Journal of Statistical Research *https://doi.org/10.3329/jsr.v59i2.88067*
2025, Vol. 59, No. 2, pp. 221-247 *ISSN 0256 - 422 X*

STATISTICAL INFERENCE IN GAMMA REGRESSION MODEL UNDER LEFT-CENSORED DATA USING R

MOHAMED TAHAR BOUKADOUM*

Laboratory of Stochastic Modeling and Data Mining
University of Science and Technology Houari Boumediene, Algiers, Algeria

Email: boukadoum.mt@gmail.com

KAMAL BOUKHETALA

Laboratory of Stochastic Modeling and Data Mining
University of Science and Technology Houari Boumediene, Algiers, Algeria

Email: kamel.boukhetala@usthb.edu.dz

JEAN-FRANÇOIS DUPUY

Mathematics Research Institute of Rennes (IRMAR, statistics group)

Email: Jean-Francois.Dupuy@insa-rennes.fr

SUMMARY

In this paper, we consider the problem of the Gamma regression model under left-censored data with covariates. The method investigated consists of solving left-censored maximum likelihood estimating equations. We show that the resulting estimates are asymptotically normal. A simulation study assesses the proposed parameters' finite-sample properties and the root mean square error estimates. An application using car insurance data is presented to estimate the covariates coefficients in calculating the provisions for claims to be paid. We examine the effect of the censoring variable on the calculation of provisions. We will employ a machine learning algorithm called Random Forest to show the impact of the presence of the censoring variable. Finally, we address financial risk management that considers the Value at Risk (VaR), the Expected Shortfall (ES), and the backtesting of the VaR.

Keywords and phrases: Gamma regression, Likelihood function, estimation, left-censored, simulation, risk, VaR

AMS Classification: 62N01, 62P05, 62J12, 91G05

* Corresponding author
© Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh.

1 Introduction

Gamma regression models are widely used in statistical analysis to model continuous positive data exhibiting skewness, and they are commonly found in biology, engineering, and economics. These models are advantageous when the response variable follows a Gamma distribution, and the mean of the response is related to a set of covariates through a link function, often the logarithmic function. Those models are well developed with McCullagh and Nelder (1989), Dunn and Smyth (2018), and Fahrmeir et al. (2013).

Data are subject to censoring in specific real-world applications, such as environmental or medical studies. Censoring occurs when the exact value of an observation is unknown, but partial information (e.g., a threshold) is available. When the event of interest has not yet occurred by the end of the study or measurement period, this is known as right-censoring, which is well treated in Klagman et al. (2019), Myers et al. (2012), Klein and Moeschberger (2003), Lawless (2003) and Jean-François Dupuy (2022).

In other applications, like insurance and financial studies, the values below a certain threshold are not observed exactly but are only known to be below that threshold. This is known as left-censoring. A left-censored Gamma distribution addresses this by adjusting the likelihood function to account for the censored observations. This ensures accurate parameter estimation and inference, which is treated in Klein and Moeschberger (2003).

In this article, the proposed left-censored gamma regression model with covariates not extensively studied in the statistical literature corrects for the bias introduced by censoring and enables robust estimation of relationships between covariates and the censored response variable. Also, we will study the case of car insurance when the insurer decides to set a threshold where all claims below this threshold will not be compensated, which is known in statistics by the left censoring.

The remainder of this paper is organized to guide the reader from theoretical development to practical application. Section 2 presents the Gamma regression model and its left-censored extension, formulating the likelihood function and deriving the maximum likelihood estimators. The asymptotic behavior of these estimators is investigated in Section 3, establishing their normality. To assess their performance in practical scenarios, Section 4 details a simulation study. The methodological framework is then applied to a real insurance dataset in Section 5, where we perform parameter estimation, provision calculation, and a Random Forest based analysis of censoring effects. Section 6 addresses financial risk management through the calculation of Value at Risk (VaR) and Expected Shortfall (ES), complemented by a Kupiec test for backtesting. A comprehensive synthesis and discussion of all findings are presented in the concluding Section 7.

2 Gamma Regression Model and its Left-Censored Extension

Let Y denote the outcome and X be a p-vector of covariates. We assume that conditionally on X, Y is Gamma distributed with probability density function as defined in Jean-françois Dupuy (2022):

$$f(Y_i \mid X_i) \;=\; \frac{1}{\Gamma(\nu)} Y_i^{\nu\,-\,1} \left(\frac{\nu}{\mu(X_i)}\right) \exp\left(-\frac{\nu Y_i}{\mu(X_i)}\right), \quad Y_i > 0, \tag{2.1}$$

where $\nu > 0$ is the shape parameter, $\Gamma(\nu) = \int_0^{+\infty} u^{\nu-1} e^{-u}\, du$ is the gamma function, and $\mu(X) = \mathbb{E}(Y \mid X)$ is the conditional expectation of Y given X. The most communly used link function in Gamma model is the log-link, defined as $\log(\mu(X)) = \beta^t X \Leftrightarrow \mu(X) = \exp(\beta^t X)$, where β is unknown p-vector regression parameters.

Now, Y_i is left censored by a random variable C. We observe

$$Y_i > C \quad \text{or} \quad Y_i \leq C.$$

We also observe an indicator $\delta = \|_{Y>C}$, the censoring indicator which tells when Y_i is observed ($\delta = 1$) and ($\delta = 0$) Y_i is censored. We define $\tilde{Y}_i = \max(Y, C)$, and assume that C and Y are independent given the covariates X, and that C is ancillary to ν and β. Assume that we observe n independent copies $(\tilde{Y}_1, \delta_1, X_1), ..., (\tilde{Y}_n, \delta_n, X_n)$ of $(\tilde{Y}, \delta, X)$. The likelihood function is :

$$\mathrm{L}(\nu, \beta) = \prod_{i=1}^{n} \left[f\left(\tilde{Y}_i \mid X_i\right)^{\delta_i} \left(F\left(\tilde{Y}_i \mid X_i\right)\right)^{1-\delta_i} \right], \tag{2.2}$$

where

$$F(y \mid X) = 1 - \frac{\Gamma(\nu, \frac{\nu y}{\mu(X)})}{\Gamma(\nu)},$$

see Jean-françois Dupuy (2022) for details, and $\mu(X) = \exp(\beta^t X)$. The log-likelihood of (ν, β) is written as

$$\begin{aligned} \log L(\nu, \beta) = \sum_{i=1}^{n} \Bigg\{ & \delta_i [(\nu - 1)\log(\tilde{Y}_i) + \nu \log(\nu) - \nu \beta^t X_i - (\tilde{Y}_i)\nu e^{-\beta^t X_i} - \log\Gamma(\nu)] \\ & + (1 - \delta_i) \log\left(1 - \frac{\Gamma(\nu, \nu \tilde{Y}_i e^{-\beta^t X_i})}{\Gamma(\nu)}\right) \Bigg\}. \end{aligned} \tag{2.3}$$

3 The Asymptotic Normality of the Estimators

The maximum likelihood estimator is consistent and asymptotically distributed as a Gaussian law.

Proof. We follow the classical steps of MLE theory. The central limit theorem for MLE applies if certain conditions are met (identifiability, regularity...). First,the maximum likelihood estimator $\hat{\beta}_{\mathrm{MLE}}$ is defined by the first order conditions (also called the score equations). In fact, the MLE estimator is the solution to the system of equations where the log-likelihood gradient concerning β equals 0 at $\hat{\beta}_{\mathrm{MLE}}$.

$$\nabla_\beta \ell(\hat{\beta}_{\mathrm{MLE}}) = 0. \tag{3.1}$$

The derivatives of the log-likelihood concerning β give the score equations, they take into account both censored and uncensored observations. To simplify the writing, we put

$$\log L(\nu, \beta) = \ell(\beta) = \sum_{i=1}^{n} [I(Y_i > c_i) \log f(Y_i \mid X_i, \beta) + I(Y_i \leq c_i) \log F(c_i \mid X_i, \beta)],$$

where $f(Y_i \mid X_i, \beta)$ is the probability density function of the Gamma distribution, $F(c_i \mid X_i, \beta)$ is the is the cumulative distribution function of the Gamma distribution, evaluated at c_i, and $I(.)$ is the indicator function.

For the uncensored observations $(Y_i > c_i)$, the probability density function of the Gamma distribution is given by :

$$f(Y_i \mid X_i, \beta) = \frac{Y_i^{\frac{1}{\nu} - 1} e^{-Y_i/(\mu_i \nu)}}{\mu_i^{1/\nu} \Gamma(1/\nu)}.$$

The derivative of the log-likelihood concerning β for an uncensored observation is given by:

$$\frac{\partial}{\partial \beta} \log f(Y_i \mid X_i, \beta) \;=\; \frac{\partial}{\partial \beta} \left[\left(\frac{1}{\nu} \;-\; 1 \right) \log Y_i \;-\; \frac{Y_i}{\mu_i \nu} \;-\; \frac{1}{\nu} \log \mu_i \right],$$

by deriving each term with respect to β, and taking into account that $\mu_i \;=\; g^{-1}(X_i^T \beta)$, we find

$$\frac{\partial}{\partial \beta} \log f(Y_i \mid X_i, \beta) \;=\; \frac{Y_i \;-\; \mu_i}{\mu_i^2 \nu} \frac{\partial \mu_i}{\partial \beta} = \; \frac{Y_i - \mu_i}{\mu_i \nu} X_i,$$

where

$$\frac{\partial \mu_i}{\partial \beta} \;=\; \mu_i \frac{\partial}{\partial \beta} (X_i^t \beta) \;=\; \mu_i X_i.$$

For the censored observations, let:

$$F(c_i \mid X_i, \beta) \;=\; P(Y_i \leq c_i \mid \mu_i, \nu),$$

the derivative of this term for β is :

$$\frac{\partial}{\partial \beta} \log F(c_i \mid X_i, \beta) \;=\; \frac{1}{F(c_i \mid X_i, \beta)} \frac{\partial}{\partial \beta} F(c_i \mid X_i, \beta),$$

using the rule of derivation for the Gamma distribution function, we obtain:

$$\frac{\partial \beta}{\partial F(c_i \mid X_i, \beta)} = f(c_i \mid X_i, \beta) \cdot \frac{\partial \beta}{\partial \mu_i},$$

thus, the contribution of censored observations to the score equations is:

$$\frac{\partial}{\partial \beta} \log F(c_i \mid X_i, \beta) \;=\; \frac{F(c_i \mid X_i, \beta)}{f(c_i \mid X_i, \beta)} \cdot \mu_i X_i.$$

Finally, the score equations are given by

$$U(\beta) \;=\; \sum_{i=1}^{n} \left[I(Y_i > c_i) \frac{\mu_i \phi(Y_i \;-\; \mu_i)}{\nu (Y_i \;-\; \mu_i)} X_i \;+\; I(Y_i \leq c_i) \frac{F(c_i \mid X_i, \beta)}{f(c_i \mid X_i, \beta)} \cdot \mu_i X_i \right]. \tag{3.2}$$

Secondly, we perform a Taylor expansion of the log-likelihood around the true parameter β_0 at order 1:

$$\nabla_\beta \ell(\hat{\beta}_{\text{MLE}}) \;=\; \nabla_\beta \ell(\beta_0) \;+\; (\hat{\beta}_{\text{MLE}} - \beta_0) \nabla_\beta^2 \ell(\beta_0) \;+\; o_p(1),$$

and $\nabla_\beta \ell(\hat{\beta}_{\text{MLE}}) = 0$, we obtain:

$$0 \;=\; \nabla_\beta \ell(\beta_0) \;+\; \nabla^2_\beta \ell(\beta_0) \;+\; o_p(1),$$

which leads to :

$$\sqrt{n}(\hat{\beta}_{\text{MLE}} - \beta_0) \;=\; -\left(\frac{1}{n}\nabla^2_\beta \ell(\beta_0)\right)^{-1} \frac{1}{\sqrt{n}} \nabla_\beta \ell(\beta_0) \;+\; o_p(1).$$

Next , we apply the law of large numbers

$$\frac{1}{n}\nabla^2_\beta \ell(\beta_0) \xrightarrow{P} I(\beta_0), \tag{3.3}$$

where $I(\beta) \;=\; -\,E\left[\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T}\right]$ is the Fisher information.

In the context of a left-censored Gamma regression model, the expression of the Fisher information matrix depends on the contributions of the censored and uncensored observations. For the uncensored observations the log-likelihood function is given by:

$$\ell_i(\beta) \;=\; (\nu^{-1} \;-\; 1)\log Y_i \;-\; \frac{\mu_i \nu}{Y_i} \;-\; \nu^{-1}\log \mu_i,$$

where the first derivative is given by:

$$\frac{\partial \ell_i(\beta)}{\partial \beta} \;=\; \frac{\nu(Y_i \;-\; \mu_i)}{\mu_i} X_i,$$

and the second derivative is given by:

$$\frac{\partial^2 \ell_i(\beta)}{\partial\beta\partial\beta^T} \;=\; -\frac{\nu}{\mu_i^2}(Y_i \;-\; \mu_i)^2 X_i X_i^t,$$

and by taking the conditional expectation for an uncensored observation, the contribution to Fisher information becomes:

$$E\left[\frac{\partial^2 \ell_i(\beta)}{\partial\beta\partial\beta^T}\right] \;=\; \frac{\mu_i^2}{\nu} X_i X_i^t.$$

For censored observations the contribution to Fisher information comes from the cumulative distribution function of Gamma distribution $F(c_i \mid X_i, \beta)$ with,

$$\frac{\partial \log F(c_i \mid X_i, \beta)}{\partial \beta} \;=\; \frac{F(c_i \mid X_i, \beta)}{f(c_i \mid X_i, \beta)} \cdot \mu_i X_i,$$

then, the contribution of censored observations for Fisher information is given by:

$$E\left[\frac{F(c_i \mid X_i, \beta)}{f(c_i \mid X_i, \beta)} \mu_i X_i\right].$$

The complete matrix of Fisher information is given by:

$$I(\beta) \;=\; \sum_{i=1}^{n}\left[I(Y_i > c_i)\frac{\mu_i^2}{\nu} X_i X_i^t \;+\; I(Y_i \le c_i)\frac{F(c_i \mid X_i, \beta)}{f(c_i \mid X_i, \beta)} \cdot \mu_i X_i X_i^t\right]. \tag{3.4}$$

The main regularity assumptions needed to apply the central limit theorem are:

1. **Identifiability**: The parameter β_0 is identifiable,the likelihood function has a unique maximum at β_0.
2. **Regularity conditions**: The likelihood function is differentiable with respect to β and the derivatives satisfy certain integrability and continuity conditions.
3. **Independent observations**: the Observations Y_i are independent.
4. **Properly handled censoring**: The left-censoring model is correctly specified, and the proportion of censored observations is neither too large nor too small to allow asymptotic estimation.

Under this assumptions the normalized scores asymptotically follow a multivariate normal distribution:

$$\frac{1}{\sqrt{n}}\nabla_\beta \ell(\beta_0) \xrightarrow{d} N(0\ ,\ I(\beta_0)). \tag{3.5}$$

Combine the previous results, we obtain the asymptotic normality of $\hat{\beta}_{\text{MLE}}$,

$$\sqrt{n}(\hat{\beta}_{\text{MLE}} - \beta_0) \xrightarrow{d} N(0\ ,\ I(\beta_0)^{-1}). \tag{3.6}$$

This means that, for a sufficiently large sample n, the maximum likelihood estimator $\hat{\beta}_{\text{MLE}}$ asymptotically follows a normal distribution centered at β_0, with a covariance matrix given by the inverse of the Fisher information $I(\beta_0)^{-1}$.

4 Simulation Study

4.1 Uncensored gamma regression model

We simulated 1,000 replications with a sample size of $n = 100$ using R programming language. After that, we assume the true values of regression parameters $\beta = (0.2, -0.1, 0.4, 0.3, 0.5)$, vectors of covariates: an intercept, X_1 follows a Normal distribution with mean $m = 2$ and variance $\sigma^2 = 5$, X_2 follows a Normal distribution with mean $m = 1$ and variance $\sigma^2 = 3$, X_3 follows a Normal distribution with mean $m = 1$ and variance $\sigma^2 = 2$, X_4 follows a Normal distribution with mean $m = 3$ and variance $\sigma^2 = 6$ and a dependent variable Y following Gamma distribution with shape parameter $\nu = 1.5$ and scale μ/ν. Now, we will estimate the regression parameters using the maxLik function from the package "maxLik" , we obtain the following results:

Table 1: Parameter estimates for $n = 100$.

parameters	estimate	std.error	t value	$Pr(\geq t)$
ν	1.41573	0.18168	7.793	6.56e-15
β_1	0.26327	0.37580	0.701	0.484
β_2	-0.08718	0.07833	-1.113	0.266
β_3	0.37742	0.19826	1.904	0.047
β_4	0.28304	0.06603	4.286	1.82e-05
β_5	0.49369	0.10185	4.847	1.25e-06

Table 1 shows that the most significant parameters in our regression model are the shape parameter ν with (p-value $< .001$) and the coefficients parameters β_3, β_4 and β_5 with p-value equal respectively to 0.047, $< .001$ and $< .001$. The coefficients associated with β_1 and β_2 are not significant (p-value ≥ 0.05), suggesting that their respective explanatory variables do not have a notable effect on the dependent variable Y. We notice that the estimated parameters are close to the true values used for the simulation study and this is due to the absence of a censoring variable and the robustness of the likelihood function. This occurs because the maximum likelihood estimator (MLE) is consistent and efficient under the correct model specification. In other words, when the model assumptions are met and no data are missing (i.e., no censoring), the MLE is expected to yield estimates that are unbiased and converge to the true parameter values as the sample size increases.

4.2 Left-censored gamma regression model

We simulated 1,000 replications of a different sample sizes ranging from $n =$ 100–1,000. A censoring variable C quantile of 30%. After that, we assume the true values of regression parameters $\beta = (0.2, -0.1, 0.4, 0.3, 0.5)$, a vector of covariates: an intercept , X_1 follows a Normal distribution with mean m=2 and variance $\sigma^2 = 5$, X_2 follows a Normal distribution with mean $m = 1$ and variance $\sigma^2 = 3$, X_3 follows a Normal distribution with mean $m = 1$ and variance $\sigma^2 = 2$, X_4 follows a Normal distribution with mean $m = 3$ and variance $\sigma^2 = 6$ and a dependent variable Y following Gamma distribution with shape parameter $\nu = 1.5$ and scale μ / ν. Now, we will estimate the regression parameters using the maxLik function from the package "maxLik" under R we obtain the following results:

Table 2: Left-censored Gamma regression results with Bias and RMSE for different sample sizes.

Sample size	**Metric**	ν	β_1	β_2	β_3	β_4	β_5
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
	Bias	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	Coefficient	1.1613	0.1677	-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
	Bias	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	Coefficient	1.4453	0.2569	-0.1490	0.4576	0.3269	0.4197
	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
	Bias	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
1,000	Coefficient	1.4544	0.2268	-0.0712	0.4435	0.3236	0.4781
	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
	Bias	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

Table 2 shows that the shape parameter ν, varies between 2.3295 (for $n = 100$) and 1.45 (for $n = 1{,}000$), it is always significant (p-value ≤ 0.05), which indicates that the Gamma distribution is

well adapted to simulated data. As the sample size increases, the estimated value of ν approaches the simulated value ($\nu = 1.5$), showing a convergence to the true value.

β_1: varies between 0.2297 (for $n = 100$) and 0.2268 for ($n = 1{,}000$). Not significant for $n = 100$ and $n = 200$ (p-value ≥ 0.05), but significant for $n = 500$ (p-value = 0.0303). The impact of the first explanatory variable is small and only detected for large samples. In statistical terms, the power of a hypothesis test, the probability of detecting a real effect increases with the sample size. A small effect contributes a weak 'signal' relative to the inherent variability in the data. In small samples, this signal can be too faint to distinguish from zero. However, as the sample size grows, the precision of the estimate improves (evidenced by the decreasing standard error in Table 2), making it possible to reliably identify that the effect, while small, is genuinely different from zero. Also the presence of censoring, which introduces additional information loss and noise, underscoring the importance of adequate sample sizes in practical applications.

- β_2: varies between -0.2286 (for $n = 500$) and -0.0712 (for $n = 1{,}000$). Not significant for $n = 100$ and $n = 200$, but significant for $n = 500$ and $n = 1{,}000$ (p-value ≤ 0.05). The impact of the second explanatory variable is negative for large samples, which corresponds to the simulated value (-0.1).
- β_3: varies between 0.4231 (for $n = 100$) and 0.4435 (for $n = 1{,}000$). Significant for $n = 200$, $n = 500$ and $n = 1{,}000$ (p-value ≤ 0.05). The impact of the third explanatory variable is positive and approaches the simulated value (0.4) as sample size increases.
- β_4: varies between 0.2851 (for $n = 100$) and 0.3236 (for $n = 1{,}000$). Always significant (p-value ≤ 0.05). The impact of the fourth explanatory variable is strong and stable, close to the simulated value (0.3).
- β_5: varies between 0.5690 (for $n = 100$) and 0.4781 (for $n = 1{,}000$). Always significant (p-value ≤ 0.05). The impact of the fifth explanatory variable is strong and close to the simulated value (0.5) for large samples.

Standard errors decrease as the sample size increases, showing greater accuracy of estimates. For example, for β_4, the standard error changes from 0.0906 (for $n = 100$) to 0.0232 (for $n = 1{,}000$). The t-values increase with sample size, reflecting greater confidence in estimates. For example, for β_5, the value of t goes from 4.084 (for $n = 100$) to 12.722 (for $n = 1{,}000$). The p-values generally decrease with increasing sample size, making the coefficients more significant. For example, for β_2, the p-value changes from 0.0678 (for $n = 100$) to 0.0247 (for $n = 1{,}000$).

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. From Table 2, the biases are relatively low, except for ν (0.1914) and β_3 (-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 $n = 1{,}000$, the biases are almost equal to zero, confirming that the estimators are unbiased. The RMSE is very low, indicating high accuracy. The results found for bias and RMSE show that estimators converge to true parameter values as sample size increases. This is consistent with the asymptotic properties of maximum likelihood estimators.

5 Application in Car Insurance Data

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 67566 policies of which 4589 had at least one claim.

1. Veh-value: Vehicle value (in 10,000 dollars).
2. Exposure : Policy exposure (scaled between 0 and 1).
3. Clm: Occurrence of a claim (binary: 0 = no claim, 1 = at least one claim).
4. Numclaims: Number of claims per policy.
5. Claimcst0: Claim amount (0 if no claim occurred).
6. Veh-body: Vehicle body type, with categories such as BUS, COUPE, SEDAN, TRUCK, etc.
7. veh-age: Age of the vehicle (discrete: 1 = youngest, up to 4 = oldest).
8. Gender: F = Female and M = Male.
9. Area: Geographic area of the policyholder.
10. 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. 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 387.88 dollars (quartile of 30%) 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 387.88 dollars, but we do not know their exact values. We use the Anderson–Darling test to verify that the dependent variable "Claimcst0" follows a Gamma distribution and the "maxLik" function under R to estimate the coefficients of all covariates.

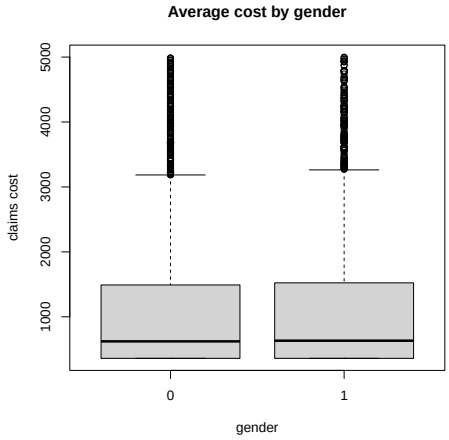
The p-value is 0.1289, which is higher than the typical significance threshold of 0.05, this means that we do not reject the null hypothesis and therefore the dependent variable "Claimcst0" follows a Gamma distribution. A test statistic of An=1.2365 is relatively small, suggesting the data is close to the hypothesized distribution.

Table 3: Coefficient Estimates, Standard Errors.

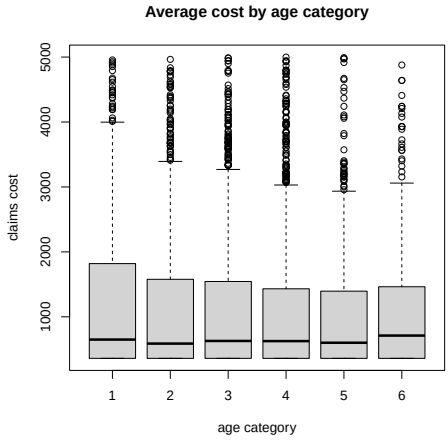
Variable	Estimate	Std. Error	t-value	$Pr \geq t$
shape	0.508482	0.011269	45.123	2.00e-16
Intercept	7.608142	0.101666	74.834	2.00e-16
Gender	0.177963	0.042356	4.202	2.65e-05
Veh-age	0.050602	0.022720	2.227	0.0259
Veh-value	0.002405	0.020056	0.120	0.9046
Agecat	-0.064805	0.014530	-4.460	8.19e-06
SEDAN	-0.095062	0.045773	-2.077	0.0378

Table 3 shows that the shape parameter $\nu = 0.50$ exhibits moderate dispersion around the average costs.

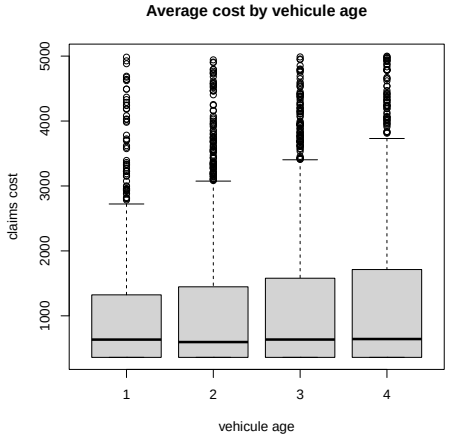
- **Gender (0.1779):** Significant positive effect (p-value 2.65e-06) indicating that male insured persons have higher average claim costs as shown in Figure 1.
- **Veh-age (0.0506):** Significant positive effect (p-value 0.0259) of vehicle age.
- **Veh-value (0.0024):** The lack of significance (p-value 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.



(a) The average Claim Costs by Gender.



(b) The average Claim Costs by Age-cat.



(c) The average Claim Costs by Vehicle age.

Figure 1: Claim Costs by Gender,Age category and Vehicle age.

5.1 Provisions for claims to be paid

The provision calculation under Gamma distribution was investigated in Mingari Scarpello, G., Ritelli, D., and Spelta, D (2006) but without taking into account the censoring variable and covariates. The left-censored Gamma model with covariates allows an accurate estimation of provisions for future costs taking into account both censored data and explanatory factors related to the characteristics of the policyholder. This approach is particularly useful for partially observed claims data.

Our model uses the following hypotheses: The dependent variable "Claimcst0" (claim costs) follows a Gamma distribution conditional on the covariates. Left censoring is applied for observations whose value is less than a threshold c. The covariates include: Gender, Veh-age, Veh-value, Agecat, and Sedan-indicator. For each observation, the provision is calculated as the conditional expectation of the remaining costs:

$$\text{Provision}_i \;=\; \frac{\nu\mu_i\left(1-F_\Gamma\left(c,\nu+1,\frac{\mu_i}{\nu}\right)\right)}{1-F_\Gamma\left(c,\nu,\frac{\mu_i}{\nu}\right)}, \tag{5.1}$$

where μ_i is the conditional mean predicted by the model for each observation, and ν is the shape parameter for the Gamma distribution, and c is the censoring threshold. $F_\Gamma\left(c,\nu,\frac{\mu_i}{\nu}\right)$ denotes the cumulative distribution function (CDF) of a Gamma distribution with shape parameter ν, and scale parameter $\frac{\mu_i}{\nu}$ evaluated at c. This equation adjusts the predicted mean μ_i to account for left-censoring. The numerator $F_\Gamma\left(c,\nu+1,\frac{\mu_i}{\nu}\right)$ modifies the mean based on the tail probability of a Gamma distribution with a slightly higher shape parameter $\nu+1$, while the denominator $F_\Gamma\left(c,\nu,\frac{\mu_i}{\nu}\right)$ normalizes this adjustment by the probability of exceeding the censoring threshold c. In essence, this provision represents an estimate of the latent (true) expected cost that corrects for the downward bias introduced by left-censoring. The results in Table 4 show some cost claims and its provisions.

Table 4: Cost Claims and Provisions.

Cost Claims	Provisions
1,250	884.76
1,452	1,017.78
1,505	1,052.10
1,830	1,258.37
2,145	1,452.97
2,300	1,547.32
2,500	1,769.85
2,950	2,089.58
3,230	2,275.32
3,500	2,455.76

Using equation (5.1) for each observation, we find that the mean of provisions equals 1,316.163 dollars which is a punctual estimation. We will perform Monte Carlo simulations to assess the sensitivity of provisions to variations in model parameters to quantify the uncertainty around the estimates and suggest robust confidence intervals for the provisions. We find that the confidence interval at 95% equal to:

$$[1,251.783, \quad 1,395.516]$$

The Monte Carlo simulation results indicate the uncertainty associated with the estimate of provisions. This credibility interval indicates that, given uncertainties on the model parameters, there is a 95% probability that the actual provision is between approximately 1,251.783 and 1,395.516 dollars.

5.2 The effect of censoring threshold on the provisions: stress scenario

The results in Table 5 show that each time the censoring threshold increases, the provisions decrease according to the value of each claim. For example, for a censoring threshold of 30%, the provisions are about 80% of the value of the cost claims. Still, from which we go to 40% of the censoring threshold, the value of the provisions is around 65 to 70% of the value of the claims. These results are due to the censoring variable affecting the model's parameter estimation.

Table 5: Censoring Threshold, Cost Claims and Provisions.

Censoring Threshold	Cost Claims	Provisions
Censoring Threshold 30% = 387.88 dollars	1,250	884.76
	1,452	1,017.78
	2,500	2,052.75
Censoring Threshold 40% = 500.00 dollars	1,250	800.92
	1,452	1,033.32
	2,500	1,711.47
Censoring Threshold 50% = 761.56 dollars	1,250	756.88
	1,452	867.99
	2,500	1,434.58

5.3 Random forest algorithm

We used Random Forest as a non-parametric benchmark because it flexibly captures non-linear interactions induced by left-censoring and provides robust variable-importance measures. These properties make it particularly suitable for isolating the impact of the censoring threshold without imposing restrictive distributional assumptions that alternative parametric methods would require.

The important output from the Random Forests model provides two key metrics for each variable:

1. %IncMSE: This measures the increase in the Mean Squared Error (MSE) of predictions when the variable is randomly permuted. A higher value indicates that the variable is more important for predicting claim costs or provisions.

2. IncNodePurity: This measures the total decrease in node impurity measured by the Gini index or variance reduction attributable to the variable across all trees. A higher value indicates greater importance.

We use the package "RandomForest" under R, to obtain the following results:

Table 6: %IncMSE and IncNodePurity.

Variables	%IncMSE	IncNodePurity
Gender	-2.374523	420,132,949
Veh-age	7.698457	724,164,082
Veh-value	7.861396	5,597,432,190
Agecat	3.755806	1,261,500,338
Sedan-indicator	-3.018224	240,584,035
Delta	82.696565	4,145,124,187

Table 6 shows that the variable Delta has the highest %IncMSE (82.69) and the second highest IncNodePurity (4,145,124,187). This indicates that the censoring indicator is the most important variable for predicting the cost claim and provisions (see Figure 2). This makes sense because delta directly indicates whether the claim cost is censored or not, which is a critical piece of information for the model.

- Agecat: this variable has a moderate %IncMSE (3.75) and a relatively high IncNodePurity (1,261,500,338), indicating that the age category of policyholder is important for predicting cost claims and provisions.
- Veh-value: this variable has the highest IncNodePurity (5,597,432,190) and a moderately high %IncMSE (7.86). This suggests that the value of the vehicle is a strong predictor of claim costs.
- Veh-age: with a %IncMSE of (7.69) and an IncNodePurity of 724,164,082, the age of the vehicle is also an important predictor.

Gender and Sedan-indicators have negative %IncMSE values and relatively low IncNodePurity scores, suggesting they contribute little to the model's predictive power.

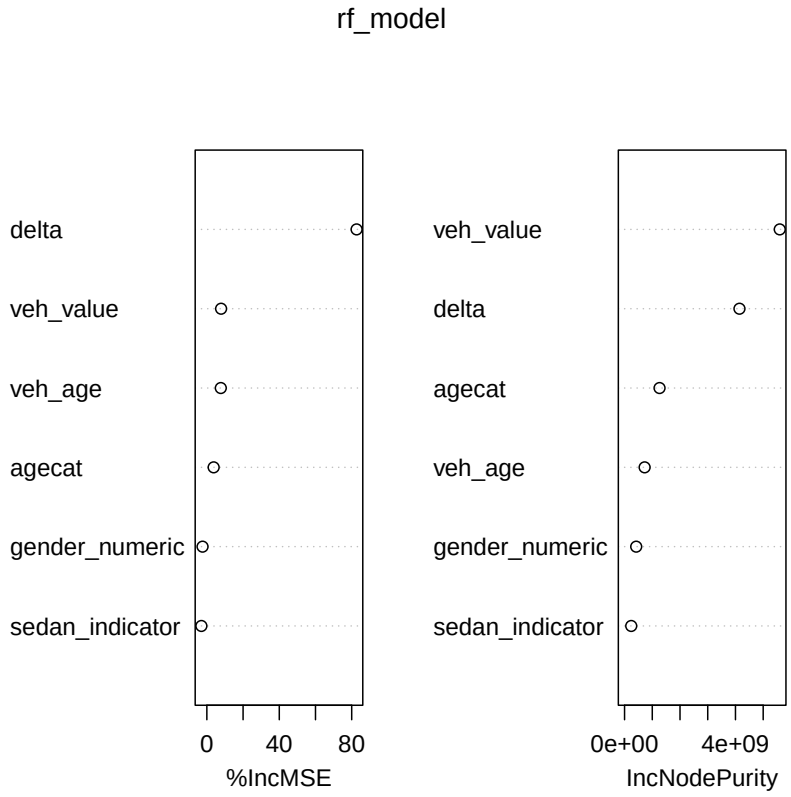


Figure 2: The most important variables.

6 Financial Risk Management

6.1 Value at risk (VaR) and expected shortfall (ES)

The Value at Risk (VaR) 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 see Rehman, M.Z., et al. (2024) for details. Value at Risk gives the probability of losing more than a given amount in a portfolio. The explicit expression for the VaR is given by:

$$\mathrm{VaR}_{\alpha}^{(i)} \;=\; \inf\left\{ x \in \mathbb{R} : \int_0^x \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu y}{\mu_i}\right) dy \geq \alpha \right\}. \tag{6.1}$$

This formulation ensures that the VaR reflects the heterogeneity induced by covariates through the scale parameter μ_i, making the risk measure observation-specific and data-driven.

The Monte Carlo simulation method for calculating the VaR is a flexible and powerful approach to modeling complex distributions and dependencies between assets. It is based on generating a large

number of future loss scenarios using an appropriate probability distribution and then extracting the VaR from these scenarios. It captures the worst case scenario in 95% of cases. Using this method under R programming language with 10,000 simulations, we find that the VaR equals to 5,092 dollars (see Figure 3). It indicates that the costliest claims can reach or exceed this amount. This reflects the variability and tail risk in our claim data. It helps us to determine the capital required to cover costly events. to suggest confidence interval for the VaR. We find that there is a 95% probability that the actual VaR is between $[4,800,\ 5,400]$. **The Expected Shortfall (ES)** is the average of losses that exceed the VaR. It provides a better idea of the magnitude of extreme events, unlike the VaR, which gives only the threshold. It is calculated using the following formula:

$$\text{ES} \ = \ \mathbb{E}[Cst|Cst > \text{VaR}],$$

where Cst, is the claim cost.

In our case, an ES of 7,580 dollars means that for claims that exceed the Var , the average cost of these extreme claims is 7,580 dollars. We are going to perform Monte Carlo simulations for a second time to suggest confidence interval for the ES. We find that there is a 95% probability that the actual ES is between $[7,200,\ 8,000]$. These measures allow us to better understand and manage the risks in our insurance portfolio.

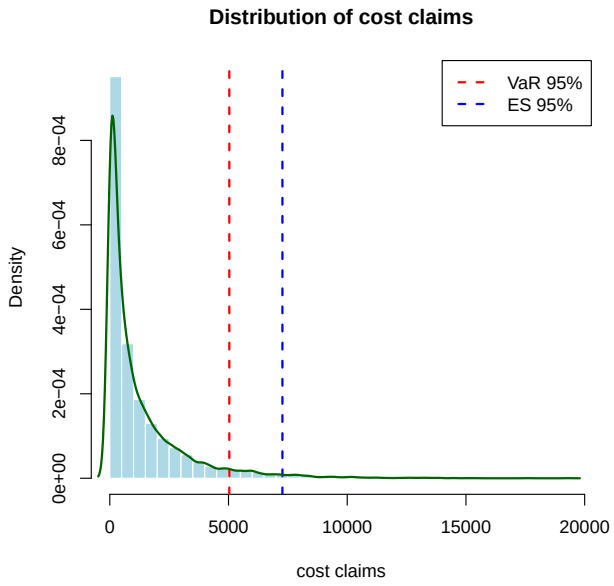


Figure 3: Distribution of Costs, VaR and ES.

6.2 Ratio VaR/mean and ratio ES/mean

Ratio VaR/mean : this ratio is calculated by dividing the Value-at-Risk (VaR) by the average of the simulated future losses. In our case it is equal to 3.97 which means that the 95th percentile (maximum loss in 95% of cases) is about 4 times higher than the average loss (1,269.585 dollars). This indicates that while most losses are relatively small, there is a significant likelihood of extreme losses being much higher than the average.

Ratio ES/mean : the ES/average ratio is calculated by dividing the Expected Shortfall (ES) by the average of the simulated future losses. In our case it is equal to 5.73 which means that in extreme cases, the average loss is almost 5.73 times higher than the overall average loss. This highlights the presence of a heavy tail in the loss distribution, indicating an extremely high risk.

6.3 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 Kupiec statistical test which uses a likelihood ratio to determine if the observed exception rate statistically deviates from the expected rate. It uses two hypotheses, H0: the model is correct (5% exceedances are expected) vs H1: the model is incorrect.

$$\mathrm{LR}_{uc} \;=\; -2\left[x\log\left(\frac{x}{n\cdot p}\right) \;+\; (n-x)\log\left(\frac{1-\frac{x}{n}}{1-p}\right)\right], \tag{6.2}$$

where x is number of exceedance, n is the number of days or observations, and $p = 0.05$. In our study, over 1,000 days, the loss exceeded the VaR 54 times (5.4%). The results are summarized in Table 7 and Figure 4. The difference of 0.4% is statistically insignificant, as confirmed by the

Table 7: The Backtesting Results.

observed rate	expected rate	the difference
5.4%	5%	0.4%

Kupiec test: LR = 0.33 and corresponding p-value = 0.56. In practice, this means that our model is valid for a confidence level of 95%. Test statistic (LR): 0.33 close to 0 low evidence against H_0. The p-value = $0.56 \geq 0.05$. Conclusion we accept H_0, the model is correct.

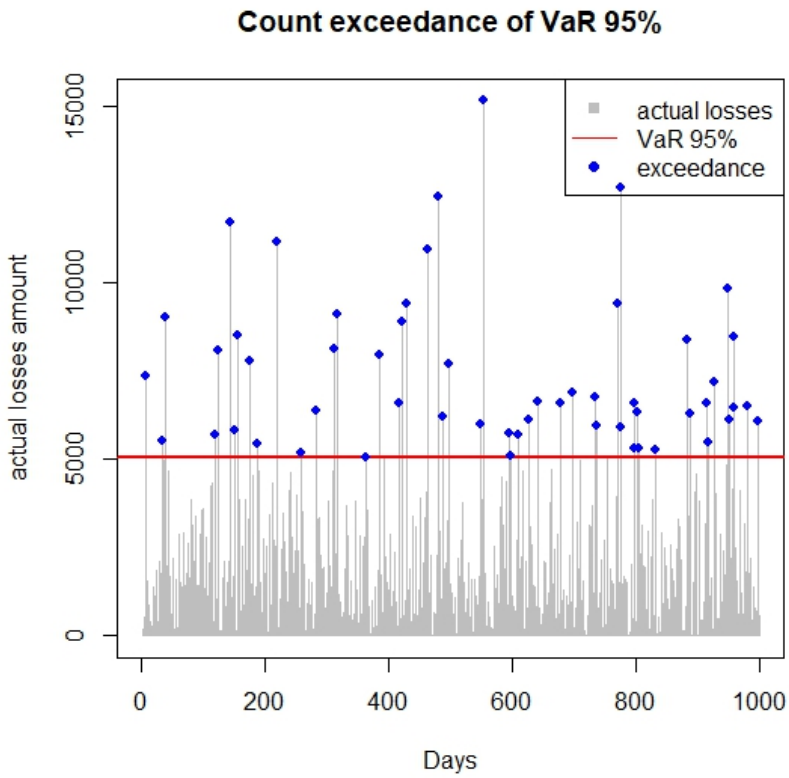


Figure 4: Exceedance Rate.

7 Discussion and Conclusion

The analysis of provisions for claims to be paid revealed that the left-censored Gamma regression model effectively adjusts for censoring bias by incorporating the conditional expectation of latent claim costs, ensuring accurate reserve estimation. The model's formula in equation (5.1) that integrates the Gamma cumulative distribution function, demonstrated that higher censoring thresholds (e.g., 30% to 50%) systematically reduce provisions by 15-30%, reflecting the downward bias introduced by unreported low-value claims. Crucially, the Random Forest algorithm identified the censoring indicator (Delta) as the most influential predictor (%IncMSE = 82.7), emphasizing its dual statistical and financial significance, it drives both the accuracy of parameter estimates and the reliability of risk metrics. In risk management, Monte Carlo simulations yielded a 95% VaR of 5,092 dollars and an ES of 7,580 dollars, capturing tail risk, and informing capital adequacy requirements. Backtesting using the Kupiec test confirmed the VaR model's validity, with a 5.4% exception rate (vs. 5% expected), statistically insignificant deviation (p-value = 0.56), and no evidence of

systematic risk underestimation. These results highlight how left-censoring complicates actuarial and financial analyses, necessitating tailored methodologies to mitigate bias in reserve calculations, capital planning, and regulatory reporting.

In this paper, we demonstrated the efficacy of a left-censored Gamma regression model for analyzing insurance claims with partial observability, validated through theoretical proofs, simulations, and empirical application. The model's maximum likelihood estimators exhibited asymptotic normality, with simulations confirming unbiasedness and accuracy as sample sizes increased. We have shown through the empirical study that the framework quantified covariate impacts, calculated provisions, and derived risk metrics while backtesting confirmed the VaR model's reliability. We used Random Forest algorithm to illustrate the critical role of censoring indicators in prediction. The approach offers insurers a robust reserve estimation and risk management tool in censored data environments, bridging statistical rigor with practical applicability. In the next study, we will see the effect of truncated variables under incomplete data on the estimation of provisions and financial risk management.

References

Scornet, E., Biau, G., and Vert, J. P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4), 1716–1741. https://doi.org/10.1214/15-AOS1321.

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

Haddad, K., Rahman, A., and Green, J. (2011). Design rainfall estimation in Australia: A case study using L-moments and generalized least squares regression. *Stochastic Environmental Research and Risk Assessment*, 25, 815–825. https://doi.org/10.1007/s00477-011-0460-1.

Kjeldsen, T. R., and Jones, D. A. (2004). Sampling variance of flood quantiles from the generalized logistic distribution estimated using the method of L-moments. *Hydrology and Earth System Sciences*, 8, 183–190. https://doi.org/10.5194/hess-8-183-2004.

Austin, P. C., Lee, D. S., and Fine, J. P. (2016). Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133, 601–609. https://doi.org/10.1161/CIRCULATIONAHA.115.017719.

Moscovici, J. L., and Ratitch, B. (2017). Combining Survival Analysis Results after Multiple Imputation of Censored Event Times. In *Proceedings of PharmaSUG 2017* - Paper SP05.

McCullagh, P., and Nelder, J. A. (1989). Generalized Linear Models.

Dunn, P., and Smyth, G. (2018). Generalized Linear Models with Examples in R. *Springer*. https://doi.org/10.1007/978-1-4419-0118-7.

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

Myers, R. H., Montgomery, D. C., and Vining, G. G. (2012).Generalized Linear Models with Applications in Engineering and the Sciences. *Wiley Series in Probability and Statistics*. https://doi.org/10.1002/9780470556986.

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

Gajewski, B. J., Nannette, N., and Widen, J. E. (2012). Predicting Hearing Threshold in Nonresponsive Subjects Using a Log-Normal Bayesian Linear Model in the Presence of Left-Censored Covariates. https://doi.org/10.1198/sbr.2009.0015.

Lawless, J. F. (2003). Statistical Models and Methods for Lifetime Data. *Wiley Series in Probability and Statistics.* https://doi.org/10.1002/9781118033005.

Dupuy, J. F. (2022). Censored Gamma Regression with Uncertain Censoring Status. *Mathematical Methods of Statistics*. https://doi.org/10.3103/S106653072004002X.

Wang, B.-B., Li, C.-G., Liu, P., Latengbaolide, A., and Yang, L. (2011). Log-normal censored regression model detecting prognostic factors in gastric cancer: A study of 3018 cases.*World Journal of Gastroenterology*, 17(23), 2867–2872. https://doi.org/10.3748/wjg.v17.i23.2867.

Mingari, S. G., Ritelli, D., and Spelta, D. (2006). Actuarial values calculated using the incomplete Gamma function. *Statistica*, 66(1), 77–84. https://doi.org/10.6092/issn.1973-2201/449.

Lauar, A., Boukhetala, K., and Sabre, R. (2024). Statistical analysis of stable distribution application in non-life insurance. *Finance: Theory and Practice*, 28(5). https://doi.org/10.26794/2587-5671-2024-28-5-146-155.

Rehman, M. Z., Nain, M. Z., Alhashim, M., and Bhat, J. A. (2024). Choice between sustainable versus conventional investments: Relative efficiency analysis. *Sustainability*. https://doi.org/10.3390/su16135340.

Levy, P. (1925). Calcul des probabilites. Gauthier-Villars, Paris.

Nguyen, V. T., and Dupuy, J. F. (2021). Asymptotic results in censored zero-inflated regression model. *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610926.2019.1676442.

Sarul, L. S., and Sahin, S. (2015). An application of claim frequency data using zero-inflated and hurdle models in general insurance. *Journal of Business, Economics and Finance*. https://doi.org/10.17261/Pressacademia.2015414539.

Bae, S., Famoye, F., Wulu, J. T., Bartolucci, A., and Singh, K. P. (2005). A rich family of generalized Poisson regression models with applications. *Mathematics and Computers in Simulation*. https://doi.org/10.1016/j.matcom.2005.02.026.

Adesina, O. S., Dare, R. J., and Famurewa, O. K. (2018). Using R for Actuarial Analysis in Valuation and Reserving. *Annals. Computer Science Series*.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). Regression: Models, Methods and Applications. *Springer*. https://doi.org/10.1007/978-3-642-34333-9..

Received: April 28, 2025

Accepted: November 3, 2025

A Appendix

A.1 R Code for Simulation study

```
# For uncenred Gamma Regression
resultat=matrix(0,nrow=1000,ncol=6)
for(i in 1:1000){   # Number of simulations (replicates)
n=100  #Sample size in each simulation
n=100
b=c(0.2,-0.1,0.4,0.3,0.5)
nu=1.5
inter=rep(1,n)
 X1=rnorm(n,2,5)
 X2=rnorm(n,1,3)
 X3=rnorm(n,1,2)
 X4=rnorm(n,3,6)
 X=rbind(inter,X1,X2,X3,X4)
mu=exp(t(b)%*%X)
Y=rgamma(n, shape = nu, scale = mu/nu)
gamma.non.cens=function(param){ # MLE calculation - no censoring
shape=param[1]
beta=param[-1]
sum((shape-1)*log(df) - shape*covariates%*%beta -shape*df*exp(-covariates
    %*%beta) + shape*log(shape) - log(gamma(shape)))
}
 gamma.non.cens=function(param){
nu=param[1]
beta=param[-1]
sum((nu-1)*log(Y) - nu*t(beta)%*%X -nu*Y*exp(-t(beta)%*%X) + nu*log(nu) -
     log(gamma(nu)))
}
## Estimate it with beta=c(1,1,1,1,1,1) as start values
```

```
ml <- maxLik(gamma.non.cens, start =c(1,1,1,1,1,1,1,1,1,1))
print(summary(ml))

resultat[i, ]=coef(maxLik(gamma.non.cens, start =c(1,1,1,1,1,1)))
}
##calculation of bias
colMeans(resultat)
##calculation of nu bias
biais_nu = mean(resultat[ ,1])-1.5
##calculation  RMSE of nu
rmse_nu= mean((resultat[ ,1]-1.5)^2)

# Left-censored Gamma regression
resultat=matrix(0,nrow=1000,ncol=6)
for(i in 1:1000){ # Number of simulations (replicates)
n=100  #Sample size in each simulation
b=c(0.2,-0.1,0.4,0.3,0.5)
nu=1.5
inter=rep(1,n)
 X1=rnorm(n,2,5)
 X2=rnorm(n,1,3)
 X3=rnorm(n,1,2)
 X4=rnorm(n,3,6)
 X=rbind(inter,X1,X2,X3,X4)
mu=exp(t(b)%*%X)
Y=rgamma(n, shape = nu, scale = mu/nu)
c=quantile(Y, 0.3)
 Yc=Y*(Y>c)+c*(Y<=c)
delta=as.integer(Y>c)
gamma.cens=function(param){ # MLE calculation - censored case
nu=param[1]
beta=param[-1]
sum(delta*((nu-1)*log(Yc) - nu*t(beta)%*%X -nu*Yc*exp(-t(beta)%*%X)
 + nu*log(nu) - log(gamma(nu))) +
(1-delta)*log(pgamma(Yc, shape = nu, scale = exp(t(beta)%*%X)/nu)))}

## Estimate it with beta=c(1,1,1,1,1,1) as start values
ml1 <- maxLik(gamma.cens, start =c(1,1,1,1,1,1))
print(summary(ml1))
resultat[i, ]=coef(maxLik(gamma.cens, start =c(1,1,1,1,1,1)))
}

##calculation bias of nu
biais_nu = mean(resultat[ ,1])-1.5
##calculation  RMSE of nu
rmse_nu= mean((resultat[ ,1]-1.5)^2)
```

A.2 R Code for Real Data Application

```
library(insuranceData)
library(maxLik)

#Load data
data("dataCar")

# Transformation of gender to numeric
dataCar$gender_numeric <- ifelse(dataCar$gender == "M", 1, 0)

# Creation of indicatrice for veh_body == "SEDAN"
dataCar$sedan_indicator <- ifelse(dataCar$veh_body == "SEDAN", 1, 0)

# selected variables
selected_vars <- c("claimcst0", "gender_numeric", "veh_age",
"veh_value", "agecat", "sedan_indicator")
dataCar_filtered <- na.omit(dataCar[selected_vars])
dataCar_filtered <- subset(dataCar_filtered, claimcst0 > 0)


# Matrice of covariates with intercept
X <- as.matrix(cbind(1, dataCar_filtered[, c("gender_numeric", "veh_age",
     "veh_value", "agecat", "sedan_indicator")]))

#  dependante variable
Y <- dataCar_filtered$claimcst0
head(Y)
# Generation of  censoring threshold
set.seed(123)
c <- quantile(Y, 0.30)

# Application of censoring variable
Yc <- pmax(ifelse(Y > c, Y, c), 1e-10)  # for log(0)
delta <- as.integer(Y > c)  # Indicator of censoring

# log-likelihood function of censored Gamma
gamma_cens <- function(param) {
  nu <- param[1]           # shape parameter
  beta <- param[-1]        # Coefficients of covariates

  # Calcul de mu
  eta <- X %*% beta
  mu <- exp(eta)

  # Verification of erreurs
  if (nu <= 0 || any(mu <= 0)) return(-1e10)
```

```
  # Calculation of log-likelihood
  logLik <- delta * ((nu - 1) * log(Yc) - nu * eta - nu * Yc / mu +
                       nu * log(nu) - lgamma(nu)) +
              (1 - delta) * log(pgamma(Yc, shape = nu, scale = mu / nu))

  return(sum(logLik, na.rm = TRUE))
}


init_params <- c(1, rep(1, ncol(X)))

# Estimation  with maxLik
ml1 <- maxLik(logLik = gamma_cens, start = init_params)

# print results
print(summary(ml1))

# Box plot visualisation for age category
boxplot(Yc ~ agecat, data = dataCar_filtered, main = "Average␣cost␣by␣
    gender",
        xlab = "gender", ylab = "claims␣cost")

# calculation of provisions
library(stats)
# given parameters for claimcst=3500$
nu <- 0.50
mu_i <- 3500
c <- 384.058
# Calculation of Gamma density functions
F_gamma_nu1 <- pgamma(c, shape = nu + 1, scale = mu_i / (nu + 1))
F_gamma_nu <- pgamma(c, shape = nu, scale = mu_i / nu)
provision <- (nu * mu_i * (1 - F_gamma_nu1)) / (1 - F_gamma_nu)
cat("Valeur␣de␣la␣provision␣:", provision, "\n")

# For stress scenario we keep the same R code and we change only the
    censoring threshold.

# Random Forest Algorithm
install.packages("randomForest")
library(randomForest)
data("dataCar")
dataCar$gender_numeric <- ifelse(dataCar$gender == "M", 1, 0)
dataCar$sedan_indicator <- ifelse(dataCar$veh_body == "SEDAN", 1, 0)
selected_vars <- c("claimcst0", "gender_numeric", "veh_age", "veh_value",
     "agecat", "sedan_indicator")
dataCar_filtered <- na.omit(dataCar[selected_vars])
dataCar_filtered <- subset(dataCar_filtered, claimcst0 > 0)
X <- dataCar_filtered[, c("gender_numeric", "veh_age", "veh_value", "
    agecat", "sedan_indicator")]
```

```
Y <- dataCar_filtered$claimcst0
set.seed(123)
c <- quantile(Y, 0.30)
Yc <- pmax(ifelse(Y > c, Y, c), 1e-10)
delta <- as.integer(Y > c)
# Add delta as a feature to X
X$delta <- delta
# Split into training and testing sets
set.seed(123)
train_indices <- sample(1:nrow(X), 0.8 * nrow(X))
X_train <- X[train_indices, ]
X_test <- X[-train_indices, ]
Y_train <- Yc[train_indices]
Y_test <- Yc[-train_indices]
# Train the Random Forest model
rf_model <- randomForest(x = X_train, y = Y_train, ntree = 500,
    importance = TRUE)
# Make predictions on the test set
rf_predictions <- predict(rf_model, X_test)
# Evaluate the Random Forest model
rf_mse <- mean((Y_test - rf_predictions)^2)
cat("Random␣Forest␣Test␣MSE:", rf_mse, "\n")
# Compare with gamma regression results
cat("Gamma␣Regression␣Log-Likelihood:", logLik(ml1), "\n")
# Variable importance plot
importance(rf_model)
varImpPlot(rf_model)

# Calculation of VaR and ES
nu_hat <- 0.50
mu_hat <- 1316       # estimated mean provision
n_sim <- 10000      # Number of simulations
scale = mu/nu :
scale_param <- mu_hat / nu_hat
sim_losses <- rgamma(n_sim, shape = nu_hat, scale = scale_param)
VaR_95 <- quantile(sim_losses, 0.95)
ES_95 <- mean(sim_losses[sim_losses > VaR_95])
cat("Value-at-Risk␣(95%)␣:", VaR_95, "\n")
cat("Expected␣Shortfall␣(95%)␣:", ES_95, "\n")

# Graphical visualization
hist(sim_losses, breaks = 50, col = "lightblue", border = "white",
     main = "Distribution␣of␣cost␣claims",
     xlab = "cost␣claims",freq= FALSE)
abline(v = VaR_95, col = "red", lwd = 2, lty = 2)
abline(v = ES_95, col = "blue", lwd = 2, lty = 2)
legend("topright", legend = c("VaR␣95%", "ES␣95%"),
       col = c("red", "blue"), lty = 2, lwd = 2)
```

```r
# Ratio VaR/moyenne et ES/moyenne
mean_loss <- mean(sim_losses)
ratio_VaR <- VaR_95 / mean_loss
ratio_ES <- ES_95 / mean_loss
cat("Ratio VaR/moyenne :", ratio_VaR, "\n")
cat("Ratio ES/moyenne :", ratio_ES, "\n"

# Backtesting of the VaR
set.seed(456)  # Different seed for real losses
n_days <- 1000  # Number of historical days
# Simulate real losses (same Gamma distribution)
actual_losses <- rgamma(n_days, shape = nu_hat, scale = scale_param)
# 1. Count VaR overruns
exceedances <- actual_losses > VaR_95
count_exceedances <- sum(exceedances)
exceedance_rate <- mean(exceedances)
# 2. Kupiec Statistical Test
test_kupiec <- function(n, k, alpha = 0.05) {
  lr <- -2 * (log((1 - alpha)^(n - k) * alpha^k) - log((1 - k/n)^(n - k)
      * (k/n)^k))
  p_value <- 1 - pchisq(lr, df = 1)
  return(list(lr = lr, p_value = p_value))
}
result <- test_kupiec(n_days, count_exceedances)
# 3. Results
cat("Backtesting results:\n")
cat("---------------------------------\n")
cat("VaR 95% calculated :", VaR_95, "\n")
cat("Observed overruns :", count_exceedances, "/", n_days, "\n")
cat("Exceedance rate :", round(exceedance_rate * 100, 2), "%\n")
cat("Statistique LR (Kupiec) :", round(result$lr, 2), "\n")
cat("p-value :", result$p_value, "\n")
# 4. Graphical visualization
plot(actual_losses, type = "h", col = "grey",
     main = "Count exceedance of VaR 95%",
     xlab = "Days", ylab = "actual losses amount")
abline(h = VaR_95, col = "red", lwd = 2)
points(which(exceedances), actual_losses[exceedances],
       col = "blue", pch = 19, cex = 0.8)
legend("topright",
       legend = c("actual losses", "VaR 95%", "exceedance"),
       col = c("grey", "red", "blue"),
       pch = c(15, NA, 19),
       lty = c(0, 1, 0))
```