**International Journal of Theoretical and Applied Mathematics**
2021; 7(6): 92-103
http://www.sciencepublishinggroup.com/j/ijtam
doi: 10.11648/j.ijtam.20210706.12
ISSN: 2575-5072 (Print); ISSN: 2575-5080 (Online)

# Stochastic Modelling of Claim Frequency in the Ethiopian Motor Insurance Corporation: A Case Study of Hawassa District in Ethiopia

**Mikiyas Gebresamuel Gebru**

Department of Mathematics, Arba Minch University, Arba Minch, Ethiopia

**Email address:**
mikiyas.gebresamuel@amu.edu.et

**To cite this article:**
Mikiyas Gebresamuel Gebru. Stochastic Modelling of Claim Frequency in the Ethiopian Motor Insurance Corporation: A Case Study of Hawassa District in Ethiopia. *International Journal of Theoretical and Applied Mathematics.* Vol. 7, No. 6, 2021, pp. 92-103.
doi: 10.11648/j.ijtam.20210706.12

**Received**: February 18, 2021; **Accepted**: November 11, 2021; **Published**: February 5, 2022

**Abstract:** The objective of this study is to model seasonal variations in claim intensities and to evaluate the dependency of covariates on claim rates. The data for this study are obtained from claimants registered during September 2009 to August 2011, both inclusive at the Ethiopian Insurance Corporation in Hawassa. We present a procedure for consistent estimation of the claim frequency for motor vehicles in the Ethiopian Insurance Corporation, Hawassa District. The seasonal variation is modeled with a non-homogeneous Poisson process with a time varying intensity function. Covariates of the policy holders, like gender and age, is corrected for in the average claim rate by Poisson regression in a GLM setting. An approximate maximum likelihood criterion is used when estimating the model parameters. Numerical simulations are carried out applying the numerical software Matlab. These simulations show the reliability of the estimator. The seasonal parameters are found to be ±25% and to be statistically significant. February has highest while August has lowest claim rate. Only age group 36 - 45 has significantly lower claim rate than age group 20 - 25. The rate is about one third. Lastly female is not found to have significantly lower claim rates than males, however, there are indications that might be slightly so.

**Keywords:** Non-Homogeneous Poisson Process, Claim Intensity, Seasonal Model, Maximum Likelihood Estimation

## 1. Introduction

An insurance company has a portfolio of customers. Some of them will never make a claim, while others might make one or multiple claims. Insurance companies are constantly looking for ways to better predict claims [4]. Overall, insurance involves the sum of a large number of individual risks of which very few will result in insurance claims being made. It is argued that insurance companies need to treat risk management as a series of related factors and events. Therefore, an insurance company has to find ways to predict claims and appropriately charge a premium to cover risks.

Ethiopia's insurance industry which was established on 1976 is relatively undeveloped and is exemplified by the sectors low penetration levels. There are indications that the insurance companies in Ethiopia are characterized by high claim frequency and cost [12]. According to Sisay Wuyu and Patrick Cerna [12], there is a well-known problem in the Ethiopian Insurance Corporation concerning the proper pricing of an insurance policy. Kanbiro and Ayneshet [13] studied on Factors Affecting Financial Performance of Insurance Companies Operating in Hawassa City Administration, Ethiopia. They concluded that financial performance of insurance companies operating in Hawassa was best explained by the explanatory variables, risks included in their model.

One of the most widely used and accepted models for claim count processes is the Poisson process model. Over the years, numerous variants of the classical Poisson model have been proposed in hopes of improving its adequacy and validity in a broader range of contexts. These variants allow the Poisson process model to be utilized for non-homogeneous Poisson Model (NHPM), seasonal data. Nelder, J. and R [9] and Antonio, K. and Beirlant, J [1-3] proposed application of Generalized Linear Models and Log-normal mixed models for claim reserving respectively. Claims have long been estimated using a pure algorithmic technique or a simple stochastic

technique [11]. These methods result in poor estimations. Huang, Zhao and Tang [10] consider a risk model in which the claim number process is treated as a Poisson model and the individual claim size is assumed to be a fuzzy random variable. Jørgensen and Souza [7] suggested a Poisson sum of Gamma random variables called Tweedie to estimate insurance risk. According to Smyth and Jørgensen [5], there is also another problem in that the proposed Tweedie model does not permit the separate estimation of probability and claim size. Claro, P, Caetano, L., Artes, R [6] used to estimate insurance claims from an auto dataset using the Tweedie and zero-adjusted inverse Gaussian (ZAIG) methods.

Some recent papers have also studied on the factors influencing profitability of insurance companies [15, 16] and on the performance of insurance companies in Ethiopia [14].

The motivation behind these proposed models is to better reflect a real-world system without sacrificing the tractability of the classical Poisson process model. In examining the nature of the risk associated with a portfolio, it is often of interest to assess the frequency of claims on each month. One approach concerns the use of seasonal variation. This quantity, referred to as insurers risk, varies in time. None of the studies seem to discuss the seasonal variation of the Claim intensities, and how models can be constructed from the historical data, apart from general statements such as suggesting that the insurance industries in Ethiopia need to treat risk management to better predict claims.

The focus of this study is to model claim frequency and to see the trend of seasonal variations in Ethiopian non-life Insurance Corporation. Specific objectives are: Formulation of stochastic model for claim intensity, to develop a scheme for how parameters are identified, to evaluate the dependency of covariates using Poisson regression on claim rates and to model seasonal variations in claim intensities which require non homogeneous Poisson model.

# 2. Data Sets and Methods

### *2.1. Data Sets*

This study is based on secondary data obtained from a motor portfolio in Ethiopian Insurance Corporation at Hawassa district, Ethiopia. It is collected during a period of 36 months starting September 2017 and ending August 2019, both inclusive. Several covariates representing the drivers have been used. The data set contains only claims on a single coverage of the comprehensive part of the motor insurance.

The following variables are considered in this study.

Independent categorical variables:

Age of driver: categorized as:

$20-25$.

$26-35$

$36-45$

$46+$

Sex of driver:

Male, Female

Continuous independent variable

Exposure time: length of the exposure measured in year.

Dependent variables

Number of claims: Mostly equals 0, but sometimes 1 or more.

Accident month of claims

The data is an extract of three years of totally $n=1553$ policies, for which the number of claims is 157 with total exposure $T_1+...+T_n=844$ portfolio years. Among the 1553 polices with 1501 customers 10.5% are involved in claims, see Table 1.

In order to compare the impact of each variable on the number of claims, the mean number of claims with respect to the levels of each variable was calculated, see Tables 2 and 3.

Note that males in age-class $26-35$ is dominating the claim numbers, and that few females produce claims, see Table 2. Moreover, most claims are reported in the dry warm season January and February, see Table 3.

***Table 1.*** *Existence of claims.*

| | Frequency | Percent |
|---|---|---|
| No claim | 1344 | 89.5 |
| One or more claim | 157 | 10.5 |
| Total | 1501 | 100.0 |

***Table 2.*** *Number of claims in age/sex classes.*

| **Age classes** | | | | | |
|---|---|---|---|---|---|
| **Sex** | **20-25** | **26-35** | **36-45** | **46+** | **Total** |
| Male | 10 | 71 | 41 | 25 | 147 |
| Female | 0 | 6 | 4 | 0 | 10 |
| Total | 10 | 77 | 45 | 25 | 157 |

The data are stored on the following form:

| covariates | claims | exposure time | acc.month |
|---|---|---|---|
| $x_{11}\dots x_{1m}$ | $n_1$ | $T_1$ | $t_{11}\dots t_{1n_1}$ |
| $x_{21}\dots x_{2m}$ | $n_2$ | $T_2$ | $t_{21}\dots t_{2n_2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_{n1}\dots x_{nm}$ | $n_n$ | $T_n$ | $t_{n1}\dots t_{nn_n}$ |

On row $j$ we have the values of the explanatory variables $x_{j1},\dots,x_{jm}$, the number of claims $n_j$, the exposure time to risk $T_j$ as proportion of year and the months of the $n_j$ accidents $t_{j1},\dots,t_{jn_j}$. This is known as the data matrix.

***Table 3.*** *Number of claims for covariates.*

| **Variables** | **Covariates** | **Number of claims** | **Percent** |
|---|---|---|---|
| Sex of driver | Male | 147 | 93.6 |
| | Female | 10 | 6.4 |
| | Total | 157 | 100.0 |
| Age of driver | 20-25 | 10 | 6.8 |
| | 26-35 | 77 | 49.0 |
| | 36-45 | 45 | 27.2 |
| | 46+ | 25 | 17.0 |
| | Total | 157 | 100.0 |

| Variables | Covariates | Number of claims | Percent |
|---|---|---|---|
| Accident month | January | 28 | 18.1 |
| | February | 18 | 11.6 |
| | March | 12 | 7.7 |
| | April | 10 | 6.5 |
| | May | 14 | 8.4 |
| | June | 9 | 5.8 |
| | July | 18 | 11.0 |
| | August | 8 | 4.8 |
| | September | 9 | 6.1 |
| | October | 12 | 8.2 |
| | November | 10 | 6.8 |
| | December | 9 | 6.1 |
| | Total | 157 | 100.0 |

### *2.2. Methods*

The variables relevant for our study are:
Number of claims in a year - $N$,
Occurrence of claims - $(t_1,...,t_N)$,
Exposure time of a customer in a year - $T$, $0<T<1$,
Covariates: driver age and gender - $x$.

#### *2.2.1. The Poisson Process*

In this section we consider the most common claim number process, the Poisson process. Recall that an integer-valued random variable $N$ is said to have a Poisson distribution with parameter $\lambda>0$, $(N\sim \text{Pois}(\lambda))$ if it has distribution

$$p(N=k)=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,\ldots \tag{1}$$

A stochastic process $N=(N(t))_{t\geq 0}$ is said to be a Poisson process if the following conditions hold:

The process starts at zero

$$N(0)=0$$

The process has independent increments: for any $t_i$, $i=1,\ldots,n,$ such that $0\leq t_1<\ldots<t_n$, the increments $N(t_{i-1},t_i)=N(t_i)-N(t_{i-1}), i=1,...,n$ are mutually independent.

There exists $\mu(t)$, $\mu(0)=0$ such that the increments $N(s,t]$ have a Poisson distribution $\text{Pois}(\mu[s,t])$ with $\mu(s,t)=\mu(t)-\mu(s)$. We call $\mu(t)$ the mean value function of $N$.

The mean value function $\mu(t)$ can be considered as an inner clock or operational time of the counting process N. Depending on the magnitude of $\mu(t)$ in the interval $(s,t]$, it determines the model for the random increment $N(s,t)$.

Since $N(0)=0$ and $\mu(0)=0$,

$$\begin{aligned} N(t) &= N(t)-N(0) \\ &\sim \text{Pois}(\mu(0,t)) \\ &= \text{Pois}(\mu(t)) \end{aligned} \tag{2}$$

In summary, the claim numbers, $N$ for policies is Poisson distributed for the time interval $(0,T]$ see (1). The parameters being linear in time provide $\mu(T)=\lambda T$.

Recall that the number of claims $n$ observed in a time interval $(0,T]$, is distributed as:

$$Pr(N=n)=\frac{(\lambda T)^n}{n!}e^{-\lambda T}.$$

If there are observations $O=[(n_1,T_1),\ldots,(n_n,T_n)]$, then the likelihood estimation will be:

$$\begin{aligned} L(O;\lambda) &= \prod_{i=1}^{n}\frac{(\lambda T_i)^{n_i}}{n_i!}e^{-\lambda T_i} \\ &= \lambda^{\sum_{i=1}^{n} n_i}\cdot\prod_{i=1}^{n}\frac{T_i^{\,n_i}}{n_i!}\cdot e^{-\lambda\sum_{i=1}^{n} T_i}, \end{aligned}$$

with log-likelihood given by:

$$\begin{aligned} L(O;\lambda) &= -\lambda\sum_{i=1}^{n}T_i+\sum_{i=1}^{n}n_i\log\lambda \\ &\quad +\sum_{i=1}^{n}\log\left(\frac{T_i^{n_i}}{n_i!}\right) \end{aligned}$$

and optimized by:

$$\frac{\partial}{\partial\lambda}\log L(O;\lambda)=-\sum_{i=1}^{n}T_i+\sum_{i=1}^{n}n_i\frac{1}{\lambda}=0$$

This implies that:

$$\hat{\lambda}=\frac{\sum_{i=1}^{n}n_i}{\sum_{i=1}^{n}T_i} \tag{3}$$

Note that:

$$E\left[\hat{\lambda}\right]=\frac{\sum_{i=1}^{n}E[N_i]}{\sum_{i=1}^{n}T_i}=\frac{\sum_{i=1}^{n}\lambda T_i}{\sum_{i=1}^{n}T_i}=\lambda \tag{4}$$

which is unbiased, and

$$Var\left[\hat{\lambda}\right]=\frac{\sum_{i=1}^{n}\text{Var}(N_i)}{\left(\sum_{i=1}^{n}T_i\right)^2}=\frac{\sum_{i=1}^{n}\lambda T_i}{\left(\sum_{i=1}^{n}T_i\right)^2}=\frac{\lambda}{\sum_{i=1}^{n}T_i}. \tag{5}$$

#### *2.2.2. The Poisson Regression Model*

Insurance companies want to solve the problem of linking risk to explanatory variables (or covariates). This helps them to understand which customers are profitable and which are not and to charge differently in different customer segments of the corporation [8].

The method used in practice is Poisson regression where the claim intensity $\mu^x$ is 'explained' by a set of observable variables $x_1,...,x_m$ through a relationship of the form:

$$\log\left(\mu^x\right)=\beta_0+\beta_1x_1+...+\beta_mx_m \tag{6}$$

with $\beta_0,...,\beta_m$ being coefficients and the link between $\mu$ and the explanatory variables $(x_1,...,x_m)$ allows us to discriminate customers according to the risk they represent. Poisson regression is a special case of Generalized Linear Models (GLM). Therefore, we can setup GLM model to estimate the mean vector $\mu$ (our expectation to the response). In doing so we have to estimate the parameters $\beta=(\beta_0,\ldots,\beta_m)$. This is done by maximum likelihood estimation and often requires numerical optimization such as the Newton-Raphson method. The likelihood function of $n$ independent identically, distributed observations is,

$$L(y_1,\ldots,y_n\mid\beta,X)=\prod_{j=1}^{n}f(y_j\mid\beta,X_j) \tag{7}$$

where $f$ is the probability density function of the response distribution. The values of $\beta=(\beta_0,\ldots,\beta_m)$ that maximizes $L$ is called the maximum likelihood estimates.

To simplify optimization, one usually takes the logarithm of $L$ to get the log likelihood,

$$\begin{aligned}l(y_1,\ldots,y_n\mid\beta,X) &=\log L(y_1,\ldots,y_n\mid\beta,X)\\ &=\sum_{j=1}^{n}\log f(y_j\mid\beta,X_j)\end{aligned}$$

As the logarithm is a monotonously increasing function the values of $\beta$ that maximizes $L$ are identical to those which maximize $l$. Methods for optimizing the log likelihood on an automated basis are implemented in most statistical software packages. The output from the GLM analysis is $\beta=(\widehat{\beta_0},...,\widehat{\beta_m})$.

#### *2.2.3. The Non-Homogeneous Poisson Regression Model*

Consider the non-homogeneous Poisson process which is defined by the mean value function $\mu(t)$:

$$\mu(s,t]=\int_s^t\mu(u)du;$$

with $s<t$, then the non-homogeneous Poisson process with covariates $x=(x_1,...,x_m)^t$ is:

$$\mu^x(s,t]=\int_s^t\mu^x(u)du, s<t\,.$$

Assume separability entailing that all covariate classes have a common seasonal component, then:

$$\mu^x(t)=\lambda_0^x\tau(t) \text{ and } \mu^x(s,t]=\lambda_0^x\int_s^t\tau(u)du\,.$$

Poisson regression makes use of relationships of the form:

$$\log\left(\lambda^x\right)=\beta_0+\beta_1x_1+...+\beta_mx_m\,. \tag{8}$$

The coefficients $\beta_0,...,\beta_m$ are usually determined by likelihood estimation.

Consider a homogeneous model with $\tau(t)=1$, it is then assumed that $n_j$ is Poisson with parameter $\mu_j=\lambda^xT_j$ where $\lambda^x$ are tied to covariates $x_{j1},...,x_{jm}$. The density function of $n_j$ is then:

$$f(n_j;T_j)=\frac{(\lambda^xT_j)^{n_j}}{n_j!}exp(-\lambda^xT_j),$$

or

$$\begin{aligned}\log f(n_j,T_j)&=n_j\log(\lambda^x)+n_j\log(T_j)\\ &\quad-\log(n_j!)-T_j\lambda^x,\end{aligned}$$

which is to be added over all $j$ for the likelihood function $L(\beta_0,...,\beta_m)$. We may drop the two middle terms $n_jT_j$ and $\log(n_j!)$ since they are not dependent on $\lambda^x$. The likelihood criterion then becomes:

$$L(\beta_0,...,\beta_m)=\sum_{j=1}^{n}\left[n_j\log(\lambda^x)-T_j\lambda^x\right]$$

Where:

$$\log(\lambda^x)=\beta_0+\beta_1x_{j1}+...+\beta_mx_{jm}\,.$$

This correspond to GLM Poisson regression and $\beta=(\beta_0,...,\beta_m)$ can be estimated by maximum likelihood estimation. The optimization is usually done by Expectation-Maximization (EM) algorithm.

#### *2.2.4. The Seasonal Intensity Model*

Claim intensities do not necessarily remain constant over time. They serve as an example of a seasonal variation that is present in many parts of the world due to changing weather conditions [8]. The time variation can be handled through a time-varying intensity function $\mu=\mu(t)$.

Consider now a non-homogeneous model and assume that we parametrize the seasonal model $\tau(t)$ by

$$\tau(t;\alpha_0,\tau_0) = 1+\alpha_0 \sin(2\pi(t-\tau_0)) \quad (9)$$

where $0 < t < 1$ represent time of normalized year. For the non-homogeneous Poisson process, consider the random variable

$[N, T_1,\ldots,T_N \mid \Delta t = (t^s,t^e)]; 0 \le t^s < T_i < t^e \le 1$ , $i = 1,\ldots,N$ , where $N$ is the number of claims observed in the registration interval $(t^s,t^e)$ and $T_1,...,T_N$ are the occurrence times. Note that both $N$ and $T_1,...,T_N$ are random variables. From the model with intensity $\mu(t)$ parametrized as $\lambda_0\tau(t;\alpha_0,\tau_0)$ :

$$\text{Prob}\left[n, t_1,...t_n \mid \Delta t = (t^s,t^e), \mu(t) = \lambda_0\tau(t;\alpha_0,\tau_0)\right] = \frac{\lambda_0^n}{n!}\prod_{i=1}^{n}\left[\tau(t_i;\alpha_0,\tau_0)exp\left\{-\lambda_0\tau_I(\Delta t;\alpha_0,\tau_0)\right\}\right]$$

for $n = 0,1,\ldots$ and $t_i \in [t^s,t^e]; i = 1,\ldots,n,$ with

$$\tau_I\left(\Delta t;\alpha_0,\tau_0\right) = \int_{t^s}^{t^e}\tau\left(t;\alpha_0,\tau_0\right)dt\,.$$

Consider now the set of observations:

$$O = \left[n_j, t_{j1},\ldots,t_{jn_j} \mid \Delta t_j = (t_j^s,t_j^e)\right];\ \ j = 1,\ldots,m,$$

the likelihood for O is:

$$L\left(\lambda_0,\alpha_0,\tau_0 \mid O\right) = \prod_{j=1}^{m}\frac{\lambda_0^{n_j}}{n_j!}\prod_{i=1}^{n_j}\tau(t_{ij};\alpha_0,\tau_0)\times\exp\left\{-\lambda_0\tau_I(\Delta t_j;\alpha_0,\tau_0)\right\}$$

$$= \frac{\lambda_0^{\sum_{j=1}^{m}n_j}}{\prod_{j=1}^{m}n_j!}\prod_{j=1}^{m}\prod_{i=1}^{n_j}\tau(t_{ji};\alpha_0,\tau_0)\times\exp\left\{-\lambda_0\sum_{j}^{m}\tau_I(\Delta t_j;\alpha_0,\tau_0)\right\}$$

with corresponding Log-Likelihood model:

$$L\left(\lambda_0,\alpha_0,\tau_0 \mid O\right) \approx \sum_{j=1}^{m}n_j\log\lambda_0 - \sum_{j=1}^{m}\log n_j! + \sum_{j}^{m}\sum_{i}^{n}\log\left[1+\alpha_0\sin\left(2\pi(t_{ij}-\tau_0)\right)\right] - \lambda_0\sum_{j}^{m}\left(\Delta t_j\right).$$

Note that for $\tau(t;\alpha_0,\tau_0) = 1+\alpha_0\sin\left(2\pi(t-\tau_0)\right)$ , we obtain:

$$\tau_I\left(\Delta t;\alpha_0,\tau_0\right) = \int_{t^s}^{t^e}\left(1+\alpha_0\sin\left(2\pi(t-\tau_0)\right)\right)dt = \left(t^e-t^s\right)+\alpha_0\int_{t^s}^{t^e}\sin\left(2\pi(t-\tau_0)\right)dt.$$

Since we do not have $(t^s,t^e)$ available, only $t^s - t^e$, we approximate the last term to 0. Hence,

$$\tau_I\left(\Delta t;\alpha_0,\tau_0\right) \approx t^e - t^s = \Delta t.$$

The approximate log-likelihood for $O$ is then:

$$L\left(\lambda_0,\alpha_0,\tau_0 \mid O\right) \approx \sum_{j=1}^{m}n_j\log\lambda_0 - \sum_{j=1}^{m}\log n_j! + \sum_{j}^{m}\sum_{i}^{n}\log\left[1+\alpha_0\sin\left(2\pi(t_{ij}-\tau_0)\right)\right] - \lambda_0\sum_{j}^{m}\left(\Delta t_j\right).$$

Note that the approximate log-likelihood is separable with respect to $\lambda_0$ and $(\alpha_0,\tau_0)$:

$$\frac{\partial\log L\left(\lambda_0,\alpha_0,\tau_0 \mid O\right)}{\partial\lambda_0} = 0, \text{ so that: } \widehat{\lambda_0} = \frac{\sum_{j}^{m}n_j}{\sum_{j}^{m}\left(\Delta t_j\right)},$$

which is independent of $(\alpha_0,\tau_0)$ and identical to the classical unbiased estimator in (6), and recall that:

$$E\left(\widehat{\lambda_0}\right) = \lambda_0 \quad \text{and} \quad \text{Var}\left(\widehat{\lambda_0}\right) = \frac{\lambda_0}{\sum_{j=1}^{m}\left(\Delta t_j\right)}.$$

The approximate MLE estimator for $(\alpha_0,\tau_0)$ is:

$$\left(\widehat{\alpha_0},\widehat{\tau_0}\right) = \arg\max_{\alpha_0,\tau_0}\left\{\sum_{j}^{m}\sum_{i}^{n_j}\log\left[1+\alpha_0\sin\left(2\pi(t_{ji}-\tau_0)\right)\right]\right\}, \quad (10)$$

which must be determined by numerical optimization. Note

however that $(\alpha_0, \tau_0)$ is constrained by $0 < \alpha_0, \tau_0 < 1$, hence a brute force grid search on a f.ex $[100 \times 100]$, grid can be used in the optimization. The MLE estimators $(\widehat{\alpha_0}, \widehat{\tau_0})$ are most likely not unbiased, but still we assume that the variance can be approximated by:

$$Var\begin{pmatrix}\widehat{\alpha_0}\\ \widehat{\tau_0}\end{pmatrix} \approx \left[ I\left(\widehat{\alpha_0}, \widehat{\tau_0}\right)\right]^{-1},$$

where the $2 \times 2$ matrix:

$$I\left(\widehat{\alpha_0}, \widehat{\tau_0}\right) = -\frac{d^2}{d\left(\alpha_0, \tau_0\right)} \log L\left(\alpha_0, \tau_0 \mid O\right)|_{\widehat{\alpha_0}, \widehat{\tau_0}}$$

is the Fisher information matrix. This matrix must be determined numerically.

# 3. Simulation Results and Discussions

The claim data is presented in Section 2. It contains observations as follows:

| covariates | claims | exposure time | acc.month |
|---|---|---|---|
| $x_{11} \ldots x_{1m}$ | $n_1$ | $T_1$ | $t_{11} \ldots t_{1n_1}$ |
| $x_{21} \ldots x_{2m}$ | $n_2$ | $T_2$ | $t_{21} \ldots t_{2n_2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_{n1} \ldots x_{nm}$ | $n_n$ | $T_n$ | $t_{n1} \ldots t_{nn_n}$. |

The claim process is modeled as a non-homogeneous Poisson process with time varying intensity dependent on the covariates:

$$\mu^x(t) = \lambda_0^x \times \tau\left(t; \alpha_0 \tau_0\right)$$

where

$$\log\left(\lambda_0^x\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$$

and

$$\tau\left(t; \alpha_0, \tau_0\right) = 1 + \alpha_0 \times \sin\left[2\pi \times \left(\frac{t - 0.5}{12} - \tau_0\right)\right].$$

The factor $\lambda_0^x$ is an average intensity which is dependent on the covariate $x = (x_1, \ldots, x_m)$ while the seasonal factor $\tau\left(t; \alpha_0, \tau_0\right)$ capture varying intensities during the year.

The parameter $\lambda_0^x$ is dependent on the parameters $\beta = \left(\beta_0, \beta_1, \ldots, \beta_m\right)$ which are regression parameters in Poisson regression, see (8). The factor $\tau\left(t; \alpha_0, \tau_0\right)$ is dependent on the parameters $(\alpha_0, \tau_0)$ where $\alpha_0$ is the amplitude of the seasonal variations and $\tau_0$ is a phase shift parameter.

Note that $t$ represent month number, and that the shift - $0.5$ is introduced to center the observations in the actual month. The challenge is to estimate the model parameters $\beta$, and $(\alpha_0, \tau_0)$ from the claim observations. In Section 2 a maximum approximate likelihood estimator is developed for $\beta$ and $\left(\alpha_0, \tau_0\right)$. The estimators are separable in the sense that $\beta$ and $\left(\alpha_0, \tau_0\right)$ can be estimated separately. We will first discuss the estimation of the seasonal component $\tau\left(t; \alpha_0, \tau_0\right)$, hence the model parameters $\alpha_0, \tau_0$, thereafter we will discuss estimation of $\lambda_0^x$ hence $\beta = \left(\beta_0, \beta_1, \ldots, \beta_m\right)$

### *3.1. Estimation of Seasonal Component*

The MLE estimator for $\left(\alpha_0, \tau_0\right)$ is developed in (10):

$$(\widehat{\alpha_0}, \widehat{\tau_0}) = \arg\max_{\alpha_0, \tau_0} \left\{ \sum_{j=1}^{m} \sum_{i=1}^{n_j} \log\left[1 + \alpha_0 \sin\left(2\pi\left(\frac{t_{ij} - 0.5}{\tau_0}\right)\right)\right]\right\}$$

The covariance matrix for $(\widehat{\alpha_0}, \widehat{\tau_0})$ can be approximated by the inverse Fisher information matrix. Both the estimate and associated covariance matrix must be determined numerically.

In order to explore the reliability of the estimator we will first conduct a small simulation study. Thereafter we will estimate $\left(\alpha_0, \tau_0\right)$ based on our data. We use a MATLAB computer code for both the simulation study and the real data.

#### *3.1.1. Test Study*

The simulation study is designed as follows:

Define model by assigning $(\alpha_0, \tau_0)$ defining:

$$\mu(t) = \lambda_0 \times \left[1 + \alpha_0 \times \sin\left(2\pi \times (t - \tau_0)\right)\right];\ 0 \le t \le 1.$$

Note that the study is invariant to average intensity, we use $\lambda_0 = 10$. The following parameter values are used.

$$\alpha_0 = 0.8, \quad \tau_0 = 0.2$$

$$\alpha_0 = 0.5, \quad \tau_0 = 0.5$$

Simulate synthetic observation - $n$ series: the algorithm for simulation from a non-homogeneous Poisson process is based on rejection sampling,

$$\left(n_j; t_{j1}, \ldots, t_{jn_j}\right); j = 1, \ldots, n.$$

Estimate $(\alpha_0, \tau_0)$ based on synthetic observation with $n = 20$ and $n = 100$ by using the MLE estimator given above.

Estimate the Fisher information matrix and compute the covariance matrix of $(\widehat{\alpha_0}, \widehat{\tau_0})$.

Simulation 1

Initial parameters $(\alpha_0, \tau_0) = (0.8, 0.2)$, and we consider two cases, one with $n = 100$ and one $n = 20$. The two cases demonstrate the dependence of the covariance matrix on the number of observations.

Figure 1 display realizations from the non-homogeneous Poisson model with $n = 100$. Recall that $\lambda_0 = 10$, hence the expected number of claims in a year is 10, moreover the parameter $\tau_0 = 0.2$ entails that the highest intensity is at $(\tau_0 + 0.25) = 0.45$. These features can be observed, the estimates of the parameters are:

$$\left(\widehat{\alpha_0}, \widehat{\tau_0}\right) = (0.81, 0.21)$$

with covariance matrix

$$\begin{pmatrix} 0.010^2 & 0.000 \\ 0.000 & 0.033^2 \end{pmatrix}.$$

The estimates with joint 0.95 confidence region is displayed in Figure 3, left display. Figure 2 display realizations from the non-homogeneous Poisson model with $n = 20$. The figure is similar to Figure 1, but the total number of observations is smaller. The estimates of the parameters are:

$$\left(\widehat{\alpha_0}, \widehat{\tau_0}\right) = (0.84, 0.22)$$

with covariance matrix

$$\begin{pmatrix} 0.020^2 & 0.000 \\ 0.000 & 0.080^2 \end{pmatrix}.$$

The estimates with joint 0.95 confidence region is displayed in Figure 3, right display.

The joint 0.95 confidence regions for $(\alpha_0, \tau_0)$ are displayed in Figure 3. Note that the region is smaller for $n = 100$ than for $n = 20$ of course. Note that the true values are $(\alpha_0, \tau_0) = (0.8, 0.2)$, which fall just on the border of the 0.95 confidence region

Simulation 2

Initial parameters $(\alpha_0, \tau_0) = (0.5, 0.5)$ and two cases with $n = 100$ and $n = 20$. Figure 4 display observations for $n = 100$ in a similar format as Figure 1, for other parameter values. In this case the maximum intensity is at $(\tau_0 + 0.25) = 0.75$ and there is less difference between max/min intensity than in Figure 1 since $\alpha_0$ is smaller.

The estimates are:

$$\left(\widehat{\alpha_0}, \widehat{\tau_0}\right) = (0.54, 0.50)$$

with covariance matrix

$$\begin{pmatrix} 0.014^2 & 0.000 \\ 0.000 & 0.040^2 \end{pmatrix}.$$

The joint 0.95 confidence region is presented in Figure 6, left display. Figure 5 display results for $n = 20$ and correspond to Figure 4.

The estimates are:

$$\left(\widehat{\alpha_0}, \widehat{\tau_0}\right) = (0.54, 0.50)$$

with covariance matrix

$$\begin{pmatrix} 0.030^2 & 0.000 \\ 0.000 & 0.090^2 \end{pmatrix}.$$

The 0.95 confidence region is larger than for $n = 100$, see Figure 6, right display. The joint 0.95 confidence region for $(\alpha_0, \tau_0)$ are displayed in Figure 6 and we observe that the true values $(\alpha_0, \tau_0) = (0.50, 0.50)$ falls well within the region for $n = 20$, but not for $n = 100$ where $\alpha_0$ falls outside the region $[0.512, 0.568]$. This may happen for one random sample.

### *3.1.2. Case Study: Ethiopian Claims Observation*

The observations are presented in Section 2, and the monthly variations are displayed in Figure 7. There appears to be higher claim intensities during January-February and lower during August-September. It is unclear whether these variations are statistically significant.

The seasonal model for claim intensity is:

$$\tau(t; \alpha_0, \tau_0) = 1 + \alpha_0 \times \sin\left(2\pi\left(\frac{t - 0.5}{12} - \tau_0\right)\right)$$

where $1 \le t \le 12$ is month number of claim, and the correction factor $0.5/12$ is used to center the observations in each month. The estimator for $(\alpha_0, \tau_0)$ is discussed in Section 2 and presented in (10). By introducing the real observations, we obtain:

$$\begin{pmatrix} \widehat{\alpha_0} \\ \widehat{\tau_0} \end{pmatrix} = \begin{pmatrix} 0.25 \\ 0.87 \end{pmatrix}$$

and

$$\mathrm{Var}\begin{pmatrix} \widehat{\alpha_0} \\ \widehat{\tau_0} \end{pmatrix} \approx \begin{pmatrix} 0.076^2 & 0.028^2 \\ 0.028^2 & 0.106^2 \end{pmatrix}.$$

The corresponding joint 0.95 confidence region for $(\alpha_0, \tau_0)$ is displayed in Figure 8. Note that for the amplitude $\alpha_0$ the approximate 0.95 confidence interval is $[0.10, 0.40]$ which does not include $\alpha_0 = 0$. Consequently, the hypothesis $\alpha_0 = 0$ versus $\alpha_0 \neq 0$ will be rejected at a 0.05 significance level. This entails that the seasonal variations are significant. The best estimate is $\widehat{\alpha_0} = 0.25$, which entails that the claim intensity varies between $0.75\lambda_0^x$ and $1.25\lambda_0^x$

during the year, hence by ±25% .

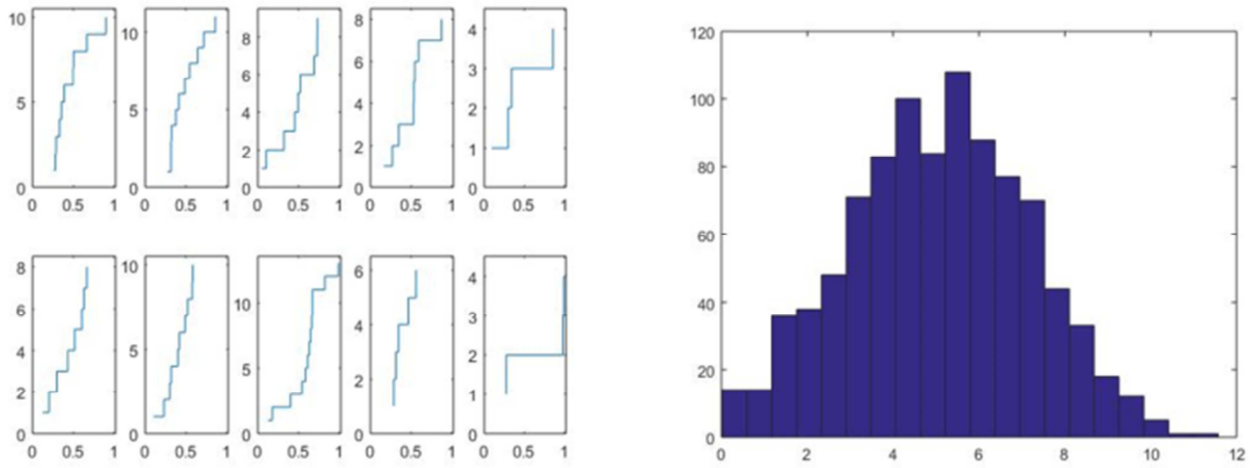

***Figure 1.*** *Parameters (0.8, 0.2) and n=100, Ten realizations (left); histogram of observations (right).*

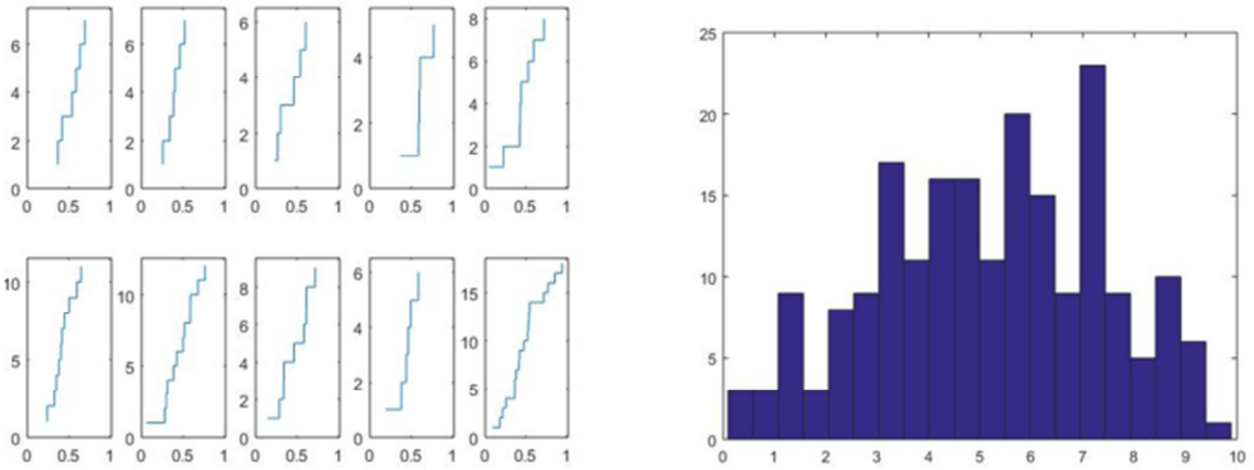

***Figure 2.*** *Parameters (0.8, 0.2) and n=20, Ten realizations (left); histogram of observations (right).*

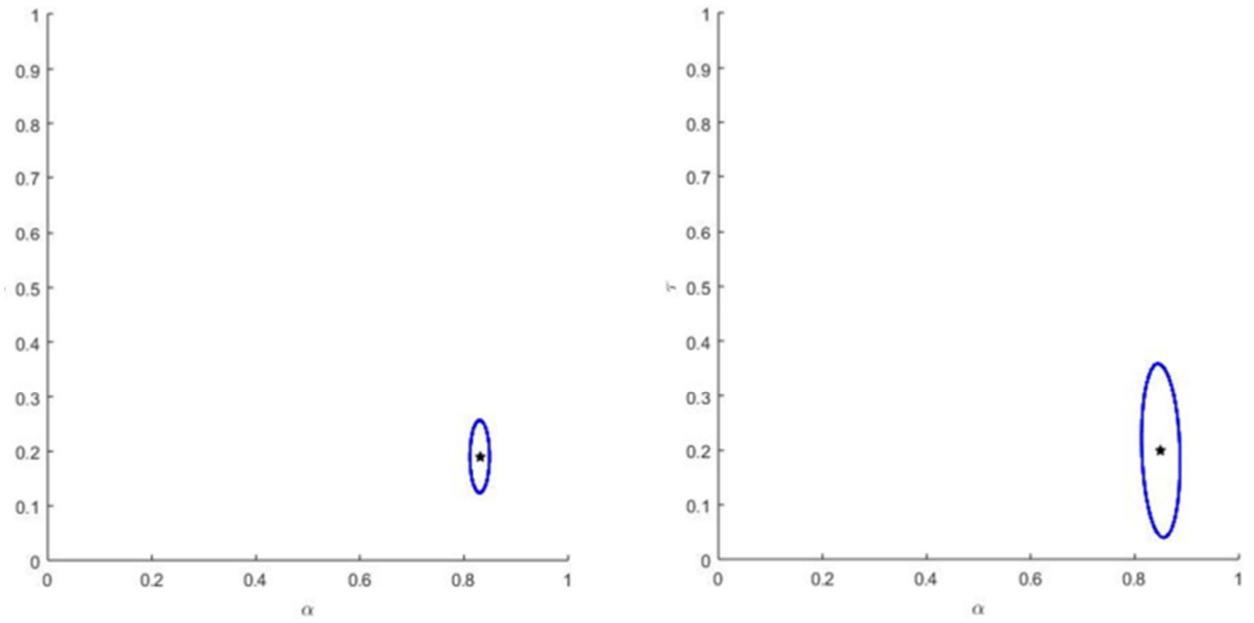


***Figure 3.*** *Joint* $(\alpha_0,\tau_0)$ *− 0.95 confidence regions with* $(\alpha_0,\tau_0)=(0.8,0.2)$, *, n=100 (left); n=20 (right).*

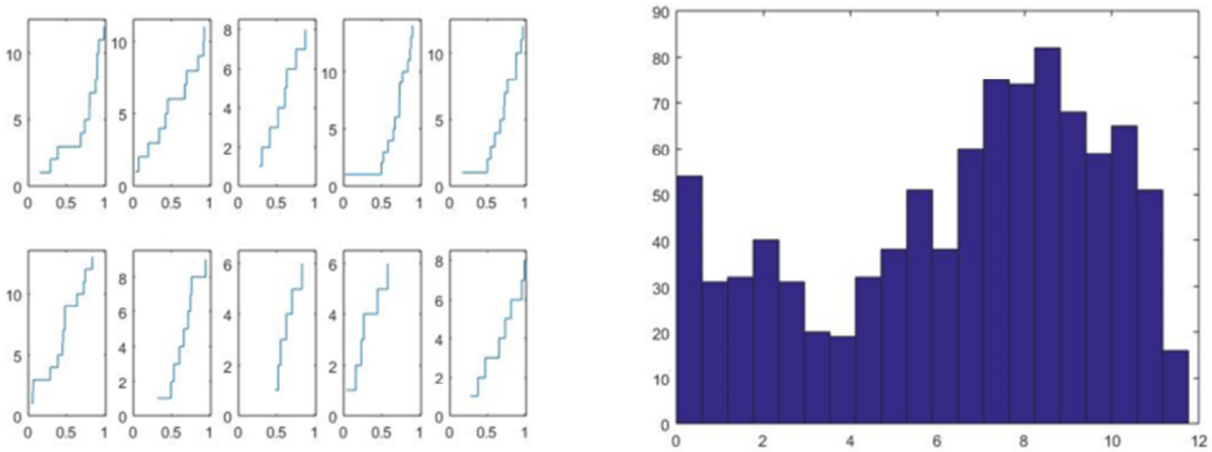

***Figure 4.*** *Parameters (0.5, 0.5) and n=100, Ten realizations (left); histogram of observation (right).*

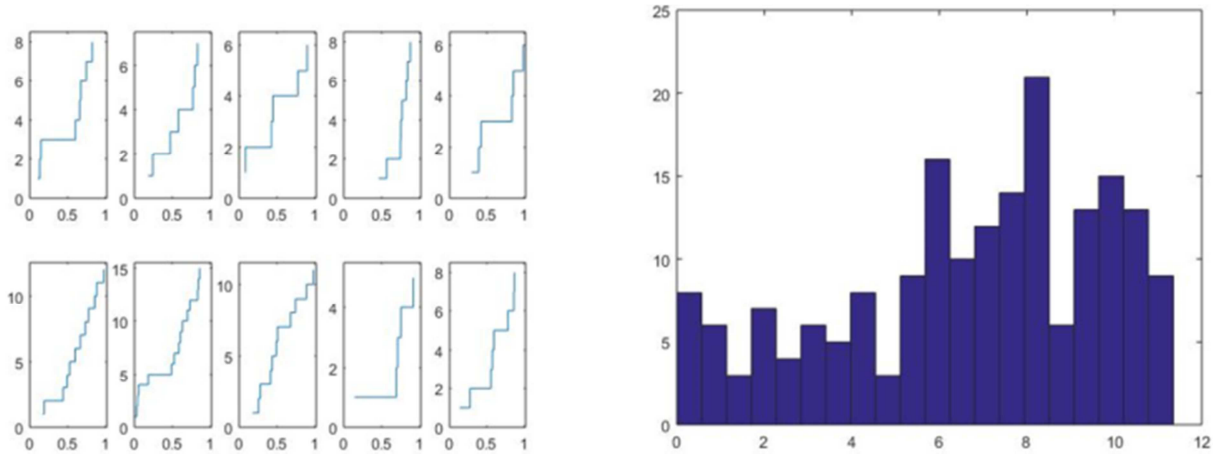

***Figure 5.*** *Parameters (0.5, 0.5) and n=20, Ten realizations (left); histogram of observation (right).*

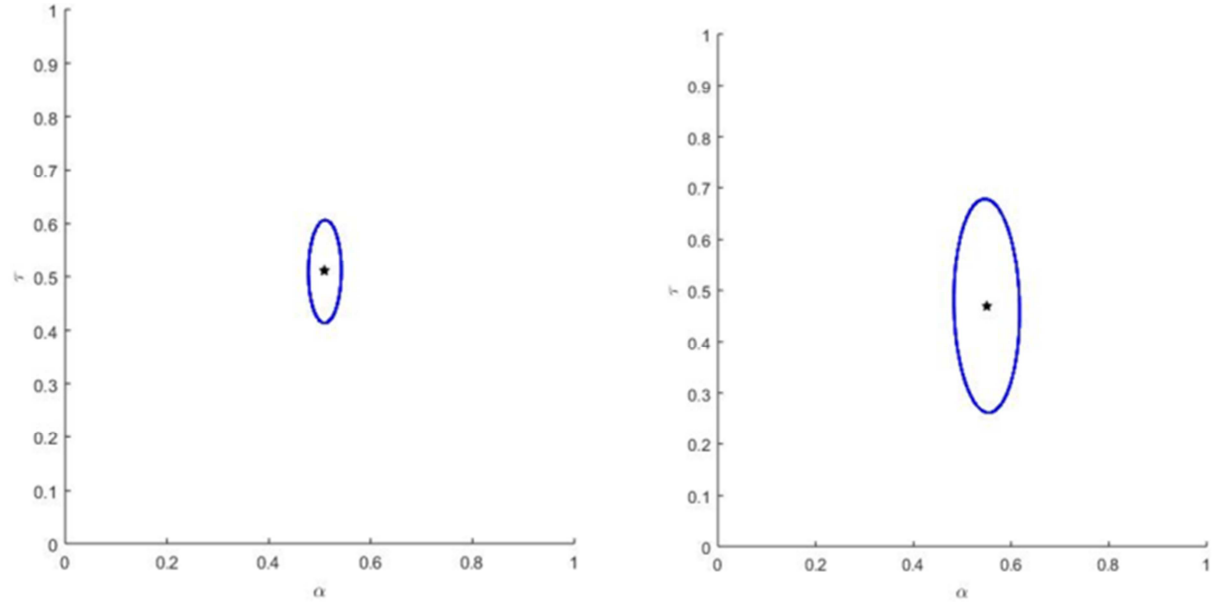

***Figure 6.*** *Joint* $(\alpha_0, \tau_0) = 0.95$, *confidence regions with* $(\alpha_0, \tau_0) = (0.5, 0.5)$, *n=100 (left); n=20 (right).*

The phase shift $\tau_0$ has approximate 0.95 confidence interval $[0.65, 1.09]$. Note that $\tau_0$ is periodic with period $[0, 1]$ due to the sinuosity, hence $\tau_0 = 0$ is inside the interval and the phase shift is not significant to the 0.95 level. The best estimate is however, $\widehat{\tau_0} = 0.87$, which entails that maximum intensity occurs according to the model,

$(-0.13+0.25) = 0.12$ into the year, that is in mid-February while minimum intensity occurs $(0.12+0.50) = 0.62$ into the year, i.e. mid-August.

These extremes can be compared with the display in Figure 7, and they are believable. The high numbers of claims in January and July complicates the picture, however.

To summarize the results for the seasonal variation: According to a sinuous yearly variational model, the maximum number of claims occurs in February, while the minimum occurs in August. The extremes vary ±25% from the average intensity. Hence it appears as sunny, hot weather with many cars, baggage. bicycles, people and animals in the streets causes more accidents than rainy, cold weather with few Traficant. Hence, Traficant intensity seems to cause more accidents than bad driving conditions.

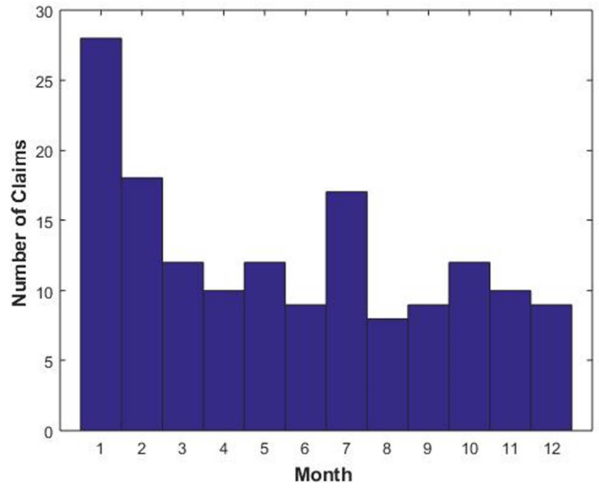


***Figure 7.*** *Histogram of Ethiopian claims observation.*

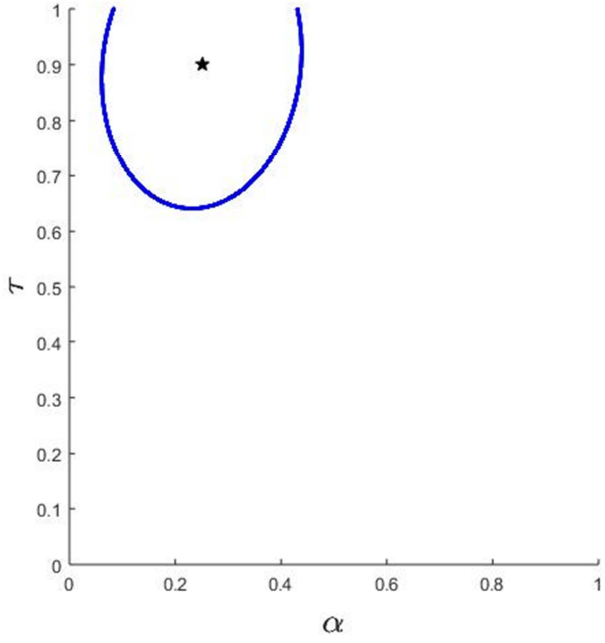


***Figure 8****. 0.95 confidence region for* $(\alpha_0, \tau_0)$ *of Ethiopian claims observations.*

### *3.2. Estimation of Covariate Dependency*

The MLE estimator for $\beta = (\beta_0, ..., \beta_m)$ is developed in Section 2.

$$\hat{\beta} = \text{argmax}_{\beta} \left\{ \sum_{j=1}^{n} \left[ n_j \log \lambda_j^x - T_j \lambda_j^x \right] \right\}$$

with

$$\log(\lambda_j^x) = \beta_0 + \beta_1 x_{j1} + ... + \beta_m x_{jm}.$$

This constitutes a GLM Poisson regression model with observation specific intensity. The GLM package in $R$ can be used for solving this optimization problem. Poisson models are usually fitted empirically by means of data which are counts of claims $n_1, ..., n_n$ from $n$ policies that have been under exposure $T_1, ..., T_n$. The data are an extract of three years for which the number of claims is 157 with total exposure $T_1 + ... + T_n$ =844 automobile years. This yields as average yearly claim intensity:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} T_i} = \frac{157}{844} = 0.186.$$

The estimate is unbiased with variance determined by:

$$\widehat{\text{Var}}(\hat{\lambda}) = \frac{\hat{\lambda}}{\sum_{i=1}^{n} T_i} = \frac{0.186}{844} = 0.014^2.$$

This is known to be the most accurate estimator and even applies when the same policy holder appears under different $T_j$.

The driver's age has four categories. Among these categories, drivers with in the age group $26-35$ are responsible for the largest number of claims $49\%$. Drivers with age $20-25$ have the smallest share in all the four measurements, $6.8\%$. There are $93.6\%$ males and $6.4\%$ females.

There are two explanatory variables, sex and age. The former is represented by $x_1$. The latter has four classes and is split into three variables $(x_2, x_3, x_4)$. The most obvious way of feeding them into regression model is to write:

$$\log(\lambda_j^x) = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3} + \beta_4 x_{j4}$$

with

$$x_{j1} = \begin{cases} 1 & \text{Female} \\ 0 & \text{else} \end{cases}, x_{j2} = \begin{cases} 1 & \text{Age 26-35} \\ 0 & \text{else} \end{cases},$$

$$x_{j3} = \begin{cases} 1 & \text{Age 36-45} \\ 0 & \text{else} \end{cases} \text{ and } x_{j4} = \begin{cases} 1 & \text{Age 46+} \\ 0 & \text{else} \end{cases}.$$

Hence, Male/Age $20-25$ is defined as the reference class. As an example of the Poisson regression model, suppose owners $j$ and $i$ are of the same age class, for example Age $20-25$, the former being a Male and the other a Female, then:

$$\frac{\lambda_i^x}{\lambda_j^x} = \frac{\exp\{\beta_0+\beta_1\}}{\exp\{\beta_0\}} = \exp\{\beta_1\}.$$

Hence, the relative increase in intensity between Male and Female is $\exp\{\beta_1\}$. By using the Poisson regression, we can estimate the model parameters $\hat{\beta} = \left(\widehat{\beta_0},\ldots,\widehat{\beta_4}\right)$, see Table 4. Based on the results of $R$, the regression equation consisting of the variables in Table 4 is given by:

$$\begin{aligned}\log(\lambda_j^x) &= -3.026 - 0.002x_{j1} - 0.629x_{j2} \\ &\quad -1.180x_{j3} - 0.178x_{j4}.\end{aligned}$$

Note that for the constant coefficient $\beta_0$ the approximate 0.95 confidence interval is $[-3.649,-2.402]$ which does not include $\beta_0 = 0$. Consequently, the hypothesis $\beta_0 = 0$ versus $\beta_0 \neq 0$ will be rejected at a $0.05$ significance level. This entails that the constant coefficient is significantly different from zero.

***Table 4.*** *Estimates of regression coefficients in Poisson regression.*

| Nclaims | Coef. | [95% Conf. Interval] |
|---|---|---|
| Sex | -0.0022 | -0.649 0. 654 |
| Age | | |
| 26-35 | -0.629 | -1.302 0. 0446 |
| 36-45 | -1.180 | -1.903 -0. 4573 |
| 46+ | 0.1776 | -0. 9125 0. 5574 |
| Intercept | -3.026 | -3.649 -2.403 |

Female drivers seem to have lower claim rates than male drivers since $\widehat{\beta_1} = -0.002$. The p-value of $\widehat{\beta_1}$ is much larger than $0.05$, hence the hypothesis $\beta_1 = 0$ versus $\beta_1 < 0$ cannot be rejected, and the lower claim rate for female drivers is not significant on $0.95$ level.

The coefficient for Age $36-45$, $\beta_3$ has approximate $0.95$ confidence interval $[-1.903,-0.457]$, which does not include $\beta_3 = 0$. Consequently, the hypothesis $\beta_3 = 0$ versus $\beta_3 \neq 0$ will be rejected at $0.05$ significance level. This entails that, $\beta_3 = -1.180$ is significant and drivers in Age $36-45$ has lower claim intensity than drivers in Age $20-25$.

When we look at this analysis, we observe that only Age $36-45$ has significantly different claim rate than Age $20-25$, since the other age groups appear as insignificant at $0.05$ level.

To summarize the influence of the covariates on the number of claims: Female and male drivers do not have significantly different claim rates. But the observations do indicate that the female rate is slightly lower. There is age-dependence in the claim rate, but only Age $36-45$ has significantly lower rate than Age $20-25$, the rate of the former is about one-third of the latter.

### *3.3. Summary of Claim Intensity Model*

From the discussions above, including evaluation of significant effects, we obtain:

$$\begin{aligned}\mu(t \mid x_1,\ldots,x_4) &= \exp\{-3.02-1.18x_3\}\times \\ &\quad \left[1+0.25\sin\left(2\pi\left(\frac{t-0.5}{12}-0.87\right)\right)\right] \\ &\quad ;t\in[1,\ldots,12]\end{aligned}$$

where $t$ is month number. This model appears as the best intensity model for a non-homogeneous Poisson process for claim occurrences for a customer with characteristics $\left(x_1,\ldots x_4\right)$.

# 4. Conclusions and Recommendations

Poisson process theory is used to analyze insurance claim data from the Ethiopian Insurance Corporation, Hawassa district. The seasonal variations are modeled through a non-homogeneous Poisson process, while customer covariates like gender and age are included through Poisson regression in a GLM setting. A factorial intensity model is used, which separate average claim intensity dependent on covariates, and seasonal variations parametrized as a sine-function. The approximate maximum likelihood criterion will also separate average claim intensity and seasonal components, hence simplify inference.

Two covariates are used: gender and age, the latter is discretized into four age classes. Relative to claim intensity for Age 20-25 only intensity for age 36-45 deviates significantly, by being only about one third as large. Relative to male claim intensity, the female intensity is not significantly smaller. There are indications that it might be slightly smaller, however and this should be studied closer.

There is a significant seasonal variation, ±25% of the year-average claim intensity. The highest claim rate is in February, while minimum rate is observed in August. It appears as if heavy traffic causes more claims than poor driving conditions.

It can be learnt from this study that in addition to the efforts being made to reduce the frequency of claims in general, special attention should be given to reduce the severity of accidents by taking the following into consideration:

Strict control and management of vehicle movement is necessary especially in the dry season with many Traficant's, for example, during festivals or other holidays. The same would be required for areas around schools and religious institutions.

Further studies should be made on the collection of claim data by considering detail and accurate information on various variables.

The insurance corporation would benefit greatly by storing

the statistics on digital form, such that statistical analysis is facilitated.

# References


[1] Antonio, K. and Beirlant, J., "Issues in Claims Reserving and Credibility," a Semi parametric Approach with Mixed Models Journal of Risk and Insurance, vol 75, pp. 643-676, 2008.

[2] Antonio, K. & Beirlant, J. ''Applications of Generalized Linear Mixed Models in Actuarial Statistics, Insurance'': Mathematics and Economics North American Actuarial Journal, vol 40, pp. 58-76, 2007.

[3] Antonio, Katrien, Jan Beirlant, Tom Hoedemakers and Robert Verlaak, 'Log-normal mixed models for reported claims reserving'. North American Actuarial Journal, Vol 10 (1): PP. 30-48, 2006.

[4] Bølviken, E. "Computation and Modelling in Insurance and Finance": International series on actuarial Science, Not yet published, 2013.

[5] Smyth, G. K., & Jørgensen, B. Fitting tweedie's compound Poisson model to insurance claims data: dispersion modeling. *Actuarial Studies in Non-life insurance (ASTIN) Bulletin, 32* (1), 143-157, 2002.

[6] Claro, P, Caetano, L., Artes, R. "Estimating Total Claim Size in the Auto Insurance Industry": a Comparison between Tweedie and Zero-Adjusted Inverse Gaussian Distribution BAR, Brazilian Administration Review. vol 8, pp. 37-47. 2011.

[7] Jørgensen, B., & Souza, M. C. P. de. "Fitting Tweedie´s Compound Poisson Model to Insurance Claims Data". Scandinavian Actuarial Journal, vol 1 (1), pp. 69-93, 1994.

[8] Mikosch, T, "Non-Life Insurance Mathematics", Springer, 2003.

[9] Nelder, J. and R. "Generalized Linear Models". Journal of the Royal Statistical Society. Series A (General), vol 135 (3), pp. 370-384, 1972.

[10] Huang, T., Zhao, R., & Tang, W. (2009). Risk model with fuzzy random individual claim amount. *European Journal of Operational Research, 192* (3), 879-890.

[11] Wüthrich, M. V., & Merz, M. (2008). Stochastic claims reserving methods in insurance. West Sussex: John Wiley & Sons.

[12] Sisay Wuyu, Patrick Cerna. "Risk Assessment Predictive Modelling in Ethiopian Insurance Industry Using Data Mining. Software Engineering". Vol. 6, No. 4, 2018, pp. 121-127. doi: 10.11648/j.se.20180604.13.

[13] Kanbiro Orkaido Deyganto, Ayneshet Agegnew Alemu, "Factors Affecting Financial Performance of Insurance Companies Operating in Hawassa City Administration, Ethiopia", Universal Journal of Accounting and Finance 7 (1): 1-10, 2019.

[14] Yuvaraj & Abate G. (2013). "A Study on the *P*erformance of insurance companies in Ethiopia", International Journal of Marketing, Financial Services & Management Research, Vol 2, No 7, July (2013).

[15] Bishnu Prasad Bhattarai, "Factors Influencing Profitability of Insurance Companies in Nepal", *International Journal of Management,* 11 (9), 2020, pp. 8-14.

[16] *Mengistu Tegegn, Leta Sera, Tesfaye Melaku Merra,* "Factors Affecting Profitability of Insurance Companies in Ethiopia: Panel Evidence", *International Journal of Commerce and Finance*, Vol. 6, Issue 1, 2020, 1-14.