DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

AUTOMOBILE INSURANCE RATEMAKING: CLASS RATING AND MERIT RATING

by
Pervin BAYLAN

February, 2024
İZMİR

AUTOMOBILE INSURANCE RATEMAKING: CLASS RATING AND MERIT RATING

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Statistics

by
Pervin BAYLAN

February, 2024
İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**AUTOMOBILE INSURANCE RATEMAKING: CLASS RATING AND MERIT RATING**" completed by **PERVIN BAYLAN** under supervision of **PROF. DR. NESLIHAN DEMİREL** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

....

Prof. Dr. Neslihan DEMİREL

Supervisor

....

Prof. Dr. Fatih TANK

Thesis Committee Member

....

Prof. Dr. Burcu HÜDAVERDİ

Thesis Committee Member

....

Prof. Dr. Gözde ULUTAGAY

Examining Committee Member

....

Prof. Dr. Ufuk YOLCU

Examining Committee Member

Prof. Dr. Okan FISTIKOĞLU
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

First of all, I am very grateful to my supervisor, Prof. Dr. Neslihan DEMİREL, for her valuable advice, guidance, continuous support, and patience during my academic life and Ph.D. study. Her personality, positivity, and openness to innovation have always inspired me.

I want to thank Prof. Dr. Fatih TANK and Prof. Dr. Burcu HÜDAVERDİ, who are on the thesis meeting committee and shared their knowledge and suggestions about my work and guided me with their contributions.

I am grateful to Prof. Dr. Serdar KURT for his unwavering support and patience during my academic life. I would also like to thank Prof. Dr. Jeffrey S. PAI for his valuable advice on the research subject. I want to thank Prof. Dr. Umay Zeynep UZUNOĞLU KOÇER for her support throughout my work. I would also like to thank Uğur BİNZAT for his support both academically and as a friend.

I also gratefully acknowledge the Canadian insurance company for providing the data used for conducting the analysis. This thesis was partially supported by the Scientific and Technological Research Council of Turkey (TUBITAK 2214-A).

I also take this opportunity to express gratitude to all of the Department of Statistics members for their help and support.

I am so grateful to my dear family, my fiance, and my friends for their invaluable support. Especially my fiance motivated me and gave me strong support. I want to thank them for their patience and encouragement. Finally, I dedicate this thesis to my mother, who does not currently exist in the world and wishes to witness my achievement.

Pervin BAYLAN

AUTOMOBILE INSURANCE RATEMAKING: CLASS RATING AND MERIT RATING

ABSTRACT

The use of statistical modeling in actuarial science enables the integration of risk factors into the premium pricing process, thereby enhancing the accuracy of insurance premiums and mitigating the financial risk for insurers. One of the aims of this thesis is to present 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-year period between 2003 and 2012, a statistical model of property damage (PD) is 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 are observed under the weights — the number of trials used to generate each success proportion — in the different classes of insureds. The performance metrics calculated by considering the class imbalance in binary outcomes, $F1$ score = 0.934, and PR AUC = 0.953, indicate that the model performs well in the classification. The other metrics also support this model's ability to accurately predict classes.

Another aim of this thesis is to evaluate the premiums by considering the claim types under third-party liability (TPL) insurance. In this thesis, a statistical analysis is presented that examines the effect of various risk factors on incurred PD and bodily injury (BI) losses in private passenger vehicle accidents. The PD and BI claims are explored via a broker-specific random intercept effect model and a valuation year-specific random intercept effect model, respectively.

The results indicate that several risk factors — class, modifier, claims history, and time — have a significant impact on the incurred losses of PD claims. For BI claims, the risk factors that are correlated with change of the incurred losses are also class,

rate modifier, gender, valuation year, and time, observed their effects under the heterogeneity of residual variances between the class groups. The performance metrics, *R-squared* = 0.7779 for the PD claims, and *R-squared* = 0.7157 for the BI claims, verify the ability of models to accurately predict the incurred losses. The other metrics also support that these models perform well in the prediction.

Over these predicted incurred losses, credibility premiums are calculated for each claim type by using the Bühlmann-Straub model. When calculating credibility premiums, the predictions in statistical modeling are weighted by earned exposures. The results obtained in the Bühlmann-Straub model indicate that the variance between claim types is much smaller than the variance within the types of claim. Furthermore, the credibility premium for BI claims is much higher than for PD claims. In addition, bonus-malus scales are designed by considering the claim types. The premiums are distributed reasonably to bonus-malus levels when the system is designed by considering the types of claims.

In summary, the statistical modeling employed in this thesis provides information about the risk characteristics of the policyholders crucial for determining the basic premium. The findings of this analysis can help insurers better understand the underlying drivers of PD and BI. In addition, these findings can support insurers in developing more accurate and effective strategies for risk mitigation. These results indicate that it is important to consider whether the claims are property or bodily in the evaluation to be made based on both the severity of claims and the number of claims.

Keywords: Credibility theory, bonus-malus scale, linear mixed models, premium evaluation, third-party liability insurance.

OTOMOBİL SİGORTASINDA AKTÜERYAL TARİFE: SINIF DEĞERLENDİRMESİ VE HASARSIZLIK İNDİRİM DEĞERLENDİRMESİ

ÖZ

Aktüerya biliminde istatistiksel modelleme kullanımı, risk faktörlerinin prim fiyatlandırma sürecine entegre edilmesini sağlayarak sigorta primlerinin doğruluğunu arttırır ve sigorta şirketleri için finansal riski azaltır. Bu tezin amaçlarından biri, özel binek araç kazalarında çeşitli risk faktörlerinin doğrudan tazmin edilen maddi hasarlar (Direct Compensation Property Damage - DCPD) üzerindeki etkisini değerlendiren istatistiksel bir analiz sunmaktır. 2003 ile 2012 yılları arasındaki on yıllık döneme ait Ontario, Kanada'daki otomobil sigortası verileri kullanılarak, genelleştirilmiş doğrusal ikili logit karma model aracılığıyla maddi (PD) hasarın istatistiksel bir modeli araştırılmış ve sigortalıların sınıfları arasındaki dengesizlik dikkate alınmıştır. Elde edilen 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. İkili sonuçlardaki sınıf dengesizliği dikkate alınarak hesaplanan performans ölçümleri, $F1$ skoru = 0,934 ve $PR\ AUC$ = 0,953, modelin sınıflandırmada iyi performans gösterdiğine işaret etmektedir. Diğer ölçümler de, bu modelin sınıfları doğru tahmin etme yeteneğini desteklemektedir.

Bu tezin bir diğer amacı da, üçüncü şahıs mali mesuliyet (TPL) sigortası kapsamındaki hasar türlerini dikkate alarak primleri değerlendirmektir. Bu tezde, özel binek araç kazalarında gerçekleşen PD ve bedensel (BI) hasarları üzerinde çeşitli risk faktörlerinin etkisini inceleyen istatistiksel bir analiz sunulmaktadır. PD ve BI hasarları, sırasıyla sigorta aracısına (broker) özgü rastgele sabit etki modeli ve değerlendirme yılına (valuation year) özgü rastgele sabit etki modeli aracılığıyla incelenmiştir.

Elde edilen sonuçlar, sınıf, modifikatör, hasar geçmişi ve zaman gibi çeşitli risk faktörlerinin gerçekleşen PD hasarları üzerinde önemli bir etkisinin olduğunu göstermektedir. BI hasarları için ise, gerçekleşen hasarların değişimi ile ilişkilendirilen risk faktörleri, sınıf, oran modifikatörü, cinsiyet, değerlendirme yılı ve zaman olup, etkileri sınıf grupları arasındaki artık varyanslarının heterojenliği altında gözlemlenmiştir. Performans ölçümleri, PD hasarları için R^2 = 0,7779 ve BI hasarları için R^2 = 0,7157, modellerin gerçekleşen hasarları doğru bir şekilde tahmin etme yeteneğini doğrulamaktadır. Diğer ölçümler de, bu modellerin tahminde iyi bir performans gösterdiğini desteklemektedir.

Tahmin edilen gerçekleşen hasarlar üzerinden, her hasar türü için Bühlmann-Straub modeli kullanılarak kredibilite primleri hesaplanmaktadır. Kredibilite primleri hesaplanırken, istatistiksel modellemedeki tahminler kazanılmış risklere (earned exposures) göre ağırlıklandırılır. Bühlmann-Straub modelinden elde edilen sonuçlar, hasar türleri arasındaki varyansın, hasar türleri içindeki varyanstan çok daha küçük olduğunu göstermektedir. Ayrıca, BI hasarları için kredibilite primi PD hasarlarına kıyasla çok daha yüksektir. Bunun yanı sıra, hasar türlerini dikkate alarak ödül-ceza ölçekleri (bonus-malus scales) tasarlanmıştır. Sistem, hasar türlerini göz önüne alarak tasarlandığında, primler ödül-ceza seviyelerine makul bir şekilde dağılmaktadır.

Özetle, bu tezde kullanılan istatistiksel modelleme, sigortalıların temel primin belirlenmesinde önemli olan risk özellikleri hakkında bilgi sağlamaktadır. Bu analizin bulguları, sigortacıların PD ve BI hasarlarının altında yatan nedenleri daha iyi anlamalarına yardımcı olabilir. Ayrıca bu bulgular, sigortacıların risk azaltımı için daha doğru ve etkili stratejiler geliştirmelerine destek olabilir. Bu sonuçlar, hem hasar şiddetlerine hem de hasar sayılarına göre yapılacak değerlendirmede, hasarların maddi mi yoksa bedensel mi olduğunun dikkate alınmasının önemli olduğunu göstermektedir.

Anahtar kelimeler: Kredibilite teorisi, ödül-ceza ölçeği, doğrusal karma modeller, prim değerlendirmesi, üçüncü şahıs mali mesuliyet sigortası.

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

ABBREVIATIONS

TPL	:	Third-Party Liability
BI	:	Bodily Injury
PD	:	Property Damage
DCPD	:	Direct Compensation Property Damage
BMS	:	Bonus-Malus System
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
LMM	:	Linear Mixed Model
RMSE	:	Root Mean Square Error
MSE	:	Mean Square Error
MAPE	:	Mean Absolute Percentage Error
MTD	:	Month-to-Date
GEE	:	Generalized Estimating Equation

CHAPTER ONE
INTRODUCTION

1.1 Statement of the Topic

In this introductory section, non-life insurance is presented, and the issue of pricing in this industry is delineated. Non-life insurance pricing is the art of setting the price of an insurance policy by considering various properties of the insured property and the policyholder. A non-life insurance policy is also an agreement between an insurance company and an insured to compensate the policyholder for certain unpredictable losses occurring during a defined period in the future against the premium. By the insurance contract, economic risk is transferred from the insured to the insurer, and the price, in other words, the premium charged to policyholders should be based on the expected loss transferred from the policyholder to the insurer. Hence, the premium serves as an estimation of the forthcoming costs associated with the insurance coverage.

A non-life insurance policy can extend coverage to damages incurred on a car, house, or other property, as well as losses resulting from bodily injury to the insured or third-party liability (TPL). Essentially, any insurance excluding life coverage falls under the non-life insurance category, commonly referred to as property and casualty insurance. Consequently, automobile insurance is one of the types of property and casualty insurance, with coverages divided into first and third-party coverage. Mandatory TPL insurance protects the policyholder deemed legally responsible for bodily injury (BI), death, or property damage (PD) to another party. This sector constitutes a significant portion of the annual non-life premium collection in developed countries. The economic significance and the accessibility of detailed information, encompassing policyholders' characteristics and claim histories, explain why a substantial portion of non-life actuarial literature is dedicated to this particular line of business.

The rise in the global number of registered motor vehicles has led to an increase in accidents, resulting in a higher number of fatalities and injuries. Consequently, this situation has not only prompted the implementation of mandatory TPL insurance to cover the adverse outcomes of accidents in many developing countries but has also confronted actuaries with the problem of determining how the rating system should be structured.

Today, the market has been deregulated in many countries. It means that if an insurance company charges too high a premium for some contracts, these policies will be lost to a competitor with a much fairer premium. This adverse selection will result in economic loss both ways for the insurance company by losing profit and gaining underpriced contracts. Therefore, in a competitive market, it is advantageous to charge a fair premium for the policyholder, that is, by fairness each insured should pay a premium corresponding to the expected losses transferred to the insurance company. The expected losses differ among policies due to variations in the accident rates of policyholders. Consequently, a driver with a higher likelihood of accidents should incur higher premiums for automobile insurance. The 2012 final report to the Insurance Bureau of Canada says that pricing adequacy in the aggregate is important for the long-run sustainability of the insurance system. In private sector insurance systems, pricing inadequacy will distort insurance supply, reducing competition, and insurance availability. Hence, pricing models that are used in the premium determination should be designed to maintain the encouragement for safe driving, treat different policyholders fairly based on their risk characteristics, and provide affordable insurance and a fair rate of return to insurers.

Because of the passing of the free tariff in traffic, not only do premiums differ from company to company, but there is also a dramatic diversity between the premiums charged by insurance companies. This situation raises unfair price competition between insurance companies due to premium inadequacy and can’t maintain to provide affordable premiums to policyholders. The lack of profitability has become a chronic problem in the sector as a consequence of extreme competition and wrong pricing.

In light of this information, the requirement regarding the solution of pricing problems in TPL insurance in the current market calls to mind the question of how fairly and accurately the existing premium assessment is done.

1.2 Motivation

The accurate calculation of the premiums, which is the keystone of the sector, plays a key role in the prevention of suffering loss in the insurance sector. The accurate premium calculation means that the premium will cover the expected losses and expenses and provide the targeted profit for the entity assuming the risk. Therefore, one of the aims of this thesis is to determine the premium for compulsory TPL insurance by estimating the losses next year.

An improvement in the premium assessment for this line of business will make the insurance business one of the most dynamic sectors of the economy. Therefore, the fact that pricing models that are used in the premium determination are designed to charge fairly policyholders premium is of vital importance in terms of constructing more efficient systems in risk evaluations and premium calculation. Another purpose of this thesis is to evaluate claims as PD and BI separately, obtain the rating factors important for each claim type through statistical models, and calculate the premiums for PD and BI claims based on the predictions obtained from these models.

Class rating and merit rating are highly effective premium calculation methods in determining insurance rates. In this research, the Bühlmann-Straub credibility model as class rating and the multi-event bonus-malus scales as merit rating will be examined. Both the class rating and the merit rating will be designed taking into account the types of claims and the results will be compared with the insurance company's written premium.

In light of this information, it is expected that in this thesis, the combination of class rating and merit rating considering the PD and BI claims separately in private passenger vehicle accidents will play an effective role in premium production.

Direct Compensation Property Damage (DCPD) claims have a significant share in TPL insurance compared to PD claims. Quantifying the impact of risk factors on the likelihood of DCPD claims versus PD claims can help insurers make more informed decisions about insurance underwriting and policy design. Another motivation for this thesis is to create a statistical model that identifies the impact of the most important risk factors on DCPD claims in private passenger vehicle accidents.

1.3 Theoretical Orientation

The classical method in literature is the Bühlmann credibility or its variants. The method can be applied to determine pure premiums in various insurance policies. In the Bühlmann model, it is assumed that the random variables representing the outcomes for the same number of observations are independently and identically distributed with identical means and variance. The requirement that the random variables for risk be identically distributed is a major assumption that is easily violated in practice since risk characteristics can change for a variety of reasons. Since the assumptions of the Bühlmann-Straub model are reasonable for practice and Bayesian credibility is not easy to apply in practice, the Bühlmann-Straub model is used in this research.

In addition, the bonus-malus system (BMS) can be seen as a commercial simplification of credibility mechanisms. All the classical BMSs are based on a single type of event. Thus, the classical BMS is commonly used in automobile insurance. In that case, the severity of the claims cannot be integrated into the premiums. Both integration of the severity of claims and recognition of the partial liability of the policyholder can be considered by designing bonus-malus scales involving different types of claims as PD and BI. Therefore, the multi-event bonus-malus scales are used in this research.

1.4 Research Questions

Since premium income plays an important role in the development of the insurance industry, the main purpose of this thesis is to contribute to the progress of the insurance sector by constructing premium evaluation in TPL insurance via an empirical study involving the class rating and the merit rating. The following primary research questions detailed into sub-questions guide us toward this objective.

1. How is taken into account separately of the PD and BI claims on the premium evaluation of private passenger vehicle accidents in TPL?
    - Can the PD and BI claims be introduced separately into both the class rating model and the merit rating model?
2. How to be classified the policyholders into homogeneous classes?
    - Which rating factors are significant in the risk classification?
3. Do premiums calculated according to class rating, taking into account claim types, give better results than written premiums of the insurance company?
4. Does the distribution of premiums to the bonus-malus levels differ according to whether bonus-malus scales take into account claim types or not?
5. What is the effect of the PD claims, the BI claims, and the combination of class rating and merit rating on premium production?
6. What is the impact of risk factors on the likelihood of DCPD claims versus PD claims covered by TPL insurance?

1.5 Outline of the Thesis

The rest of this thesis is organized as follows:

In chapter two, the impact of rating factors on the DCPD claims in Canadian automobile insurance is examined via a generalized linear binary logit model. The methodological framework used in this study is given by introducing the basic concepts of generalized linear mixed models (GLMMs). Each section of this chapter mentions the structure of Canadian automobile insurance data, statistical analysis of binary outcomes, the performance measure metrics used in this study, and the results of the developed model.

In chapter three, a premium evaluation is made using the Bühlmann-Straub credibility model and the multi-level bonus malus scales for private passenger vehicle accidents in Canadian TPL insurance. Some background information about the models used in this study is given by introducing the basic concepts. The data is described in detail. The results of the developed models in both statistical modeling and actuarial modeling are mentioned.

Chapter four provides a comprehensive conclusion derived from the findings of this thesis and proposes potential approaches for future research.

CHAPTER TWO

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

2.1 Introduction

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 PD and not BI claims occurring in a car accident; while enabling the repair of damage on the vehicle of the policyholders faster, without the delays and complications that 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 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 that 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 that 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.

This chapter is structured as follows: Section 2.2 gives the literature background of GLMMs. In Section 2.3, the methodological framework used in this study is described with the basic concepts of GLMMs. Section 2.4 mentions the structure of automobile insurance data provided by a Canadian insurance company. Section 2.5 includes a statistical analysis of binary outcomes such as DCPD claims and PD claims covered by TPL insurance, and how risk factors are identified. In addition, in Section 2.6, the performance metrics used in this study are explained in detail. Section 2.7 presents the results of the model developed for estimating the likelihood of DCPD claims. Section 2.8 introduces the main conclusions of this study.

2.2 Literature Review

The use of 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 & Wedderburn (1972) and McCullagh & 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 & Renshaw (1996) provides an overview of the applications of GLMs in actuarial science and shows that GLMs are not limited to models for automobile insurance premiums. Embrechts & Wüthrich (2022) in the case of non-life insurance demonstrates 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 & 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 & 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.

2.3 Methodology

2.3.1 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 & 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:

$$g(\mu_{ij}) = \text{logit}(\mu_{ij}) = \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{b}_i, \quad i = 1, ..., n, \quad j = 1, ..., t_i \tag{2.1}$$

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

$$g(\mu_{ij}) = \text{logit}(\mu_{ij}) = \log_e \left[\frac{\mu_{ij}}{1 - \mu_{ij}}\right] \tag{2.2}$$

where μ_{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.,

$$\mu_{ij} = E(y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}) = P(y_{ij} = 1|\mathbf{b}_i, \mathbf{x}_{ij}) \tag{2.3}$$

(Hedeker, 2005). Assuming that the random effects are mutually independent and identically distributed completes the specification of the GLMM. Furthermore, a correlation between observations on the same subject occurs since they share the same random effects $\mathbf{b}_i$ (Antonio & Beirlant, 2007).

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

2.4 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.

The analysis is performed on 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 2.1 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. At the same time, 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.

Table 2.1 Variables in the dataset

Variable	Definition
Age	Age of policyholder at the time of claim
Territory	Residential area 0: Urban; 1: Rural
Usage	Vehicle usage 0: Work/Business; 1: Pleasure
Time	Accident year $j = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$, corresponding to values of 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011 and 2012, respectively
Class	The code of the class 0: Vehicle used for pleasure or having vehicle usage restrictions for commuting to work one way and the 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 the driver is 25 years of age or over; 2: Vehicle not having the vehicle usage restrictions and the driver is under 21 years of age; 3: Vehicle not having vehicle usage restrictions, and the driver is under 25 years of age, but not under 21 years of age
Driver record	The number of claims-free years for each policy (in the last 6 years) 0 (No claims-free years), 1 (One claims-free year), 2, 3, 4, 5, 6
Claims history	The number of claims the risk has had in the last 6 years before the policy was rated 0 (Number of chargeable claims is zero), 1, 2 (Number of chargeable claims is two or more)
Claims-free years	The number of years since the risk had a claim 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 0 (Zero year), 1, 2, 3, 4, 5, 6, 7, 8, 9 (Nine or more years)
Training	Driving education provided to all candidates 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	Losses reported to the insurer but are still in the process of settlement
Incurred loss	The amount paid in losses during a specified time
Property damage type	0: PD covered by liability insurance 1: DCPD covered by liability insurance

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 2.2.

Table 2.2 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	620.50	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 2.3 presents the mean of the outstanding and incurred losses used in the forthcoming estimations for each of the six years.

Table 2.3 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 2.4.

Table 2.4 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 commute 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 2.5 shows the correlation between the independent variables in this fitted model.

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

Table 2.5 Correlation matrix of independent variables in the model

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

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%).

2.5 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:

$$g(\mu_{ijk}) = \beta_0 + \beta_1 x_{ijk}^{(3)} + \beta_2 x_{ijk}^{(4)} + \beta_3 x_{ijk}^{(10)} + \beta_4 x_{ijk}^{(11)} + \beta_5 x_{ijk}^{(12)} + \beta_6 x_{ijk}^{(13)} + b_{0k}, \quad i = 1, ..., n, \quad j = 1, ..., t_i, \quad k = 1, ..., m \tag{2.4}$$

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, μ_{ijk} is the probability of a claim on policy i (i=1,...,942) at time j (j=1,...,6) for broker k (k=1,...,179).

In the fixed-effects part of the model, the parameters β_0, β_1, and β_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 β_3 and β_4, for a given random intercept, respectively. Additionally, β_5 and β_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 random intercept.

In the random-effects part of the model, the term b_{0k} in Equation 2.4 denotes a broker-specific random intercept. The random intercept b_{0k} is a subject-specific deviation from the fixed intercept β_0. The results of the panel data generalized linear binary logit mixed model are summarized in Table 2.6.

Table 2.6 Generalized linear binary logit mixed model estimation results

Variable	Estimated Coefficients	Std. Error	z-value	Pr(>|z|)	Exp(β)
Intercept	- 4.691	0.694	- 6.761	< 0.001 ***	0.009
Usage1	- 1.083	0.268	- 4.038	< 0.001 ***	0.339
Time	0.157	0.085	1.844	0.065 .	1.170
Training1	1.215	0.464	2.621	0.009 **	3.372
Training2	3.411	1.235	2.763	0.006 **	30.279
Gender1	- 0.463	0.275	- 1.680	0.093 .	0.629
Outstanding loss	0.187	0.063	2.948	0.003 **	1.205
Incurred loss	0.946	0.075	12.665	< 0.001 ***	2.576
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

2.6 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 the likelihood ratio test (*LRT*) are used. If the number of observations (N) is large enough, $v < (N/40)$, *AIC* is defined as

$$AIC = -2\ln(\hat{L}) + 2v \tag{2.5}$$

where v represents the number of estimated parameters in the fitted model and $\ln(\hat{L})$ is the maximum log-likelihood value (Portet, 2020). Equation 2.5 is used to calculate the *AIC* value since $v = 9$ 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 2.6.

$$LRT = 2\{\text{logLik(reference)} - \text{logLik(nested)}\} \tag{2.6}$$

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 & 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(χ_1^2>64.058). From the chi-square table, we can conclude that Pr(χ_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*), $F1$ 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 2.7 (Hossin & Sulaiman, 2015).

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,

Table 2.7 Confusion matrix for the binary classification

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

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 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 2.7 – 2.12.

$$SN = \frac{TP}{TP+FN} \tag{2.7}$$

$$SP = \frac{TN}{TN+FP} \tag{2.8}$$

$$P = \frac{TP}{TP+FP} \tag{2.9}$$

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \tag{2.10}$$

$$BA = \frac{SN+SP}{2} \tag{2.11}$$

$$F1 \text{ score} = \frac{2TP}{2TP+FP+FN} \tag{2.12}$$

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 $F1$ 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 the 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 & Goadrich, 2006).

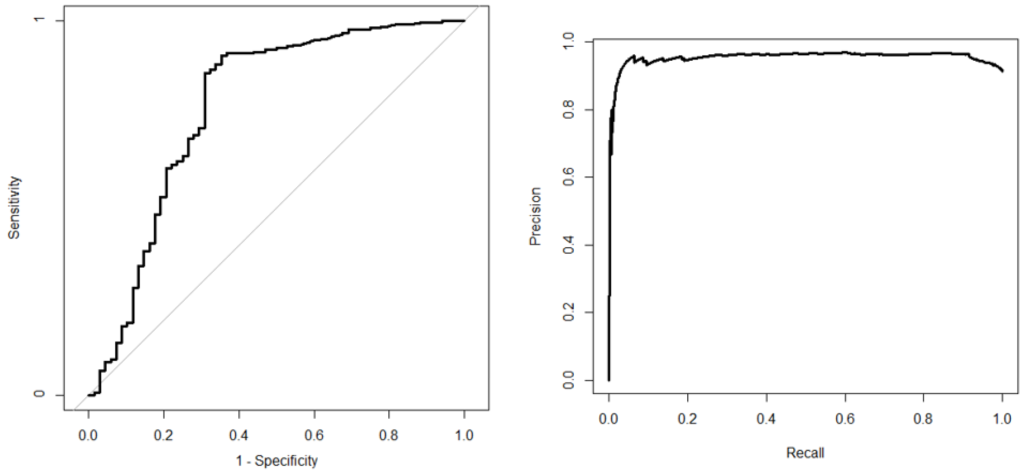
A Precision-Recall (PR) curve, on the other hand, evaluates the fraction of true positives among positive predictions. By offering valuable insights into the effectiveness of the classification model in capturing and correctly labeling minority class instances, the PR curve can provide an accurate prediction 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 & 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 the *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 2.8

Table 2.8 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 2.1.

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, the *PR AUC*, calculated as 0.953, describes that the



(a) (b)

Figure 2.1 Predictive performance of the fitted model:
(a) the ROC curve and (b) the PR curve

fitted model performed very well in predicting the positive class.

2.7 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 2.6. 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 outstanding 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 another driver caused the accident 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.

2.8 Summary

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 proposed 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 highlights 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 of 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.

CHAPTER THREE
DEVELOPING A PURE PREMIUM MODEL TO REDUCE RISK OF SECTOR FAILURE FOR THE THIRD-PARTY LIABILITY INSURANCE

3.1 Introduction

When insurers are faced with as accurately as possible a premium calculation problem for each policyholder in automobile insurance, posteriori ratemaking mechanisms, which are provided by experience rating systems, play an important role in the sector. Partitioning policies according to their risk characteristics as a priori ratemaking is crucial in automobile ratemaking. The idea behind risk classification within the pricing process is that the insurer wants to optimally group the risks in the portfolio. Thus, the policyholders with a similar risk profile pay the same premium rate. However, the fact that risk classes maintain heterogeneity due to posteriori risk characteristics leads to insurers experience rating.

The premium adjustment relies on the policyholder's historical data to mitigate unobserved risk characteristics, thereby restoring fairness among policyholders. Credibility theory is an appropriate model for premium adjustment because it allows an insurer to perform prospective experience rating adjusting future premiums based on the individual claim experience.

In addition, the allowance for the severity of claims is as important as taking the number of claims into account in the premium evaluation. The evaluation only by the number of claims leads to the same premium surcharge for the policyholders who have the same claim frequency although they have different claim severity, thus this situation results in injustice for the policyholder who hasn't severe damage. In this chapter, a credibility model is developed by subdividing the claims into two categories, PD and BI, for third-party liability automobile insurance. The idea behind the determined model is to calculate future premiums based on both collective and individual information. In real-life automobile insurance, the numbers of exposure units and the distribution of claim sizes differ across past policy years. The Bühlmann-Straub model accommodates these diversities; therefore, in this chapter, a Bühlmann-Straub model reflecting pricing is determined under one-dimensional credibility assumptions.

The objective of this chapter is to develop a statistical model that assesses the incurred losses of different claim types separately. The statistical model is examined using a linear mixed-effect model (LMM) for the PD and BI claims, revealing various risk factors that have a significant impact on the incurred losses associated with different claim types. Subsequently, credibility premiums are obtained for the PD and BI claims by weighting the predicted losses through earned exposures, and then the bonus-malus scales are designed by taking into account the types of claims.

The remainder of this chapter is organized as follows: The next section discusses literature, focusing on possible alternatives for evaluating premiums via credibility theory and bonus-malus scales. A methodology is presented as the theoretical approaches for credibility theory and bonus-malus scales, and then a section describing data follows it. Subsequently, the results of the analysis are contrasted and a summary section is presented.

3.2 Literature Review

Credibility has a long history in actuarial science, with fundamental contributions dating back to (Mowbray, 1914). Mowbray wants to distinguish between situations when large employers with substantial information would use their own experience and when small employers with limited experience would use external sources. Subsequently, Whitney (1918) introduces the concept of using a weighted average of claims from the risk class (the group's claims experience) and claims over all risk classes (a manual rate) to predict future expected claims. Credibility theory has been used for more than 50 years in insurance pricing before it is placed on a firm mathematical foundation by Bühlmann (1967). Bühlmann hypothesizes the existence of unobserved characteristics of the group.

Traditional approaches of credibility theory consider one unobserved risk parameter, for each policyholder and treat one policy or coverage. Thus, these approaches are based on the ideas of Bichsel (1967), Bühlmann (1967), and more recently Lemaire (1995). A large number of extensions related to credibility theory have been derived by Jewell (1974), Hachemeister (1975), Sundt (1979, 1981), and Zehnwirth (1985) after the approaches are presented by Bühlmann (1967) and Bühlmann et al. (1970). A concise review of credibility theory can be found in (Norberg, 2004). Hachemeister (1975) contributes to the credibility context by introducing a regression model.

Evolutionary models are not new in credibility theory. This approach is introduced in the 1970s for one-dimensional credibility models by Gerber & Jones (1975a), Gerber & Jones (1975b), De Vijlder (1976), De Vylder (1977), Sundt (1981), Kremer (1982), and Pinquet et al. (2001). In Sundt (1983), the generalized Bühlmann-Straub model is proposed with consecutive error terms assumed to follow AR(1) dependences. Pinquet et al. (2001) show that the date of the claim does matter because the effects of a claim on the risk evaluation diminish over time. Much of the work on the time-dependent models focuses on credibility formulas of the updating type. Pinquet et al. (2001) and Bolancé et al. (2003) present empirical studies performed on panel data. Furthermore, these studies support time-varying random effects.

Sundt (1988) has studied experience rating for motor insurance based on credibility estimators with geometric weights within the simple Bühlmann model. He has discussed how to find optimal weights. He has also compared the estimators with geometric weights to the traditional credibility estimators and he has shown to be more robust against a certain type of violations against the model assumptions, but he could not go any further into practical modifications of his scheme in this paper.

The multidimensional generalization of the credibility approach to more than one line of business is first introduced by Jewell (1973, 1974). Pinquet (1997, 1998) also include examples that are given of experience rating that incorporate additional claim information, from the number and the cost of claims and claims at fault and not at fault within the same line of business. In brief, with the aid of multi-equation Poisson models with random effects, Pinquet (1998) designs an optimal credibility model for different types of claims. Pinquet (1997) also allows for the costs of the claims. Autoregressive specifications of the error structure in the credibility context have been proposed by Bolancé et al. (2003).

Englund et al. (2008) have investigated a concept of multivariate pricing, which includes claim history for more than one line of business and is a generalization of the Bühlmann-Straub model. They have extended the multivariate credibility model to allow for the age of claims to influence the estimation of future claims. They have also applied this model to data from a portfolio of commercial lines of business.

In 2008, Couret and Venter considered the estimation of the percentage of loss that is over high deductibles. By Couret & Venter (2008), a key element of the excess percentage is the frequency of loss by injury type, and the vector of claim frequency by injury type can be estimated by class of business using multidimensional credibility techniques. They have shown that credibility estimation by class produces additional information in comparison to large groups of classes in workers compensation.

Englund et al. (2009) considers insurance business lines for the pricing methodology. They have compared the multivariate credibility approach with the classical one-dimensional credibility theory. They have concluded that the multivariate approach is capable of improving the quality of estimation. They have also contributed to this theory with new robust estimation methods for time (in)dependency. They have concluded that adding a time effect will improve prediction if the average duration of information is long enough.

As to the various theoretical approaches regarding merit rating, Pinquet et al. (2001) have purposed to use the age of claims in the prediction of risks. They have presented a dynamic random effects model on longitudinal count data since an optimal BMS estimated from short histories and applied to a longer duration will overestimate the individual credibilities.

Brouhns et al. (2003) have proposed a computer-intensive methodology to build bonus-malus scales in automobile insurance. They have compared four different credibility models based on real data: static versus dynamic heterogeneity, with and without recognizing a priori risk classification.

In 1997, Taylor considered the operation of a BMS, superimposed on a premium system involving several other rating variables. He has discussed the issue that to the extent that good risks are rewarded in their base premiums, through the other rating variables, the size of the bonus they require for equity is reduced. He has considered the extent to which the BMS differentiates the risk classes over time, in other words, the extent to which the BMS differentiates individuals over time. As to methods incorporating data from different business lines on the same policyholder, the study by Desjardins et al. (2001) mentions that the BMSs for fleets of vehicles are derived from the claims or safety offenses history. Pitrebois et al. (2003) has also proposed an analytic analog to the simulation procedure described in Taylor (1997). They have discussed the interaction between a priori and a posteriori ratemakings.

In 1976, Norberg developed optimal credibility premium scales for Markovian BMSs, with given transition rules, under an infinite horizon approach, and assumed the minimization of the expected squared difference between the true net premium and the premium paid by the policyholder. The premium relativities are traditionally computed with the help of a quadratic loss function in Norberg (1976). Borgan 1 & Hoem (1981) have generalized this result to a finite horizon approach. Andrade e Silva & Centeno (2005) underline some potential problems of the linear scales, and they propose the use of geometric scales as a possible solution. Gilde & Sundt (1989) assume that the BMS forms a first-order Markov chain and introduce the optimal bonus scales.

Pitrebois et al. (2006) have offered an alternative approach to traditional bonus-malus scales. They have designed bonus-malus scales involving different types of claims as bodily injuries and material damage instead of considering one type of claim to integrate the severity of the claims and to recognize the partial liability of the policyholder. Different bonus-malus scale models for each type of insured are proposed by using recursive partitioning methods instead of classic past claim rating models in Boucher (2022).

3.3 Methodology

3.3.1 Linear Mixed Models

LMMs extend the capabilities of the linear model by accommodating clustered or longitudinal data.

The linear mixed model may be expressed as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, ..., n \tag{3.1}$$

where

- $\boldsymbol{\beta}$ (dx1) describes the d fixed effects in the model. These are fixed unknown regression parameters and they are common to all subjects.
- $\mathbf{u}_i$ (fx1) is the vector with the random effects for the ith subject in the data. The use of random effects reflects that there is heterogeneity among subjects for a subset of the regression coefficients in $\boldsymbol{\beta}$.
- $\mathbf{X}_i$ (r_ixd) and $\mathbf{Z}_i$ (r_ixf) are the design matrices for the d fixed and f random effects where r_i describes the repeated observations for the ith subject over time.
- $\boldsymbol{\varepsilon}_i$ (r_ix1) contains the residual components for subject i.

$$\mathbf{u}_i \sim N(0, \sigma_u^2)$$
$$\boldsymbol{\varepsilon}_i \sim N(0, \sigma_\varepsilon^2)$$

Following the fitting of an LMM and preceding any inferences drawn from it, it is crucial to verify the fulfillment of model assumptions. The principal distributional assumption pertains to the normality of the residual errors, commonly assessed through a quantile-quantile plot (q-q plot). Substantial deviations from linearity in observations or asymmetrical scales serve as indicators of a departure from normality in the residuals.

Ensuring that the model is not unduly influenced by individual observations or a small subset is crucial. Such influence could suggest overfitting or sensitivity to specific observations included in the model. Assessing leverage and influence in mixed models can be challenging, with limited tools available; commonly used tools include leverage and/or Cook's distance for linear mixed models.

Other main distributional assumptions pertain to the normality of the random effects. Practically, the assessment of the normality assumption for $\mathbf{u}_i$ should rely on comparing the outcomes derived from an LMM with and without assuming normality. For more details, see Gałecki & Burzykowski (2013).

3.3.2 Bühlmann-Straub Credibility Model

The basic idea of credibility theory is to use claims experience and additional information to develop a pricing formula. In this way, the premiums charged using credibility theory take the driving capabilities of both individual and whole portfolios into account.

In the context of the one-dimensional Bühlmann-Straub model, the observation vector associated with risk *i* in year *t* is represented by the claims ratio or average claim size, denoted as $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{it_i})' \quad i = 1, ..., n, \quad j = 1, ..., t_i$. Moreover, the associated known weights are denoted as w_{ij}.

An individual risk is a member of a larger population and the risk has an associated risk parameter that distinguishes the individual's risk characteristics (Dean, 2005). Although these characteristics are unobserved features of the group, they are common to all observations in the group. The risk i is characterized by its individual risk profile θ_i, which is itself the realization of a random variable Θ_i. The model assumptions for the one-dimensional Bühlmann-Straub model are as follows:

1. Conditionally, given Θ_i, the X_{ij} $j = 1, ..., t_i$ are independent with

$$E[X_{ij}|\Theta_i] = \mu(\Theta_i)$$
$$Var(X_{ij}|\Theta_i) = \frac{\sigma^2(\Theta_i)}{w_{ij}}$$

2. The pairs (Θ_i, X_i) $i = 1, 2, ..., n$ are independent, and $\Theta_1, \Theta_2, ..., \Theta_n$ are independent and identically distributed.

The credibility estimator for the individual risk premium, $\mu(\Theta_i)$, depends only on the observations from the ith risk and all data. Because the fact that $E[X_i] = \mu_0$, the credibility estimator is of the form,

$$\widehat{\mu(\Theta_i)} = z_i X_i + (1 - z_i)\mu_0, \tag{3.2}$$

where μ_0=$E[\mu(\Theta_i)]$ is collective premium and $z_i = w_{i.}/(w_{i.} + \frac{\sigma^2}{\tau^2})$ is a credibility factor.

Average variance within individual risk and variance between individual risk premiums are σ^2=$E[\sigma^2(\Theta_i)]$ and τ^2=$Var[\mu(\Theta_i)]$, respectively. For more details, see Bühlmann & Gisler (2005).

3.3.3 Bonus-Malus Scale Model

With some types of insurance, especially automobile insurance, charging a premium based exclusively on factors known a priori, such as age, sex, occupation of the main driver, region of residence, type of car, and so on is insufficient. Since the values of such variables can be determined before the insured starts to drive, they are called *a priori rating variables*. The main purpose for their use is to subdivide insureds into homogeneous classes, but heterogeneous driving behaviors, such as state of health, reflexes, accident proneness, and so on are still observed in each tariff cell, and such variables are called *a posteriori rating variables*. The insurers can base prices on the above-mentioned unobservable characteristics by taking into account the prior claims experience of the policyholder. *Merit rating* which modifies premiums with claim history penalizes the policyholders responsible for one or more accidents by an additional premium or *malus*, and rewards claim-free insureds by awarding a discount or *bonus*. Such systems on the one hand use premiums based on a priori factors, on the other hand, they adjust these premiums by use of merit rating. In this way, premiums reflect the exact driving capabilities of the driver much better. This situation can be modeled as a Markov chain.

BMSs are special cases of Markov processes. In such processes, one goes from one state to another in time. The Markov property says that the process is in a sense memoryless: the probability of such transitions does not depend on how one arrived in a particular state (Kaas et al., 2002). It means that the knowledge of the present class and the number of claims for the year suffices to determine next year's class.

The number of claims reported during the year, N, is Poisson distributed with expected claim frequency, λ. The N claims may be classified into h categories, according to a multinomial partitioning scheme with probabilities $q_1, q_2, ..., q_h$ that the claim is of type $1, 2, ..., h$, respectively, for a policyholder in risk class.

In this study, the claims reported by the policyholders are distinguished into two separate categories PD and BI. Each type of claim induces a specific penalty for the policyholder. The number of the types of claims is denoted as h, and h is equal to 2 for the above-mentioned distinction.

For this model, the following assumption is considered:

The number of claims of type c, N_c, has a Binomial distribution with parameters n and q_c. The expected value of N_c is nq_c. Then the random variables $N_1, ..., N_h$ are independent and Poisson distributed with respective parameters $\lambda q_1, ..., \lambda q_h$. The joint probability mass function is as follows:

$$\Pr[N_1 = n_1, ..., N_h = n_h] = \prod_{c=1}^{h} \exp(-\lambda q_c) \frac{(\lambda q_c)^{n_c}}{n_c!} \tag{3.3}$$

Let $p_{l_1,l_2}(\lambda; \mathbf{q})$ be the probability of moving from level l_1 to level l_2 for an insured with annual expected claim frequency λ and vector probability $\mathbf{q} = (q_1, ..., q_h)^{\mathrm{T}}$; q_h is the probability that the claim be of type h. The transition matrix consisting of q_hs, $\mathbf{P}(\lambda; \mathbf{q})$, is also the one-step transition matrix for s levels and is represented below:

$$\mathbf{P}(\lambda; \mathbf{q}) = \begin{pmatrix} p_{0,0}(\lambda; \mathbf{q}) & \cdots & p_{0,s}(\lambda; \mathbf{q}) \\ \vdots & \ddots & \vdots \\ p_{s,0}(\lambda; \mathbf{q}) & \cdots & p_{s,s}(\lambda; \mathbf{q}) \end{pmatrix}$$

The Markov chain possesses a stationary distribution

$$\boldsymbol{\pi}(\lambda; \mathbf{q}) = (\pi_0(\lambda; \mathbf{q}), \pi_1(\lambda; \mathbf{q}), ..., \pi_s(\lambda; \mathbf{q}))^T,$$

where $\pi_l(\lambda; \mathbf{q})$ is the stationary probability for a policyholder with expected frequency λ to be in level l, $l = 1, ..., s$. For more details, see Lemaire (1995) and Denuit et al. (2007).

3.4 Data Description

Data about only private passenger automobiles are provided from the automobile portfolio of an active insurance company in Ontario, Canada. The dataset does not contain information about the policyholders who transferred their policy to a different insurance company at any time and insurance details for the policy year in which no claim was filed. These policies 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. There are no missing observations in the data. There are two types of claims in the data set: PD and BI. The dataset about PD claims comprises insurance details for a total of 156 observations across 147 policies, whereas the dataset associated with BI claims encompasses information for a total of 134 observations of 124 policies.

The analysis is performed on the company's liability insurance claim experience for 2003–2012. The data comprise outstanding loss, which only includes zero and positive claim amounts, incurred loss, which only includes positive claim amounts, and several rating factors for each policy that consist of age, territory, usage, time, valuation year, class, modifier, rate modifier, driver record, claims history, claims-free years, experience, training, and gender.

Driver characteristics also involve the date of birth of the policyholders, while the claim profiles include information on the type of coverage regarding PD and BI, policy effective and expiry date, claim identification number, and accident date.

In the analysis, territory, usage, class, modifier, rate modifier, training, and gender are treated as factor covariates while age, time, valuation year, driver record, claims history, claims-free years, experience, outstanding loss, and incurred loss are treated as continuous covariates in the model. In the following section, the best-fitted models for PD claims and BI claims are presented. In the PD model, the significant explanatory variables are class ($x^{(1)}$), modifier1 ($x^{(2)}$), modifier2 ($x^{(3)}$), claims history ($x^{(4)}$) and time ($x^{(5)}$) while the independent variables are class ($x^{(1)}$), rate modifier1 ($x^{(2)}$), rate modifier2 ($x^{(3)}$), gender ($x^{(4)}$), valuation year ($x^{(5)}$) and time ($x^{(6)}$) in the BI model. Table 3.1 and Table 3.2 give detailed information about significant rating factors in the PD and BI models, respectively. In addition, the descriptive statistics of continuous covariates in these models are summarized as shown in Table 3.3.

Table 3.1 Variables in the PD model

Variable	Definition
Time	Accident year j = 1, 2, 3, 4, 5, 6, corresponding to values of 2003, 2004, 2005, 2006, 2007 and 2008, respectively
Class	0: Pleasure use only; driver is 25 years of age or over; No male driver under 25 years of age; maximum annual distance driven 18,000 km 1: Pleasure or commute use over 20 km one way; driver is 25 years of age or over; No male driver under 25 years of age; No annual distance driven restrictions
Modifier	A single value or the rating groups used to adjust a policy's premium based on a set of risk characteristics (Three groups : 0; 1; 2)
Claims history	0 (Number of chargeable claims is zero), 1, 2 (Number of chargeable claims is two or more)

Table 3.2 Variables in the BI model

Variable	Definition
Time	Accident year j = 1, 2, 3, 4, 5, 6, corresponding to values of 2003, 2004, 2005, 2006, 2007 and 2008, respectively
Class	0: Pleasure use only; driver is 25 years of age or over; No male driver under 25 years of age; maximum annual distance driven 18,000 km 1: Pleasure or commute use over 20 km one way; driver is 25 years of age or over; No male driver under 25 years of age; No annual distance driven restrictions
Rate modifier	A single value or the rating groups used to adjust a policy's premium based on a set of risk characteristics (Three groups: 0; 1; 2)
Gender	0: Male 1: Female
Valuation year	Loss and expense information is provided by valuation year j = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, corresponding to values of 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, and 2013, respectively

Table 3.3 Summary statistics of the variables in the models

	Variable	Mean	Std.Dev.	Minimum	Maximum
	Time	3.432	1.468	1.000	6.000
Property Damage	Claims history	0.046	0.259	0.000	2.000
	Incurred loss	878.420	1912.304	26.840	9961.710
	Time	3.658	1.543	1.000	6.000
Bodily Injury	Valuation year	6.165	2.705	1.000	10.000
	Incurred loss	49874.300	80664.390	5.000	492488.900

The models of PD and BI claims are fitted using the claims for the years 2003–2008, and their predictive ability is evaluated using the claims from 2009–2012. The PD data for 2003–2008 consists of 88 observations on 84 policies for 62 brokers, and each observation includes the claim experience at the individual policy level. The BI data for 2003–2008 consists of 79 observations on 75 policies for 10 valuation years, and each observation includes the claim experience at the individual policy level. Of the 167 observations, 88 (52.69%) have PD claims covered by liability insurance and 79 (47.31%) have BI claims.

The analysis herein predicts the incurred losses for PD and BI claims covered by TPL insurance. Table 3.4 presents the mean of the incurred losses of PD and BI claims used in the forthcoming estimations for each of the six years.

Table 3.4 Distribution of losses by years

Year	Property Damage Mean ($)	Bodily Injury Mean ($)
2003	223.492	71375.340
2004	971.011	18915.500
2005	1123.777	51496.350
2006	602.406	42494.930
2007	822.963	56468.520
2008	225.507	65004.330

The models are built considering the correlation between the variables to prevent multicollinearity issues. Table 3.5 and Table 3.6 show the correlation between the explanatory variables in the fitted models for the PD and BI claims, respectively. As a result, the models presented below do not exhibit any multicollinearity issues.

Table 3.5 Correlation matrix of independent variables in the PD model

	$\boldsymbol{x}^{(1)}$	$\boldsymbol{x}^{(2)}$	$\boldsymbol{x}^{(4)}$	$\boldsymbol{x}^{(5)}$
$\boldsymbol{x}^{(1)}$	1.000	0.059^{a}	-0.134^{b*}	0.202^{b*}
$\boldsymbol{x}^{(2)}$	0.059^{a}	1.000	-0.204^{b*}	0.141^{b*}
$\boldsymbol{x}^{(4)}$	-0.134^{b*}	-0.204^{b*}	1.000	0.086^{c}
$\boldsymbol{x}^{(5)}$	0.202^{b*}	0.141^{b*}	0.086^{c}	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]Point-biserial correlation coefficient
[c]Spearman correlation coefficient
(Khamis, 2008)

Table 3.6 Correlation matrix of independent variables in the BI model

	$\boldsymbol{x}^{(1)}$	$\boldsymbol{x}^{(2)}$	$\boldsymbol{x}^{(4)}$	$\boldsymbol{x}^{(5)}$	$\boldsymbol{x}^{(6)}$
$\boldsymbol{x}^{(1)}$	1.000	0.033^{a}	0.029^{a}	0.168^{b*}	0.134^{b*}
$\boldsymbol{x}^{(2)}$	0.033^{a}	1.000	0.055^{a}	0.248^{b*}	0.285^{b*}
$\boldsymbol{x}^{(4)}$	0.029^{a}	0.055^{a}	1.000	-0.089^{b}	-0.160^{b}
$\boldsymbol{x}^{(5)}$	0.168^{b*}	0.248^{b*}	-0.089^{b}	1.000	0.393^{c}
$\boldsymbol{x}^{(6)}$	0.134^{b*}	0.285^{b*}	-0.160^{b}	0.393^{c}	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]Point-biserial correlation coefficient
[c]Spearman correlation coefficient
(Khamis, 2008)

3.5 Statistical Modeling

3.5.1 Fitted Model for Property Damage Claims

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

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk}^{(1)} + \beta_2 x_{ijk}^{(2)} + \beta_3 x_{ijk}^{(3)} + \beta_4 x_{ijk}^{(4)} + \beta_5 x_{ijk}^{(5)} + b_{0k} + \varepsilon_{ijk}, \quad i = 1, ..., n, \quad j = 1, ..., t_i, \quad k = 1, ..., m \tag{3.4}$$

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, y_{ijk} is the incurred loss on policy *i* (*i*=1,...,84) at time *j* (*j*=1,...,6) for broker *k* (*k*=1,...,62).

In the fixed-effects part of the model, for a given random intercept, the parameter β_0 defines an overall intercept. β_1, β_2, and β_3 express the change in the incurred loss for class, modifier1, and modifier2, respectively. The change caused by a one-year change in claims history and time is specified in parameters β_4 and β_5, respectively.

In the random-effects part of the model, the term b_{0k} in Equation 3.4 denotes a broker-specific random intercept, while ε_{ijk} is a residual random error. The random intercept b_{0k} is a subject-specific deviation from the fixed intercept β_0. The variance of the residuals is constant (homogeneous) across the groups of the categorical variables in the model.

In this model, the Box-Cox transformation is applied to incurred losses. The residuals are checked with the Shapiro-Wilk test to ensure that the normality assumption is validated. As a result, the residuals are normally distributed according to this test (p-value = 0.1873). Due to the panel data structure, the Durbin-Watson test, which can be applied to unbalanced data, is used to evaluate whether the residuals are independent. According to the Durbin-Watson test (DW = 2.1684; p-value = 0.2500), the residuals appear to be independent and not autocorrelated. When the outliers detected according to Cook's distance are removed from the data, there is no significant change in the parameters of the model. Since there are no influential observations among the outliers, no observations are removed from the data. The results of the broker-specific random intercept effect model for PD claims are summarized in Table 3.7.

Table 3.7 The broker-specific random intercept effect model estimation results

Variable	Estimated Coefficients	Std. Error	DF	t-value	p-value
Intercept	0.7298	0.0006	61	1269.6987	0.0000 ***
Class1	– 0.0018	0.0008	21	– 2.1841	0.0404 *
Modifier1	0.0009	0.0004	21	2.1550	0.0429 *
Modifier2	0.0015	0.0005	21	3.1515	0.0048 **
Claims history	0.0013	0.0007	21	1.8970	0.0717 .
Time	0.0000	0.0001	21	0.2624	0.7956
Random parameter					
Std. dev. of broker	0.0012				
Log-likelihood	399.413				
AIC	– 782.826				
BIC	– 763.572				

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

3.5.2 Fitted Model for Bodily Injury Claims

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

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk}^{(1)} + \beta_2 x_{ijk}^{(2)} + \beta_3 x_{ijk}^{(3)} + \beta_4 x_{ijk}^{(4)} + \beta_5 x_{ijk}^{(5)} + \beta_6 x_{ijk}^{(6)} + b_{0k} + \varepsilon_{ijk}, \quad i = 1, ..., n, \quad j = 1, ..., t_i, \quad k = 1, ..., m \tag{3.5}$$

where *n* is the total number of different policies; *m* is the total number of different valuation years; 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, y_{ijk} is the incurred loss on policy *i* (*i*=1,...,75) at time *j* (*j*=1,...,6) for valuation year *k* (*k*=1,...,10).

In the fixed-effects part of the model, for a given random intercept, the parameter β_0 defines an overall intercept. β_1, β_2, β_3, and β_4 express the change in the incurred loss for class, rate modifier1, rate modifier2, and gender, respectively. The change caused by a one-year change in valuation year and time is specified in parameters β_5 and β_6, respectively.

In the random-effects part of the model, the term b_{0k} in Equation 3.5 denotes a valuation year-specific random intercept, while ε_{ijk} is a residual random error. The random intercept b_{0k} is a subject-specific deviation from the fixed intercept β_0. The variance of the residuals is not constant (heterogeneous) across the class groups in this model. The variance of the residuals for the class0 and class1 is denoted as σ_0^2 and σ_1^2, respectively.

In this model, the logarithmic transformation is applied to incurred losses. The residuals are normally distributed according to the Shapiro-Wilk test (p-value = 0.1874). The Durbin-Watson test (DW = 2.1529; p-value = 0.5080) shows that the residuals appear to be independent and not autocorrelated. When the outliers detected according to Cook's distance are removed from the data, there is no significant change in the parameters of the model. Since there are no influential observations among the outliers, no observations are removed from the data. The results of the valuation year-specific random intercept effect model for BI claims are summarized in Table 3.8.

3.5.3 Performance Metrics

In this study, the LMM approach is applied to unbalanced panel data to determine which factors have a significant impact on the incurred losses of PD and BI claims that policyholders will make next year and to predict the losses for next year. To inform model selection, the AIC and LRT are used.

The evaluation metrics used in this analysis are R-squared, Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Percentage Error (MAPE). These performance metrics are the most common metrics used to measure the prediction accuracy of a model.

Table 3.8 The valuation year-specific random intercept effect model estimation results

Variable	Estimated Coefficients	Std. Error	DF	t-value	p-value
Intercept	4.5138	1.0835	64	4.1660	0.0001 ***
Class1	0.7321	0.3757	64	1.9483	0.0558 .
Rate modifier1	0.9184	0.3923	64	2.3410	0.0224 *
Rate modifier2	– 2.7084	0.5033	64	– 5.3817	0.0000 ***
Gender1	– 1.0485	0.3455	64	– 3.0349	0.0035 **
Valuation year	0.8981	0.1627	8	5.5212	0.0006***
Time	– 0.2519	0.1367	64	– 1.8432	0.0699 .
Random parameter					
Std. dev. of valuation year	1.281				
Log-likelihood	– 154.775				
AIC	329.550				
BIC	352.317				
σ_0^2	1.000				
σ_1^2	0.263				

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

R-squared is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. R-squared is shown in Equation 3.6.

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{3.6}$$

where *n* is the number of observations, $\hat{y}_i$ is the predicted value for the *i*th observation in the dataset, and y_i is the observed value for the *i*th observation. Generally, a higher R-squared indicates more variability which is explained by the model. It takes values between 0 and 1.

RMSE measures the average difference between predicted values by a model and the actual values. The model with a lower RMSE is considered a better model. RMSE is shown in Equation 3.7.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{3.7}$$

MSE measures the amount of error in statistical models. This metric assesses the average squared difference between the observed and predicted values. The model with a lower MSE is considered a better model. MSE is shown in Equation 3.8.

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{3.8}$$

MAPE measures the prediction accuracy in statistical models. Lower MAPE values indicate better model performance. It usually expresses the accuracy as a ratio defined by Equation 3.9.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}\text{x100} \tag{3.9}$$

These performance assessment measures are acquired as presented in Table 3.9.

Table 3.9 Performance evaluation metrics

	R-squared	**RMSE**	**MSE**	**MAPE**
Property Damage	0.7779	0.0009	9.834e^{-7}	36.0236
Bodily Injury	0.7157	1.5408	2.3739	2.5105

Regarding the R-squared of the PD and BI models, the fitted model's R-squared of 0.7779 and 0.7157, respectively. 78% of the variation in the incurred losses for PD claims is accounted for by the values of the explanatory variables in the model while 72% of the variation in the incurred losses for BI claims is explained by the values of the variables in the model. In other words, these R-squared values show that the models fit the data well. Additionally, the small RMSE and MSE values support these results.

A MAPE of 36.0236% implies that, on average, the predictions of the incurred PD losses are 36.0236% off the actual values. This metric describes that the fitted model reasonably predicts the incurred PD losses. In the predictions of the incurred BI losses, a MAPE of 2.5105% shows that the fitted model predicts the incurred BI losses highly accurately.

3.6 Actuarial Modeling

3.6.1 Bühlmann-Straub Credibility Model

In the class rating part of this study, the predicted incurred losses for PD and BI claims in the previous sections are weighted with the earned exposures. Earned exposure is one of the most commonly used exposure statistics. In the context of automobile insurance, ”earned exposure” typically refers to the exposure or risk that an insurance company has accumulated over a specific period for a policyholder. This exposure is considered ”earned” because it corresponds to the time for which the insurance coverage has been provided and is in force. Earned exposure is often used in the calculation of insurance premiums. The insurance companies use earned exposure calculations to assess the amount of risk they have assumed during a specific period and to determine the appropriate premium to charge for providing coverage. This approach helps insurers align premiums with the actual level of risk exposure over time. The average earned exposures by year are given in Table 3.10 as month-to-date (MTD).

Using the one-dimensional Bühlmann-Straub model, the credibility premium calculated for each claim type is specified in Table 3.11.

A loss ratio is a fundamental financial metric used in the insurance industry to measure the profitability of an insurance company. The loss ratio is calculated as incurred losses divided by earned premiums over a specified period. The results in Table 3.12 are obtained by weighting the loss ratios with the earned exposure.

Table 3.10 Distribution of average earned exposures by years

Year	Property Damage Mean (MTD)	Bodily Injury Mean (MTD)
2003	0.1321	0.1132
2004	1.3861	1.4564
2005	1.8988	1.8957
2006	1.6241	1.7432
2007	1.4833	1.8336
2008	1.7581	2.2039
2009	1.7568	1.4249
2010	1.8278	1.9929
2011	1.7430	2.0971
2012	1.9982	1.8609

Table 3.11 Bühlmann-Straub credibility model estimation results for incurred losses

Claim Type	Individual Mean	Weight	Credibility Factor	Credibility Premium
Property Damage	1356.796	132.305	0.9753	12711.3
Bodily Injury	791957.440	128.986	0.9747	894315.1
Structure parameters estimators				
Collective premium (μ_0)		453513.2		
Between claim types variance (τ^2)		3.047112e^5		
Within claim types variance (σ^2)		1.020628e^{12}		

The average inflation rate of the six years in the data that we use in the statistical modeling is 0.1437. The premiums are calculated considering this ratio.

3.6.2 Bonus-Malus Scale Model

In the merit rating part of this study, bonus-malus scales are examined for two scenarios: one claim type and two claim types (PD and BI).

For both scenarios, the scale with seven levels (numbered from 0 to 6) is used. The policyholders enter the system in the sixth level and move between levels based on

Table 3.12 Bühlmann-Straub credibility model estimation results for loss ratio

Claim Type	Individual Mean	Weight	Credibility Factor	Credibility Premium
Property Damage	3.591	132.305	0.834	21.420
Bodily Injury	186.039	128.986	0.830	195.015
Structure parameters estimators				
Collective premium (μ_0)		108.217		
Between claim types variance (τ^2)		13846.230		
Within claim types variance (σ^2)		365396.200		

whether they have a claim or not. As they incur claims, they progress from level 6 to level 0.

Scenario 1: $-1/+2$

The total number of claims, N, is Poisson distributed with annual mean claim frequency, $\lambda = 0.1856$. The observed and fitted distribution of the number of claims is given in Table 3.13.

Table 3.13 Observed and fitted distribution of number of claims

	a	n_a	p_a	np_a
	0	744	0.8306	617.9707
	1	145	0.1542	22.3533
	2	11	0.0143	0.1574
	≥ 3	0	0.0009	0
Total		**900**		**640.4813**

In this bonus-malus scale, only one claim type is considered. The claims with property damage only result in a penalty of two levels. If no claims are reported, the policyholder moves up by one level. The transition matrix for a policyholder with annual mean claim frequency λ is as below:

$$
\mathbf{P}(\lambda) = \begin{pmatrix} 0.8306 & 0 & 0.1542 & 0 & 0.0143 & 0 & 0.0009 \\ 0.8306 & 0 & 0 & 0.1542 & 0 & 0.0143 & 0.0009 \\ 0 & 0.8306 & 0 & 0 & 0.1542 & 0 & 0.0152 \\ 0 & 0 & 0.8306 & 0 & 0 & 0.1542 & 0.0152 \\ 0 & 0 & 0 & 0.8306 & 0 & 0 & 0.1694 \\ 0 & 0 & 0 & 0 & 0.8306 & 0 & 0.1694 \\ 0 & 0 & 0 & 0 & 0 & 0.8306 & 0.1694 \end{pmatrix}
$$

Scenario 2: $-1/+2/+3$

The total number of claims, N, is Poisson distributed with annual mean claim frequency, $\lambda = 0.1751$. The probability that each claim is classified in one of the two possible categories, PD or BI, is $q_1 = 0.1729$ and $q_2 = 0.1236$, respectively. The number of PD claims, N_1, is Poisson distributed with annual mean claim frequency, $\lambda q_1 = 0.0303$, while the number of BI claims, N_2, is Poisson distributed with annual mean claim frequency, $\lambda q_2 = 0.0216$. The observed and fitted distribution of the number of claims is given in Table 3.14.

Table 3.14 Observed and fitted distribution of number of claims

	a	n_a	p_a	np_a
	0	789	0.8394	662.2654
	1	163	0.1469	23.9568
	2	2	0.0129	0.0257
	≥ 3	0	0.0008	0
Total		**954**		**686.2479**

In this bonus-malus scale, instead of considering one type of claim, the claims are penalized differently as claims with PD only and BI. The claims with PD only result in a penalty of two levels while the claims with BI are penalized by three levels. If n_1 claims with PD only and n_2 claims with BI are reported during the year, the policyholders move $2n_1+3n_2$ levels down. If no claims are reported, the policyholder moves up by one level. The transition matrix for a policyholder with annual mean claim frequency λ and vector probability $\mathbf{q} = (q_1, q_2)^T$ is as below:

$$\mathbf{P}(\lambda;\mathbf{q}) = \begin{pmatrix} 0.9494 & 0 & 0.0287 & 0.0205 & 0.0004 & 0.0006 & 0.0004 \\ 0.9494 & 0 & 0 & 0.0287 & 0.0205 & 0.0004 & 0.0010 \\ 0 & 0.9494 & 0 & 0 & 0.0287 & 0.0205 & 0.0014 \\ 0 & 0 & 0.9494 & 0 & 0 & 0.0287 & 0.0219 \\ 0 & 0 & 0 & 0.9494 & 0 & 0 & 0.0506 \\ 0 & 0 & 0 & 0 & 0.9494 & 0 & 0.0506 \\ 0 & 0 & 0 & 0 & 0 & 0.9494 & 0.0506 \end{pmatrix}$$

For both scenarios mentioned above, the stationary probabilities for a policyholder with mean frequency λ to be in level l are obtained. The collective premium obtained over predicted incurred losses for 2010 in the previous section and the written premium of the insurance company for 2010 are distributed to the bonus-malus levels for both scenarios by the following Equation 3.10 and 3.11.

For $\mathbf{-1/+2}$;

$$\text{Premium} = \text{P} * \pi_l \tag{3.10}$$

and for $\mathbf{-1/+2/+3}$;

$$\text{Premium} = \text{P} * \pi_l * w_1 + \text{P} * \pi_l * w_2 \tag{3.11}$$

where P is collective premium or written premium, π_l is the stationary probabilities for a policyholder to be in level l. w_1 = 0.5269 and w_2 = 0.4731 are weights of PD and BI claims in the portfolio, respectively.

Written premium denotes the overall amount customers are obligated to pay for insurance coverage on policies issued by a company during a specific period. The total written premium of the insurance company in 2010 is obtained at \$15226.08 (with inflation rate) as month-to-date. The calculated collective premium for the incurred losses in 2010 is $453513.2 (with inflation rate).

The distribution of the collective premium and the written premium to the levels are given in Table 3.15 and Table 3.16, respectively.

Table 3.15 Results for the bonus-malus systems -1/+2/+3 and -1/+2 for collective premium

	-1/+2/+3		-1/+2	
Level *l*	π_l	**Premium ($)**	π_l	**Premium ($)**
0	0.1601%	726.08	2.1519%	9759.15
1	0.3779%	1713.83	2.9279%	13278.41
2	0.6283%	2849.42	5.4255%	24605.36
3	2.5092%	11379.55	6.3112%	28622.13
4	4.8740%	22104.23	14.0913%	63905.91
5	4.6274%	20985.87	11.7042%	53080.09
6	86.8231%	393754.20	57.3880%	260262.20

Table 3.16 Results for the bonus-malus systems -1/+2/+3 and -1/+2 for written premium(MTD)

	-1/+2/+3		-1/+2	
Level *l*	π_l	**Premium ($)**	π_l	**Premium ($)**
0	0.1601%	24.38	2.1519%	327.65
1	0.3779%	57.54	2.9279%	445.80
2	0.6283%	95.67	5.4255%	826.09
3	2.5092%	382.05	6.3112%	960.95
4	4.8740%	742.12	14.0913%	2145.55
5	4.6274%	704.57	11.7042%	1782.09
6	86.8231%	13219.75	57.3880%	8737.94

3.7 Results

This study aims to evaluate premiums using the Bühlmann-Straub credibility model and the bonus-malus scales by considering the claim types as PD and BI in the TPL insurance. LMMs are employed to predict PD and BI incurred losses separately, given the unbalanced nature of the data structure with repeated measures. Within the LMM framework, random effects not only ascertain the correlation structure among observations from the same subject but also account for subject-specific heterogeneity arising from unobserved characteristics.

In LMM analysis for the broker-specific random intercept effect homogeneous model of PD claims, the risk factors having a significant impact on the incurred losses are “class”, “modifier”, and “claims history”. The ”time” is included in the model even if it is thought to be ineffective and a one-year change in time causes increased incurred losses of PD claims. The most important feature that distinguishes class0 and class1 is that there are no annual distance-driven restrictions in class1. The class1 provides a mitigating effect on the incurred claims compared to class0. The reason for this may be that since drivers in class1 are both pleasure and commute users and are in traffic much more than drivers in class0, they have gained experience and hence tend to make claims less. The modifier groups, modifier1 and modifier2, determined by the insurance company have an increasing effect on the incurred losses compared to modifier0. In addition, as claims history increases, the incurred losses of PD claims also increase.

In LMM analysis for the valuation year-specific random intercept effect heterogeneous model of BI claims, the risk factors having a significant impact on the incurred losses are “class”, “rate modifier”, “gender”, “valuation year”, and “time”. The class1 provides an increasing effect on the incurred losses compared to class0. The rate modifier groups, rate modifier1, and rate modifier2, determined by the insurance company are rating factors linked to the modifier. The rate modifier1 has an increasing effect on the incurred losses compared to rate modifier0 whereas the drivers in rate modifier2 tend to make fewer BI claims. Female drivers make fewer BI claims than male drivers. Being a long-distance driver increases the risk of BI claims. Because male drivers drive more intercity than female drivers, they may tend to make much more BI claims. A one-year change in the accident year causes decreased incurred losses of BI claims while a one-year change in the valuation year causes increased incurred losses. In other words, as the accident year decreases, the incurred losses of BI also decrease. If the valuation year increases after the claim is reported, the process may be taking longer because there is a BI claim.

The statistical models mentioned above both provide us with information about the risk characteristics of the policyholders used in the basic premium calculation and give predictions of the incurred losses according to the claim types. These predicted incurred losses by weighting with earned exposures, credibility premiums are calculated for each claim type.

The results obtained in the Bühlmann-Straub model indicate that the variance between claim types is much less than the variance within the types of claim. Furthermore, the credibility premium for BI is much higher than PD when these results are evaluated on a portfolio basis. The credibility factor represents that the policyholders' claims are weighted more than the average of the portfolio (μ_0) in the premium evaluation made based on the incurred losses. When these results are compared with ones obtained on loss ratios, in the premium evaluation based on loss ratios the overall mean is weighted much more than the evaluation made on the incurred losses.

Transition probabilities at the bonus-malus scales are examined under two scenarios. The first is to evaluate the number of claims only as PD and the other is to evaluate the claims according to whether they are PD or BI. When the distribution of the collective premium and the written premium to the levels is examined, the collective premium and the written premium are distributed more reasonably between the levels in the system taking into account the claim types. Most of the policyholders are expected to concentrate at the starting level in the long term and thus, the intensity of the levels decreases towards zero level. The results of scenario 2 (-1/+2/+3) support this expectation.

3.8 Summary

The main purpose of this study is to make a fair premium evaluation by considering the claim types under TPL insurance in private passenger vehicle accidents in Ontario, Canada. LMMs are consistently employed for modeling actuarial data, offering a distinct advantage in analyzing unbalanced panel data. This methodology mitigates the additional complexity stemming from the inherent correlation within each policy.

In this study, the developed model for PD claims analyzes the change of the incurred losses in the context of a broker-specific random intercept effect homogeneous model by dealing with unbalanced panel data. The estimation results obtained from the model demonstrate that the broker, which is a time-varying factor, has a significant influence on the change of the incurred losses of PD claims as a random parameter. In addition, rating factors such as class, modifier, and claims history have been found to correlate with the change of the incurred losses as fixed effects.

The constructed model for BI claims examines the variation in incurred losses within the framework of a valuation year-specific random intercept effect heterogeneous model, addressing the challenges posed by unbalanced panel data. In addition, it considers the difference in the variances of the residuals across classes. The estimation results obtained from the model reveal a notable impact of the valuation year treated as a time-varying factor, on the variation in incurred losses for BI claims as a random parameter. Furthermore, the fixed-effects part of the model indicates that the rating factors including class, rate modifier, gender, valuation year, and time are correlated with changes in incurred losses. Moreover, by considering different performance evaluation metrics, this study represents the models that fit the data well and hence provides a more reliable interpretation of the results.

The rating factors obtained in the statistical modeling show that these risk characteristics are important in premium evaluation. The predictions obtained from these statistical models are weighted by earned exposures and credibility premiums are calculated for each claim type. The fact that the premium for BI claims is much more than the premium for PD claims emphasizes the importance of considering claim types when evaluating premiums. In the bonus-malus scale, the reasonable distribution of premiums to BMS levels over the transition matrix prepared by considering the types of claims shows how important it is to take the number of claims into account whether they are PD or BI.

In this study, the emphasis has been placed on a fairer premium assessment over the severity of claims and the number of claims by taking into account the PD and BI claims, separately. In this way, the sector will be prevented from suffering losses and a more fair determination will be made to the policyholders.

CHAPTER FOUR
CONCLUSIONS AND FUTURE WORKS

The use of statistical modeling in actuarial science enables the integration of risk factors into the premium pricing process, thereby enhancing the accuracy of insurance premiums and mitigating the financial risk for insurers. One of the aims of this study is to present a statistical analysis assessing the impact of various risk factors on DCPD claims in private passenger vehicle accidents. Using automobile insurance data in Ontario, Canada for the decade-year period between 2003 and 2012, a statistical model of PD is 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 are observed under the weights — the number of trials used to generate each success proportion — in the different classes of insureds. The GLMMs analysis provides a powerful tool for quantifying the impact of risk factors on binary outcomes, which are called DCPD claims and PD claims covered by 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 model's ability to accurately predict classes. The $F1$ score, an evaluation metric to measure the performance of classification, is calculated as 0.934. In addition, the *PR AUC*, which is the area under the PR curve, is computed as 0.953. These high scores indicate that the model performs well in the classification. The other metrics also support the classification accuracy of the 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.

Another aim of this study is to evaluate the premiums by considering the claim types under TPL insurance. In this study, a statistical analysis is presented that examines the effect of various risk factors on incurred PD and BI losses in private passenger vehicle accidents. The statistical models of the PD and BI claims are explored via a subject-specific random intercept effect. The estimation results obtained from the models demonstrate that the broker, which is a time-varying factor, has a significant influence on the change of the incurred losses of PD claims as a random parameter whereas the predictions reveal a notable impact of the valuation year treated as a time-varying factor, on the variation in incurred losses for BI claims as a random parameter. The results indicate that several risk factors, class, modifier, claims history, and time, have a significant impact on the incurred losses of PD claims. For BI claims, the risk factors that are correlated with change of the incurred losses are class, rate modifier, gender, valuation year, and time, observed their effects under the heterogeneity of residual variances between the class groups. The LMM analysis provides information about the risk characteristics of the policyholders used in the basic premium calculation and gives predictions of the incurred losses according to the claim types. The LMMs are usually used in automobile insurance since actuaries have frequently repeated measurements over time in the insurance industry.

The performance metrics verify the model's ability to accurately predict the incurred losses. The *R-squared*, an evaluation metric to measure the variation of the incurred losses, is calculated, for PD and BI claims, as 0.7779 and 0.7157, respectively. These high scores indicate that the models perform well in the prediction. The other metrics also support the prediction accuracy of these models.

Over these predicted incurred losses, credibility premiums are calculated for each claim type by using the Bühlmann-Straub model. When calculating credibility premiums, the predictions in statistical modeling are weighted by earned exposures. The results obtained in the Bühlmann-Straub model indicate that the variance between claim types is much less than the variance within the types of claim. Furthermore, the credibility premium for BI claims is much higher than for PD claims. In addition, bonus-malus scales are designed by considering the claim types. The premiums are distributed reasonably to bonus-malus levels when the system is designed by considering the types of claims. These results indicate that it is important to consider whether the claims are property or bodily in the evaluation to be made based on both the severity of claims and the number of claims. The severity of claims plays a role as important as the frequency information of claims in the premium evaluation. In BMS, the severity information is indirectly reflected model by distinguishing the claims into two categories, PD and BI, although the system is designed based on the number of claims.

In future work, under the linear credibility approach, the multidimensional credibility model may be considered as the Bühlmann-Straub model with time dependence. Thanks to adding the time effect to the credibility model, it is assumed that the ability of drivers is not constant over time, and this assumption can raise the quality of the estimator in the case of being adequate claim history. The interpretation of this approach is that new claim information will affect the claim prediction more than old claim information. Furthermore, the premiums may be evaluated by taking into account the claim modifications. In additional work, a new bonus-malus scale theory could explore how insurers might develop different bonus-malus scales to account for the differences in a priori risk and the possible transitions between different scale models because the a priori risk can change over time. Transition rules within the classic bonus-malus scale models do not depend on the a priori risk, therefore; this experience rating model generates the same surcharges and the same discounts for all policyholders, Hence, this system may appear unfair to many policyholders. In addition, the assumption of independence between observations is retained in classical linear models. This assumption fails when correlation prevails among the observations. To extend the application of classical linear models to the case that longitudinal data with measurements for a subject taken at different time points are considered, a generalized estimating equation (GEE) may be used instead of GLMM and LMM.

REFERENCES

Andrade e Silva, J. M., & Centeno, M. d. L. (2005). A note on bonus scales. *Journal of Risk and Insurance*, *72*(4), 601–607.

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

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

Bichsel, F. (1967). Experience rating in subsets of risks. *ASTIN Bulletin: The Journal of the IAA*, *4*(3), 210–227.

Bolancé, C., Guillén, M., & Pinquet, J. (2003). Time-varying credibility for frequency risk models: Estimation and tests for autoregressive specifications on the random effects. *Insurance: Mathematics and Economics*, *33*(2), 273–282.

Borgan 1, Ø., & Hoem, J. M. (1981). A nonasymptotic criterion for the evaluation of automobile bonus systems. *Scandinavian Actuarial Journal*, *1981*(3), 165–178.

Boucher, J.-P. (2022). Multiple bonus–malus scale models for insureds of different sizes. *Risks*, *10*(8), 152.

Brouhns, N., Guillén, M., Denuit, M., & Pinquet, J. (2003). Optimal bonus-malus scales in segmented tariffs. *Journal of Risk and Insurance*.

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin: The Journal of the IAA*, *4*(3), 199–207.

Bühlmann, H., Straub, E., & Brooks, C. (1970). Glaubwürdigkeit für schadensätze (credibility for loss ratios). *Mitteilungen der Vereinigung Schweizerischer Versicherungsmathematiker*, *70*, 111–133.

Bühlmann, H., & Gisler, A. (2005). *A course in credibility theory and its applications*. Berlin: Springer.

Couret, J., & Venter, G. (2008). Using multi-dimensional credibility to estimate class frequency vectors in workers compensation. *ASTIN Bulletin: The Journal of the IAA*, *38*(1), 73–85.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, (233–240).

De Jong, P., & Heller, G. Z. (2008). *Generalized linear models for insurance data*. International Series on Actuarial Science. Cambridge University Press.

De Vijlder, F. (1976). Geometrical credibility. *Scandinavian Actuarial Journal*, *1976*(3), 121–149.

De Vylder, F. (1977). Iterative credibility. *Bulletin of Swiss Association of Actuaries*, *77*, 25–33.

Dean, C. (2005). *Topics in credibility theory*. Education and Examination Committee of the Society of Actuaries, Construction, and Evaluation of Actuarial Models Study Note.

Denuit, M., Maréchal, X., Pitrebois, S., & Walhin, J.-F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.

Desjardins, D., Dionne, G., & Pinquet, J. (2001). Experience rating schemes for fleets of vehicles. *ASTIN Bulletin: The Journal of the IAA*, *31*(1), 81–105.

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

Englund, M., Guillén, M., Gustafsson, J., Nielsen, L. H., & Nielsen, J. P. (2008). Multivariate latent risk: A credibility approach. *ASTIN Bulletin: The Journal of the IAA*, *38*(1), 137–146.

Englund, M., Gustafsson, J., Nielsen, J. P., & Thuring, F. (2009). Multidimensional credibility with time effects: An application to commercial business lines. *Journal of Risk and Insurance*, *76*(2), 443–453.

Frees, E. (2010). *Regression modeling with actuarial and financial applications*. International Series on Actuarial Science. Cambridge University Press.

Gałecki, A., & Burzykowski, T. (2013). *Linear mixed-effects model*. Springer New York.

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

Gerber, H. U., & Jones, D. A. (1975a). Credibility formulae with geometric weights. *Transaction of the Society of Actuaries*, *27*, 39–52.

Gerber, H. U., & Jones, D. A. (1975b). Credibility formulas of the updating type. *Credibility: Theory and Applications. Edited by P. M. Kahn. New York: Academic Press, Inc*, 89–105.

Gilde, V., & Sundt, B. (1989). On bonus systems with credibility scales. *Scandinavian Actuarial Journal*, *1989*(1), 13–22.

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

Hachemeister, C. (1975). Credibility for regression models with application to trend (reprint). *Credibility: Theory and Applications. Edited by P. Kahn. New York: Academic Press, Inc*, 129–163.

Hedeker, D. (2005). Generalized linear mixed models. *Encyclopedia of Statistics in Behavioral Science*.

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

Jewell, W. S. (1973). Multidimensional credibility. *Operations Research Center*, *73*(7).

Jewell, W. S. (1974). Exact multidimensional credibility. *Bulletin of Swiss Association of Actuaries*, *74*(1), 193–214.

Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2002). *Modern actuarial risk theory*. Kluwer Academic Publishers.

Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2008). *Modern actuarial risk theory: using R*, volume 128. Springer Science & Business Media.

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

Kremer, E. (1982). Credibility theory for some evolutionary models. *Scandinavian Actuarial Journal*, *1982*(3-4), 129–142.

Lemaire, J. (1995). *Bonus-malus systems in automobile insurance*. Boston: Kluwer Academic Publisher.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability 37. Chapman and Hall, London, New York.

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

Mowbray, A. H. (1914). How extensive a payroll exposure is necessary to give a dependable pure premium? In *Proceedings of the Casualty Actuarial Society*, volume 1, (24–30).

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *135*(3), 370–384.

Norberg, R. (1976). A credibility theory for automobile bonus systems. *Scandinavian Actuarial Journal*, *1976*(2), 92–107.

Norberg, R. (2004). Credibility theory. *Encyclopedia of Actuarial Science. Edited by B. Sundt and J. Teugels. New York: Academic Press, Inc*.

Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*. EAA Series. Springer Berlin Heidelberg.

Pai, J. S., & Walch, A. (2020). *ACTEX study manual for exam MAS-II*. ACTEX Learning/SRBooks. Greenland, NH.

Pinquet, J. (1997). Allowance for the cost of claims in bonus-malus systems. *ASTIN Bulletin: The Journal of the IAA*, *27*(1), 33–57.

Pinquet, J. (1998). Designing optimal bonus-malus systems from different types of claims. *ASTIN Bulletin: The Journal of the IAA*, *28*(2), 205–220.

Pinquet, J., Guillén, M., & Bolancé, C. (2001). Allowance for the age of claims in bonus-malus systems. *ASTIN Bulletin: The Journal of the IAA*, *31*(2), 337–348.

Pitrebois, S., Denuit, M., & Walhin, J.-F. (2003). Setting a bonus-malus scale in the presence of other rating factors: Taylor's work revisited. *ASTIN Bulletin: The Journal of the IAA*, *33*(2), 419–436.

Pitrebois, S., Denuit, M., & Walhin, J.-F. (2006). Multi-event bonus-malus scales. *Journal of Risk and Insurance*, *73*(3), 517–528.

Portet, S. (2020). A primer on model selection using the akaike information criterion. *Infectious Disease Modelling*, *5*, 111–128.

Saito, T., & 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.

Sundt, B. (1979). A hierarchical credibility regression model. *Scandinavian Actuarial Journal*, *1979*(2-3), 107–114.

Sundt, B. (1981). Recursive credibility estimation. *Scandinavian Actuarial Journal*, *1981*(1), 3–21.

Sundt, B. (1983). Parameter estimation in some credibility models. *Scandinavian Actuarial Journal*, *1983*(4), 239–255.

Sundt, B. (1988). Credibility estimators with geometric weights. *Insurance: Mathematics and Economics*, *7*(2), 113–122.

Taylor, G. (1997). Setting a bonus-malus scale in the presence of other rating factors. *ASTIN Bulletin: The Journal of the IAA*, *27*(2), 319–327.

Whitney, A. W. (1918). The theory of experience rating. In *Proceedings of the Casualty Actuarial Society*, volume 4, (274–292).

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

Zehnwirth, B. (1985). Linear filtering and recursive credibility estimation. *ASTIN Bulletin: The Journal of the IAA*, *15*(1), 19–35.