***Yayın Geliş Tarihi (Submitted):*** **29/11/2023**

***Yayın Kabul Tarihi (Accepted):*** **17/01/2024**

***Makele Türü (Paper Type):*** **Araştırma Makalesi – Research Paper**

**Please Cite As/Atıf için:**
Baylan, P. and Demirel, N. (2024), Quantifyıng the impact of risk factors on direct compensation property damage in canadian automobile insurance, *Nicel Bilimler Dergisi*, 6(1), 103-127. *doi: 10.51541/nicel.1397941*

# QUANTIFYING THE IMPACT OF RISK FACTORS ON DIRECT COMPENSATION PROPERTY DAMAGE IN CANADIAN AUTOMOBILE INSURANCE

Pervin Baylan[1] and Neslihan Demirel[2]

## ABSTRACT

This study presents a statistical analysis assessing the impact of various risk factors on direct compensation property damage (DCPD) claims in private passenger vehicle accidents. Using automobile insurance data in Ontario, Canada for the decade years period between 2003 and 2012, a statistical model of property damage was explored via a generalized linear binary logit mixed model and considered the imbalance between the classes of insureds. The results indicate that several risk factors have a significant impact on the likelihood of DCPD claims, including usage, training, outstanding loss, and incurred loss. The effects of these risk factors were observed under the weights — the number of trials used to generate each success proportion — in the different classes of insureds. The generalized linear mixed models (GLMMs) analysis provides a powerful tool for quantifying the impact of risk factors on binary outcomes, which are called DCPD claims and property damage (PD) claims covered by third-party liability (TPL) insurance. These models can also inform insurance underwriting and policy design, focusing on identifying the most significant risk factors. The performance metrics calculated by considering the class imbalance in binary outcomes verify the resulting model's ability to accurately predict classes. The *F*1 score, an evaluation metric to measure the performance of classification, was calculated as 0.934. In addition, *PR AUC*, which is the

[1] Corresponding Author, Research Assistant, Department of Statistics, Faculty of Science, Dokuz Eylul University, Izmir, Turkiye. ORCID ID: https://orcid.org/0000-0003-2660-3814
[2] Prof.Dr., Department of Statistics, Faculty of Science, Dokuz Eylul University, Izmir, Turkiye. ORCID ID: https://orcid.org/0000-0002-5394-4721

area under the Precision-Recall (PR) curve, was computed as 0.953. These high scores indicate that the resulting model performs well in the classification. The other metrics also support the classification accuracy of this model. The findings of the analysis can help insurers better understand the underlying drivers of property damages and develop more accurate and effective strategies for risk mitigation. Furthermore, this study highlights the importance of developing class-specific risk assessment models to account for the imbalance across different classes.

**Keywords:** Binary Logit Model, Direct Compensation Property Damage, Generalized Linear Mixed Model, Third-Party Liability Insurance, Unbalanced Panel Data.

# KANADA OTOMOBİL SİGORTASINDA RİSK FAKTÖRLERİNİN DOĞRUDAN TAZMİN EDİLEN MADDİ HASAR ÜZERİNDEKİ ETKİSİNİN DEĞERLENDİRİLMESİ

## ÖZ

Bu çalışma, özel binek araç kazalarında çeşitli risk faktörlerinin doğrudan tazmin edilen maddi hasarlar (Direct Compensation Property Damage - DCPD) üzerindeki etkisini değerlendiren istatistiksel bir analiz sunmaktadır. 2003 ile 2012 yılları arasındaki on yıllık döneme ait Ontario, Kanada'daki otomobil sigortası verileri kullanılarak, genelleştirilmiş doğrusal ikili logit karma model aracılığıyla maddi hasarın istatistiksel bir modeli araştırılmış ve sigortalıların sınıfları arasındaki dengesizlik dikkate alınmıştır. Sonuçlar, kullanım amacı, sürücü eğitimi, muallak hasar ve gerçekleşen hasar dahil olmak üzere çeşitli risk faktörlerinin DCPD hasarlarının olasılığı üzerinde önemli bir etkiye sahip olduğunu göstermektedir. Bu risk faktörlerinin etkileri, farklı sigortalı sınıflarındaki ağırlıklar — her bir başarı oranını oluşturmak için kullanılan deneme sayısı — altında gözlemlenmiştir. Genelleştirilmiş doğrusal karma modeller (GLMMs) analizi, risk faktörlerinin üçüncü şahıs sorumluluk (TPL) sigortası kapsamındaki DCPD hasarları ve maddi hasarlar (PD) olarak adlandırılan ikili sonuçlar üzerindeki etkisinin değerlendirilmesinde güçlü bir araçtır. Bu modeller, en önemli risk faktörlerini belirlemeye odaklanarak sigorta risk değerlendirmesine ve poliçe tasarımına da bilgi sağlayabilir. İkili sonuçlardaki sınıf dengesizliği dikkate alınarak hesaplanan performans ölçümleri, elde edilen modelin sınıfları doğru tahmin etme yeteneğini doğrulamaktadır. Sınıflandırma performansını ölçmeye yönelik değerlendirme ölçümü olan $F1$ skoru 0,934 olarak hesaplanmıştır. Ayrıca, Kesinlik-Duyarlılık (Precision-Recall (PR))

eğrisinin altında kalan alan olan *PR AUC* ise 0,953 olarak elde edilmiştir. Bu yüksek skorlar, elde edilen modelin sınıflandırmada iyi performans gösterdiğine işaret etmektedir. Diğer ölçümler de, bu modelin sınıflandırma doğruluğunu desteklemektedir. Analizin bulguları, sigortacıların maddi hasarların altında yatan nedenleri daha iyi anlamalarına ve risk azaltımı için daha doğru ve etkili stratejiler geliştirmelerine yardımcı olabilir. Ayrıca bu çalışma, farklı sınıflar arasındaki dengesizliği hesaba katmak için sınıfa özgü risk değerlendirme modellerinin geliştirilmesinin önemini vurgulamaktadır.

**Anahtar Kelimeler:** İkili Logit Model, Doğrudan Tazmin Edilen Maddi Hasar, Genelleştirilmiş Doğrusal Karma Model, Üçüncü Şahıs Sorumluluk Sigortası, Dengesiz Panel Veri

## 1. INTRODUCTION

Direct compensation property damage (DCPD) is a type of automobile insurance coverage that is designed to provide compensation to policyholders for damages to their vehicles caused by another driver in an accident. Under DCPD coverage, the policyholders' own insurer handles the claim and pays for the damages up to the limit of their coverage in cases where the accident was caused by another driver and was not their own fault; instead of seeking compensation from the other driver's insurance company. This coverage involves only property damage (PD) and not bodily injury claims occurring in a car accident; while enabling the repair of damage on the vehicle of the policyholders faster, without the delays and complications which might arise when dealing with another driver's insurer. Therefore, being an efficient and fair approach to insurance claims and vehicle repairs, DCPD coverage is available in several provinces in Canada, including Ontario, Quebec, Nova Scotia, New Brunswick and Prince Edward Island. If the policyholders are at fault for the accident, they will need to rely on other types of coverage, such as collision or liability insurance, to cover the cost of damages.

One of the major problems facing actuaries in third-party liability (TPL) insurance is the building of an accurate mathematical model to calculate insurance premiums. This is because it is essential to strike a balance between charging premiums which are affordable for policyholders and generating enough revenue to cover the costs of potential claims and provide a profit for the insurer. To develop an accurate mathematical model, actuaries should consider various risk factors that might influence the likelihood and cost of claims. Accurate

assessment of risk factors is a complex process which involves analyzing historical claims data. The improper models built in the analysis of the historical claims data lead to the premiums being determined lower than they should be, and thus increase the risk of sector failure. Overall, the accurate assessment of risk factors and the development of predictive models that estimate the likelihood of an insured event are crucial components for insurers in the automobile insurance sector, in terms of effectively managing their risk and providing their policyholders with affordable coverage. Actuaries typically use statistical models to calculate insurance premiums; considering the estimated risk of an insured and the potential cost of a claim. By using statistical models to price insurance premiums that reflect the true risk of potential claims, actuaries can help insurers to provide affordable coverage to policyholders; while also ensuring the long-term stability and success of the insurance industry.

Various problems in actuarial science rely on the creation of a mathematical model that can be used in premium pricing. The accurate calculation of premiums for compulsory TPL insurance is particularly important because this type of insurance has a significant impact on the non-life premium income of insurers. By improving the premium evaluation for this line of business, the potential financial losses of the insurance sector can be prevented. DCPD is a mandatory component of automobile insurance in Ontario and is included in all basic auto policies along with TPL insurance. Therefore, it has a considerable share of the yearly non-life premium income. Quantifying the impact of risk factors on the likelihood of DCPD claims versus PD claims covered by TPL insurance can help insurers make more informed decisions about insurance underwriting and policy design. By taking these risk factors into account, the actuaries can calculate insurance premiums appropriate for the level of risk being assumed by the insurer, so that identifying the most significant risk factors leads to a more efficient and effective insurance market.

The use of generalized linear mixed models (GLMMs) in actuarial science allows for the incorporation of risk factors into the premium pricing process, improving the accuracy of insurance premiums and reducing the risk of financial losses for insurers. Most actuarial pricing techniques in use today are based on the generalized linear model research of Nelder and Wedderburn (1972) and McCullagh and Nelder (1989). Over the last 30 years, generalized linear models (GLMs) have been one of the most commonly used statistical tools for modeling actuarial data in actuarial work. In an actuarial context, Haberman and Renshaw (1996) provide an overview of the applications of GLMs in actuarial science and show that

GLMs are not limited to models for automobile insurance premiums. Embrechts and Wüthrich (2022) in the case of non-life insurance demonstrate how combining traditional statistical methods, such as GLMs with neural networks, improves comprehension and interpretation of actuarial data.

Many actuarial problems have a data structure that includes repeated measurements, especially panel data, which are characterized by a tendency to correlate repeated observations on a group of subjects over time. This correlation between observations on the same subject leads to extra difficulties during the analysis. Since the assumption of independence is not fulfilled in GLMs due to this correlation, GLMMs, which are extensions of GLMs, can be used for correlated data. Statistical techniques are considered for modeling panel data within the framework of GLMs in Antonio and Beirlant (2007). They also discuss the advantages of the GLM approach and represent the usage of GLMMs in actuarial mathematics. Miao (2018), using a hierarchical generalized linear model, shows that GLMMs can more effectively reflect the differences between distinct risk individuals as well as the heterogeneity and correlation of risk individual loss over multiple insurance periods.

The GLMM approach has been frequently used to model actuarial data and provides a useful approach in the analysis of unbalanced panel data. This approach procures extra flexibility in estimating the model and helps eliminate the extra complexity resulting from the internal correlation of each subject. Yau et al. (2003) consider the application of the GLMM approach to the analysis of repeated claim frequency data in motor insurance. All of these mentioned features also make GLMMs a powerful tool for identifying risk factors. Antonio and Valdez (2012) present a risk classification based on GLMs in insurance. Garrido et al. (2016) explore how the assumption that claim counts and amounts are independent in non-life insurance can be relaxed via GLMs while incorporating rating factors into the model.

The motivation of this study is to create a statistical model within the framework of GLMs that identifies the impact of the most important risk factors on DCPD claims in private passenger vehicle accidents. For this purpose, the paper is structured as follows. Section 2 describes the methodological framework used in this study. Each subsection of this part mentions the basic concepts of GLMMs, the structure of automobile insurance data provided by a Canadian insurance company, statistical analysis of binary outcomes such as DCPD claims and PD claims covered by TPL insurance, and how risk factors are identified. In addition, the performance metrics used in this study are explained in detail. Section 3 presents the results of the model developed for estimating the likelihood of DCPD claims. Section 4

introduces the main conclusions of this study. The acronyms used in this study are listed in Table 1.

**Table 1.** A list of acronyms and definitions used throughout the paper

| **Acronym** | **Definition** |
|---|---|
| DCPD | Direct Compensation Property Damage |
| PD | Property Damage |
| TPL | Third-Party Liability |
| GLMM | Generalized Linear Mixed Model |
| GLM | Generalized Linear Model |
| AIC | Akaike Information Criterion |
| BIC | Bayesian Information Criterion |
| LRT | Likelihood Ratio Test |
| SN | Sensitivity |
| SP | Specificity |
| P | Precision |
| ACC | Accuracy |
| BA | Balanced Accuracy |
| AUC | Area Under the ROC Curve |
| TP | True Positive |
| TN | True Negative |
| FP | False Positive |
| FN | False Negative |
| ROC | Receiver Operating Characteristic |
| PR | Precision-Recall |
| PR AUC | Area Under the PR Curve |

## 2. MATERIAL AND METHODS

### 2.1. Generalized Linear Mixed Models

A logistic regression model that can be viewed as a GLM is generally used to model binary or more than two categories under the assumption of independence. However, in many actuarial problems, observations on the same subject over time are often correlated. In these circumstances, the logistic GLM might not be appropriate to model repeated observations due to the structure of correlation between observations of the same subject. GLMs are extended to GLMMs by including random effects in the linear estimator that determine the inherent

correlation between observations on the same subject. Thus, the random effect also accounts for unobserved heterogeneity between subjects due to unobserved characteristics.

GLMM provides a more flexible approach in terms of normality and homoscedasticity assumptions since it is extended to distributions from the exponential family. In addition, in GLMM, the additive effect of independent variables is modeled on a transformation of the mean (Antonio and Beirlant, 2007).

Here, the model is extended to include random effects since the focus will be on longitudinal design, which is repeated observations on a group of subjects over time. We consider a model where the conditional distribution of **y**, a vector of the outcome variable $y_{ij}$, given the random effects, follows a binomial distribution such that the property damage type of the *i*th subject in time *j*. A GLMM for binary data with logit-link, which is the link function $g(\mu_{ij})$ determining how the mean is related to the independent variables **x**, is written in the form:

$$g(\mu_{ij}) = \text{logit}(\mu_{ij}) = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{b}_i, \;\; i = 1, \dots, n, \; j = 1, \dots, t_i \quad (1)$$

where $\boldsymbol{\beta}$ ($p$ x 1) is a vector of fixed effect parameters; $\boldsymbol{b}_i$ ($q$ x 1) is a vector of random effects which represent the influence of subject *i* on its repeated observations, having dimension *n*; $\boldsymbol{x}_{ij}$ ($p$ x 1) is a vector of independent variables associated with the *ij*th observation; and $\boldsymbol{z}_{ij}$ ($q$ x 1) is a vector of variables having random effects (Antonio and Beirlant, 2007). GLMM utilizes the logit-link for the analysis of dichotomous data, namely

$$g(\mu_{ij}) = \text{logit}(\mu_{ij}) = log \quad _e\left[\frac{\mu_{ij}}{1-\mu_{ij}}\right] \quad (2)$$

where $\mu_{ij}$ is the probability of an event on subject *i* in time *j*. Here, the conditional expectation equals the conditional probability of a response given the random effects and covariate values, i.e.,

$$\mu_{ij} = E(y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}) = P(y_{ij} = 1|\boldsymbol{b}_i, \boldsymbol{x}_{ij}) \quad (3)$$

(Hedeker, 2005). Assuming that the random effects are mutually independent and identically distributed completes the specification of the GLMM. Furthermore, the correlation between

observations on the same subject occurs since they share the same random effects $\boldsymbol{b}_i$ (Antonio and Beirlant, 2007).

For more information on the theory and application of GLMs, see McCullagh and Nelder (1989), De Jong and Heller (2008), Kaas et al. (2008), Frees (2010), and Ohlsson and Johansson (2010).

### 2.2. Other Traditional Methods

The random parameter approach has been the most widely used to account for unobserved heterogeneity. Alternative approaches for addressing heterogeneity and panel effects include grouped random parameter (Fountas et al., 2019; Pantangi et al., 2019), correlated random parameter (Balusu et al., 2018; Fountas et al., 2019; Tran et al., 2015), bivariate/multivariate random parameter (Barua et al., 2015, 2016; Dong et al., 2014; Gong et al., 2022; Pantangi et al., 2019), and mixed generalized models (Anarkooli et al., 2017; Balusu et al., 2018; Chen et al., 2018; Eluru et al., 2008). The random parameter model, under the concept of hierarchical modeling, is also the most widely used technique (Bakhshi and Ahmed, 2021; Fountas and Anastasopoulos, 2017; Kim et al., 2017; Lord and Mannering, 2010). Mannering et al. (2016) summarize the methodological approaches accounting for unobserved heterogeneity.

Because the data structure consisting of unbalanced repeated measures and panel data can be problematic to analyze, GLMMs are suitable for this purpose. In the GLMM context, in addition to determining the structure of correlation between observations on the same subject, the random effects also consider heterogeneity among subjects due to unobserved features (Antonio and Beirlant, 2007). Since the mixed-effects logistic regression model is the most popular GLMM, the data are analyzed by means of the logistic GLMM in this study (Hedeker, 2005). In order to model the risk factors having a major impact on the likelihood of DCPD claims and to account for unobserved heterogeneity when addressing these risk factors, a random parameter approach is used in the framework of GLMMs in this study. This model is used to analyze binary data from TPL insurance, specifically DCPD and PD, by assuming the same random parameter mean and variance for all observations. In addition, a GLMM with independent, homoscedastic residual errors is specified for the likelihood of DCPD claims from the Canadian automobile insurance dataset in this study.

### 2.3. Data Description

Data about only private passenger automobiles are provided from the automobile portfolio of an active insurance company in Canada. The dataset includes insurance information about a total of 1,946 observations for 1,397 policies that have been in the portfolio for ten complete years, each of which consists of the claim experience for several rating factors and a given calendar year. The data do not contain insurance details for the policy year in which no claim was filed.

**Table 2.** Variables in the dataset

| **Variable** | **Definition** |
|---|---|
| Age | Age of policyholder at the time of claim |
| Territory | Residential area<br>0 (Urban), 1(Rural) |
| Usage | Vehicle usage<br>0 (Work/Business), 1 (Pleasure) |
| Time | Accident year<br>$j$ = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, corresponding to values of 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011 and 2012, respectively |
| Class | Code of class<br>0 (Vehicle used for pleasure or having vehicle usage restrictions for commuting to work one way, and driver is 25 years of age or over), 1 (Vehicle used for pleasure and business or not having vehicle usage restrictions for commuting to work one way, and driver is 25 years of age or over), 2 (Vehicle not having vehicle usage restrictions, and driver is under 21 years of age), 3 (Vehicle not having vehicle usage restrictions, and driver is under 25 years of age, but not under 21 years of age) |
| Driver record | Number of claims-free years for each policy (in the last 6 years)<br>0 (No claims-free years), 1 (One claims-free year), 2, 3, 4, 5, 6 |
| Claims history | Number of claims the risk has had in the last 6 years before the policy was rated<br>0 (Number of chargeable claims is zero), 1, 2 (Number of chargeable claims is two or more) |
| Claims-free years | Number of years since the risk had a claim<br>0 (Zero year), 1, 2, 3, 4, 5, 6, 7, 8, 9 (Nine or more years) |
| Experience | Number of years the driver has been licensed<br>0 (Zero year), 1, 2, 3, 4, 5, 6, 7, 8, 9 (Nine or more years) |
| Training | Driving education provided to all candidates<br>0 (Drivers have taken the course in Ontario), 1 (Drivers have taken the course, but maybe a different jurisdictionally specific one), 2 (Drivers have not taken the course) |
| Gender | 0 (Female), 1 (Male) |
| Outstanding loss | Loss reported to the insurer but is still in the process of settlement |
| Incurred loss | Amount actually paid in loss during a specified time |

The analysis is performed to the company's liability insurance claim experience for 2003–2012. The data comprise outstanding loss ($x^{(12)}$), which only includes zero and positive claim amounts, incurred loss ($x^{(13)}$), which only includes positive claim amounts, and several rating factors for each policy that consist of age ($x^{(1)}$), territory ($x^{(2)}$), usage ($x^{(3)}$), time ($x^{(4)}$), class ($x^{(5)}$), driver record ($x^{(6)}$), claims history ($x^{(7)}$), claims-free years ($x^{(8)}$), experience ($x^{(9)}$), training ($x^{(10)}$), and gender ($x^{(11)}$). Table 2 gives detailed information about the rating factors of the policy.

In the following analysis, territory ($x^{(2)}$), usage ($x^{(3)}$), class ($x^{(5)}$), training ($x^{(10)}$), and gender ($x^{(11)}$) are treated as factor covariates while age ($x^{(1)}$), time ($x^{(4)}$), driver record ($x^{(6)}$), claims history ($x^{(7)}$), claims-free years ($x^{(8)}$), experience ($x^{(9)}$), outstanding loss ($x^{(12)}$), and incurred loss ($x^{(13)}$) are treated as continuous covariates in the model.

Driver characteristics also involve the date of birth of the policyholders, while the claim profiles include information on the type of coverage regarding property damage, such as 0 (PD covered by liability insurance) and 1 (DCPD), policy effective and expiry date, claim identification number, and accident date.

The model is fitted using the claims for the years 2003–2008, and its predictive ability is evaluated using the claims from 2009–2012. The data for 2003–2008 consist of 1,169 observations on 942 policies for 179 brokers, and each observation includes the claim experience at the individual policy level. Of the 1,169 observations, 88 (7.5%) have PD covered by liability insurance and 1,081 (92.5%) have DCPD. These observations are summarized as shown in Table 3.

**Table 3.** Summary statistics of the data

| **Variable** | **Mean** | **Std.Dev.** | **Minimum** | **Maximum** |
|---|---|---|---|---|
| Age | 45.05 | 13.33 | 18.42 | 85.10 |
| Time | 3.78 | 1.54 | 1.00 | 6.00 |
| Driver record | 5.54 | 1.33 | 0.00 | 6.00 |
| Claims history | 0.08 | 0.28 | 0.00 | 2.00 |
| Claims-free years | 7.69 | 2.60 | 0.00 | 9.00 |
| Experience | 8.37 | 1.78 | 0.00 | 9.00 |
| Outstanding loss | 620.5 | 1356.19 | 0.00 | 5550.00 |
| Incurred loss | 3625.20 | 4129.49 | 26.84 | 43539.90 |

The analysis herein focuses on estimating the model using the property damages that occurred during each individual year to examine the likelihood of DCPD claims versus PD claims covered by TPL insurance. Table 4 presents the mean of the outstanding and incurred losses used in the forthcoming estimations for each of the six years.

**Table 4.** Mean of outstanding and incurred loss distribution by years

| Year | Outstanding Loss | Incurred Loss |
|---|---|---|
| 2003 | 1172.18 | 2502.82 |
| 2004 | 307.43 | 3498.63 |
| 2005 | 603.57 | 3612.98 |
| 2006 | 431.30 | 3735.09 |
| 2007 | 814.31 | 3695.84 |
| 2008 | 778.56 | 3966.12 |

To optimize the merits of the variables in the model, a transformation is applied to both outstanding and incurred losses. The Yeo-Johnson transformation handles both positive and negative values, whereas the Box-Cox transformation only handles positive values. Because outstanding loss only includes zero and positive claim amounts, the Yeo-Johnson transformation is made for outstanding loss. Incurred loss, on the other hand, only includes positive claim amounts. Therefore, the Box-Cox transformation is applied for incurred loss.

In the insurance portfolio, these observations are handled as separate classes. The frequency table of the classes is given in Table 5.

**Table 5.** Frequency table of the class

| Variable | Group | Number of Observations | Percent (%) |
|---|---|---|---|
| Class | 0 | 1,002 | 85.71 |
| | 1 | 126 | 10.78 |
| | 2 | 31 | 2.65 |
| | 3 | 10 | 0.86 |
| Total | | 1,169 | 100.00 |

Of the 1,169 observations, 1,002 (85.71%) include the drivers who are 25 years of age or over and use their vehicle for pleasure or have vehicle usage restrictions for commuting to work one way, 126 (10.78%) consist of drivers who are 25 years of age or over and use their vehicle for pleasure and business or not have vehicle usage restrictions for commuting to

work one way, 31 (2.65%) contain drivers who are under 21 years of age and not have vehicle usage restrictions, and 10 (0.86%) comprise drivers who are under 25 years of age, but not under 21 years of age and not have vehicle usage restrictions. Because the observations in the data are not distributed in a balanced way among the categories of class ($x^{(5)}$) from the factor covariates, the weights on class ($x^{(5)}$), which are the number of trials, are used to generate each success proportion. As a result, since the dataset is unbalanced, using weights allows us to consider the relative importance of various possible target values and to better fit the model.

Among the rating variables in the dataset, claims-free years ($x^{(8)}$) and experience ($x^{(9)}$) are highly correlated. The models are built considering the correlation between these variables and then compared to one another to determine the best model. In the following analysis, the best fitted model is presented. Table 6 shows the correlation between the independent variables in this fitted model.

**Table 6.** Correlation matrix of independent variables in the model

| | $x^{(3)}$ | $x^{(4)}$ | $x^{(10)}$ | $x^{(11)}$ | $x^{(12)}$ | $x^{(13)}$ |
|---|---|---|---|---|---|---|
| $x^{(3)}$ | 1.000 | – 0.012[c] | 0.007[a] | 0.033[b] | 0.016[c] | – 0.027[c] |
| $x^{(4)}$ | – 0.012[c] | 1.000 | 0.059[c*] | – 0.018[c] | 0.064[d] | 0.106[d] |
| $x^{(10)}$ | 0.007[a] | 0.059[c*] | 1.000 | 0.037[a] | – 0.058[c*] | 0.030[c*] |
| $x^{(11)}$ | 0.033[b] | – 0.018[c] | 0.037[a] | 1.000 | 0.016[c] | – 0.014[c] |
| $x^{(12)}$ | 0.016[c] | 0.064[d] | – 0.058[c*] | 0.016[c] | 1.000 | 0.051[d] |
| $x^{(13)}$ | – 0.027[c] | 0.106[d] | 0.030[c*] | – 0.014[c] | 0.051[d] | 1.000 |

[*] The greatest correlation between the discrete or continuous variable and all possible pairs of levels of the nominal variable

[a] Goodman and Kruskal's Lambda

[b] Phi coefficient

[c] Point-biserial correlation coefficient

[d] Spearman correlation coefficient

(Khamis, 2008)

As a result, the model presented below does not exhibit any multicollinearity issue. Within this model, 650 (55.6%) of the 1,169 observations use the vehicle for work and business, while 519 (44.4%) use it for pleasure. 20 (1.7%) of the observations include the drivers who have taken the course in Ontario, whereas 1,130 (96.7%) consist of those who have taken it in another jurisdiction. 19 (1.6%) of the observations also comprise the drivers

who have not taken the course. Female drivers make up 524 (44.8%) of the observations, while male drivers add up to 645 (55.2%).

### 2.4. Fitted Model

A random intercept effect model is a type of GLMM that allows for the inclusion of individual-specific random effects in addition to more general risk factors. This model can help to account for unobserved heterogeneity in the data, which can have a significant impact on the likelihood of claims. By incorporating random intercepts into the model, the effect of unobserved heterogeneity can be accounted for, resulting in more accurate estimates of risk and more appropriate insurance premiums.

This study aims to determine how the most significant risk factors affect DCPD claims under TPL insurance. Two categories are addressed to model the property damage coverage type following a traffic accident: DCPD or PD covered by liability insurance. The GLMM described in Section 2.1 is fitted using the glmer function in R with logit-link.

Using GLMM analysis for the subject-specific random intercept effect model, the best-fitting random intercept effect model is specified as follows:

$$g(\mu_{ijk}) = \beta_0 + \beta_1 x_{ijk}^{(3)} + \beta_2 x_{ijk}^{(4)} + \beta_3 x_{ijk}^{(10)} + \beta_4 x_{ijk}^{(11)} + \beta_5 x_{ijk}^{(12)} + \beta_6 x_{ijk}^{(13)} + b_{0k}, \quad i = 1, \dots, n,\ j = 1, \dots, t_i,\ k = 1, \dots, m \qquad (4)$$

where $n$ is the total number of different policies; $m$ is the total number of different brokers; $t_i$ is the number of repeated observations in policy $i$. $t_i$ is the same for all policies in balanced panel data, but conversely, the panel data structure here is unbalanced. In addition, $\mu_{ijk}$ is the probability of a claim on policy $i$ $(i = 1, \dots, 942)$ at time $j$ $(j = 1, \dots, 6)$ for broker $k$ $(k = 1, \dots, 179)$.

In the fixed-effects part of the model, the parameters $\beta_0$, $\beta_1$, and $\beta_2$ define an overall intercept, the change in the expected log odds of DCPD claims for vehicle usage, and the change caused by a one-year change in time, for a given the random intercept, respectively. The change in the expected log odds of DCPD claims for driving education and gender are expressed in parameters $\beta_3$ and $\beta_4$, for a given the random intercept, respectively. Additionally, $\beta_5$ and $\beta_6$ describe how the expected log odds of DCPD claims have changed due to a unit increase in both outstanding loss and incurred loss for a given the random intercept.

In the random-effects part of the model, the term $b_{0k}$ in Equation (4) denotes a broker-specific random intercept. The random intercept $b_{0k}$ is a subject-specific deviation from the fixed intercept $\beta_0$. The results of the panel data generalized linear binary logit mixed model are summarized in Table 7.

**Table 7.** Generalized linear binary logit mixed model estimation results

| Parameter | Variable | Estimated Coefficients | Std. Error | z-value | Pr(>\|z\|) | Exp($\beta$) |
|---|---|---|---|---|---|---|
| $\beta_0$ | Intercept | - 4.691 | 0.694 | - 6.761 | < 0.001 *** | 0.009 |
| $\beta_1$ | Usage1 | - 1.083 | 0.268 | - 4.038 | < 0.001 *** | 0.339 |
| $\beta_2$ | Time | 0.157 | 0.085 | 1.844 | 0.065 . | 1.170 |
| $\beta_3$ | Training1 | 1.215 | 0.464 | 2.621 | 0.009 ** | 3.372 |
| | Training2 | 3.411 | 1.235 | 2.763 | 0.006 ** | 30.279 |
| $\beta_4$ | Gender1 | - 0.463 | 0.275 | - 1.680 | 0.093 . | 0.629 |
| $\beta_5$ | Outstanding loss | 0.187 | 0.063 | 2.948 | 0.003 ** | 1.205 |
| $\beta_6$ | Incurred loss | 0.946 | 0.075 | 12.665 | < 0.001 *** | 2.576 |
| | | | | | | |
| $b_{0k}$ | Random parameter | | | | | |
| | Std. dev. of broker | 0.632 | | | | |
| | - 2 Log-likelihood | 462.0 | | | | |
| | AIC | 480.1 | | | | |
| | BIC | 525.7 | | | | |

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## 2.5. Performance Metrics

In this study, the GLMM approach is applied to unbalanced panel data to determine which factors have a significant impact on the likelihood of DCPD claims that policyholders will make next year. To inform model selection, the Akaike information criterion (*AIC*) and likelihood ratio test (*LRT*) are used. If the number of observations (*N*) is large enough, $k < (N/40)$, *AIC* is defined as

$$AIC = -2\,ln(\hat{L}) + 2k \qquad (5)$$

where $k$ represents the number of estimated parameters in the fitted model and $ln(\hat{L})$ is the maximum log-likelihood value (Portet, 2020). In this study, Equation (5) is used to calculate the *AIC* value since $k = 9$ is smaller than $N/40 = 29.225$ for $N = 1169$, and the model with a lower *AIC* value is preferred.

The reference model, which includes weights (the number of trials used to generate each success proportion) in the different classes of insureds, is compared to the nested model, which is reduced to a model without weights, using likelihood ratio tests to determine which is statistically preferable. The likelihood ratio test is shown in Equation (6).

$$LRT = 2\{\text{logLik(reference)} - \text{logLik(nested)}\} \quad (6)$$

where logLik (reference) and logLik (nested) are the log-likelihood of the generalized linear mixed model with weights (under the alternative hypothesis) and the generalized linear mixed model without weights (under the null hypothesis) for the same dataset, respectively. With degrees of freedom equal to the difference in the number of parameters between the two models, the test statistic is a chi-square distribution (Pai and Walch, 2020). The chi-square value of the test is 64.058 with one degree of freedom. The corresponding *p*-value is$(0.5)\ Pr(\chi_1^2 > 64.058)$. From the chi-square table, we can conclude that $Pr(\chi_1^2 > 7.88) = 0.005$ and hence the *p*-value is significantly lower than 0.0025. The model under the alternative hypothesis is chosen since the *p*-value is much less than 0.05. In other words, the random-effects model with weights is preferred because it significantly differs from the random-effects model without weights.

The evaluation metrics used in this analysis include measures of sensitivity (recall) (*SN*), specificity (*SP*), precision (*P*), accuracy (*ACC*), balanced accuracy (*BA*), *F*1 score, and area under the curve (*AUC*) to assess the performance of each model and to determine which model is most effective for predicting the likelihood of DCPD claims. These measures are defined based on a confusion matrix, as shown in Table 8 (Hossin and Sulaiman, 2015).

**Table 8.** Confusion matrix for the binary classification

| | **Prediction** | |
|---|---|---|
| **Actual** | **DCPD** | **PD** |
| **DCPD** | *TP* | *FN* |
| **PD** | *FP* | *TN* |

In this confusion matrix, *TP* (true positive) and *TN* (true negative) denote the number of positive (classifying the claim as DCPD) and negative (classifying the claim as PD covered by liability insurance) claims that are correctly classified, respectively. Additionally, *FP* (false positive) and *FN* (false negative) represent the number of positive and negative claims that are incorrectly classified, respectively. In other words, *TP* and *TN* indicate DCPD claims correctly identified as DCPD and PD claims correctly identified as PD, respectively. *FP* stands for PD claims incorrectly identified as DCPD, whereas *FN* implies DCPD claims incorrectly identified as PD. The performance evaluation metrics used in this analysis are generated as shown in Equations (7) – (12).

$$SN = \frac{TP}{TP+FN} \tag{7}$$

$$SP = \frac{TN}{TN+FP} \tag{8}$$

$$P = \frac{TP}{TP+FP} \tag{9}$$

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \tag{10}$$

$$BA = \frac{SN+SP}{2} \tag{11}$$

$$F1\ \text{score} = \frac{2TP}{2TP+FP+FN} \tag{12}$$

Precision and recall are employed as the evaluation metrics in this study since the developed model aims to predict 1 as accurately as feasible and to identify as many actual 1 as possible. In classification issues, accuracy is one of the most frequently used evaluation metrics. It is helpful when the target class is well-balanced but not a suitable option when the classes are unbalanced. This study assesses the target classes that are to be applied to a severely unbalanced dataset in which positives greatly outnumber negatives. Balanced accuracy is chosen as a performance measure in this analysis because it is a better metric when dealing with imbalanced data, and it also accounts for both positive and negative classes and avoids data imbalances that could be misleading. Additionally, the *F*1 score is a commonly employed evaluation metric to measure the performance of binary classification and outperforms accuracy in enhancing the target classes for binary classification problems. Therefore, it is used in this analysis as a performance measure rather than an accuracy measure.

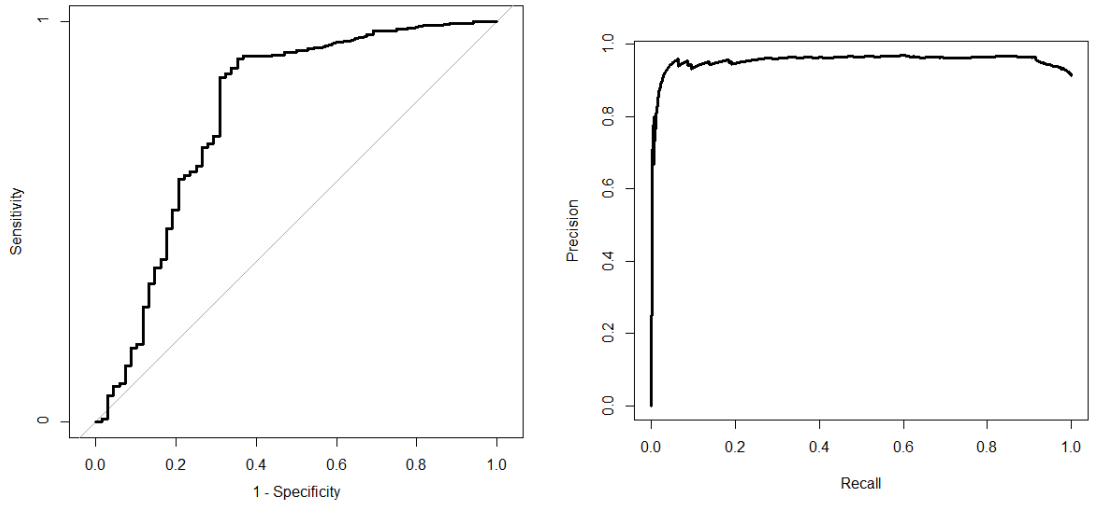
Another evaluation metric is the Receiver Operating Characteristic (ROC) curve, which assesses the predictive performance of the fitted model. The ROC curve is a plot of the true positive rate (*SN*) versus the false positive rate (1-*SP*), which shows how the number of correctly classified positive instances varies with the number of incorrectly classified negatives when evaluating binary decision problems. The ROC curve captures the trade-off between these performance measure parameters for different possible thresholds. The resulting score known as *AUC* is the area under the ROC curve and illustrates the model's ability to accurately predict classes. A higher score indicates a higher probability of making correct predictions and can be viewed as a measure of accuracy (Davis and Goadrich, 2006).

A Precision-Recall (PR) curve, on the other hand, evaluates the fraction of true positives among positive predictions. By offering valuable insights into the effectiveness of the classification model in capturing and correctly labeling minority class instances, the PR curve can provide an accurate prediction of future classification performance. The PR curve outperforms the ROC curve in terms of both information and power when dealing with binary classes on unbalanced datasets (Saito and Rehmsmeier, 2015). Due to class imbalance in this analysis, presenting results by considering only the ROC curve could be misleading about the reliability of classification performance. In this study, as well as the ROC curve, the PR curve is also considered to evaluate the classification performance because the PR curve can explicitly reveal claim differences in imbalanced cases. The resulting score known as *PR AUC* is the area under the PR curve and emphasizes the performance of the model for predicting the positive class. A high *PR AUC* means that the model performs better in predicting the positive class. These performance assessment measures are acquired as presented in Table 9.

**Table 9.** Performance evaluation metrics

| **Sensitivity** | **Specificity** | **Precision** | **Accuracy** |
|---|---|---|---|
| 0.906 | 0.647 | 0.964 | 0.883 |
| | | | |
| **Balanced Accuracy** | ***F*1 Score** | ***AUC*** | ***PR AUC*** |
| 0.776 | 0.934 | 0.776 | 0.953 |
| | | | |

The fitted model's $F1$ score of 0.934, which is regarded as a very good value, indicates that it can both capture positive classes and accurately predict the classes it does capture. Regarding the balanced accuracy, it has a value of 0.776, indicating that the fitted model performs well at predicting whether policyholders will make DCPD claims. Due to the imbalanced classes in this analysis, the balanced accuracy gives us a more realistic picture of how well the model classifies both groups correctly. To evaluate the predictive performance of the fitted model, the ROC and PR curves are plotted as shown in Figure 1.




(a) (b)

**Figure 1.** Predictive performance of the fitted model: (a) the ROC curve and (b) the PR curve

For the fitted model using different probability thresholds, the ROC curve highlights the trade-off between the true positive rate and the false positive rate. The fitted model provides a good fit to the data according to the computed *AUC* of 0.776. For the fitted model employing different probability thresholds, the PR curve highlights the trade-off between the true positive rate and the positive predictive value. Compared to the ROC curve, the PR curve is preferable to the ROC curve for imbalanced datasets. Due to the class imbalance in this analysis, *PR AUC*, calculated as 0.953, describes that the fitted model performed very well in predicting the positive class.

## 3. RESULTS

This paper describes a generalized linear binary logit mixed model considering the imbalance between the classes of policyholders using automobile insurance data. This model

assesses the impact of various risk factors on DCPD claims in private passenger vehicle accidents. The risk factors having a significant impact on the likelihood of DCPD claims are the independent variables named “usage”, “time”, “training”, “gender”, “outstanding loss”, and “incurred loss” estimated in unbalanced longitudinal data.

Gender and time are included in the model even if they are thought to be ineffective. However, these two variables are significant at the 0.10 level, as shown in Table 7. The results of these variables indicate that female drivers are 1.59 times more likely to make a DCPD claim than male drivers, and that the risk of a DCPD claim occurring is 1.17 times higher when time increases by 1 year.

As for other significant variables, usage, training, outstanding loss, and incurred loss have a significant effect on the likelihood of DCPD claims. For policyholders who use their vehicles for work or business, the risk of making a DCPD claim is 2.95 times greater than for those who use them for pleasure. Since drivers who commute to work or use the vehicle for business are far more likely to be in traffic than those who drive for pleasure, this result is meaningful and the vehicle usage has a quite significant effect on DCPD claims.

Driver training is of vital importance in preventing traffic accidents. Even if most drivers in Ontario have taken courses, some have not taken any training. Given that Ontario is one of the provinces with the highest number of immigrants, many drivers have taken driver training in various jurisdictions, whereas some have taken it in Ontario. According to the results of the training variable in the model, policyholders who have taken the driver training in a separate jurisdiction are 3.37 times more likely to make a DCPD claim than those who have taken it in Ontario; whereas policyholders who have not taken courses are 30.28 times more likely to make a DCPD claim than those with driver training in Ontario. These results indicate that drivers who have taken courses in a different jurisdiction or have not taken any training pose a risk in traffic and support the importance of driver training in preventing traffic accidents.

For insurers to manage their claims liabilities, determine appropriate premium rates, and evaluate their overall financial circumstances, outstanding loss and incurred loss are crucial. The claim reported to the insurance company but has not yet been paid is known as an outstanding loss. This claim is an estimate of the insurer's future financial obligations. Incurred loss, also called paid loss, is the actual loss that the insurance company has paid or became obligated to pay during a specific period. The results of these two variables in the model demonstrate that the risk of a DCPD claim occurring is 1.21 times higher when the outstand-

ing loss increases by $1 and that the risk of a DCPD claim occurring is 2.58 times greater when the incurred loss increases by $1.

DCPD claims are one of the most common types of damage insurance companies incur. DCPD coverage under TPL insurance provides compensation to policyholders for damages by the policyholders' own insurer in cases where the accident was caused by another driver and was not their own fault. It can indeed be advantageous to consider these rating factors which significantly affect the likelihood of DCPD claims for evaluating insurance premiums and enhancing the financial stability of an insurance company. By incorporating these factors into the premium evaluation process, insurers can more accurately estimate the risk associated with each policyholder and price premiums accordingly.

It is recommended that the above rating factors having a significant impact on the likelihood of DCPD claims be considered in the premium evaluation since it is believed to help the financial stability of the insurance company. The financial stability of the company could potentially be affected if the insurance company pays more compensation than it collects in premiums.

## 4. CONCLUSIONS

The purpose of this study is to develop a statistical model that identifies the impact of the most important risk factors on DCPD claims under TPL insurance in private passenger vehicle accidents in Ontario, Canada. GLMMs are approaches that are constantly used to model actuarial data and provide an advantage in the analysis of unbalanced panel data. This approach eliminates the extra complexity resulting from the internal correlation of each policy. Therefore, the developed model in this study analyzes the likelihood of DCPD claims in the context of a generalized linear binary logit mixed model by dealing with unbalanced panel data, and also, the imbalance between the classes of insureds is considered in this model.

As a type of data application, the data in this study include many factors associated with the driver and claim characteristics found critical to the likelihood of DCPD claims. The estimation results from the model demonstrate that the broker, which is a time-varying factor, has a significant influence on the likelihood of DCPD claims as a random parameter. In addition, rating factors such as usage patterns, driver training, outstanding loss and incurred loss have been found to correlate with the likelihood of DCPD claims as fixed effects. Observing the effects of these risk factors under the weights in the different classes of policyholders high-

lights the importance of developing class-specific risk assessment models. Moreover, by considering the performance evaluation metrics in detail, this study ensures a comprehensive assessment that accounts for the potential challenges posed by imbalanced datasets and provides a more reliable interpretation of the results.

Taking these factors into account during premium evaluation helps insurers maintain financial stability by ensuring that premiums are adequately priced based on the associated risks. This, in turn, helps the company avoid potential financial instability caused by underpricing policies or facing a higher volume of claims than anticipated.

Ultimately, incorporating rating factors that have a significant impact on the likelihood of DCPD claims in premium evaluation promotes a fair and sustainable insurance pricing strategy, benefiting both the insurance company and its policyholders.

## ACKNOWLEDGMENTS

In addition to the editor and the anonymous referees for their constructive suggestions, the authors would like to thank Dr. Jeffrey S. Pai from the University of Manitoba and the Canadian insurance company for providing the data used for conducting the analysis. This work was partially supported by the Scientific and Technological Research Council of Turkey (TUBITAK 2214-A).

## ETHICAL DECLARATION

In the writing process of the study titled “Quantifying the impact of risk factors on direct compensation property damage in Canadian automobile insurance”, there were followed the scientific, ethical and the citation rules; was not made any falsification on the collected data and this study was not sent to any other academic media for evaluation.

## REFERENCES

Anarkooli, A. J., Hosseinpour, M. and Kardar, A. (2017), Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model, *Accident Analysis and Prevention*, 106, 399-410.

Antonio, K. and Beirlant, J. (2007), Actuarial statistics with generalized linear mixed models, *Insurance: Mathematics and Economics*, 40(1), 58-76.

Antonio, K. and Valdez, E. A. (2012), Statistical concepts of a priori and a posteriori risk classification in insurance, *AStA Advances in Statistical Analysis*, 96, 187-224.

Bakhshi, A. K. and Ahmed, M. M. (2021), Practical advantage of crossed random intercepts under Bayesian hierarchical modeling to tackle unobserved heterogeneity in clustering critical versus non-critical crashes, *Accident Analysis and Prevention*, 149, 105855.

Balusu, S. K., Pinjari, A. R., Mannering, F. L. and Eluru, N. (2018), Non-decreasing threshold variances in mixed generalized ordered response models: A negative correlations approach to variance reduction, *Analytic Methods in Accident Research*, 20, 46-67.

Barua, S., El-Basyouny, K. and Islam, M. T. (2015), Effects of spatial correlation in random parameters collision count-data models, *Analytic Methods in Accident Research*, 5, 28-42.

Barua, S., El-Basyouny, K. and Islam, M. T. (2016), Multivariate random parameters collision count data models with spatial heterogeneity, *Analytic Methods in Accident Research*, 9, 1-15.

Chen, F., Chen, S. and Ma, X. (2018), Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data, *Journal of Safety Research*, 65, 153-159.

Davis, J. and Goadrich, M. (2006), The relationship between Precision-Recall and ROC curves, In: *Proceedings of the 23rd International Conference on Machine Learning – ICML '06*, 233-240.

De Jong, P. and Heller, G. Z. (2008), *Generalized Linear Models for Insurance Data*, In: International Series on Actuarial Science, Cambridge University Press.

Dong, C., Clarke, D. B., Yan, X., Khattak, A. and Huang, B. (2014), Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections, *Accident Analysis and Prevention*, 70, 320-329.

Eluru, N., Bhat, C. R. and Hensher, D. A. (2008), A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes, *Accident Analysis and Prevention*, 40(3), 1033-1054.

Embrechts, P. and Wüthrich, M. V. (2022), Recent challenges in actuarial science, *Annual Review of Statistics and Its Application*, 9, 119-140.

Frees, E. W. (2010), *Regression Modeling with Actuarial and Financial Applications*, In: International Series on Actuarial Science, Cambridge University Press.

Fountas, G. and Anastasopoulos, P. C. (2017), A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities, *Analytic Methods in Accident Research*, 15, 1-16.

Fountas, G., Pantangi, S. S., Hulme, K. F. and Anastasopoulos, P. C. (2019), The effects of driver fatigue, gender, and distracted driving on perceived and observed aggressive driving behavior: A correlated grouped random parameters bivariate probit approach, *Analytic Methods in Accident Research*, 22, 100091.

Garrido, J., Genest, C. and Schulz, J. (2016), Generalized linear models for dependent frequency and severity of insurance claims, *Insurance: Mathematics and Economics*, 70, 205-215.

Gong, H., Fu, T., Sun, Y., Guo, Z., Cong, L., Hu, W. and Ling, Z. (2022), Two-vehicle driver-injury severity: A multivariate random parameters logit approach, *Analytic Methods in Accident Research*, 33, 100190.

Haberman, S. and Renshaw, A. E. (1996), Generalized linear models and actuarial science, *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(4), 407-436.

Hedeker, D. (2005), Generalized linear mixed models, In: B. Everitt, D. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*, John Wiley & Sons, New York, 729-738.

Hossin, M. and Sulaiman, M. N. (2015), A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining and Knowledge Management Process*, 5(2), 1-11.

Kaas, R., Goovaerts, M., Dhaene, J. and Denuit, M. (2008), *Modern Actuarial Risk Theory: Using R*, Second Edition, Springer Berlin, Heidelberg.

Khamis, H. (2008), Measures of association: How to choose?, *Journal of Diagnostic Medical Sonography*, 24(3), 155-162.

Kim, M., Kho, S. Y. and Kim, D. K. (2017), Hierarchical ordered model for injury severity of pedestrian crashes in South Korea, *Journal of Safety Research*, 61, 33-40.

Lord, D. and Mannering, F. (2010), The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives, *Transportation Research Part A: Policy and Practice*, 44(5), 291-305.

Mannering, F. L., Shankar, V. and Bhat, C. R. (2016), Unobserved heterogeneity and the statistical analysis of highway accident data, *Analytic Methods in Accident Research*, 11, 1-16.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, In: Monographs on Statistics and Applied Probability 37, Second Edition, Chapman and Hall, London, New York.

Miao, G. M. (2018), Application of hierarchical model in non-life insurance actuarial science, *Modern Economy*, 9(3), 393-399.

Nelder, J. A. and Wedderburn, R. W. M. (1972), Generalized linear models, *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370-384.

Ohlsson, E. and Johansson, B. (2010), *Non-life Insurance Pricing with Generalized Linear Models*, In: EAA Series Textbook, Springer Berlin, Heidelberg.

Pai, J. S. and Walch, A. H. (2020), *ACTEX Study Manual for Exam MAS-II*, ACTEX Learning/SRBooks, Inc., Greenland, NH.

Pantangi, S. S., Fountas, G., Sarwar, M. T., Anastasopoulos, P. C., Blatt, A., Majka, K., Pierowicz, J. and Mohan, S. B. (2019), A preliminary investigation of the effectiveness of high visibility enforcement programs using naturalistic driving study data: A grouped random parameters approach, *Analytic Methods in Accident Research*, 21, 1-12.

Portet, S. (2020), A primer on model selection using the Akaike Information Criterion, *Infectious Disease Modelling*, 5, 111-128.

Saito, T. and Rehmsmeier, M. (2015), The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PloS One*, 10(3), e0118432.

Tran, V., Liu, D., Pradhan, A. K., Li, K., Bingham, C. R., Simons-Morton, B. G. and Albert, P. S. (2015), Assessing risk-taking in a driving simulator study: Modeling longitudinal semi-continuous driving data using a two-part regression model with correlated random effects, *Analytic Methods in Accident Research*, 5, 17-27.

Yau, K., Yip, K. and Yuen, H. K. (2003), Modelling repeated insurance claim frequency data using the generalized linear mixed model, *Journal of Applied Statistics*, 30(8), 857-865.