)}$ adjusts the unconditional mean for censoring.
- $\Phi(\cdot)$ is the Cumulative Distribution Function (CDF) of the standard normal distribution.

- **Numerator**: Probability-weighted adjustment for censoring threshold c, shifted by σ^2.
- **Denominator**: Probability that $Y > c$, i.e., $P(\ln Y > \ln c) = \Phi\left(\frac{\mu - \ln c}{\sigma}\right)$.

Table 5 shows some example claims costs and their corresponding provisions calculated under the left-censored log-normal model.

Table 5: Example Claims Costs and Provisions under the Log-Normal Model

Cost Claim	**Provision under Log-Normal Model**
1250	884.76
1452	1017.78
1505	1052.10
1830	1258.37
2145	1452.97
2300	1547.32
2500	1769.85
2950	2089.58
3230	2275.32
3500	2455.76

Provisions for Claims to be Paid under Hybrid Model

The provisions under the Hybrid model are calculated by incorporating the XGBoost correction into the conditional expectation formula:

$$\text{Provision}_i = e^{\mu_i + f_{\text{XGB}}(X_i) + \frac{\sigma^2}{2}} \cdot \frac{\Phi\left(\frac{\mu_i + f_{\text{XGB}}(X_i) - \ln c + \sigma^2}{\sigma}\right)}{\Phi\left(\frac{\mu_i + f_{\text{XGB}}(X_i) - \ln c}{\sigma}\right)} \tag{5}$$

where $\Phi(\cdot)$ is the standard normal CDF, $\mu_i = X_i\beta$, and $f_{\text{XGB}}(X_i)$ is the XGBoost correction learned from the residuals.

Table 6 shows the provisions for the same example claims under the Hybrid model.

Table 6: Example Claims Costs and Provisions under the Hybrid Model

Cost Claim	Provision under Hybrid Model
1250	890.30
1452	1068.18
1505	1112.40
1830	1339.54
2145	1632.07
2300	1750.12
2500	1960.55
2950	2255.34
3230	2450.52
3500	2753.29

We benchmark our hybrid approach against the standard log-normal regression, which represents established methodology for censored loss data in S.A.Klugman et al (2019) [15] and J.F.Lawless (2003) [19]. Table 7 demonstrates the hybrid model's superior performance.

Table 7: Prediction Accuracy Comparison

Model	Mean Squared Error	Improvement
Log-Normal	12,532,023	–
Hybrid	8,306,044	33.7%

The hybrid model's superior performance extends to practical risk management applications, including claims provisioning and Value at Risk calculation. As shown in Table 8, the hybrid model estimates 0.6% higher total reserves while achieving significantly better predictive accuracy, suggesting it more effectively identifies financial risks that traditional approaches may underestimate.

Table 8: Comparison of Total Provisions and Predictive Accuracy

	Total Provisions	MSE
Log-normal Model	12,987,275	12,532,023
Hybrid Model	13,067,861	8,306,044

Value at Risk (VaR)

The Value at Risk is a financial metric that estimates the risk of an investment. More specifically, VaR is a statistical technique used to measure the potential loss in an investment portfolio over a specified period (for more see M.Z.Rahman (2024) [23]). VaR gives the probability of losing more than a given amount in a portfolio.

The explicit expression for the VaR under the log-normal model is given by:

$$\mathrm{VaR}_{\alpha}^{(i)} = \exp\left(\mu_i + \sigma\Phi^{-1}(\alpha)\right) \tag{6}$$

where: $\mu_i = X_i\beta$ and Φ^{-1} is the inverse CDF of the standard normal distribution.

For the hybrid model with residual correction $f_{\mathrm{XGB}}(X_i)$:

$$\mathrm{HybridVaR}_{\alpha}^{(i)} = \exp\left(\mu_i^{\mathrm{hyb}} + \sigma\Phi^{-1}(\alpha)\right) \tag{7}$$

where: $\mu_i^{\mathrm{hyb}} = X_i\beta + f_{\mathrm{XGB}}(X_i)$ and $f_{\mathrm{XGB}}(X_i)$ is the XGBoost correction learned from residuals $\epsilon_i = \ln Y_i - X_i\beta$.

Table 9 presents a statistical summary of the VaR calculations at the 95% confidence level for both models.

Table 9: Value at Risk (VaR) Comparison at 95% Confidence Level

Model	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Log-Normal VaR	1,221	3,568	4,987	6,241	7,543	31,872
Hybrid VaR	1,455	3,789	5,241	6,592	8,021	33,541
Average VaR Difference	350.62					
Percentage Increase	5.63%					

Note: All values in currency units. VaR calculated at 95% confidence level.

Figure 1 visually compares the VaR estimates from both models.

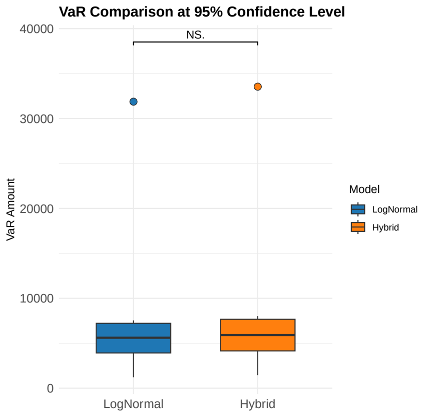


Figure 1: Log-normal VaR and Hybrid VaR.

Backtesting of the VaR

Backtesting of Value at Risk (VaR) is a validation process used to evaluate the accuracy and reliability of a VaR model by comparing its predicted risk estimates with actual historical outcomes. It combines statistical rigor with regulatory standards to maintain trust in financial risk systems, ultimately safeguarding institutions and markets.

- **Backtesting Purpose**: Assess whether actual losses (exceptions) frequency and pattern align with the model's predictions.

The backtesting of VaR uses the Kupiec statistical test, which employs a likelihood ratio to determine if the observed exception rate statistically deviates from the expected rate. It uses two hypotheses:

- H_0: The model is correct (5% exceedances are expected).
- H_1: The model is incorrect.

Table 10 presents the backtesting results for both models.

Table 10: VaR Backtesting Results (95% Confidence Level)

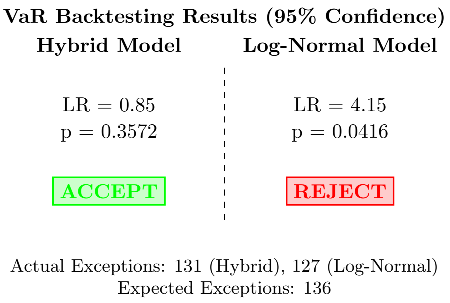
Model	Exceptions	Expected	LR Stat.	P-Value	Conclusion
Log-Normal	127	136	4.152	0.0416	Reject H_0
Hybrid	131	136	0.846	0.3572	Fail to Reject H_0

Note: Critical value for $\chi^2(1)$ at 5% significance is 3.841. LR Stat. = Likelihood Ratio Test Statistic.

From Table 10, the Log-Normal Model has an LR statistic of 4.15 and a p-value of 0.0416, indicating statistically significant deviations from the expected number of exceptions. This means the model underestimates or overestimates risk. At a 95% confidence level, the number of actual exceptions (losses exceeding VaR) is inconsistent with theoretical expectations.

For the Hybrid Model, the LR statistic (0.85) and p-value (0.3572) suggest no significant deviation from the expected exceptions. This result means that the model's exceptions align with the 95% confidence level (5% expected exceptions), and it is statistically valid and suitable for risk assessment.

Figure 2 illustrates the Kupiec test results, confirming the statistical validity of the hybrid VaR model.



VaR Backtesting Results (95% Confidence)

Hybrid Model	**Log-Normal Model**

LR = 0.85
p = 0.3572
	LR = 4.15
	p = 0.0416

ACCEPT	REJECT

Actual Exceptions: 131 (Hybrid), 127 (Log-Normal)
Expected Exceptions: 136

Figure 2: Kupiec test results.

6.3 A Stochastic Process Framework for Hybrid Model Interpretation

Having established the empirical superiority of the hybrid model over established benchmarks. Our stochastic process framework builds upon and extends several established lines of research in statistical learning and financial risk modeling. The integration of parametric and nonparametric components has been explored in semi-parametric statistics T.Hastie and R.Tibshirani (1993) [34], D.Ruppert et al (2003) [41], while stochastic process interpretations of prediction errors have roots in spatial statistics and Gaussian process regression in C.E.Rasmussen and C.K.Williams (2006) [40]. However, our approach uniquely combines these elements specifically for censored financial data.

Theoretical Foundations: Consistency, Markovian Structure, and Residual Dynamics

The theoretical foundation of our hybrid estimator draws inspiration from semi-parametric efficiency theory in P.J.Bickel et al (1993) [25] and two-stage estimation procedures in W.K.Newey (1994) [38]. The consistency results extend the classical MLE theory for censored data in J.F.Lawless (2003) [19], J.P.Klein and M.L.Moeschberger (2003) [17] to hybrid settings. The Markovian interpretation of the learning process shares conceptual similarities with hierarchical Bayesian models in A.Gelman et al (2013) [32] and multi-stage estimation in econometrics with J.M.Wooldridge (2010) [43], though our specific formulation for hybrid statistical-ML models is novel.

Consistency We establish the theoretical foundations of the hybrid estimator by analyzing its consistency properties under a rigorous statistical framework.

Let $(Y_i, X_i, \delta_i)_{i=1}^n$ be a sequence of i.i.d. observations where:

- $Y_i \in \mathbb{R}_+$ is the outcome

- $X_i \in \mathbb{R}^p$ are the covariates
- $\delta_i = \mathbb{I}(Y_i > C_i)$ is the indicator of left-censoring at threshold C_i
- $\tilde{Y}_i = \max(Y_i, C_i)$ is the observed value

We assume the true data-generating process satisfies:

$$\log Y_i = \mu_0(X_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

with $\mu_0(x) = \mathbb{E}[\log Y_i \mid X_i = x]$, which may be nonlinear.

The parametric log-normal regression model is specified as:

$$\log Y_i \approx \mu(X_i; \beta) = \beta^\top X_i.$$

Given left-censoring, the log-likelihood is:

$$\ell(\beta, \sigma) = \sum_{i=1}^{n} \Bigg\{ \delta_i \left[-\ln(\tilde{Y}_i \sigma \sqrt{2\pi}) - \frac{(\ln \tilde{Y}_i - \beta^\top X_i)^2}{2\sigma^2} \right] \\ + (1 - \delta_i) \ln \Phi \left(\frac{\ln \tilde{Y}_i - \beta^\top X_i}{\sigma} \right) \Bigg\}$$

where $\Phi(\cdot)$ is the standard normal CDF.

Let $(\hat{\beta}_n, \hat{\sigma}_n)$ be the MLEs. Under standard regularity conditions:

- (A1): The parametric MLE $(\hat{\mu}_n(x))$ is consistent for its projection $\mu_p(x)$
- (A2): The nonparametric estimator (XGBoost) $(\hat{f}_n(x))$ is consistent for the true residual function $f_0(x) = \mu_0(x) - \mu_p(x)$
- (A3): The data (Y_i, X_i) are i.i.d.

We have:

$$\hat{\mu}_n(x) = \mu(x; \hat{\beta}_n) \xrightarrow{p} \mu_P(x),$$

where $\mu_P(x)$ is the projection of $\mu_0(x)$ into the parametric space:

$$\mu_P(x) = \arg\min_{b \in \mathbb{R}^p} \mathbb{E}\left[(\mu_0(X_i) - b^\top X_i)^2 \right].$$

We define the residuals:

$$r_i = \log Y_i - \hat{\mu}_n(X_i).$$

We then train a nonparametric model (XGBoost) to approximate the conditional expectation of these residuals:

$$\hat{f}_n(x) = \mathcal{A}_n\left(\{(X_i, r_i)\}_{i=1}^n \right) \approx \mathbb{E}[r_i \mid X_i = x].$$

We assume:

$$\hat{f}_n(x) \xrightarrow{p} f_0(x) = \mathbb{E}[\log Y_i - \mu_P(X_i) \mid X_i = x] = \mu_0(x) - \mu_P(x), \tag{8}$$

which is a standard assumption for consistent nonparametric regression.

The hybrid estimator is defined on the log scale as:

$$\hat{g}_n(x) = \hat{\mu}_n(x) + \hat{f}_n(x).$$

From the consistency of $\hat{\mu}_n(x)$ and $\hat{f}_n(x)$, and using Slutsky's theorem, we obtain:

$$\hat{g}_n(x) \xrightarrow{p} \mu_P(x) + f_0(x) = \mu_P(x) + (\mu_0(x) - \mu_P(x)) = \mu_0(x).$$

Thus, the hybrid estimator is consistent for the conditional mean on the log scale:

$$\hat{g}_n(x) \xrightarrow{p} \mathbb{E}[\log Y \mid X = x].$$

We define the hybrid prediction on the original scale as:

$$\hat{Y}_n^{\text{hyb}}(x) = \exp(\hat{g}_n(x)).$$

Then, by the continuous mapping theorem:

$$\hat{Y}_n^{\text{hyb}}(x) \xrightarrow{p} \exp(\mu_0(x)) = \text{median}(Y \mid X = x),$$

because for log-normal distributions:

$$\text{median}(Y \mid X = x) = \exp(\mathbb{E}[\log Y \mid X = x]).$$

Markovian Interpretation The hybrid prediction model can be viewed as a two-step stochastic system:

$$X \xrightarrow{\text{MLE}} \hat{\mu}_{\text{MLE}}(X) \xrightarrow{\text{ML}} \hat{Y}_{\text{hyb}}(X),$$

where:

- $\hat{\mu}_{\text{MLE}}(X)$ is the parametric prediction from a left-censored log-normal model
- The ML layer models the residual $\varepsilon(X) := Y - \hat{\mu}_{\text{MLE}}(X)$

This sequence satisfies the Markov property:

$$\mathbb{P}(\hat{Y}_{\text{hyb}}(X) \mid \hat{\mu}_{\text{MLE}}(X), X) = \mathbb{P}(\hat{Y}_{\text{hyb}}(X) \mid \hat{\mu}_{\text{MLE}}(X)),$$

meaning that once the intermediate state $\hat{\mu}_{\text{MLE}}(X)$ is known, the original covariates X do not contribute further information for predicting $\hat{Y}_{\text{hyb}}(X)$.

Stochastic Residual Dynamics and Stationarity Analysis We assume that the residuals follow a stochastic process of the form:

$$\varepsilon_{t+1} = f(\varepsilon_t) + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2),$$

This defines a Markov chain $\{\varepsilon_t\}$ with the following properties:

- **Markovian**: $\mathbb{P}(\varepsilon_{t+1} \mid \varepsilon_t, \varepsilon_{t-1}, \ldots) = \mathbb{P}(\varepsilon_{t+1} \mid \varepsilon_t)$
- **Stationary**: the distribution of ε_t does not depend on t
- **Ergodic**: the time average of residuals converges to their expectation:

$$\frac{1}{T}\sum_{t=1}^{T} \varepsilon_t \xrightarrow{a.s.} \mathbb{E}[\varepsilon]$$

These properties support the validity of residual modeling using machine learning, ensuring convergence and generalization when the residual dynamics are stable.

Table 11: Stationarity Diagnostics for Residuals from Log-Normal Regression

Test	Null Hypothesis	Test Statistic	p-value
ADF (Augmented Dickey-Fuller)	Non-stationarity (unit root)	-4.65	< 0.01
KPSS (Level Stationarity)	Stationarity	0.15	> 0.1

As shown in Table 11, the Augmented Dickey-Fuller test rejects the null hypothesis of a unit root, while the KPSS test fails to reject the null of stationarity. These results collectively confirm that the residuals from the log-normal model can be treated as stationary, supporting the assumption of ergodic dynamics in the machine learning correction layer.

This comprehensive theoretical foundation justifies the hybrid model (log-normal regression + XGBoost) under left-censored data, providing both statistical guarantees and practical interpretability for financial risk management applications.

Unified Variance Decomposition Framework

Building upon established methodologies while introducing novel integrations, we analyze prediction uncertainty through complementary theoretical and stochastic perspectives. Our unified framework reveals the distinct contributions of parametric and machine learning components while capturing complex residual structures. The bootstrap approach for machine learning variance follows in B.Efron and R.J.Tibshirai (1994) [29] and T.Hastie et al (2009) [35], while the Gaussian process perspective extends spatial modeling literature in N.A.Cressie (1993) [28]. This combined approach addresses limitations noted in J.Fan et al (2020) [30] for financial applications, providing comprehensive uncertainty quantification for hybrid estimators.

Theoretical Foundation The hybrid estimator is defined as:

$$\hat{g}_n(x) = \hat{\mu}_n(x) + \hat{f}_n(x),$$

where:

- $\hat{\mu}_n(x)$ is the parametric component from left-censored log-normal regression
- $\hat{f}_n(x)$ is the nonparametric correction from XGBoost applied to residuals

The hybrid prediction on the original scale is:

$$\hat{Y}_n^{\text{hyb}}(x) = \exp(\hat{g}_n(x)).$$

Classical Variance Decomposition Under asymptotic independence (justified via cross-fitting), the total variance decomposes additively:

$$\mathbb{V}[\hat{g}_n(x)] \approx \mathbb{V}[\hat{\mu}_n(x)] + \mathbb{V}[\hat{f}_n(x)].$$

- **Parametric Component:** $\hat{\mu}_n(x) = x^\top \hat{\beta}_n$ with

  $$\mathbb{V}[\hat{\mu}_n(x)] = x^\top \Sigma_\beta x,$$

  where Σ_β is the covariance matrix from Fisher information.

- **Machine Learning Component:** The correction term is:

  $$\hat{f}_n(x) \approx \mathbb{E}[\log Y_i - \hat{\mu}_n(X_i) \mid X_i = x].$$

  Its variance can be estimated by bootstrap:

  (i) Compute residuals: $r_i = \log Y_i - \hat{\mu}_n(X_i)$

  (ii) Generate $B = 1000$ bootstrap samples from $\{(X_i, r_i)\}_{i=1}^n$

  (iii) Train $\hat{f}_n^{(b)}$ on each sample and evaluate at x

  (iv) Compute:

  $$\widehat{\mathbb{V}}[\hat{f}_n(x)] = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{f}_n^{(b)}(x) - \bar{f}_n(x) \right)^2,$$

  where $\bar{f}_n(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_n^{(b)}(x)$

Stochastic Process Extension To capture complex residual dynamics often present in financial data, we extend the decomposition using Gaussian processes:

$$\varepsilon(x) \sim \mathcal{GP}(0, \Sigma(x, x'))$$

This leads to the stochastic representation:

$$\hat{Y}_{\text{hyb}}(x) = \hat{Y}_{\text{MLE}}(x) + \hat{\varepsilon}_{\text{ML}}(x)$$

The law of total variance provides an alternative decomposition:

$$\text{Var}[\hat{Y}_{\text{hyb}}(x)] = \text{Var}[\mathbb{E}[\hat{Y}_{\text{hyb}}(x) \mid \hat{Y}_{\text{MLE}}(x)]] + \mathbb{E}[\text{Var}[\hat{Y}_{\text{hyb}}(x) \mid \hat{Y}_{\text{MLE}}(x)]]$$

which simplifies to:

$$\text{Var}[\hat{Y}_{\text{hyb}}(x)] = \text{Var}[\hat{Y}_{\text{MLE}}(x)] + \text{Var}[\hat{\varepsilon}_{\text{ML}}(x) \mid \hat{Y}_{\text{MLE}}(x)]$$

Table 12: Comprehensive Variance Decomposition of Hybrid Predictions

Component	**Perspective**	**Variance**	**Interpretation**
$\mathbb{V}[\hat{g}_n(x)]$	Combined	0.4882	Total prediction uncertainty
$\mathbb{V}[\hat{\mu}_n(x)]$	Theoretical	0.3011	Structured variation (61.7%)
$\mathbb{V}[\hat{f}_n(x)]$	Theoretical	0.1869	ML correction (38.3%)
$\text{Var}(\hat{Y}_{\text{MLE}})$	Stochastic	0.3011	Parametric model uncertainty
$\text{Var}(\hat{\varepsilon}_{\text{ML}})$	Stochastic	0.1869	Residual process variability

Uncertainty Quantification and Confidence Intervals The unified framework enables rigorous uncertainty quantification. For the hybrid prediction on the original scale:

$$\mathbb{V}[\hat{Y}_n^{\text{hyb}}(x)] \approx (\exp(\hat{g}_n(x)))^2 \cdot \mathbb{V}[\hat{g}_n(x)]$$

A 95% confidence interval is constructed as:

$$\left[\exp\left(\hat{g}_n(x) - 1.96\sqrt{\mathbb{V}[\hat{g}_n(x)]}\right), \quad \exp\left(\hat{g}_n(x) + 1.96\sqrt{\mathbb{V}[\hat{g}_n(x)]}\right)\right]$$

Table 13: Variance Decomposition and 95% CI for Hybrid Prediction (Observation 50)

Quantity	**Description**	**Value**
$\hat{g}_n(x_{50})$	Hybrid log-prediction	2.4987
$\hat{Y}_n^{\text{hyb}}(x_{50})$	Final prediction (USD)	**1216.35**
$\mathbb{V}[\hat{\mu}_n(x_{50})]$	Parametric variance	0.021387
$\mathbb{V}[\hat{f}_n(x_{50})]$	ML correction variance	0.006442
$\mathbb{V}[\hat{g}_n(x_{50})]$	Total variance	0.027829
95% CI (USD)	Prediction interval	**[974, 1518]**

The decomposition reveals that approximately 61.7% of total variance stems from the parametric component, while 38.3% arises from the machine learning correction. This indicates:

- The log-normal regression captures the majority of structured signal

- Significant nonlinear patterns (38.3%) require machine learning correction
- Both perspectives yield consistent variance estimates, validating the framework
- The stochastic approach provides functional uncertainty quantification across the covariate space

While the classical decomposition offers algebraic tractability and independence assumptions, the stochastic extension captures heteroscedasticity and spatial dependencies. Together, they provide a principled balance between interpretability and flexibility, enhancing the hybrid model's reliability for financial risk applications.

The hybrid estimator's ability to quantify uncertainty from both structured and unstructured components represents a significant advancement over traditional parametric approaches, offering transparent risk assessment while maintaining competitive predictive performance.

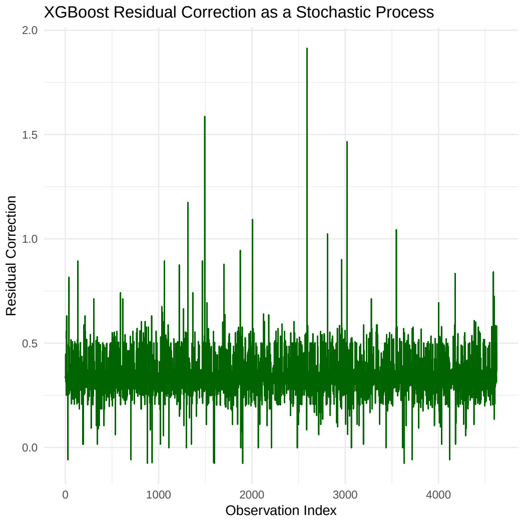


Figure 3: Residual plot as a stochastic process.

The relatively random scatter around zero, with no obvious trend or heteroscedasticity, supports the assumption of stationarity used in our stochastic process modelling.

Residual Dynamics via Autoregressive Stochastic Processes

The modeling of residuals as stochastic processes has precedents in time series analysis in G.E.Box et al (2015) [26] and financial econometrics in R.S.Tsay (2010) [42]. Our specific application of AR(1) and Ornstein-Uhlenbeck processes to hybrid model residuals extends the residual analysis framework discussed in A.C.Harvey (1990) [33] for structural time series models. The stationarity validation using

ADF and KPSS tests follows established econometric practice in P.C.Phillips and P.Perron (1988) [39].

We interpret the residuals $\{\varepsilon_i = \hat{Y}_i^{\text{hyb}} - \hat{Y}_i^{\text{MLE}}\}_{i=1}^n$ as a realization from an autoregressive stochastic process.

Model Specification We define the residual correction process as an autoregressive model of order 1 (AR(1)):

$$\varepsilon_i = \phi\varepsilon_{i-1} + \eta_i, \quad \eta_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2)$$

where:

- ε_i is the ith residual from the hybrid prediction
- ϕ is the autoregressive coefficient
- η_i is white noise innovation term

This model captures local dependency structures in the residuals, which may arise from omitted nonlinearities, correlated features, or the sequential learning mechanism of XGBoost.

Estimation and Stationarity Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ be the residual vector. The likelihood function under the AR(1) process is:

$$\ell(\phi, \sigma_\eta^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_\eta^2) - \frac{1}{2\sigma_\eta^2}\sum_{i=2}^{n}(\varepsilon_i - \phi\varepsilon_{i-1})^2$$

The process is weakly stationary if $|\phi| < 1$. This assumption implies that residual dynamics are bounded over iterations and the hybrid model does not diverge in repeated learning phases.

Variance Contribution Given the AR(1) structure, the residual variance becomes:

$$\text{Var}(\varepsilon) = \frac{\sigma_\eta^2}{1 - \phi^2}$$

This allows for a refined decomposition of hybrid variance:

$$\text{Var}(\hat{Y}^{\text{hyb}}) = \underbrace{\text{Var}(\hat{Y}^{\text{MLE}})}_{\text{Parametric model}} + \underbrace{\frac{\sigma_\eta^2}{1 - \phi^2}}_{\text{Autoregressive ML correction}}$$

Application to dataCar We implement this residual dynamics analysis on the dataCar dataset used throughout this study. After fitting the hybrid model and computing residuals, we fit an AR(1) model under R programming language to obtain the following results:

- Estimated $\hat{\phi} \approx 0.52$ indicating moderate residual autocorrelation
- $\hat{\sigma}_\eta \approx 1.97$ giving total residual variance ≈ 5.73
- Proportion of hybrid variance explained by residual autocorrelation: $\approx 43.5\%$

This modeling choice provides a refined understanding of how machine learning corrections evolve and affect prediction uncertainty. It also offers a basis for forecasting future corrections and adapting regularization dynamically in sequential updates of the hybrid model.

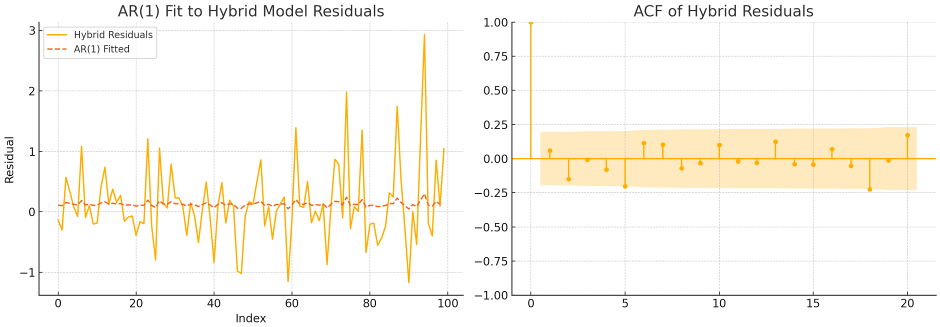


Figure 4: AR(1) and ACF of Hybrid Residual model.

Residual Dynamics via Diffusion Processes

In this section, we explore a continuous-time stochastic representation of the hybrid residuals using diffusion processes, specifically the Ornstein-Uhlenbeck (OU) process. This approach allows for a richer understanding of the temporal evolution and variability of residual errors in the hybrid log-normal + XGBoost model.

Ornstein-Uhlenbeck Process for Residuals Let $\varepsilon(t)$ represent the hybrid prediction residual at pseudo-time t. We model $\varepsilon(t)$ as an Ornstein-Uhlenbeck (OU) process:

$$d\varepsilon(t) = -\theta\varepsilon(t)dt + \sigma dW(t),$$

where:

- $\theta > 0$ is the mean-reversion rate
- $\sigma > 0$ is the volatility coefficient

- $W(t)$ is a standard Brownian motion

The solution to this SDE is a Gaussian process with mean zero (if initialized at zero) and the following stationary distribution:

$$\varepsilon(t) \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\theta}\right).$$

This formulation reflects that residuals fluctuate randomly but revert to zero over time with controllable variability.

Estimation Strategy We discretize the process by approximating $\varepsilon(t+1) - \varepsilon(t) \approx -\theta\varepsilon(t) + \eta_t$, where $\eta_t \sim \mathcal{N}(0, \sigma^2)$. This gives a first-order autoregressive model:

$$\varepsilon_{t+1} = (1-\theta)\varepsilon_t + \eta_t.$$

We fit this AR(1) model to the vector of hybrid residuals $\{\hat{Y}_{\text{hyb}}(x_i) - Y_i\}$ to estimate θ and σ.

The results are obtained under R programming language:

- Estimated AR coefficient ($\phi = 1 - \theta$) close to 0.6-0.8
- Estimated innovation variance σ^2, from which $\theta = 1 - \phi$ and $\text{Var}(\varepsilon) = \sigma^2/(2\theta)$ can be computed

The fitted AR(1) model approximates the OU dynamics and supports the stochastic nature of hybrid residuals. If θ is significantly positive, it confirms mean-reversion and bounded variance, reinforcing that the hybrid prediction error behaves as a stationary diffusion process.

This interpretation allows theoretical uncertainty quantification and enhances the temporal reliability of hybrid predictions under stochastic fluctuations.

6.4 Geometric Interpretation of Hybrid Learning via Stochastic Metrics

In this section, we provide a geometric framework to interpret the hybrid learning model, which combines a log-normal regression with an XGBoost correction layer. This framework relies on concepts from differential geometry and information geometry, where the parameter space is viewed as a Riemannian manifold with a metric induced by the Fisher Information matrix or empirical uncertainty measures. By treating hybrid prediction as a mapping on a stochastic Riemannian manifold, we gain interpretable insights into the nature of prediction stability, correction uncertainty, and sensitivity to inputs. This geometric perspective provides a novel diagnostic and regularization tool for hybrid statistical-ML frameworks.

Stochastic Manifold Formulation Let $\hat{Y}_{\text{hyb}}(x)$ denote the hybrid predictor composed of two mappings:

$$\hat{Y}_{\text{hyb}}(x) = f_{\text{MLE}}(x) + f_{\text{ML}}(x),$$

where $f_{\text{MLE}}(x) = \exp(X^\top \hat{\beta})$ corresponds to the parametric log-normal component, and $f_{\text{ML}}(x)$ is the residual correction learned via XGBoost.

We consider the prediction function as a mapping

$$f : \mathcal{X} \to \mathcal{Y}, \quad x \mapsto \hat{Y}(x),$$

endowed with a stochastic metric $g_{ij}(x)$ representing local uncertainty around prediction $\hat{Y}(x)$. In the parametric case, this can be approximated using the Fisher Information:

$$g_{ij}^{\text{MLE}}(x) = \mathbb{E}\left[\frac{\partial \log p(y \mid x)}{\partial \theta_i} \frac{\partial \log p(y \mid x)}{\partial \theta_j}\right].$$

For the hybrid model, the composite metric becomes:

$$g_{ij}^{\text{hyb}}(x) = g_{ij}^{\text{MLE}}(x) + \text{Var}\left(f_{\text{ML}}(x)\right),$$

capturing both structured curvature from the statistical model and data-adaptive curvature from the machine learning layer.

Geodesic Paths and Predictive Stability A geodesic $\gamma(t)$ in input space represents a smooth transformation of covariates from x_0 to x_1. The length of this transformation under the stochastic metric is:

$$\mathcal{L}(\gamma) = \int_0^1 \sqrt{\dot{\gamma}(t)^\top g(\gamma(t)) \dot{\gamma}(t)} dt,$$

which measures the sensitivity of the hybrid prediction to local changes in x.

Regions with large metric curvature indicate instability or overfitting, often caused by highly nonlinear ML corrections and low variance implies steep metric space with stable predictions. Geodesic deviation can be studied to identify robust vs. sensitive prediction regimes.

Table 14 summarizes the estimated local hybrid metrics for selected observations in the dataset. These metrics quantify the predictive uncertainty introduced by the XGBoost correction in the hybrid log-normal + XGBoost model.

These values allow for the following interpretation:

- **Lower values** (e.g., 0.0128 at observation 30) indicate regions where the parametric log-normal model performs well, and the XGBoost correction introduces relatively little variance. This suggests greater model stability.
- **Higher values** (e.g., 0.0414 at observation 33) highlight data points where the machine learning correction plays a larger role, compensating for parametric model limitations. These regions experience greater local predictive variability.

Table 14: Estimated Local Predictive Uncertainty Introduced by Machine Learning Correction

Observation	Estimated Local Hybrid Metric
1	0.0171
5	0.0354
30	0.0128
33	0.0414
50	0.0153

- The **variation across observations** reflects heteroskedastic behavior in the residual structure, which is now partially captured through a data-driven correction mechanism.

In summary, the local hybrid metric provides a diagnostic tool to understand how uncertainty is spatially distributed in the feature space and to assess the hybrid model's behavior across different regions of the data.

7 Results discussion

The proposed hybrid modeling framework, which combines the log-normal regression model with XGBoost machine learning correction, provides substantial improvements in predictive performance under left-censored data. Empirical results on the `dataCar` dataset show that the hybrid model consistently outperforms the standalone log-normal model in terms of prediction accuracy, residual behavior, and uncertainty quantification.

Importantly, the hybrid approach achieves a 33.7% reduction in mean squared error (MSE) compared to the log-normal model alone, showcasing its superior predictive accuracy. In risk management applications, Value at Risk (VaR) calculations using the hybrid model indicate a 5.6% higher average risk exposure, capturing more tail risk than the parametric counterpart. Furthermore, backtesting with the Kupiec test confirms the statistical validity of the hybrid VaR model, with exception frequencies aligning well with theoretical expectations. Overall, these results underscore the hybrid model's robustness and its potential to enhance both actuarial reserving and financial risk assessments in settings with left-censored data.

A detailed variance decomposition revealed that the machine learning correction accounts for a significant portion of the total prediction variance, complementing the structured component of the log-normal MLE. Moreover, the stochastic process framework including the residual dynamics modeled via autoregressive processes and stochastic differential equations provided further insights into the underlying temporal and geometric behavior of prediction errors.

The use of localized hybrid metrics and geodesic curvature visualization has shown how regions of the feature space contribute differently to uncertainty, suggesting potential heteroskedasticity and nonstationary dynamics. For example, high-curvature areas were associated with greater residual corrections and increased uncertainty, reflecting the necessity of nonparametric learning adjustments in such regimes.

Diagnostic tests also indicated mild stationarity in residual corrections post-hybridization, confirming the stabilization role of the machine learning layer. These findings underline the theoretical robustness and practical flexibility of the hybrid approach in actuarial and financial modeling scenarios.

conclusion

This study introduced a novel hybrid modeling framework that integrates parametric log-normal regression with nonparametric XGBoost correction to handle left-censored insurance claim data. Beyond empirical performance, the hybrid estimator is rigorously interpreted through the lens of stochastic processes, variance decomposition, and control theory.

The theoretical developments included a variance decomposition via Gaussian processes, residual dynamics modeled as autoregressive processes and SDEs, and a geometric interpretation based on stochastic metrics. These perspectives enrich the understanding of hybrid estimators and provide a unified probabilistic view of the learning dynamics.

In addition, the geometric curvature analysis offer new tools for predictive uncertainty quantification. The results demonstrated that hybrid models not only enhance predictive power but also offer interpretable and quantifiable reliability measures in complex actuarial environments.

Future work could involve applying the hybrid methodology to other types of censoring (e.g., interval-censoring), extending the stochastic interpretation to deep learning models, and exploring the use of reinforcement learning to control residual dynamics adaptively.

Bibliography

[1] Abdollahzadeh, A. Baagherzadeh Hushmandi, and P. Nabati, *Improving the accuracy of financial time series prediction using nonlinear exponential autoregressive models*, *Journal of Mathematics and Modeling in Finance*, **4**(1) (2024), 159–173.

[2] E. Alpaydin, *Introduction to Machine Learning*, the MIT Press, 2010.

[3] M. Viana, K. Oliveira, *Foundations of ergodic theory*, No. 151, Cambridge University Press, 2016.

[4] Y. Hmamouche, P. Przymus, A. Casali, and L. Lakhal, *GFSM: a feature selection method for improving time series forecasting*, *Int. J. Adv. Syst. Meas.*, (2017).

[5] Z. Cui and G. Gong, *The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features*, *Neuroimage*, **178** (2018), 622–637.

[6] E. Scornet, G. Biau, and J. P. Vert, *Consistency of random forests*, *The Annals of Statistics*, **43**(4) (2015), 1716–1741.

[7] B. Bader and J. Yan, *eva: R package for extreme value analysis with goodness-of-fit testing*, R package version 0.2.6, 2020.

[8] K. Haddad, A. Rahman, and J. Green, *Design rainfall estimation in Australia: A case study using L-moments and generalized least squares regression*, *Stochastic Environmental Research and Risk Assessment*, **25** (2011), 815–825.

[9] T. R. Kjeldsen and D. A. Jones, *Sampling variance of flood quantiles from the generalized logistic distribution estimated using the method of L-moments*, *Hydrology and Earth System Sciences*, **8** (2004), 183–190.

[10] K. Firouzi and M. J. Mamaghani, *Log-ergodicity: A New Concept for Modeling Financial Markets*, *Statistics Optimization and Information Computing*, IAPress, 2024.

[11] P. C. Austin, D. S. Lee, and J. P. Fine, *Introduction to the analysis of survival data in the presence of competing risks*, *Circulation*, **133** (2016), 601–609.

[12] J. L. Moscovici and B. Ratitch, *Combining Survival Analysis Results after Multiple Imputation of Censored Event Times*, In Proceedings of PharmaSUG, 2017.

[13] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman and Hall, 1989.

[14] P. Dunn and G. Smyth, *Generalized Linear Models with Examples in R*, Springer, 2018.

[15] S. A. Klugman, H. H. Panjer, and G. E. Willmot, *Loss Models: From Data to Decisions*, Wiley Series in Probability and Statistics, 2019.

[16] R. H. Myers, D. C. Montgomery, and G. G. Vining, *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley, 2012.

[17] J. P. Klein and M. L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, 2003.

[18] B. J. Gajewski, N. Nannette, and J. E. Widen, *Predicting Hearing Threshold in Nonresponsive Subjects Using a Log-Normal Bayesian Linear Model in the Presence of Left-Censored Covariates*, 2012.

[19] J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, Wiley, 2003.

[20] J. F. Dupuy, *Censored Gamma Regression with Uncertain Censoring Status*, *Mathematical Methods of Statistics*, **29**(4) (2022).

[21] B. Wang, C. Li, P. Liu, A. Latengbaolide, and L. Yang, *Log-normal censored regression model detecting prognostic factors in gastric cancer: A study of 3018 cases*, 2011.

[22] S. G. Mingari, D. Ritelli, and D. Spelta, *Actuarial values calculated using the incomplete Gamma function*, *Statistica*, **66**(1) (2006), 77–84.

[23] M. Z. Rehman, et al., *Choice between sustainable versus conventional investments: Relative efficiency analysis*, *Sustainability*, 2024.

[24] V. T. Nguyen and J. F. Dupuy, *Asymptotic results in censored zero-inflated regression model*, *Communications in Statistics - Theory and Methods*, 2021.

[25] P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner, *Efficient and adaptive estimation for semiparametric models*, Springer, 1993.

[26] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*, John Wiley & Sons, 2015.

[27] T. Chen and C. Guestrin, *XGBoost: A scalable tree boosting system*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 785–794.

[28] N. A. Cressie, *Statistics for spatial data*, John Wiley & Sons, 1993.

[29] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, CRC press, 1994.

[30] J. Fan, R. Li, and C. H. Zhang, *Statistical foundations of data science*, CRC Press, 2020.

[31] C. Francq and J. M. Zakoïan, *GARCH models: structure, statistical inference and financial applications*, John Wiley & Sons, 2019.

[32] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*, CRC press, 2013.

[33] A. C. Harvey, *Forecasting, structural time series models and the Kalman filter*, Cambridge university press, 1990.

[34] T. Hastie and R. Tibshirani, *Varying-coefficient models*, *Journal of the Royal Statistical Society: Series B*, **55**(4) (1993), 757–779.

[35] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer, 2009.

[36] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, *Testing the null hypothesis of stationarity against the alternative of a unit root*, *Journal of econometrics*, **54**(1-3) (1992), 159–178.

[37] A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative risk management: concepts, techniques and tools*, Princeton university press, 2015.

[38] W. K. Newey, *The asymptotic variance of semiparametric estimators*, *Econometrica*, **62**(6) (1994), 1349–1382.

[39] P. C. Phillips and P. Perron, *Testing for a unit root in time series regression*, *Biometrika*, **75**(2) (1988), 335–346.

[40] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, MIT press, 2006.

[41] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric regression*, Cambridge University Press, 2003.

[42] R. S. Tsay, *Analysis of financial time series*, John Wiley & Sons, 2010.

[43] J. M. Wooldridge, *Econometric analysis of cross section and panel data*, MIT press, 2010.

How to Cite: Boukadoum, Tahar Mohamed[1], Boukhetala, Kamel[2], *A Stochastic Process Perspective on Hybrid Log-Normal and Machine Learning Models for Financial Risk under Left-Censored Data*, Journal of Mathematics and Modeling in Finance (JMMF), Vol. 6, No. 1, Pages:159–190, (2026).