Integration of Traditional and Telematics data for Efficient Insurance Claims Prediction

by

Hashan Peiris

B.Sc., University of Sri Jayawardenepura, 2018

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2023

Declaration of Committee

Name:	**Hashan Peiris**
Degree:	**Master of Science**
Thesis title:	**Integration of Traditional and Telematics data for Efficient Insurance Claims Prediction**
Committee:	**Chair:** Yi Lu Professor, Statistics and Actuarial Science **Himchan Jeong** Supervisor Assistant Professor, Statistics and Actuarial Science **Gary Parker** Committee Member Associate Professor, Statistics and Actuarial Science **Joan Hu** Examiner Professor, Statistics and Actuarial Science

Abstract

While driver telematics has gained attention for risk classification in auto insurance, the scarcity of observations with telematics features has been problematic, which could be owing to either privacy concerns or adverse selection compared to the data points with traditional features.

To handle this issue, we propose a data integration technique based on calibration weights. It is shown that the proposed technique can efficiently integrate the so-called traditional data and telematics data and also cope with possible adverse selection issues on the availability of telematics data. Our findings are supported by a simulation study and empirical analysis on a synthetic telematics dataset.

Keywords: Adverse selection; Automobile insurance; Data integration; Driver telematics; Missing data analysis

Dedication

For those who paved the path to achieving my dreams, especially my parents and family.

Acknowledgements

I have been fortunate to have an excellent supervisor over my graduate studies at SFU. I am thankful forever and pay my gratitude to Dr. Himchan Jeong for his immense guidance, patience, and continuous support in helping me through my studies and our project. And I would like to thank Dr. Jae-Kwang Kim for supporting me through the research work. I wish to thank SFU for providing facilities to conduct this research successfully and thank the CIA for funding the research. I am thankful for all the faculty members and staff in our department that made my days at SFU as comfortable as home. Last but not least, thank you for the love and support from my family, relatives friends, and my wife, Kavisha.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Insurers have been modeling data to set prices for automobile insurance products for decades. Since the insurance industry consists of competitors with various products with minimal differences related to automobile insurance, fair and attractive pricing of products is beneficial for both the companies and the customers. Assessing the risk of new customers would help insurers grow in market share while meeting the profit margins as it enables them to set competitive premiums for policies. Over time the insurance industry has been upgraded with different products which use new measurements about drivers and vehicles. By now, the data collection is based not only on traditional self-reported forms and bills but also on devices that generate data about different measurements in high frequency as telematics devices on vehicles. Thus, these different datasets are available for the operational purposes of the company.

As a result, insurers tend to analyze these generated datasets with large dimensions during their operations to set the prices of their products. Hence two types of datasets are recognized, traditional and telematics, based on the generation process, for the ratemaking. And also analyzing these data can help to manage the risk of operations; efficiently screen cases, evaluate those cases with accuracy, and make accurate cost predictions. Therefore such methods used to model claim counts and the insurance products bounded with telematics data are discussed in this chapter.

1.1 Modeling Claim Counts

Interestingly, predicting the number of claims as assessing the risk in motor insurance is beneficial in premium calculation. There are different approaches for insurance ratemaking and risk assessment in the actuarial literature. Among them, Generalized Linear Models (GLMs) are widely in use as the successor of classical linear models (Haberman and Renshaw (1996)). One can interpret the effects of each variable using these models without being limited to predictions. The progression of studies shows the applicability of different linear modeling methods in insurance ratemaking as it is explained briefly to introduce random effects into the Poisson model by Boucher and Denuit (2006). They compare the parameter estimates with the fixed effects and examined the interpretation of those estimates. Furthermore, Denuit et al. (2007) provides a review of possible models that are used in modeling claim counts. Later, we can recognize some studies that have introduced alternative modeling approaches as Klinker (2010), which evaluates the use of Generalized Linear Mixed Models (GLMMs) in ratemaking, and Heras et al. (2018), which evaluates the use of quantile regression to estimate the aggregate claim amount.

In the actuarial literature, a set of studies tries to improve ratemaking by accounting for different effects caused by the data. As Chapados et al. (2001) describes, modeling claim counts accurately is a challenge when considering many features while the distribution of the number of claims is skewed and has a point mass at zero. The study of Boucher et al. (2007) shows approaches to model claim counts while considering the same point and Zamani and Ismail (2013) considers the discrepancies between the observed and predicted claim numbers with different modeling approaches. With the intuition of modeling claim count data by estimating possibly nonlinear effects of continuous risk factors and assessing the spatial risk variation, Denuit and Lang (2004) and Klein et al. (2014) use Generalized Additive Model (GAMs) in a Bayesian framework which relaxes the assumption of the linearity in the systematic component in GLM. The dependency between claim count and the claim amount is studied to introduce a method to model aggregate claims, relying on GLMs, by Garrido et al. (2016). This method is discussed further by Jeong et al. (2021) to fit GLMs and GLMMs accounted for dependency between claim count and amount. Furthermore, Jeong and Valdez (2018) uses the Least Absolute Shrinkage and Selection Operator (LASSO) with a proposed penalty in a Bayesian framework within the same setting. Also, Bermúdez and Karlis (2011) review multivariate methods for insurance ratemaking with dependency between types of claims and use the Bayesian approach to fit models. Later, Fuzi et al. (2016) use Bayesian approaches to model claim counts in a quantile regression setting as a method to study the entire distribution of claims.

With the advancements of data collection and computation techniques, data mining or statistical learning techniques are also being used for insurance ratemaking similar to the explanation of Smith et al. (2000). Improving the prediction accuracy and recognizing the variable importance are among the common motives of the researchers who have used these techniques. The comprehensive description of risk classification and predicting claim cost by Yeo et al. (2001) shows the transition of the ratemaking process for large data sources in the insurance industry. From the beginning of the new millennium, most researchers tend to evaluate common approaches and introduce statistical learning methods such as decision trees, Support Vector Machine (SVM), boosting or even clustering methods for insurance ratemaking and risk classification similar to Chapados et al. (2001), Yeo et al. (2001) and Guelman (2012). Further, Sakthivel and Rajitha (2017) uses artificial neural network in comparing the performance with zero-truncated Poisson regression model and hurdle model. A compact but comprehensive note on statistical learning approaches that are used in the actuarial literature is given in Vermet (2018). Ferrario et al. (2020) provides a complete guide in using neural network models to evaluate claim counts. Thus, it is observed that methods to model claim counts are presented within a broad spectrum.

1.2 Telematics in Insurance

By now insurers have gained access to data that relate to driving habits and drivers' behaviours on roads. Telematics is a technology that can provide data that contain such additional information since it generates data related to many variables characterized for each driver including but not limited to total miles driven, number of breaks or accelerations, and at what time they are behind the wheel. With the technological advancements in the automobile industry with driver telematics, as telematics service providers offer information to users for navigation, traffic control, and location-based information, customers who use those services can be attracted to telematics-based insurance products if it is beneficial to them. Thus, the insurers can add new features to the databases in addition to the traditional features that can be used in claim predictions and risk classifications in a unified frame.

For example, usage-based insurance (UBI) is an innovative product in the insurance industry based on these technological advances to assess the risk profile of a driver. Tselentis et al. (2017) provides a comprehensive review of frequently used UBI methodologies, Pay-as-you-drive (PAYD) and Pay-how-you-drive (PHYD), showing the varieties of applicability of telematics in insurance products. Generally, it allows drivers to be rewarded based on driving habits as it leads to lower premiums when they meet requirements in the respective insurance contract. Husnjak et al. (2015) states that UBI is a beneficial option for insurance companies to deal with revenue loss due to inconsistent pricing with individual risk and competitive pricing conditions in the market as well as being beneficial to the users. Thus, telematics allows insurers to closely observe driving behaviours and set insurance premium rates. According to Bolderdijk et al. (2011), when a discount on premium is applied, the number of times that a driver violates the defined speed limit in the contract has been reduced. By considering a distance-based insurance product similar to Litman (2011), it leads to a lower premium when lower mileage is achieved. Moreover, Denuit et al. (2019) recognizes UBI products as a mechanism that leads to efficient risk evaluation and acknowledges safer driving habits. Though there are pros and cons related to these particular UBI products, these types of products can obliquely improve road safety. Moreover, as it is expected, UBI products create social, economical, and environmental benefits.

1.2.1 Uses of Telematics Data

Past studies have shown that the extra value of information derived from telematics can provide improved claims predictions, risk classification, and premium assessments. Ayuso et al. (2014) compares driving behaviours of novice and experienced young drivers with pay-as-you-drive policies using a few telematics variables as well as traditional variables. Here, they have compared the vehicle usage and driving pattern by the level of experience. Further, they analyze the relationship between the number of days and the number of kilometers to the first accident. Furthermore, Ayuso et al. (2016) examines gender discrimination in the risk of accidents using the same dataset. In a study to demonstrate the use of modeling techniques for customizing insurance products based on information, Baecke and Bocca (2017) illustrates the use of telematics variables to relate the premium to the risk. They state that at least three months of data are enough to obtain efficient risk estimates. In the study of Verbelen et al. (2018) which depicts the importance of telematics variables based on driving habits and styles in predicting claim frequency, they discuss the possibilities of PHYD policies. Gao et al. (2019) shows the predictive power of telematics covariates extracted from speed-acceleration heat maps in claim frequency modeling and supports the use of telematics features for insurance pricing. Accordingly, Pesantez-Narvaez et al. (2019) tries to improve a method to recognize the effect of risk factors on the probability of claim as a tool for regulation using the same dataset from the studies of Ayuso et al. (2014) and Ayuso et al. (2016).

As it is mentioned in section 1.1, similar approaches are followed by researchers to model data that contains both traditional and telematics variables. Due to the high-dimensionality of the available data, data mining and machine learning methods are also used in predicting claims. When briefly going through the list of studies in the literature, we can recognize a few studies with different modeling approaches with data that contains both telematics and traditional features. Among these studies, Ma et al. (2018) uses GLMs, Gao et al. (2022) uses GLMs and neural network models, Verbelen et al. (2018) uses GAMs while Gao et al. (2019) has applied GAMs but uses transformed telematics variables using Principal Components and bottleneck neural network approaches.

Apart from such models, Denuit et al. (2019) proposes multivariate mixed models to explore the joint distribution of telematics data and claim frequencies which can be utilized in predicting some policy characteristics. One can incorporate insights from unsupervised learning methods to study driving styles without discussing the risk profile of drivers as in Wüthrich (2017) which uses k-means clustering. All of these studies have shown the importance of telematics features in decision-making. Hence, those studies provide motives to introduce more procedures to model claim count efficiently.

1.2.2 Challenges

As it is mentioned earlier, insurance companies have large datasets related to policyholders that contain traditional features, including but not limited to the driver's age, gender, and vehicle characteristics. On the other hand, a telematics dataset can have a smaller number of data points with respect to the traditional dataset when the number of policyholders who owns a policy related to telematics is low. Meyers and Hoyweghen (2020) describes a study about a telematics insurance product that had failed to achieve the target number of participants despite offering a discount on premium for participation. The proposed method of Castignani et al. (2015) tries to improve the availability of telematics data by using smartphone solutions as a remedy to the reluctance of drivers to install telematics devices which results in an initial cost. And Ma et al. (2018) mentions that the lack of telematics data availability is a challenge in identifying the factors of policyholder behaviour within the actuarial pricing methods incorporated with traditional data that is easily available. Additionally Husnjak et al. (2015) states that a method to maintain the quality of data can be questionable as the collection process can be interrupted due to technical or environmental impacts. Meantime, Guillen et al. (2021) uses a modeling approach for insurance ratemaking using both traditional and telematics data but is limited to a small number of features as available data is limited. In terms of risk classification, Tuna and Cengiz (2021) recommends using telematics data that are observed for a predetermined short period. Furthermore, they have mentioned that collecting less telematics information helps to meet the recommendations of insurance regulators.

Indeed, consideration and collection of telematics data are relatively recent and there are still ongoing concerns about privacy issues, which make many policyholders reluctant to agree on the provision of their telematics data to the insurers. Duri et al. (2002) and Duri et al. (2004) state that the success of telematics applications depends on the awareness of service providers that the data is received accurately while assuring the privacy of end users. And also Milanović et al. (2020) states an idea in terms of driver's acceptance of telematics technology which is shown to be dependent on privacy concerns. Similarly, Buxbaum (2006) states that the policyholders willing to provide telematics data tend to have less concern about privacy issues. With a wider perspective, Jaisingh et al. (2016) discusses the flow of telematics data while recognizing the privacy and security issues and suggesting actions to overcome those. Thus, all of them have proposed a framework to protect data that is based on privacy and security technology. But the study of Eling and Kraft (2020) which provides a compact description of the insurability of risk using telematics data, also highlights some recommendations from the literature that can improve the number of telematics-based policyholders.

Apart from the above concerns, one can recognize that the information from available telematics data may depend more on the type of customer. The study of Denuit et al. (2019)

states that drivers with low risk would favor telematics insurance products. Further Duval et al. (2022) mentions that the attraction of safer drivers is beneficial for the insurer as it could lower the claim cost. In conclusion, all those studies have raised the fact that claim frequency can be reduced due to the awareness of drivers about being monitored. Although it is an economic benefit for the insurer, the insured, and society, this situation results in missing some insights about more risky drivers in terms of an analytical point of view. But using the explanation in the study of Cohen and Siegelman (2010) about adverse selection in insurance, one can think of a situation where a buyer knows that he can have a lower premium by achieving a lower claim count as a case weekly related to the adverse selection. Thus we can consider the aforementioned situations as challenges to having a telematics dataset as informative as the finite population of drivers.

1.3 Motivation

In this regard, it is natural to expect that an insurer need to deal with two types of datasets; traditional datasets with fewer features and a larger number of observations, and telematics datasets with more features and a smaller number of observations. Data integration techniques enable to combine information in some data sources into one. According to Yang and Kim (2020), using this concept in survey sampling leads to incorporating information from different samples to obtain efficient estimators under finite population inference. Husnjak et al. (2015) recognizes the use of telematics variables and identification of the driver's behaviour as beneficial for precise ratemaking while stating the integration of telematics data with traditional data as a way to realize the full potential of telematics data.

If a model can be formed representing all the information in these two datasets, then it will help the insurer to evaluate the claim counts behaviour under all available features. At the same time such a model is useful for regulators as the effects of each risk factor on the claim counts are available. As a possible solution for considering both traditional and telematics features simultaneously, Ayuso et al. (2019) propose an approach where a GLM with telematics variables is improved by incorporating information in traditional variables. Also, Gao et al. (2022) propose a two-step approach that utilizes telematics features to boost a regression model fitted only with traditional ratemaking factors. While these approaches are straightforward and readily available, they might be problematic when the availability of telematics features depends on the riskiness of the policyholders due to possible adverse selection. One can expect that due to the information asymmetry between the insurer and the policyholder, those with less risky driving behaviours are more likely to agree with the provision of telematics data for possible premium discounts, which is equivalent to a non-ignorable sampling mechanism of the observations with telematics features.

Hence, the data integration techniques can be used in similar studies, not only as an emerging research area in survey sampling but also it improves the efficiency of the resulting estimator. Thus, here we focus on using the propensity score estimation method proposed by Wang and Kim (2021a) which can be used as a unified tool for combining information from multiple datasets with the consideration of possible non-ignorable sampling bias. This method is also similar to the data integration method given in Wang et al. (2022) where the data integration is done by regression models using propensity score, as proposed in Wang and Kim (2021b), to incorporate information from different data sources. It suggests a model calibration technique to get partial information from external data sources (traditional dataset in our case) and integrate the data sources at once.

1.4 Summary

In summary, the following objectives of the study are recognized, which will be described appropriately in later chapters.

- Demonstrate a need for a flexible model that can deal with multiple sources of data in insurance ratemaking due to the scarcity of telematics data compared to the traditional data and possible adverse selection regarding the availability of the telematics data.

- Develop an algorithm to integrate a traditional insurance claims dataset and a telematics dataset, using the data integration method proposed by Wang and Kim (2021a).

- Test the validity and applicability of the proposed method via a simulation study and empirical analysis of a synthetic telematics dataset. Here, our goal is to evaluate the performance of the proposed method in possible selection biases in the availability of telematics data.

- Compare the performance of the proposed method with selected pre-existing methods within the same context of selection biases.

Consequently, we expect that the proposed method can help insurance companies to effectively utilize multiple sources of data for better risk classification and tarification.

The remaining part of this work is organized as follows. Chapter 2 provides a detailed description of the problem and the corresponding data structure with the missing mechanism. In Chapter 3, the proposed method for data integration is developed based on a calibration equation approach with information projection. Chapter 4 provides a simulation study to assess the effects of the proposed method compared to the existing approaches. Chapter 5 conducts an empirical analysis with a synthetic telematics data portfolio, to assess the applicability of the proposed method in practice. Chapter 6 concludes the study with some remarks.

Chapter 2

Data structure and problem description

As this study focuses on two data sources, describing the structure of the overall dataset is needed since the proposed method would use these two data sources differently when estimating parameters. Though our main objective is to propose a method to integrate these two sources of data, we also discuss whether the proposed method is affected by the adequacy of information in telematics data that are used to fit the model. Before introducing the modeling technique and the evaluation methods, the data structure and the problems that can be addressed are described briefly in this chapter.

2.1 Data Structure

When both traditional and telematics data are used for ratemaking, we can expect the structure of the dataset to be asymmetric due to the higher number of policyholders who provide traditional information than those who provide both traditional and telematics information.

Let's define the two datasets and the population as,

- $\mathcal{S}_0$ be a small dataset with M_0 number of policyholders that contains both traditional and telematics features.
- $\mathcal{S}_1$ be a large dataset with M_1 number of policyholders that contains only traditional features.
- The finite population $\mathcal{S}$ consists of $\mathcal{S}_0$ and $\mathcal{S}_1$ and the total number of policyholders in $\mathcal{S}$ is M.

We are interested in finding the relationship of the features in $\mathcal{S}_0$ with the claim counts of a certain policy by improving estimation efficiency of the model estimates through information integrated with the features in $\mathcal{S}_1$.

Now the features in those two datasets are denoted as,

- $\mathbf{x}_{i1}$ be the traditional features of the i^{th} policyholder. Those are available in both $\mathcal{S}_0$ and $\mathcal{S}_1$.
- $\mathbf{x}_{i2}$ be the telematics features of the i^{th} policyholder. Those are only available in $\mathcal{S}_0$.
- All available features in the study can be denoted as a vector, $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2})$.

A summary of the description of data is given in Figure 2.1 which depicts the asymmetric structure of the full dataset in reality.

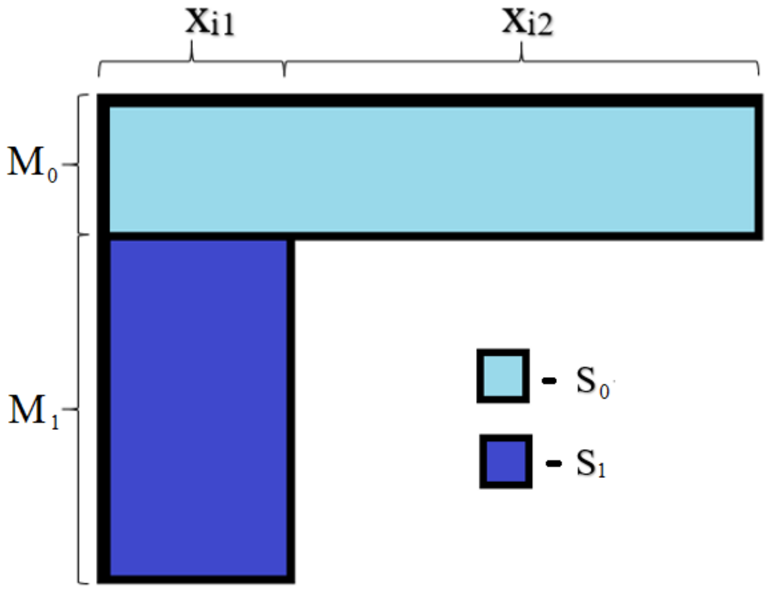


Figure 2.1: Pictorial visualization of $\mathcal{S}_0$, $\mathcal{S}_1$, $\mathbf{x}_{i1}$, and $\mathbf{x}_{i2}$

2.2 Problem Description

As in Chapter 1, our study focuses on introducing a method to efficiently model claim count data which uses all information in $\mathcal{S}$ through a novel data integration technique. And we evaluate the performance of this method and compare it with some benchmark models. Therefore, it opens up a discussion about the available data which can be stated as the problem of not being a representative sample of the finite population of customers that generate telematics data.

Note that observability of $\mathbf{x}_{i2}$ might depend on the risk profile of the i^{th} policyholder, which could make the sampling mechanism of $\mathcal{S}_0$ from the population is not completely at random. As mentioned in the previous chapter, there have been possible concerns of providing telematics data such as privacy and security issues so that it is natural to expect that a policyholder might not be willing to provide their telematics data to the insurer unless the expected benefits from the provision outweigh the possible concerns. Therefore, one can think about the following conjectures:

- Those who are younger tend to agree provide telematics data more since they are less reluctant to technology, which implies the sampling probability of an observation in $\mathcal{S}_0$ from the population is inversely proportional to the driver's age.
- Those who are less risky tend to agree to provide telematics data more so that the accessibility of $\mathbf{x}_{i2}$ is prone to adverse selection, which implies the sampling probability of an observation in $\mathcal{S}_0$ from the population is inversely proportional to the number of claims (n_i).

While our main task is not to detect possible selection biases in the availability of telematics features and prove such conjectures, we consider the situations where such conjectures do hold and discuss the benefits of the proposed method compared to pre-existing benchmark methods in various situations.

Chapter 3

Methodology

With the intuition of introducing the data integration approach, we describe necessary steps to fit a model. The general framework that we follow to estimate the model parameters using the proposed data integration method is briefly described in this chapter.

3.1 Generalized Linear Models

The Generalized Linear Models (GLMs) enable one to fit a model by specifying the form of the linear relationship between a function of the mean of a response variable, which is not always from a normal distribution, and the predictor variables. The model contains three components.

The distribution of the response variable (Y_i), which is assumed to be in the natural exponential family, is called the random component. Lets denote $E(Y_i) \equiv \mu_i$. Then, the systematic component is a linear predictor which is formed with covariates and regression coefficients as

$$\gamma_i = \sum_j \beta_j^* x_{ij}.$$

Here, x_{ij} is the value of the j^{th} covariate of the i^{th} person where $x_{i0} = 1$ and β_0 is the intercept. Moreover, the link function describes the relationship between random and systematic components. Thus it links the μ_i to γ_i by $\gamma_i = g(\mu_i)$ where $g(\cdot)$ is a monotonic and differentiable function.

Let us assume that we need to model the claim counts. Since it is a non-negative discrete random variable where possible values are integers, it is assumed to follow a Poisson distribution with mean μ. Let assume that $\mathbf{N}$ is such a random variable. If $k \in \mathbf{N}$, then $k \in \{0, 1, 2, ...\}$ in general. The probability mass function of $\mathbf{N}$ is,

$$Pr[\mathbf{N} = k] = p(k|\mu) = \frac{e^{-\mu}\mu^k}{k!}$$

where $p(k|\mu) \geq 0$ for all k and $\sum_{k=0}^{\infty} p(k|\mu) = 1$.

In general, we are interested in estimating $\boldsymbol{\beta}$ in the regression model $E(N_i \mid \mathbf{x}) = m(\mathbf{x}'\boldsymbol{\beta})$, where $m(\cdot)$ is a known function and $\boldsymbol{\beta}$ is an unknown parameter while N_i is the observed number of claims for policyholder i.

The census estimating equation for $\boldsymbol{\beta}$ can be written as

$$\sum_{i=1}^{M}\{n_i - m(\mathbf{x}_i'\boldsymbol{\beta})\}h(\mathbf{x}_i'\boldsymbol{\beta}) = 0,$$

for some $h(\cdot)$ such that the solution to this equation exists uniquely (almost everywhere) and we can define

$$U(\boldsymbol{\beta}; \mathbf{x}_i, n) = \{n - m(\mathbf{x}_i'\boldsymbol{\beta})\}h(\mathbf{x}).$$

The proposed method is independent of the distribution of N_i as it is based on general calibration and estimating equations. But we assume that N_i's are independently distributed with the Poisson distribution with mean μ_i in order to focus on the impact of the proposed data integration approach. It enables one to compare the performance of selected models in one platform. Since the Poisson is in the natural exponential family with the natural parameter $\log(\mu_i)$, this can be used as the link function where $\log(\mu_i)$ is strictly monotonic and differentiable over the range of μ_i. Using this canonical link function as it is given in Agresti (2003), we can express the model which is linear in terms of regression coefficients as

$$\log(\mu_i) = \sum_j \beta_j^* x_{ij} = \mathbf{x}_i\boldsymbol{\beta}^*.$$

When t_i is an exposure variable associated with the i^{th} claim count and η_i is the average number of claims per year, we can redefine the model in terms of η_i as,

$$\log(\eta_i) = \sum_j \beta_j x_{ij} = \mathbf{x}_i\boldsymbol{\beta}.$$

Thus, we can write this model using the definition $\mu_i = t_i\eta_i$ as,

$$\log(\mu_i) = \log(t_i) + \sum_j \beta_j x_{ij} = \log(t_i) + \mathbf{x}_i\boldsymbol{\beta}. \tag{3.1}$$

where t_i is the duration.

Likewise, under the Poisson regression model where an exposure variable is used, we have $m(\mathbf{x}_i'\boldsymbol{\beta}) = t_i \exp(\mathbf{x}_i'\boldsymbol{\beta})$ and $h(\mathbf{x}_i) = \mathbf{x}_i$. Therefore, using model (3.1), the census estimating equation for $\boldsymbol{\beta}$ can be written as

$$\sum_{i=1}^{M}\{n_i - t_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\}\mathbf{x}_i = 0. \tag{3.2}$$

Unfortunately, we do not observe $\mathbf{x}_i$ throughout the population. The structure of the dataset in Figure 2.1 depicts that $\mathbf{x}_{i2}$ are not available for $\mathcal{S}_1$ which persuades to use the proposed method.

3.2 Proposed Method

Keeping the GLM framework with the assumption on the distribution of N_i, we now adopt the single sample problem in the study of Wang and Kim (2021a) into the data structure in Section 2.1.

3.2.1 Estimation of Parameters

Let $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2})$ where $\mathbf{x}_{i2}$ is subject to missingness. We can define the estimating function for $\boldsymbol{\beta}$ as

$$U(\boldsymbol{\beta}; \mathbf{x}, n) = \{n - t\exp(\mathbf{x}\boldsymbol{\beta})\}\mathbf{x}.$$

To follow the setup of Wang and Kim (2021a), we assume that $\mathcal{S}_1$ with $(\mathbf{x}_{i1}, n_i)$ are observed throughout the finite population. In the sample $\mathcal{S}_0$, we observe $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, n_i)$. In this case, we wish to construct the propensity weight $\omega_i = \omega(\mathbf{x}_{i1}, n_i)$ in $\mathcal{S}_0$ such that

$$\sum_{i \in \mathcal{S}_0} \omega_i U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) = \sum_{i=1}^{M} \left[\delta_i U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) + (1-\delta_i)\bar{U}(\boldsymbol{\beta}; \mathbf{x}_{i1}, n_i)\right], \tag{3.3}$$

where $\delta_i = \mathbb{I}(i \in \mathcal{S}_0)$ and $\bar{U}(\boldsymbol{\beta}; \mathbf{x}_{i1}, n_i) = E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i1}, n_i\}$. Here a weighted complete case estimator is taken through an expected estimating equation approach that computes the conditional expectation of $U(\boldsymbol{\beta}; \mathbf{x}_i, n_i)$. The weight is often called the propensity score and defined as $\omega_i = 1/Pr(\delta_i = 1|\mathbf{x}_i, n_i)$.

The propensity score (PS) estimating equation satisfying (3.3) is called self-efficient, as it leads to efficient estimation of $\boldsymbol{\beta}$ as long as the conditional expectation $E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i1}, n_i\}$ is correct.

Here, we assume that the sampling mechanism for $\mathcal{S}_0$ is missing at random (MAR) in the sense of Rubin (1976). That is, we assume

$$\delta \perp \mathbf{x}_2 \mid (n, \mathbf{x}_1).$$

To find ω_i satisfying (3.3), we first find the basis functions satisfying

$$E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i1}, n_i\} \in \operatorname{span}\{b_1(\mathbf{x}_{i1}, n_i), \ldots, b_L(\mathbf{x}_{i1}, n_i)\}. \tag{3.4}$$

Here, we expect to represent the conditional expectation by a combination of basis functions that are formed only using traditional features.

Now, using the basis functions in (3.4), we impose

$$\sum_{i\in\mathcal{S}_0} \omega_i[1, b_{1i}, \cdots, b_{Li}] = \sum_{i=1}^{M}[1, b_{1i}, \cdots, b_{Li}] \tag{3.5}$$

as a constraint for propensity weights ω_i, where $b_{li} = b_l(\mathbf{x}_{i1}, n_i)$ for the i^{th} individual. To be specific, we take $[1, b_{1i}, \cdots, b_{Li}] = [1, x_{i1}, N_i, N_i * x_{i1}]$ in here. Therefore, $L = 2 * v + 1$ where v is the number of features in x_{i1}. Constraint (3.5) is often called the covariate-balancing property (Imai and Ratkovic, 2014) or calibration property (Deville and Särndal, 1992).

Now, as long as (3.5) is satisfied, we can express

$$\begin{aligned}\sum_{i\in S_0} \omega_i U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) &= \sum_{i=1}^{M} \delta_i\omega_i U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) + \sum_{i=1}^{M} (1-\delta_i\omega_i) \sum_{k=0}^{L} \alpha_k b_{ki} \\ &= \sum_{i=1}^{M} \left\{ \delta_i U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) + (1-\delta_i) \sum_{k=0}^{L} \alpha_k b_{ki} \right\} \\ &\quad + \sum_{i=1}^{M} \delta_i(\omega_i - 1) \left\{ U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) - \sum_{k=0}^{L} \alpha_k b_{ki} \right\}\end{aligned}$$

for any $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_L)$. Thus, for the choice of $\hat{\boldsymbol{\alpha}}$ satisfying

$$\sum_{i=1}^{M} \delta_i(\omega_i - 1) \left\{ U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) - \sum_{k=0}^{L} \hat{\alpha}_k b_{ki} \right\} = 0, \tag{3.6}$$

we can obtain

$$\sum_{i\in S_0} \omega_i U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) = \sum_{i=1}^{M} \left\{ \delta_i U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) + (1-\delta_i) \sum_{k=0}^{L} \hat{\alpha}_k b_{ki} \right\}. \tag{3.7}$$

Furthermore, the condition in (3.6) under model (3.4) implies that $\sum_{k=0}^{L} \hat{\alpha}_k b_{ki}$ is an estimator of $E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i1}, n_i\}$. Thus, we can see that (3.7) shows the self-efficiency in (3.3). That is, the calibration condition (3.5) on the basis functions in (3.4) is a sufficient condition for self-consistency.

Now, to uniquely determine ω_i, we can use the information projection of Wang and Kim (2021a) under constraint (3.5) to get

$$\omega_i = 1 + \frac{M_1}{M_0} \exp\left\{\phi_0 + \phi_1 b_{1i} + \cdots + \phi_L b_{Li}\right\}, \tag{3.8}$$

where $M_0 = \sum_{i=1}^{M} \delta_i$, $M_1 = M - M_0$ and $\boldsymbol{\phi} = (\phi_0, \cdots, \phi_L)$ is an unknown parameter. The exponent is the smoothed density ratio function from the information projection technique. The parameters are estimated by solving the calibration equation in (3.5).

Once $\phi_0, \cdots, \phi_L$ are estimated by (3.5), we can use

$$\hat{\omega}_i = 1 + \frac{M_1}{M_0} \exp\left\{\hat{\phi}_0 + \hat{\phi}_1 b_{1i} + \cdots + \hat{\phi}_L b_{Li}\right\}$$

as the final propensity weights for estimating $\boldsymbol{\beta}$ using (3.9):

$$\sum_{i \in \mathcal{S}} \delta_i \hat{\omega}_i(\boldsymbol{\phi}) U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) = 0, \tag{3.9}$$

where $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1$ is the combined sample. Because the propensity weights satisfy the calibration equation in (3.5), this equation satisfies the self-efficiency without estimating the regression coefficients $\hat{\boldsymbol{\alpha}}$ in the regression model

$$E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i1}, n_i\} = \sum_{k=1}^{L} \alpha_k b_l(\mathbf{x}_{i1}, n_i).$$

3.2.2 Standard Errors of Estimates

We are also interested in studying the standard errors of the estimates. The calculation of the standard errors of the estimates is done according to Wang and Kim (2021a). As it is described by Wang and Kim (2021a), there are two models in this method. One is the PS model (with parameter $\boldsymbol{\phi}$) and the other is the regression outcome model (with parameter $\boldsymbol{\beta}$).

We can construct two estimating functions for estimating two parameters by defining $\delta_i = \mathbb{I}(i \in S_0)$ as follows:

$$\begin{aligned} \hat{U}_1(\boldsymbol{\phi}) &= \sum_{i \in \mathcal{S}} \{\delta_i \omega_i(\boldsymbol{\phi}) - 1\} \mathbf{b}_i, \\ \hat{U}_2(\boldsymbol{\phi}, \beta) &= \sum_{i \in \mathcal{S}} \delta_i \hat{\omega}_i(\boldsymbol{\phi}) U(\boldsymbol{\beta}; \mathbf{x}_i, n_i), \end{aligned}$$

where $\mathbf{b}_i = (1, b_{1i}, \cdots, b_{Li})'$ and

$$\omega_i(\boldsymbol{\phi}) = 1 + \frac{M_1}{M_0} \exp\{\phi_0 + \phi_1 b_{1i} + \cdots + \phi_L b_{Li}\}.$$

The final estimator $\hat{\boldsymbol{\beta}}$ is the solution to the joint estimating equations:

$$\hat{U}_1(\boldsymbol{\phi}) = 0 \quad \text{and} \quad \hat{U}_2(\boldsymbol{\phi}, \boldsymbol{\beta}) = 0.$$

We can treat $\boldsymbol{\theta}' = (\boldsymbol{\phi}', \boldsymbol{\beta}')$ and define

$$\hat{U}(\boldsymbol{\theta}) = \begin{pmatrix} \hat{U}_1(\boldsymbol{\phi}) \\ \hat{U}_2(\boldsymbol{\phi}, \boldsymbol{\beta}) \end{pmatrix}.$$

The variance estimation for $\hat{\boldsymbol{\theta}}$ can be implemented using the Sandwich formula. That is,

$$V(\hat{\boldsymbol{\theta}}) = \tau^{-1} V(\hat{U}) \tau^{-1'} \qquad \text{where } \tau = E\left\{ \frac{\partial}{\partial \boldsymbol{\theta}'} \hat{U}(\boldsymbol{\theta}) \right\}.$$

One can use an empirical estimate of $V(\hat{\boldsymbol{\theta}})$ as follows:

$$\tilde{\tau} = \frac{\partial}{\partial \theta'} \hat{U}(\theta) \Big|_{\theta = \hat{\theta}} \qquad \text{and } \tilde{V}(\hat{U}) = \sum_{i \in \mathcal{S}} (\tilde{U}_i - \overline{\tilde{U}}_i)(\tilde{U}_i - \overline{\tilde{U}}_i)'$$

as a proxy of τ and $V(\hat{U})$, respectively, where $\hat{\theta}' = (\hat{\phi}', \hat{\boldsymbol{\beta}}')$ is the solution of the joint estimating equation and

$$\tilde{U}_i = \begin{pmatrix} \left\{ \delta_i \omega_i(\hat{\phi}) - 1 \right\} \mathbf{b}_i \\ \\ \delta_i \hat{\omega}_i(\hat{\phi}) U(\hat{\boldsymbol{\beta}}; \mathbf{x}_i, y_i) \end{pmatrix}, \quad \overline{\tilde{U}}_i = \frac{1}{M} \sum_{i \in \mathcal{S}} \tilde{U}_i.$$

3.3 Estimation scheme

Now, the estimation scheme for the study is listed orderly according to the requirement of the estimation process at each step.

1. Find $\mathcal{H} = \text{span}\{b_1(\mathbf{x}_{i1}, n_i), \ldots, b_L(\mathbf{x}_{i1}, n_i)\}$ such that

   $$E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i1}, n_i\} \in \mathcal{H},$$

   where $U(\boldsymbol{\beta}; \mathbf{x}_i, n_i)$ is the estimating function for $\boldsymbol{\beta}$.

2. Obtain $\hat{\boldsymbol{\phi}}$ by solving

   $$\sum_{i \in \mathcal{S}_0} \left\{ 1 + \frac{M_1}{M_0} \exp(\phi_0 + \phi_1 b_{1i} + \cdots + \phi_L b_{Li}) \right\} [1, b_{1i}, \cdots, b_{Li}] = \sum_{i=1}^{M} [1, b_{1i}, \cdots, b_{Li}],$$

3. Obtain $\hat{\boldsymbol{\beta}}$ by solving

   $$\sum_{i \in \mathcal{S}} \delta_i \hat{\omega}_i(\phi) U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) = 0$$

   $$\text{where} \qquad \hat{\omega}_i = 1 + \frac{M_1}{M_0} \exp\left\{ \hat{\phi}_0 + \hat{\phi}_1 b_{1i} + \cdots + \hat{\phi}_L b_{Li} \right\}.$$

Chapter 4

Simulation study

Now, we focus on assessing the applicability of the proposed method and compare this method with some feasible methods in claim count modeling using simulated datasets. Here we use different hypothetical sampling schemes to incorporate the problem regarding the availability of information in telematics data. Components of the simulation study in generating data, applying a model, and evaluating the performance are described in this chapter.

4.1 The Finite Population

By considering the conjectures based on the possible selection biases of the availability of telematics data in Section 2.2, we assume three hypothetical scenarios on the sampling mechanism of observations in $\mathcal{S}_0$. It will open up a more general platform to compare the performances of the models.

Also the dataset is formed as the traditional features are fully available and the telematics features are partially available as in Section 2.1.

First, we generate a finite population of size $100,000$ which contains three independent variables and the number of claims with the following specification:

$$\begin{aligned}
&N_i \sim \mathcal{P}(\lambda_i)\\
&\log \lambda_i = \mathbf{x}_{i1}\boldsymbol{\beta}_1 + \mathbf{x}_{i2}\boldsymbol{\beta}_2,\\
&\boldsymbol{\beta}_1 = (\beta_0, \beta_{A1}, \beta_{A2}, \beta_G) \text{ and } \boldsymbol{\beta}_2 = \beta_T,\\
&\mathbf{x}_{1i} = (1, x_{Ai}, x_{Ai}^2, x_{Gi}) \text{ and } \mathbf{x}_{2i} = x_{Ti},\\
&x_{Ai} \sim \mathcal{U}(0.18, 0.81),\ x_{Gi} \sim \mathcal{B}er(0.6) \text{ and } x_{Ti} \sim \mathcal{N}(0,1),\\
&\beta_0 = -1.3,\ \beta_{A1} = -4,\ \beta_{A2} = 3.4,\ \beta_G = 0.1 \text{ and } \beta_T = 0.5
\end{aligned}$$

where $\mathcal{P}$, $\mathcal{U}$, $\mathcal{B}er$ and $\mathcal{N}$ refer to Poisson, uniform, Bernoulli, and normal distributions, respectively.

Note that, we set the duration to be 1 for all individuals to have a simplified form of (3.1) which resulted in a model without an offset.

Furthermore, we recognize the variables as below.

- x_{Ai} refers to a traditional continuous variable with quadratic effect (e.g., driver's age)
- x_{Gi} refers to a traditional binary variable (e.g., gender)
- x_{Ti} refers to a telematics variable of significant impact on the risk profile

Note that the driver's age has been taken as an example of the quadratic effect by expecting the fact that both young and old drivers can be risky due to their driving behaviours being affected by other biological and social facts similar to Guillen et al. (2019).

Here, each feature contains 100,000 data points. Therefore, let $\mathcal{S}^*$ be the generated finite population in accordance with the notation used in Section 2.1.

4.2 Estimation Procedure

Once a finite population is generated ($\mathcal{S}^*$), we use it to have a training set with the form of the data structure of $\mathcal{S}$ and a test set with the form of $\mathcal{S}_0$.

At each round of simulation, the following scheme is applied to split the data appropriately:

Step 1: Firstly, 10% of data points are set aside at random as $\mathcal{T}$ for out-of-sample validation, where $\{N_i, \mathbf{x}_{i1}, \mathbf{x}_{i2}\}$ are all available.

Step 2: After that, 10% of data points are set aside where $\{N_i, \mathbf{x}_{i1}, \mathbf{x}_{i2}\}$ are all available, which is equivalent to $\mathcal{S}_0$ in Section 2.1. Depending on the assumption about the availability of the telematics information in Section 2.2, we applied the following three sampling schemes of $\mathcal{S}_0$:

- **Random selection**: The data points assigned to $\mathcal{S}_0$ are chosen at random,
- **Age selection**: Each data point assigned to $\mathcal{S}_0$ is chosen with the sampling probability proportional to $1/(1 + \exp(3x_{Ai}))$, which means those at younger ages are more likely to provide the telematics information due to their lower reluctance to new technologies. In this case, the sampling mechanism is MAR or ignorable and $\delta \perp N|\mathbf{x}_1$.
- **Adverse selection**: Each data point assigned to $\mathcal{S}_0$ is chosen with the sampling probability proportional to $1/(1 + \exp(2N_i))$, which means those with less risky behaviours are more likely to provide the telematics information. In this case, the sampling mechanism is still MAR or ignorable but $\delta \not\perp N|\mathbf{x}_1$.

Step 3: The remaining 80% of data points are used as a large dataset but only with traditional features $\{N_i, \mathbf{x}_{i1}\}$, which is equivalent to $\mathcal{S}_1$ in Section 2.1.

These data are appropriately used to fit different models. Here, we consider the following models to estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ where the appropriate partitions of the generated population, that are used in each model, are indicated in Figure 4.1:

- **Naive model**: Fit a usual Poisson GLM using the data points in $\mathcal{S}_0$.
- **Traditional model**: Fit a usual Poisson GLM using only the traditional features and the response variable $\{N_i, \mathbf{x}_{i1}\}$ in $\mathcal{S}_0 \cup \mathcal{S}_1$. As such, this model does not allow to use the telematics information at all in the risk classification.
- **Full model**: It uses all the data points in $\mathcal{S}^* - \mathcal{T}$ to estimate the regression coefficients of a usual Poisson GLM, so that it is expected to provide the best estimation performance. Note that $\mathcal{S}^*$ might not be available in practice.
- **Boosting model**: It uses the same estimates of $\boldsymbol{\beta}_1$ from the traditional model and compute $\hat{\eta}_i = \exp(\mathbf{x}_{i1}\hat{\boldsymbol{\beta}}_1)$ for each observation i in $\mathcal{S}_0$. After that, another Poisson GLM is fitted with $\mathcal{S}_0$ where the telematics information, $\mathbf{x}_{i2}$, is the only regressor and $\log \hat{\eta}_i$ is used as an offset, to further estimate $\hat{\boldsymbol{\beta}}_2$ which is mentioned as the boosting approach in Ayuso et al. (2019).
- **Proposed model**: It follows the estimation procedures of parameters and standard errors described in Section 3.

After all models are fitted, the regression estimates from these models were used to find the predictive value $\hat{N}_i$ for i^{th} policyholder in the out-of-sample validation set $\mathcal{T}$. Since these models use the generated data differently, a summary of estimating and predicting for each model is given in Table 4.1.

4.3 Evaluation Procedure

Note that generation of the finite population ($\mathcal{S}^*$), data split, estimation, and prediction are repeated $R = 1,000$ times with different random seeds. This enables us to evaluate and compare these methods comprehensively. First, we discuss about in-sample estimation using Bias, root mean-squared error (RMSE) and 90% confidence interval coverage (CI) of $\boldsymbol{\beta}_j$. Let these statistics be defined as follows:

$$\text{Bias}_j = \frac{1}{R}\sum_{r=1}^{R}(\beta_j - \hat{\beta}_j^{(r)}),$$

$$\text{RMSE}_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\beta_j - \hat{\beta}_j^{(r)})^2},$$

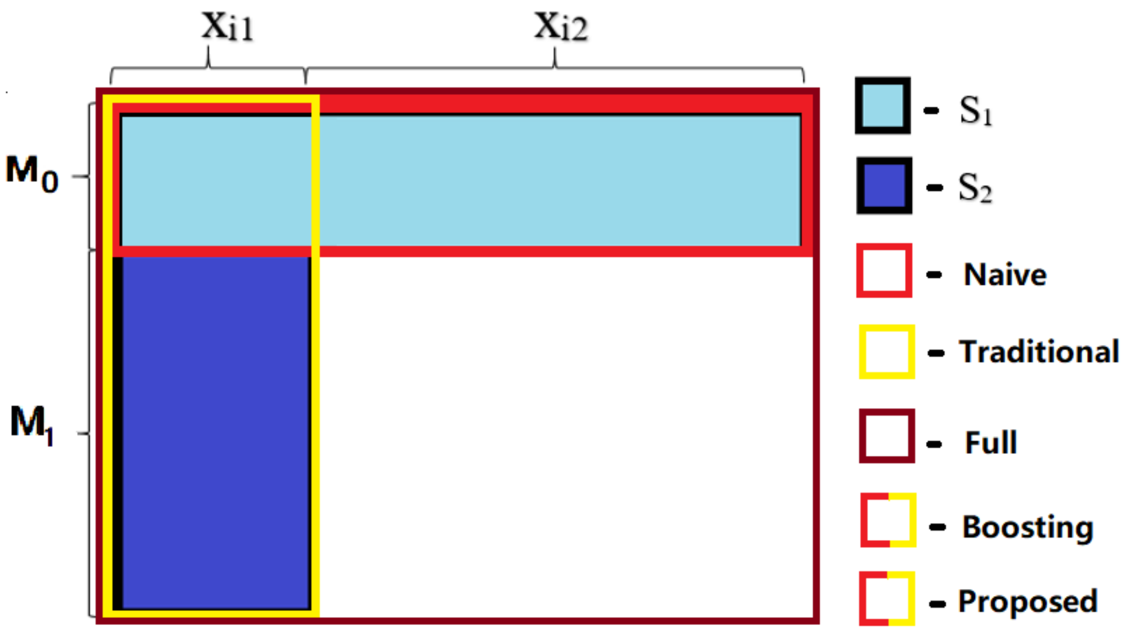


Figure 4.1: Pictorial visualization of data structure considered in each the model

$$\mathrm{CI}_j = \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}_{\{|\beta_j - \hat{\beta}_j^{(r)}| < 1.645\cdot \mathrm{SE}(\hat{\beta})_j^{(r)}\}},$$

where $\hat{\beta}_j^{(r)}$ is the estimate of β_j for the r^{th} simulation, and $\mathrm{SE}(\hat{\beta})_j^{(r)}$ is the estimated standard error of $\hat{\beta}_j^{(r)}$.

After the estimation performance of each model is assessed, we use the out-of-sample validation set $\mathcal{T}_r$ for each $r = 1, \ldots, R$ to compare their predictive performance. The out-of-sample validation is based on the prediction RMSE (pRMSE) and the Poisson deviance statistic (DEV), defined as follows:

$$\begin{aligned}
\mathrm{Avg\_pRMSE}^{(k)} &= \frac{1}{R}\sum_{r=1}^{R} \mathrm{pRMSE}^{(r;k)}, \\
\mathrm{pRMSE}^{(r;k)} &= \sqrt{\frac{1}{|\mathcal{T}_r|}\sum_{i\in\mathcal{T}_r} (N_i^{(r)} - \hat{N}_i^{(r;k)})^2}, \\
\mathrm{Avg\_DEV}^{(k)} &= \frac{1}{R}\sum_{r=1}^{R} \mathrm{DEV}^{(r;k)} \\
\mathrm{DEV}^{(r;k)} &= \frac{2}{|\mathcal{T}_r|}\sum_{i\in\mathcal{T}_r} \left[N_i^{(r)} \log(N_i^{(r)}/\hat{N}_i^{(r;k)}) + (N_i^{(r)} - \hat{N}_i^{(r;k)})\right],
\end{aligned} \tag{4.1}$$

where $|\mathcal{T}_r|$ is the number of observations in $\mathcal{T}_r$ and the predicted value $\hat{N}_i^{(r;k)}$ is generated using model k with the r^{th} simulation sample.

Model	Parameter estimation equation	Prediction
Naive	$\sum_{i\in\mathcal{S}_0} U(\boldsymbol{\beta};\mathbf{x}_i,n_i)=0$	$\hat{N}_i=\exp(\mathbf{x}_{i1}\hat{\boldsymbol{\beta}}_1+\mathbf{x}_{i2}\hat{\boldsymbol{\beta}}_2)$
Traditional	$\sum_{i\in\mathcal{S}} U(\boldsymbol{\beta}_1;\mathbf{x}_{i1},n_i)=0$	$\hat{N}_i=\exp(\mathbf{x}_{i1}\hat{\boldsymbol{\beta}}_1)$
Full	$\sum_{i\in\mathcal{S}^*} U(\boldsymbol{\beta};\mathbf{x}_i,n_i)=0$	$\hat{N}_i=\exp(\mathbf{x}_{i1}\hat{\boldsymbol{\beta}}_1+\mathbf{x}_{i2}\hat{\boldsymbol{\beta}}_2)$
Boosting	$\sum_{i\in\mathcal{S}} U(\boldsymbol{\beta}_1;\mathbf{x}_{i1},n_i)=0$	$\hat{N}_i=\hat{\eta}_i\exp(\mathbf{x}_{i2}\hat{\boldsymbol{\beta}}_2)$
	$\sum_{i=1}^{M_0}\{n_i-\hat{\eta}_i\exp(\mathbf{x}'_{i2}\boldsymbol{\beta}_2)\}\mathbf{x}_{i2}=0$	where $\hat{\eta}_i=\exp(\mathbf{x}_{i1}\hat{\boldsymbol{\beta}}_1)$
	where $\hat{\eta}_i=\exp(\mathbf{x}_{i1}\hat{\boldsymbol{\beta}}_1)$: i in $\mathcal{S}_0$	and i in $\mathcal{T}$
Proposed	$\sum_{i\in\mathcal{S}}\delta_i\hat{\omega}_i(\boldsymbol{\phi})U(\boldsymbol{\beta};\mathbf{x}_i,n_i)=0$	$\hat{N}_i=\exp(\mathbf{x}_{i1}\hat{\boldsymbol{\beta}}_1+\mathbf{x}_{i2}\hat{\boldsymbol{\beta}}_2)$
	where $\hat{\omega}_i(\boldsymbol{\phi})$ - propensity weights	
	and $\delta_i=\mathbb{I}(i\in\mathcal{S}_0)$	

Table 4.1: Summary of models

4.4 Results

As a summary, Table 4.2 shows estimation results of the regression coefficients under different model specifications and sampling schemes. Here **N**, **T**, **B**, **F**, and **P** refer to Naive, Traditional, Boosting, Full, and Proposed models, respectively.

It is clearly observed that if the sampling mechanism of $\mathcal{S}_0$ is purely random, then the use of the naive model is innocuous in terms of estimation performance. While the full model shows the best performance in the estimation performance followed by the proposed model, the boosting model (and correspondingly the traditional model) suffers from the biases in $\hat{\beta}_0$ and $\hat{\beta}_T$. But the naive model is less efficient compared to the full and proposed models, by having a larger RMSE of the estimated regression coefficients.

When the sampling mechanism is age selection, it is shown that the naive model is still unbiased and less efficient compared to the full and proposed models. Beyond the changes of bias of the naive model, one can hardly observe a noticeable difference in the performance of models within both random and age selection results. On the other hand, if the sampling mechanism of $\mathcal{S}_0$ is prone to adverse selection, then the differences in the estimation performance are more dramatic. Unlike the random sampling case, naive model severely suffers from lack of fit and biases in the estimates, since $\mathcal{S}_0$ is not a representative sample of the finite population anymore.

Furthermore, it is observed that only the full and proposed models provide acceptable ranges of estimates as the CI values are near 0.9 more often. Note that the efficiency gain in the estimation of $\boldsymbol{\beta}_2 = \beta_T$ via the proposed model is no better than the naive model unlike in the cases of $\boldsymbol{\beta}_1 = (\beta_0, \beta_{A1}, \beta_{A2}, \beta_G)$. It is reasonable since there is no information to borrow from $\mathcal{S}_1$ to better estimate $\boldsymbol{\beta}_2$ in the proposed model. And also it is obvious that the proposed model performs better than the boosting model.

	Bias					RMSE					CI				
	N	**T**	**B**	**F**	**P**	**N**	**T**	**B**	**F**	**P**	**N**	**T**	**B**	**F**	**P**
	Random selection														
β_0	0.009	-0.121	-0.121	0.005	0.005	0.208	0.140	0.140	0.071	0.083	0.915	0.474	0.474	0.890	0.892
β_{A1}	-0.019	-0.016	-0.016	-0.019	-0.018	0.944	0.316	0.316	0.313	0.359	0.897	0.897	0.897	0.911	0.904
β_{A2}	0.017	0.017	0.017	0.019	0.018	0.962	0.320	0.320	0.316	0.361	0.891	0.896	0.896	0.907	0.905
β_G	-0.001	-0.001	-0.001	-0.001	-0.001	0.056	0.020	0.020	0.020	0.023	0.925	0.896	0.896	0.905	0.895
β_T	0.000		0.049	0.000	0.000	0.030		0.055	0.010	0.030	0.899		0.384	0.899	0.898
	Age selection														
α_0	0.013	-0.121	-0.121	0.005	0.004	0.194	0.140	0.140	0.071	0.082	0.898	0.474	0.474	0.890	0.890
α_1	-0.040	-0.016	-0.016	-0.019	-0.014	0.929	0.316	0.316	0.313	0.359	0.896	0.897	0.897	0.911	0.900
β_{A2}	0.034	0.017	0.017	0.019	0.014	0.998	0.320	0.320	0.316	0.367	0.909	0.896	0.896	0.907	0.910
β_G	-0.001	-0.001	-0.001	-0.001	-0.001	0.060	0.020	0.020	0.020	0.024	0.894	0.896	0.896	0.905	0.898
β_T	-0.001		0.048	0.000	-0.001	0.028		0.053	0.010	0.030	0.902		0.350	0.899	0.907
	Adverse selection														
β_0	1.515	-0.121	-0.121	0.005	-0.008	1.581	0.140	0.140	0.071	0.097	0.027	0.474	0.474	0.890	0.856
β_{A1}	-0.271	-0.016	-0.016	-0.019	-0.018	2.054	0.316	0.316	0.313	0.383	0.902	0.897	0.897	0.911	0.870
β_{A2}	0.257	0.017	0.017	0.019	0.017	2.069	0.320	0.320	0.316	0.381	0.898	0.896	0.896	0.907	0.870
β_G	-0.002	-0.001	-0.001	-0.001	0.000	0.125	0.020	0.020	0.020	0.025	0.911	0.896	0.896	0.905	0.848
β_T	0.005		0.351	0.000	0.031	0.059		0.351	0.010	0.073	0.910		0.000	0.899	0.820

Table 4.2: Estimation performance with the simulated data

Then, the predictive performance of these models can be discussed based on the out-of-sample validation using prediction RMSE (pRMSE) and the Poisson deviance statistic (DEV). Table 4.3 showcases the out-of-sample validation performance of the aforementioned models. As in Table 4.3, the uses of naive and boosting models are more vulnerable when the availability of telematics information is prone to adverse selection. It depicts that the proposed model is as good as the full model. It is also shown that the predictive performance of the traditional model is usually inferior to the other models since it completely ignores the impacts of the available telematics information. It highlights an advantage of using telematics features in modeling claim counts.

	Naive	Traditional	Boosting	Full	Proposed
Random selection					
Avg_pRMSE	0.3435	0.3493	0.3437	0.3434	0.3435
Avg_DEV	0.4965	0.5255	0.4977	0.4960	0.4961
Age selection					
Avg_pRMSE	0.3435	0.3493	0.3437	0.3434	0.3435
Avg_DEV	0.4965	0.5255	0.4977	0.4960	0.4962
Adverse selection					
Avg_pRMSE	0.3588	0.3493	0.3466	0.3434	0.3436
Avg_DEV	0.6586	0.5255	0.5106	0.4960	0.4967

Table 4.3: Out-of-sample validation performance with the simulated data

Thus the above results provide these insights from the results by comparing the proposed model with the benchmark models in both in-sample estimation and out-of-sample prediction performance. If the drivers selected UBI products randomly, then one can analyze $\mathcal{S}_0$ for ratemaking purposes. Otherwise, the use of the proposed model is beneficial in recognizing the impacts of risk factors on the claim experience.

Chapter 5

Data analysis

We evaluated and compared the proposed method with four other methods using a hypothetical finite population where all observations have the same distribution. In this chapter, we use a synthetic dataset to evaluate and compare the same set of methods by drawing bootstrap samples to assure that observations have the same empirical distribution. We obey a similar procedure as in Chapter 4 and proffer the description within this chapter.

5.1 Data description

To assess the validity and applicability of the proposed method, we use a synthetic dataset, which is emulated using an actual dataset, from the study of So et al. (2021) with 11 traditional features, 39 telematics features, and the response variable. While the available features in the dataset are already in summarized forms compared to the raw data directly obtained from the telematics device, it is still high-dimensional. For example, one of the "traditional" features is `Region`, which is a categorical variable with 55 categories. Therefore, one needs a statistical method that can deal with such high dimensionality.

The proposed data integration approach, however, is based on estimating equations and GLMs so it lacks the ability to handle high-dimensionality on its own, unlike neural network models or tree-based models. Also the interpretation of parameter estimates would become difficult due to the existence of variables that are alike. In this regard, some of the available features were pre-processed in Jeong (2022) by using the territorial embedding and principle component analysis (PCA) for telematics variables where the approach is discussed in here concisely.

First, a feed-forward neural network (FNN) was used with three hidden layers, which depends on traditional features and the response variable for territorial embedding. It eventuated a score to represent all 55 categories, which was named as `TerritoyEmb`. Then, four clusters of telematics variables were recognized based on definitions and PCA was applied to each cluster to recognize possible representative variables from each cluster. As a result `Pct.drive.rush.am`, `Pct.drive.rush.pm`, `Accel.06miles`, `Brake.06miles`, `Left.turns` and `Right.turns` were selected to represent three clusters out of four (i.e. two features for each cluster) while all the variables in the remaining cluster were used without representatives.

After the data pre-processing, the study retained the following variables which are used in our analysis and described in Table 5.1. Hence this is a dataset with reduced dimensions. For more details about the data pre-processing, see Jeong (2022).

Type	Variable	Description
Traditional	`Duration`	Duration of the insurance coverage of a given policy, in days
	`Insured.age`	Age of insured driver, in years
	`Insured.sex`	Sex of insured driver (Male/Female)
	`Car.age`	Age of vehicle, in years
	`Marital`	Marital status (Single/Married)
	`Car.use`	Use of vehicle: Private, Commute, Farmer, Commercial
	`Credit.score`	Credit score of insured driver
	`Region`	Type of region where driver lives: rural, urban
	`Annual.miles.drive`	Annual miles expected to be driven declared by driver
	`Years.noclaims`	Number of years without any claims
	`TerritoryEmb`	Embedded value from the territorial location of vehicle
Telematics	`Annual.pct.driven`	Annualized percentage of time on the road
	`Total.miles.driven`	Total distance driven in miles
	`Pct.drive.xxx`	Percent of driving day xxx of the week: mon/tue/. . . /sun
	`Pct.drive.rush.am`	Percent of driving during am rush hours
	`Pct.drive.rush.pm`	Percent of driving during pm rush hours
	`Avgdays.week`	Mean number of days used per week
	`Accel.06miles`	Number of sudden acceleration 6mph/s per 1000miles
	`Brake.06miles`	Number of sudden brakes 6mph/s per 1000miles
	`Acbr.others`	Total number of sudden acceleration and brakes 8/9/. . . /14 mph/s per 1000miles
	`Left.turns`	Number of left turn per 1000miles with intensity greater than equal to 8
	`Right.turns`	Number of right turn per 1000miles greater than equal to 8
Response	`NB_Claim`	Number of observed claims

Table 5.1: Variable names and descriptions of the pre-processed dataset

5.2 Estimation and Evaluation

The results from the data analysis are summarized appropriately to depict the fact that the proposed model is an improved version of the naive model for a dataset with both traditional and telematics features, but a relatively small number of observations, which incorporates the information in the traditional dataset.

5.2.1 Estimation

Here we treat the pre-processed dataset as the finite population and use bootstrap samples for the analysis to ensure each observation has the same empirical distribution as the finite population. Thus, we take a bootstrap sample, $\mathcal{T}$, of size 100,000 for out-of-sample validation at random. Another bootstrap sample of size 100,000 is taken as $\mathcal{S}_0$ subjected to sampling probabilities that are listed in Chapter 4. After that, a bootstrap sample of size 800,000 is taken as $\mathcal{S}_B$ subjected to the complement of respective sampling probabilities of $\mathcal{S}_0$. Finally, $\mathcal{S}_1$ is taken by eliminating telematics features from $\mathcal{S}_B$.

For the convenience of calculations, we use sampling schemes with a slight change as follows:

- **Random selection**: The data points assigned to $\mathcal{S}_0$ are chosen at random.
- **Age selection**: Each data point assigned to $\mathcal{S}_0$ is chosen with the sampling probability proportional to $1/(1+\exp(0.03\texttt{Insured.age}_i))$.
- **Adverse selection**: Each data point assigned to $\mathcal{S}_0$ is chosen with the sampling probability proportional to $1/(1+\exp(\texttt{NB\_Claim}_i))$.

Likewise, we repeat the process of fitting and testing these five models as in Chapter 4 for $R = 500$ times to compare the estimation and predictive performance under each sampling method. Note that all data points in $\mathcal{S}_0 \cup \mathcal{S}_B$ are used to estimate the coefficients in the full model.

5.2.2 Evaluation

To assess the in-sample estimation performance, we compare the estimated regression coefficients from each method and sampling scheme with the estimated regression coefficients obtained from the finite population as in Chapter 4. More specifically, bias, root mean-squared error (RMSE), and 90% confidence interval coverage (CI) of the regression coefficients are defined as follows:

$$\text{Bias}_j = \frac{1}{R}\sum_{r=1}^{R}(\tilde{\beta} - \hat{\beta}_j^{(r)}),$$

$$\text{RMSE}_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\tilde{\beta} - \hat{\beta}_j^{(r)})^2},$$

$$\text{CI}_j = \frac{1}{R}\sum_{r=1}^{R}\mathbb{1}_{\{|\tilde{\beta}_j - \hat{\beta}_j^{(r)}| < 1.645\cdot\text{SE}(\hat{\beta})_j^{(r)}\}},$$

where $\tilde{\beta}_j$ and $\hat{\beta}_j^{(r)}$ are the estimates of β_j using the finite population and the r^{th} bootstrap sample, respectively. $\text{SE}(\hat{\beta})_j^{(r)}$ is the estimated standard error of $\hat{\beta}_j^{(r)}$. Note that we prefer a method with biases closer to 0, smaller RMSEs, and/or CIs closer to the theoretical benchmark, 90%.

On top of the estimation performance, out-of-sample validation performance are assessed using

$$\begin{aligned}
\text{Avg\_pRMSE}^{(k)} &= \frac{1}{R}\sum_{r=1}^{R} \text{pRMSE}^{(r;k)}, \\
\text{Prop\_pRMSE}^{(k)} &= \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}\left(\text{pRMSE}^{(r;k)} > \text{pRMSE}^{(r;proposed)}\right), \\
\text{Avg\_DEV}^{(k)} &= \frac{1}{R}\sum_{r=1}^{R} \text{DEV}^{(r;k)}, \\
\text{Prop\_DEV}^{(k)} &= \frac{1}{R}\sum_{r=1}^{R} \mathbb{I}\left(\text{DEV}^{(r;k)} > \text{DEV}^{(r;proposed)}\right),
\end{aligned}$$

where $\text{pRMSE}^{(r;k)}$ and $\text{DEV}^{(r;k)}$ are defined in equation 4.1. Based on the above definition, we prefer a model with lower Avg_pRMSE, Prop_pRMSE, Avg_DEV, and/or Prop_DEV.

Finally, the results are visualized using proportional improvements of pRMSE and DEV which are calculated by comparing the proposed models to the naive model. One can use these values to compare and contrast the modeling approaches in three different sampling schemes by favouring methods with larger and positive improvements. Here the proportional improvements in pRMSE or DEV with r^{th} bootstrap sample are defined as:

$$\text{Prop\_Imp\_pRMSE} = 100\left(1 - \frac{\text{pRMSE}^{(r;proposed)}}{\text{pRMSE}^{(r;naive)}}\right),$$

$$\text{Prop\_Imp\_DEV} = 100\left(1 - \frac{\text{DEV}^{(r;proposed)}}{\text{DEV}^{(r;naive)}}\right).$$

5.3 Results

With the intuition of discussing the in-sample estimation performance, Tables A.1, A.2 and A.3 in appendix A is used which shows the estimation results of the regression coefficients under different model specifications and sampling schemes of the bootstrap samples from the pre-processed synthetic data. These values are also visualized in Figures 5.1, 5.2, and 5.3, where a model with biases closer to 0, smaller RMSEs, and/or CIs closer to 90% is given a higher rank for each covariate using an ordinal scale of 1, 2, 3 and 4. Note that the estimated coefficients from the traditional model were omitted as they are only available for the traditional features and are identical to those from the boosting model.

Implications of the in-sample estimation results, under each sampling mechanism, with the actual data are as follows:

- In the case of random selection, only the boosting model suffers from the biases of the regression coefficients and there are no big differences in the estimation performance between the naive and proposed models. It implies that as long as the sampling mechanism of $\mathcal{S}_0$ (a small dataset with both traditional and telematics features) from the finite population is purely random, one can ignore $\mathcal{S}_1$ (a large dataset only with traditional features) and analyze $\mathcal{S}_0$ for ratemaking purpose.

- In the case of age selection, the naive model is more biased in the estimation of the traditional covariates (especially the intercept term), compared to the proposed model. It implies that if the observability of the telematics features depends on the traditional features, then the proposed approach might be helpful to better understand the underlying impacts of the covariates on the claim experience.

- Lastly, in the case of adverse selection, the proposed model is no more unbiased but the naive model is still more biased in the estimation of the regression coefficients. Thus, if the accessibility of the telematics features is affected by adverse selection, then it is recommended to integrate two data sources to handle the missingness of the telematics features.

Such differences are also visualized in Figures 5.1, 5.2, and 5.3. It is consistently observed that in the case of either age or adverse selection, the proposed model is the second-best, following the full model that is unattainable in practice. Furthermore, it is visible that the proposed model is performing better than the boosting model despite the sampling mechanism. Though CI values do not suggest a clear winner in the comparison of the ranges of estimates, the boosting model tends to have more out of range estimates than other models.

To discuss the out-of-sample validation performance, we use Table 5.2 which shows that the proposed model is the only comparable model to the full model in terms of pRMSE and DEV on average, especially when the observability of telematics features is prone to adverse selection. It is also observed that the naive, traditional, and boosting models do not outperform the proposed model in most of the bootstrap samples as shown in the values of Prop_pRMSE and Prop_DEV, regardless of the selection scheme. Therefore, the proposed approach is a reasonable alternative in the absence of a finite population with both traditional and telematics features or the sample with both types of features is not representative. And also it depicts that the out-of-sample performance of the traditional model does not change with sampling schemes. It is a sign of the importance of information that telematics data care about the risk of a driver.

	Naive	Traditional	Boosting	Full	Proposed
Random selection					
Avg_pRMSE	0.2117	0.2162	0.2119	0.2116	0.2116
Prop_pRMSE	0.666	1.000	1.000	0.302	-
Avg_DEV	23.8921	26.7414	24.0736	23.8700	23.8811
Prop_DEV	0.914	1.000	1.000	0.048	-
Age selection					
Avg_pRMSE	0.2117	0.2162	0.2120	0.2116	0.2116
Prop_pRMSE	0.558	1.000	1.000	0.268	-
Avg_DEV	23.9139	26.7414	24.0748	23.8701	23.8828
Prop_DEV	0.982	1.000	1.000	0.040	-
Adverse selection					
Avg_pRMSE	0.2138	0.2162	0.2137	0.2116	0.2121
Prop_pRMSE	0.804	1.000	0.878	0.176	-
Avg_DEV	25.3604	26.7415	25.3503	23.8702	24.2426
Prop_DEV	0.890	1.000	0.850	0.012	-

Table 5.2: Out-of-sample validation performance with bootstrapping from the actual data

Furthermore, Figure 5.4 showcases the distributions of proportional improvements of pRMSE and DEV. It is shown that while the proportional improvements of pRMSE are symmetric and centered at 0 with the random or age selection, the proportional improvements of DEV are symmetric and almost centered at 0. But they are clearly positive with the adverse selection which also supports the usefulness of the proposed method upon the existence of adverse selection in the provision of telematics features.

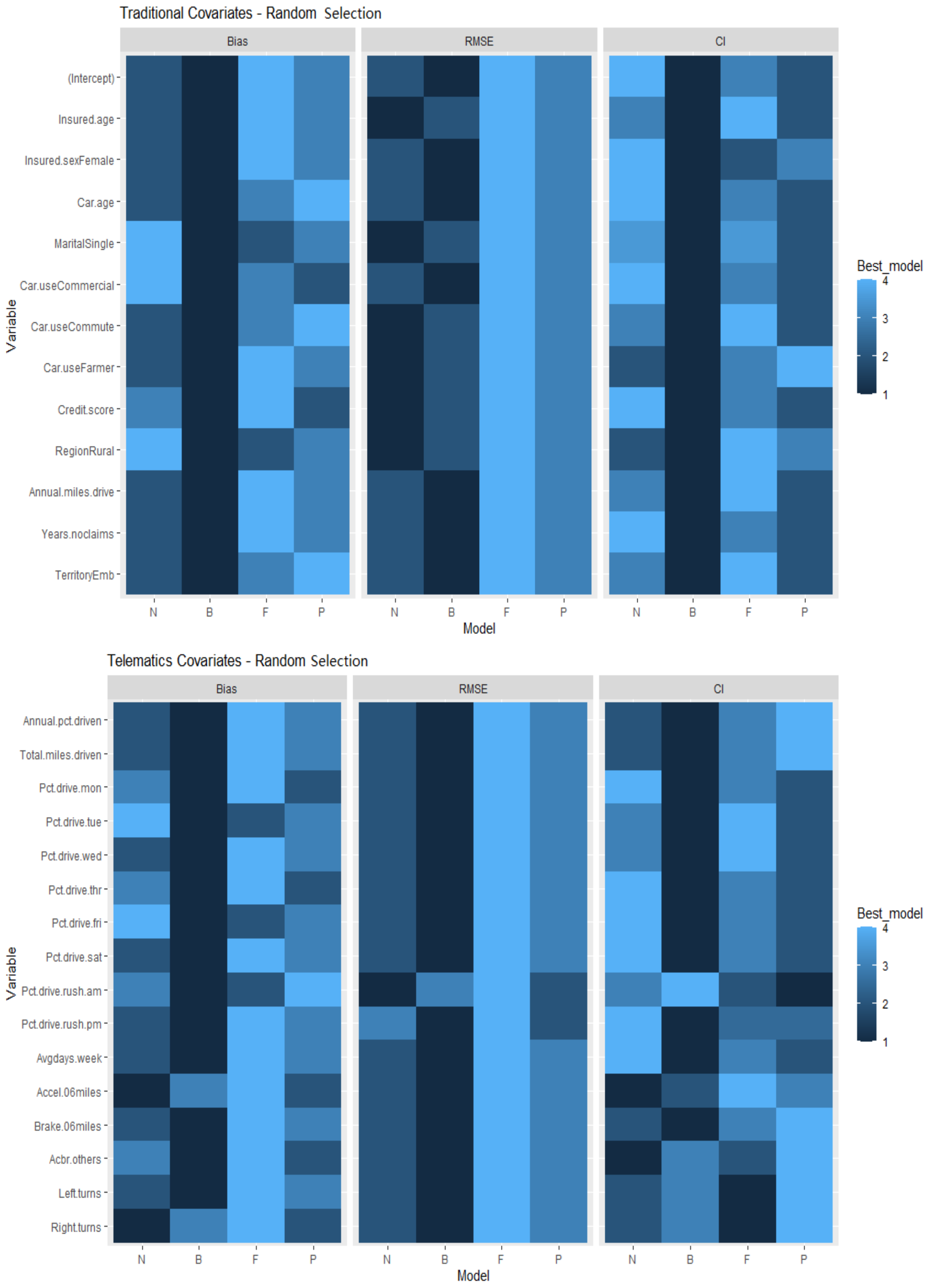


Figure 5.1: Ranks of the models in estimation performance with random selection

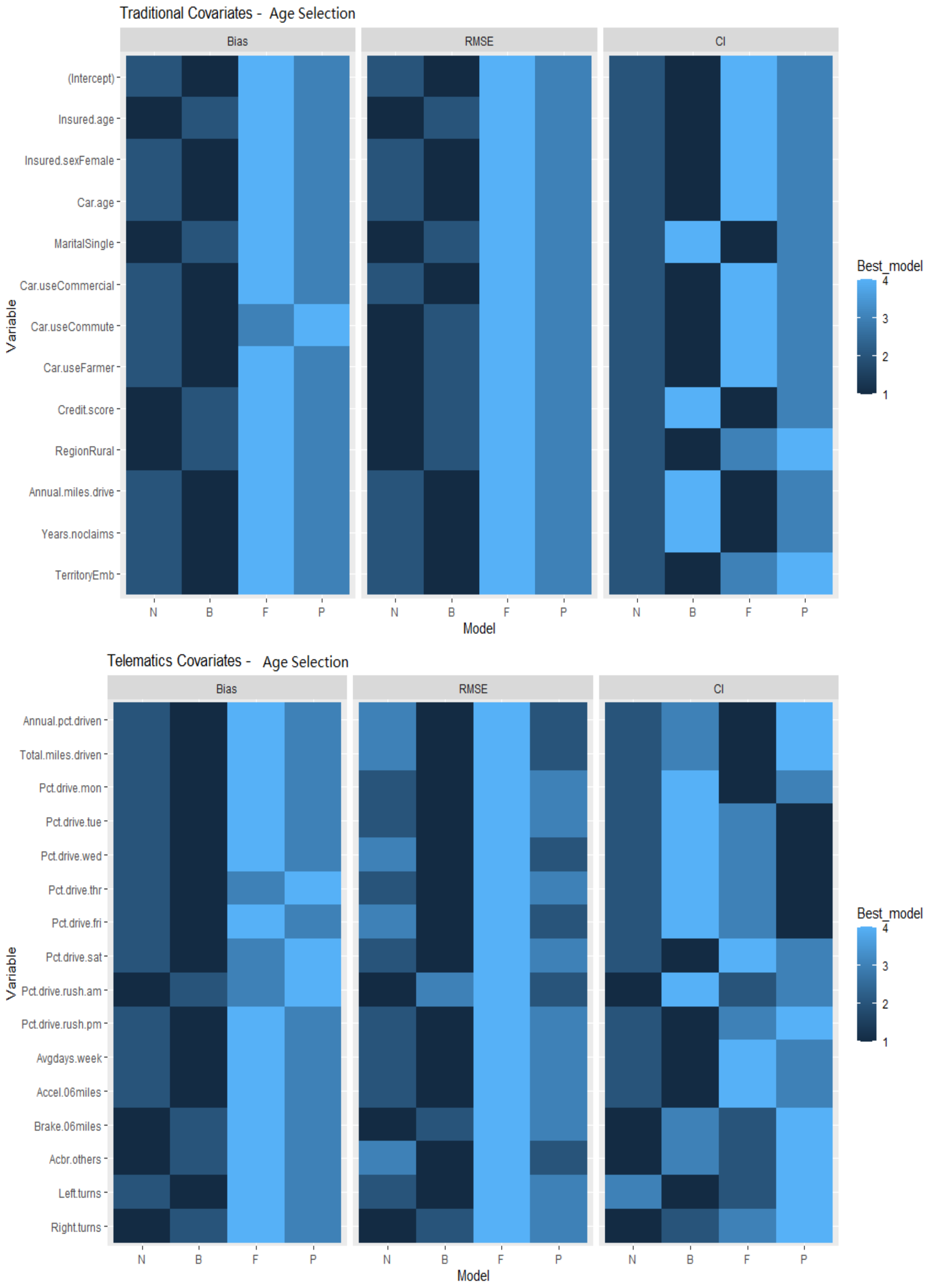


Figure 5.2: Ranks of the models in estimation performance with age selection

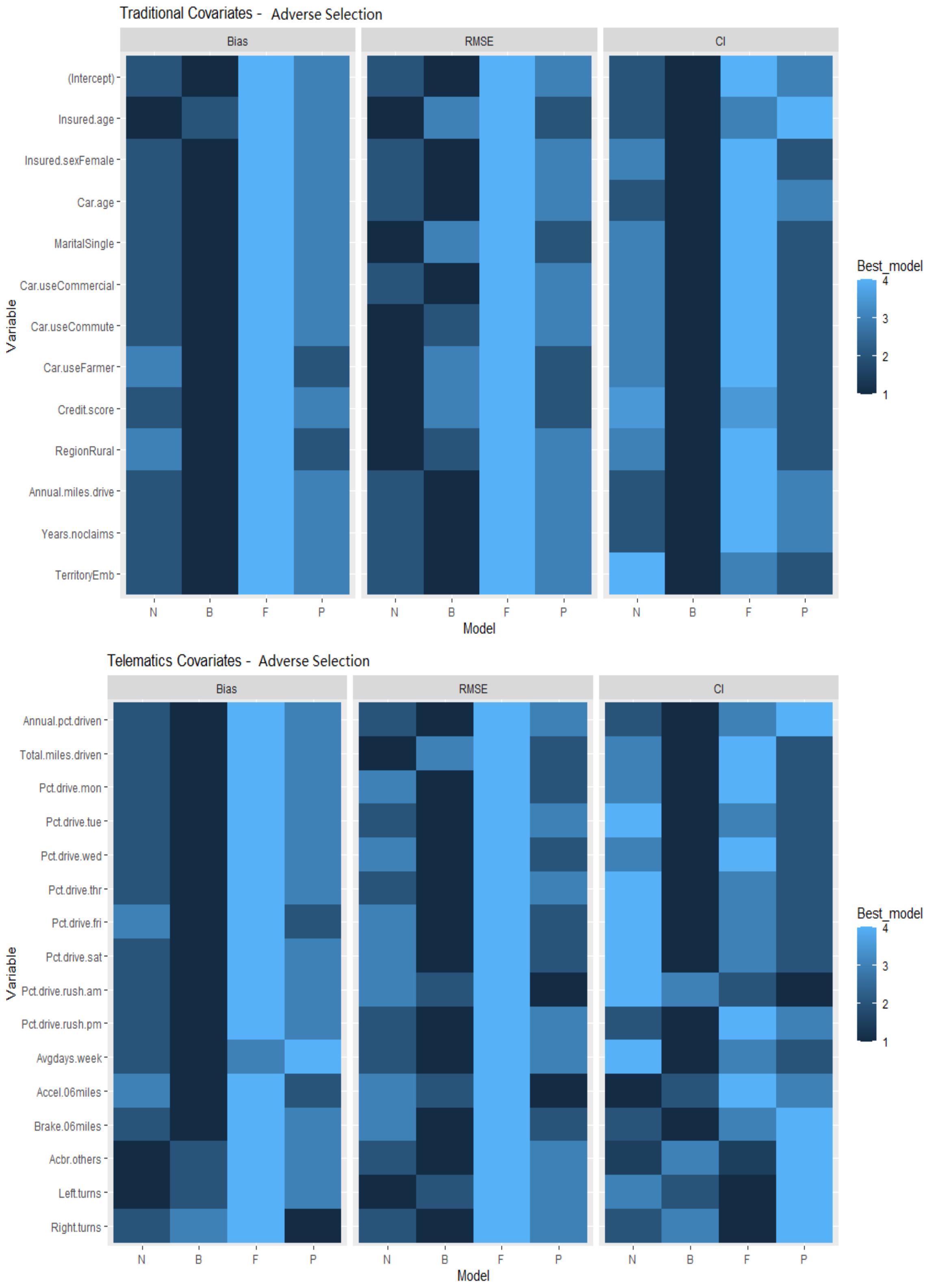
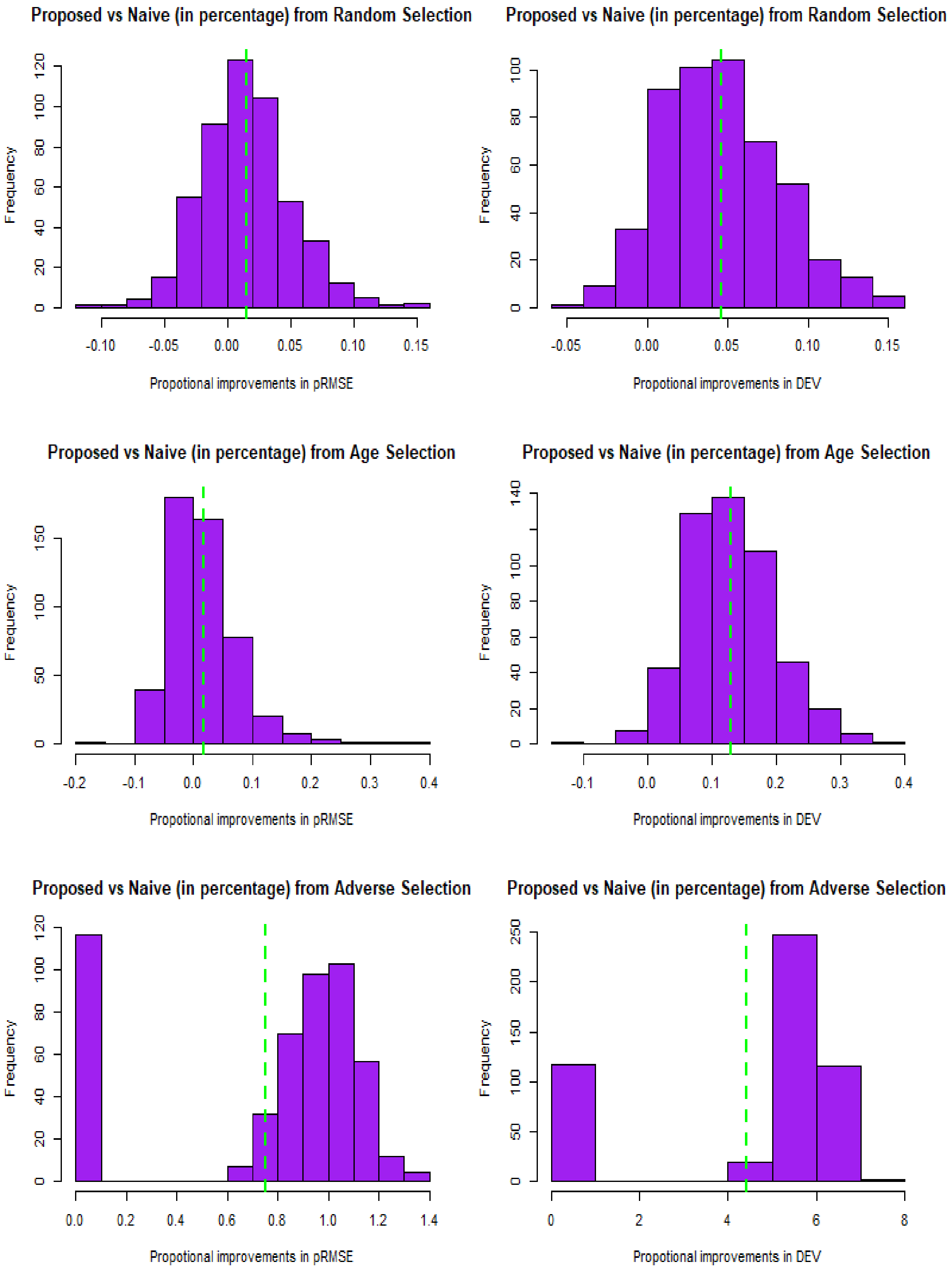


Figure 5.3: Ranks of the models in estimation performance with adverse selection



Proposed vs Naive (in percentage) from Random Selection

Proposed vs Naive (in percentage) from Random Selection

Proposed vs Naive (in percentage) from Age Selection

Proposed vs Naive (in percentage) from Age Selection

Proposed vs Naive (in percentage) from Adverse Selection

Proposed vs Naive (in percentage) from Adverse Selection

Figure 5.4: Proportional improvements in pRMSEs and DEVs

Chapter 6

Conclusions

Insurers have been using models to set vehicle insurance rates which enable them to evaluate the risk profile of drivers. Though using datasets with traditionally collected features is a common practice, using telematics features increases the efficiency of these models. Due to privacy and resistance to the new technology, the availability of telematics features in automobile insurance datasets is often limited and the insurers need to deal with two types of claim datasets where one dataset contains only traditional features with more observations and the other one contains both traditional and telematics features but with fewer of observations.

To handle such an issue, we proposed a data integration approach to effectively consider both telematics and traditional data, where the availability of telematics features for a policyholder is not completely at random. This method is independent of the distribution of the response variable. It turns out that the proposed method could achieve satisfactory performance both on in-sample estimation and out-of-sample prediction, compared to the existing benchmarks for automobile insurance ratemaking practices. Especially, the proposed method has performed well when the availability of telematics data is subjected to adverse selection which occurs due to the tendency of subscribing to UBI products by less risky drivers. Moreover, it can also be useful when there is evidence that being a subscriber depends on their age.

In conclusion, though the proposed method has preformed to the expectation, it can possibly be extended in two-fold. Firstly, the proposed data integration approach relies on the assumption in (3.4) so it might not work well if the basis function of $E\{U(\boldsymbol{\beta}; \mathbf{x}_i, n_i) \mid \mathbf{x}_{i1}, n_i\}$ is not correctly specified. To tackle such an issue, one can implement a doubly robust calibration approach that only requires either the basis function of the outcome variable or the propensity score to be correctly specified. Secondly, the proposed approach can be extended to data integration for mixed effects models where a policyholder is observed over a period of time, so that the proposed framework also considers random effects for experience ratemaking, as well as the fixed effects.

Bibliography

A. Agresti. *Categorical data analysis.* John Wiley & Sons, 2003.

M. Ayuso, M. Guillén, and A. M. Pérez-Marín. Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis and Prevention*, 73:125–131, 2014.

M. Ayuso, M. Guillen, and A. M. Pérez-Marín. Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4 (2):10, 2016.

M. Ayuso, M. Guillen, and J. P. Nielsen. Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. *Transportation*, 46(3):735–752, 2019.

P. Baecke and L. Bocca. The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, 98:69–79, 2017.

L. Bermúdez and D. Karlis. Bayesian multivariate Poisson models for insurance ratemaking. *Insurance: Mathematics and Economics*, 48(2):226–236, 2011.

J. W. Bolderdijk, J. Knockaert, E. Steg, and E. T. Verhoef. Effects of pay-as-you-drive vehicle insurance on young drivers' speed choice: Results of a dutch field experiment. *Accident Analysis & Prevention*, 43(3):1181–1186, 2011.

J.-P. Boucher and M. Denuit. Fixed versus random effects in Poisson regression models for claim counts: A case study with motor insurance. *ASTIN Bulletin: The Journal of the IAA*, 36(1):285–301, 2006.

J.-P. Boucher, M. Denuit, and M. Guillén. Risk classification for claim counts: a comparative analysis of various zeroinflated mixed Poisson and hurdle models. *North American Actuarial Journal*, 11(4):110–131, 2007.

J. Buxbaum. Mileage-based user fee demonstration project: Potential public policy implications of pay-as-you-drive leasing and insurance products. Technical report, 2006. URL `https://trid.trb.org/view/1401951`.

G. Castignani, T. Derrmann, R. Frank, and T. Engel. Driver behavior profiling using smartphones: A low-cost platform for driver monitoring. *IEEE Intelligent Transportation Systems Magazine*, 7(1):91–102, 2015.

N. Chapados, Y. Bengio, P. Vincent, J. Ghosn, C. Dugas, I. Takeuchi, and L. Meng. Estimating car insurance premia: A case study in high-dimensional data inference. *Advances in Neural Information Processing Systems*, 14, 2001.

A. Cohen and P. Siegelman. Testing for adverse selection in insurance markets. *Journal of Risk and Insurance*, 77(1):39–84, 2010.

M. Denuit and S. Lang. Non-life rate-making with Bayesian GAMs. *Insurance: Mathematics and Economics*, 35(3):627–647, 2004.

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

M. Denuit, M. Guillen, and J. Trufin. Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. *Annals of Actuarial Science*, 13(2):378–399, 2019.

J. C. Deville and C. E. Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992.

S. Duri, M. Gruteser, X. Liu, P. Moskowitz, R. Perez, M. Singh, and J.-M. Tang. Framework for security and privacy in automotive telematics. In *Proceedings of the 2nd International Workshop on Mobile Commerce*, pages 25–32, 2002.

S. Duri, J. Elliott, M. Gruteser, X. Liu, P. Moskowitz, R. Perez, M. Singh, and J.-M. Tang. Data protection and data sharing in telematics. *Mobile Networks and Applications*, 9(6): 693–701, 2004.

F. Duval, J.-P. Boucher, and M. Pigeon. Enhancing claim classification with feature extraction from anomaly-detection-derived routine and peculiarity profiles. *arXiv preprint arXiv:2209.11763*, 2022.

M. Eling and M. Kraft. The impact of telematics on the insurability of risks. *Journal of Risk Finance*, 21(2):77–109, 2020.

A. Ferrario, A. Noll, and M. V. Wüthrich. Insights from inside neural networks. *Available at SSRN 3226852*, 2020.

M. F. M. Fuzi, A. A. Jemain, and N. Ismail. Bayesian quantile regression model for claim count data. *Insurance: Mathematics and Economics*, 66:124–137, 2016.

G. Gao, S. Meng, and M. V. Wüthrich. Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, 2019(2):143–162, 2019.

G. Gao, H. Wang, and M. V. Wüthrich. Boosting Poisson regression models with telematics car driving data. *Machine Learning*, 111(1):243–272, 2022.

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

L. Guelman. Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, 39(3):3659–3667, 2012.

M. Guillen, J. P. Nielsen, M. Ayuso, and A. M. Pérez-Marín. The use of telematics devices to improve automobile insurance rates. *Risk analysis*, 39(3):662–672, 2019.

M. Guillen, J. P. Nielsen, and A. M. Pérez-Marín. Near-miss telematics in motor insurance. *Journal of Risk and Insurance*, 88(3):569–589, 2021.

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

A. Heras, I. Moreno, and J. L. Vilar-Zanón. An application of two-stage quantile regression to insurance ratemaking. *Scandinavian Actuarial Journal*, 2018(9):753–769, 2018.

S. Husnjak, D. Peraković, I. Forenbacher, and M. Mumdziev. Telematics system in usage based motor insurance. *Procedia Engineering*, 100:816–825, 2015.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:243–263, 2014.

K. Jaisingh, K. El-Khatib, and R. Akalu. Paving the way for intelligent transport systems (its): Privacy implications of vehicle infotainment and telematics systems. In *Proceedings of the 6th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications*, pages 25–31, 2016.

H. Jeong. Dimension reduction techniques for summarized telematics data. *Journal of Risk Management*, 33(4), 2022.

H. Jeong and E. A. Valdez. Ratemaking application of Bayesian LASSO with conjugate hyperprior. *Available at SSRN 3251623*, 2018.

H. Jeong, E. A. Valdez, J. Y. Ahn, and S. Park. Generalized linear mixed models for dependent compound risk models. *Variance*, 14(1), 2021.

N. Klein, M. Denuit, S. Lang, and T. Kneib. Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. *Insurance: Mathematics and Economics*, 55:225–249, 2014.

F. Klinker. Generalized linear mixed models for ratemaking: a means of introducing credibility into a generalized linear model setting. In *Casualty Actuarial Society E-Forum, Winter 2011 Volume 2*, 2010.

T. Litman. Distance-based vehicle insurance. *Canadian Electronic Library*, 2011. doi: 20.500.12592/z6dqns. URL `https://policycommons.net/artifacts/1189025/distance-based-vehicle-insurance/1742147/`.

Y.-L. Ma, X. Zhu, X. Hu, and Y.-C. Chiu. The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice*, 113:243–258, 2018.

G. Meyers and I. V. Hoyweghen. 'Happy failures': Experimentation with behaviour-based personalisation in car insurance. *Big Data & Society*, 7(1):2053951720914650, 2020.

N. Milanović, M. Milosavljević, S. Benković, D. Starčević, and Ž. Spasenić. An acceptance approach for novel technologies in car insurance. *Sustainability*, 12(24):10331, 2020.

J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz. Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks*, 7(2):70, 2019.

D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

K. Sakthivel and C. Rajitha. A comparative study of zero-inflated, hurdle models with artificial neural network in claim count modeling. *International Journal of Statistics and Systems*, 12(2):265–276, 2017.

K. A. Smith, R. J. Willis, and M. Brooks. An analysis of customer retention and insurance claim patterns using data mining: A case study. *Journal of the operational research society*, 51(5):532–541, 2000.

B. So, J.-P. Boucher, and E. A. Valdez. Synthetic dataset generation of driver telematics. *Risks*, 9(4):58, 2021.

D. I. Tselentis, G. Yannis, and E. I. Vlahogianni. Innovative motor insurance schemes: A review of current practices and emerging challenges. *Accident Analysis and Prevention*, 98:139–148, 2017.

G. Tuna and K. Cengiz. Telematics and mobile internet: current situation and 5G networks. *Principles and Applications of Narrowband Internet of Things (NBIoT)*, pages 373–396, 2021.

R. Verbelen, K. Antonio, and G. Claeskens. Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1275–1304, 2018.

F. Vermet. Statistical learning methods. *Big Data for Insurance Companies*, 1:43–82, 2018.

H. Wang and J. K. Kim. Information projection approach to propensity score estimation for handling selection bias under missing at random. *arXiv e-prints*, pages arXiv–2104, 2021a.

H. Wang and J.-K. Kim. Propensity score estimation using density ratio model under item nonresponse. *arXiv preprint arXiv:2104.13469*, 2021b.

Z. Wang, J.-K. Kim, and H. J. Kim. Survey data integration for regression analysis using model calibration. *Survey Methodology*, Forthcoming, 2022.

M. V. Wüthrich. Covariate selection from telematics car driving data. *European Actuarial Journal*, 7(1):89–108, 2017.

S. Yang and J. K. Kim. Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3(2):625–650, 2020.

A. C. Yeo, K. A. Smith, R. J. Willis, and M. Brooks. Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry. *Intelligent Systems in Accounting, Finance & Management*, 10(1):39–50, 2001.

H. Zamani and N. Ismail. Score test for testing zero-inflated Poisson regression against zero-inflated generalized Poisson alternatives. *Journal of Applied Statistics*, 40(9):2056–2068, 2013.

Appendix A

Results

	Bias				RMSE				CI			
	N	**B**	**F**	**P**	**N**	**B**	**F**	**P**	**N**	**B**	**F**	**P**
Random selection												
(Intercept)	-0.0405	5.0136	-0.0024	-0.0282	0.4781	5.0138	0.1548	0.4568	0.926	0.000	0.928	1.000
Insured.age	0.0001	-0.0011	0.0000	0.0000	0.0020	0.0013	0.0006	0.0008	0.852	0.444	0.878	0.992
Insured.sexFemale	0.0012	0.0639	-0.0002	0.0003	0.0309	0.0647	0.0096	0.0123	0.900	0.000	0.926	0.918
Car.age	0.0001	-0.0079	0.0000	0.0000	0.0041	0.0081	0.0014	0.0018	0.900	0.000	0.902	0.960
MaritalSingle	-0.0001	-0.0111	-0.0008	-0.0004	0.0333	0.0158	0.0108	0.0138	0.884	0.754	0.916	0.994
Car.useCommercial	-0.0003	0.1796	0.0006	-0.0008	0.0851	0.1818	0.0272	0.0390	0.902	0.000	0.904	0.916
Car.useCommute	0.0014	0.0340	0.0006	0.0000	0.0382	0.0361	0.0126	0.0184	0.914	0.114	0.912	0.916
Car.useFarmer	-0.0038	-0.0966	0.0007	0.0033	0.2240	0.1230	0.0755	0.0811	0.890	0.618	0.892	0.900
Credit.score	0.0000	-0.0001	0.0000	0.0000	0.0002	0.0001	0.0001	0.0001	0.900	0.588	0.922	1.000
RegionRural	0.0003	-0.0374	-0.0008	-0.0005	0.0454	0.0399	0.0145	0.0202	0.880	0.146	0.900	0.904
Annual.miles.drive	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.928	0.000	0.916	0.972
Years.noclaims	-0.0001	-0.0042	0.0000	0.0000	0.0018	0.0042	0.0006	0.0008	0.874	0.000	0.868	0.980
TerritoryEmb	0.0018	0.0964	0.0008	0.0007	0.0527	0.0982	0.0185	0.0239	0.914	0.000	0.888	0.924
Annual.pct.driven	-0.0016	-0.2028	-0.0004	-0.0014	0.0635	0.2114	0.0212	0.0633	0.932	0.062	0.922	0.920
Total.miles.driven	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.924	0.712	0.912	0.908
Pct.drive.mon	0.0422	-4.8503	0.0013	0.0425	0.5805	4.8601	0.1906	0.5759	0.928	0.000	0.942	1.000
Pct.drive.tue	0.0023	-4.2238	0.0052	0.0024	0.5058	4.2324	0.1770	0.5011	0.950	0.000	0.944	1.000
Pct.drive.wed	0.0582	-3.9782	0.0083	0.0581	0.5863	3.9932	0.1997	0.5848	0.936	0.000	0.934	1.000
Pct.drive.thr	0.0267	-4.2842	0.0020	0.0274	0.5169	4.2923	0.1652	0.5152	0.950	0.000	0.954	1.000
Pct.drive.fri	0.0002	-4.3588	0.0046	-0.0007	0.5889	4.3733	0.1849	0.5844	0.928	0.000	0.930	1.000
Pct.drive.sat	0.0606	-6.5212	0.0119	0.0597	0.7305	6.5303	0.2349	0.7269	0.912	0.000	0.926	1.000
Pct.drive.rush.am	0.0016	0.0384	-0.0031	0.0011	0.2363	0.2321	0.0732	0.2355	0.908	0.904	0.924	0.940
Pct.drive.rush.pm	0.0092	-0.4859	0.0012	0.0088	0.2622	0.5506	0.0886	0.2627	0.896	0.386	0.888	0.912
Avgdays.week	0.0010	-0.0817	-0.0001	0.0009	0.0163	0.0829	0.0052	0.0160	0.928	0.000	0.942	1.000
Accel.06miles	0.0000	0.0000	0.0000	0.0000	0.0003	0.0003	0.0001	0.0003	0.936	0.928	0.918	0.920
Brake.06miles	0.0000	-0.0003	0.0000	0.0000	0.0002	0.0004	0.0001	0.0002	0.930	0.624	0.908	0.904
Acbr.others	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	0.0001	1.000	0.996	0.998	0.912
Left.turns	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.980	0.976	0.986	0.930
Right.turns	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.948	0.936	0.956	0.926

Table A.1: Estimation performance with the bootstrap samples of the actual data in random selection

	Bias				RMSE				CI			
	N	**B**	**F**	**P**	**N**	**B**	**F**	**P**	**N**	**B**	**F**	**P**
Age selection												
(Intercept)	0.7642	5.0119	0.0006	0.0132	0.8879	5.0121	0.1477	0.4860	0.520	0.000	0.922	1.000
Insured.age	-0.0015	-0.0011	0.0000	-0.0001	0.0025	0.0012	0.0006	0.0009	0.794	0.428	0.902	0.990
Insured.sexFemale	-0.0074	0.0640	0.0000	0.0004	0.0301	0.0649	0.0105	0.0129	0.880	0.000	0.902	0.918
Car.age	-0.0058	-0.0080	0.0000	-0.0001	0.0073	0.0081	0.0014	0.0019	0.586	0.000	0.932	0.972
MaritalSingle	0.0189	-0.0103	0.0001	0.0007	0.0352	0.0150	0.0106	0.0131	0.838	0.746	0.916	0.998
Car.useCommercial	-0.0485	0.1778	-0.0002	-0.0018	0.0982	0.1801	0.0282	0.0401	0.820	0.000	0.896	0.940
Car.useCommute	0.0194	0.0338	0.0009	0.0009	0.0423	0.0358	0.0126	0.0178	0.864	0.134	0.902	0.926
Car.useFarmer	0.0835	-0.0978	-0.0009	-0.0030	0.2614	0.1223	0.0738	0.1055	0.838	0.638	0.904	0.916
Credit.score	-0.0001	-0.0001	0.0000	0.0000	0.0002	0.0001	0.0001	0.0001	0.818	0.586	0.922	1.000
RegionRural	-0.0378	-0.0361	0.0006	0.0010	0.0573	0.0386	0.0140	0.0207	0.744	0.152	0.892	0.892
Annual.miles.drive	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.854	0.000	0.910	0.956
Years.noclaims	-0.0001	-0.0042	0.0000	-0.0001	0.0019	0.0042	0.0006	0.0008	0.882	0.000	0.890	0.990
TerritoryEmb	-0.0398	0.0949	-0.0003	-0.0008	0.0665	0.0966	0.0176	0.0244	0.796	0.000	0.924	0.918
Annual.pct.driven	-0.0031	-0.1495	0.0002	0.0004	0.0633	0.1609	0.0210	0.0680	0.934	0.226	0.940	0.924
Total.miles.driven	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.920	0.774	0.930	0.916
Pct.drive.mon	-0.5396	-5.0628	0.0000	0.0040	0.7701	5.0713	0.1833	0.5986	0.830	0.000	0.952	1.000
Pct.drive.tue	-0.3614	-4.2563	-0.0058	-0.0068	0.6034	4.2634	0.1554	0.5145	0.882	0.000	0.962	1.000
Pct.drive.wed	-0.2120	-3.9631	0.0033	0.0316	0.5749	3.9763	0.1943	0.5880	0.922	0.000	0.938	1.000
Pct.drive.thr	-0.3349	-4.3216	0.0091	0.0027	0.5905	4.3301	0.1593	0.5242	0.898	0.000	0.960	1.000
Pct.drive.fri	0.1410	-3.8190	0.0112	-0.0169	0.5937	3.8345	0.1819	0.6254	0.906	0.000	0.944	1.000
Pct.drive.sat	-0.6626	-6.6750	0.0087	-0.0077	0.9514	6.6827	0.2300	0.7456	0.776	0.000	0.934	1.000
Pct.drive.rush.am	-0.1887	-0.0888	-0.0104	-0.0061	0.2802	0.2224	0.0754	0.2231	0.804	0.916	0.898	0.964
Pct.drive.rush.pm	-0.1683	-0.4243	-0.0039	-0.0096	0.2845	0.4831	0.0892	0.2701	0.826	0.426	0.880	0.898
Avgdays.week	-0.0440	-0.1032	-0.0003	-0.0005	0.0468	0.1040	0.0056	0.0172	0.158	0.000	0.930	0.998
Accel.06miles	0.0001	0.0001	0.0000	0.0000	0.0003	0.0003	0.0001	0.0003	0.880	0.876	0.900	0.910
Brake.06miles	0.0001	-0.0001	0.0000	0.0000	0.0002	0.0002	0.0001	0.0002	0.820	0.890	0.924	0.900
Acbr.others	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	0.0001	1.000	0.994	1.000	0.922
Left.turns	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.958	0.948	0.982	0.918
Right.turns	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.586	0.600	0.944	0.878

Table A.2: Estimation performance with the bootstrap samples of the actual data in age selection

	Bias				RMSE				CI			
	N	**B**	**F**	**P**	**N**	**B**	**F**	**P**	**N**	**B**	**F**	**P**
Adverse selection												
(Intercept)	-0.7202	5.0047	-0.0210	-0.1607	0.9774	5.0049	0.1532	0.7302	0.730	0.000	0.934	1.000
Insured.age	-0.0018	-0.0011	0.0000	-0.0005	0.0032	0.0012	0.0006	0.0017	0.808	0.424	0.886	0.898
Insured.sexFemale	-0.0186	0.0640	-0.0002	-0.0037	0.0472	0.0649	0.0105	0.0241	0.854	0.000	0.888	0.840
Car.age	-0.0040	-0.0079	0.0000	-0.0011	0.0073	0.0081	0.0014	0.0038	0.812	0.000	0.900	0.870
MaritalSingle	0.0041	-0.0105	-0.0001	0.0001	0.0427	0.0151	0.0105	0.0257	0.930	0.748	0.924	0.932
Car.useCommercial	-0.0278	0.1780	0.0000	-0.0051	0.1287	0.1802	0.0279	0.0731	0.868	0.000	0.902	0.830
Car.useCommute	0.0047	0.0331	0.0000	0.0029	0.0528	0.0353	0.0125	0.0320	0.890	0.156	0.906	0.832
Car.useFarmer	-0.0296	-0.0986	-0.0016	-0.0965	0.3330	0.1224	0.0726	0.2635	0.882	0.632	0.916	0.778
Credit.score	0.0000	-0.0001	0.0000	0.0000	0.0002	0.0001	0.0001	0.0001	0.912	0.616	0.912	1.000
RegionRural	0.0014	-0.0356	0.0011	0.0021	0.0625	0.0383	0.0143	0.0346	0.880	0.174	0.882	0.786
Annual.miles.drive	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.836	0.000	0.912	0.862
Years.noclaims	0.0006	-0.0042	0.0000	0.0001	0.0025	0.0042	0.0006	0.0015	0.876	0.000	0.894	0.908
TerritoryEmb	-0.0145	0.0948	-0.0002	-0.0024	0.0767	0.0964	0.0170	0.0445	0.894	0.000	0.932	0.812
Annual.pct.driven	0.1005	-0.1345	0.0019	0.0361	0.1369	0.1589	0.0213	0.1090	0.744	0.566	0.946	0.860
Total.miles.driven	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.862	0.806	0.916	0.838
Pct.drive.mon	0.1079	-5.6398	0.0125	0.0158	0.7713	5.6574	0.1859	0.7892	0.952	0.000	0.932	1.000
Pct.drive.tue	0.2735	-4.6950	0.0068	0.0874	0.7326	4.7089	0.1630	0.7013	0.942	0.000	0.958	1.000
Pct.drive.wed	0.1171	-4.6281	0.0080	0.0654	0.7818	4.6530	0.1965	0.7957	0.934	0.000	0.926	1.000
Pct.drive.thr	0.3401	-4.7322	0.0149	0.0989	0.7467	4.7442	0.1642	0.6974	0.936	0.000	0.952	1.000
Pct.drive.fri	-0.0218	-5.1114	0.0157	-0.0569	0.7816	5.1338	0.1834	0.8020	0.930	0.000	0.942	1.000
Pct.drive.sat	0.2341	-7.4253	0.0244	0.0884	1.0088	7.4406	0.2410	1.0133	0.904	0.000	0.924	1.000
Pct.drive.rush.am	-0.0411	0.0900	-0.0089	-0.0256	0.3083	0.3084	0.0736	0.3269	0.908	0.890	0.924	0.936
Pct.drive.rush.pm	-0.2780	-0.7399	-0.0012	-0.0994	0.4381	0.8164	0.0895	0.3984	0.816	0.284	0.878	0.868
Avgdays.week	-0.0018	-0.0875	0.0005	0.0003	0.0236	0.0895	0.0054	0.0233	0.908	0.004	0.926	1.000
Accel.06miles	0.0000	-0.0001	0.0000	-0.0001	0.0004	0.0004	0.0001	0.0004	0.958	0.952	0.908	0.910
Brake.06miles	-0.0001	-0.0004	0.0000	-0.0001	0.0003	0.0005	0.0001	0.0003	0.944	0.690	0.910	0.900
Acbr.others	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	0.0001	1.000	0.998	1.000	0.922
Left.turns	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.946	0.962	0.984	0.914
Right.turns	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.920	0.916	0.962	0.888

Table A.3: Estimation performance with the bootstrap samples of the actual data in adverse selection

Appendix B

Code

Code for Simulation Study

```
library(Matrix)
library(nleqslv)
#function for deviance
Poisson.Deviance <- function(pred, obs){
  200*(sum(pred)-sum(obs)+sum(log((obs/pred)^(obs))))/length(pred) }
###########data
J = 1000 # of total sampling
RMSEs <- matrix(nrow=J, ncol=6)
MAEs  <- matrix(nrow=J, ncol=6)
DEVs  <- matrix(nrow=J, ncol=6)

colnames(RMSEs) <- c("naive-10K","traditional-90K","full-90K","proposed-10K","boosting-10K")
colnames(MAEs)  <- c("naive-10K","traditional-90K","full-90K","proposed-10K","boosting-10K")
colnames(DEVs)  <- c("naive-10K","traditional-90K","full-90K","proposed-10K","boosting-10K")

naive_coef <- matrix(nrow=J, ncol=5)
prop1_coef <- matrix(nrow=J, ncol=5)
prop2_coef <- matrix(nrow=J, ncol=5)
full_coef  <- matrix(nrow=J, ncol=5)
boost_coef <- matrix(nrow=J, ncol=5)
trad_coef  <- matrix(nrow=J, ncol=5)

naive_stde <- matrix(nrow=J, ncol=5)
prop1_stde <- matrix(nrow=J, ncol=5)
prop2_stde <- matrix(nrow=J, ncol=5)
full_stde  <- matrix(nrow=J, ncol=5)
boost_stde <- matrix(nrow=J, ncol=5)
trad_stde  <- matrix(nrow=J, ncol=5)

system.time(
  for (j in 1:1000) {
    print(j)
    set.seed(j)
    I <- 100000
    x1        <- runif(I,0.18,0.81)
    x2        <- x1^2
    x3        <- rbinom(I, 1, 0.6)
```

```
x4        <- rnorm(I)
lambda    <- exp(-1.3-4*x1 + 3.4*x2 + 0.1*x3 + 0.5*x4)
NB_Claim  <- rpois(I, lambda)
Duration  <- rep(1, I)
fdata     <- as.data.frame(cbind(x1, x2, x3, x4, Duration, NB_Claim))

#for testing
set.seed(j+1000)
test_ind <- sample(1:nrow(fdata), 10000)
forr.data<-fdata[ test_ind,]
trtt.data<-fdata[-test_ind,]
#sampling-  when using a specific sampling method, comment other two sampling sections.
###################################################RS
set.seed(j+2000)
tele_ind <- sample(1:90000, nrow(fdata)*0.1)
ntr <- length(tele_ind)
####################################################NIS(advsel)
#set.seed(j+2000)
#dz <- 1/(1+exp(2*trtt.data$NB_Claim))
#dz <- dz/mean(dz)/9
#dzz <- rbinom(90000, 1, dz)
#tele_ind <- (1:90000)*(dzz==1)
#rm(dz, dzz)
#tele_ind <- tele_ind[tele_ind!=0]
#ntr <- length(tele_ind)
####################################################MAR(agesel)
#set.seed(j+2000)
#dz <- 1/(1+exp(3*trtt.data$x1))
#dz <- dz/mean(dz)/9
#dzz <- rbinom(90000, 1, dz)
#tele_ind <- (1:90000)*(dzz==1)
#rm(dz, dzz)
#tele_ind <- tele_ind[tele_ind!=0]
#ntr <- length(tele_ind)
#####################################################
S0  <- trtt.data[ tele_ind,            ]
# A small dataset that contains both telematics and traditional features
S1  <- trtt.data[-tele_ind, c(1:3, 5:6)]
# A large dataset that contains only traditional features
S1s <- trtt.data[-tele_ind,            ]
# A large dataset that contains both telematics and traditional features
S   <- rbind(S0,S1s)
rm(trtt.data)
#### Preliminary analysis ####
#naive model
glm.freq.naive <- glm(NB_Claim ~ .-Duration,  data=S0, family=poisson())
#full model
glm.freq.full  <- glm(NB_Claim ~ .-Duration,  data=S , family=poisson())
#traditional model
glm.freq.trad  <- glm(NB_Claim ~ .-Duration, data=S[, c(1:3, 5:6)], family=poisson())

#update coefficients and SE from model summaries
naive_coef[j,] <- summary(glm.freq.naive)$coefficients[,1]
naive_stde[j,] <- summary(glm.freq.naive)$coefficients[,2]
full_coef[j,]  <- summary(glm.freq.full)$coefficients[,1]
full_stde[j,]  <- summary(glm.freq.full)$coefficients[,2]

#basis fuctions
```

```
b_S0 <- as.matrix(cbind(S0[,c(5,1:3)], S0[,6]*S0[,c(5,1:3)]))
b_S  <- as.matrix(cbind(S[ ,c(5,1:3)], S[ ,6]*S[ ,c(5,1:3)]))

#function for optimize using nleqslv()
cal_eqn <- function(parm) {
  result <- colSums(as.vector(1+nrow(S1)/nrow(S0)*exp(parm%*%t(b_S0)))*b_S0)-colSums(b_S)
  return(result)  }

#find for parameters of basis functions
fit2 <- nleqslv(rep(0,8), cal_eqn)

#calculate weights from information projection
w3   <- 1+nrow(S1)/nrow(S0)*exp(b_S0 %*% fit2$x)
#############################
#combine weights to S0
SS6<-cbind(S0,w3)

#fitted the model with ws
glm.freq.S3 <- glm(NB_Claim ~ .-Duration-w3,offset=log(Duration),
                   weights= w3,  data=SS6, family=poisson())
x_S0 <- model.matrix(glm.freq.S3)

#coef of proposed model
prop2_coef[j,] <- summary(glm.freq.S3)$coefficients[,1]

# sandwich formula for variance estimation
Ui <- cbind(c(as.vector(w3)-1, rep(-1, nrow(S1)))* b_S,
  c(w3*(SS6$NB_Claim-fitted(glm.freq.S3)), rep(0, nrow(S1)))*as.matrix(S[,c(5,1:4)]))

Ui <- Ui - rep(colMeans(Ui), each=nrow(S))
V_U  <- (  t(Ui) %*% Ui)
tau  <- rbind(cbind(t(b_S0) %*% (as.vector(w3-1)*b_S0),
        matrix(0, ncol=ncol(x_S0), nrow=ncol(b_S0))),
  cbind(t(x_S0) %*% (as.vector(w3-1)*(SS6$NB_Claim-fitted(glm.freq.S3))*b_S0),
        -t(x_S0) %*% (as.vector(w3*fitted(glm.freq.S3))*x_S0) ))
invtau <- solve(tau)
prop2_stde[j,]  <- sqrt(diag(invtau %*% V_U %*% invtau))[-(1:ncol(b_S0))]
##########################boosting model
glm.freq.boost  <- glm(NB_Claim ~ x4-1,data=S0,
            offset=log(Duration)+predict(glm.freq.trad, S0), family=poisson())
#coefficients and SE
boost_coef[j,] <- c(summary(glm.freq.trad )$coefficients[,1],
                    summary(glm.freq.boost)$coefficients[,1])
boost_stde[j,] <- c(summary(glm.freq.trad )$coefficients[,2],
                    summary(glm.freq.boost)$coefficients[,2])
################################try forecast
pred.naive <- predict(glm.freq.naive, newdata = forr.data, type="response")
pred.full  <- predict(glm.freq.full , newdata = forr.data, type="response")
pred.S3    <- exp(as.matrix(forr.data[,c(5,1:4)]) %*% prop2_coef[j,])
pred.trad  <- predict(glm.freq.trad , newdata = forr.data, type="response")
pred.boost <- pred.trad * exp(coef(glm.freq.boost)*forr.data$x4)

#remove datasets for this split
rm(tele_ind, test_ind)

#RMSE
RMSEs[j,-4] <- sqrt(c(
  mean((forr.data$NB_Claim-pred.naive)^2),
      mean((forr.data$NB_Claim-pred.trad )^2),
      mean((forr.data$NB_Claim-pred.full )^2),
      mean((forr.data$NB_Claim-pred.S3   )^2),
      mean((forr.data$NB_Claim-pred.boost)^2)))

    #MAE
    MAEs[j,-4] <- c(
      mean(abs(forr.data$NB_Claim-pred.naive)),
      mean(abs(forr.data$NB_Claim-pred.trad )),
      mean(abs(forr.data$NB_Claim-pred.full )),
      mean(abs(forr.data$NB_Claim-pred.S3   )),
      mean(abs(forr.data$NB_Claim-pred.boost)))

    #DEV
    DEVs[j,-4] <- c(
      Poisson.Deviance(pred.naive, forr.data$NB_Claim),
      Poisson.Deviance(pred.trad , forr.data$NB_Claim),
      Poisson.Deviance(pred.full , forr.data$NB_Claim),
      Poisson.Deviance(pred.S3   , forr.data$NB_Claim),
      Poisson.Deviance(pred.boost, forr.data$NB_Claim))

  })

#summarizing the outputs
colMeans(RMSEs)
colMeans(MAEs)
colMeans(DEVs)

#true coeffients used for data generation
true_coef      <- c(-1.3, -4, 3.4, 0.1, 0.5)

#bias of each estimator
bias_naive <- true_coef - colMeans(naive_coef)
bias_trad  <- true_coef - colMeans(trad_coef)
bias_prop2 <- true_coef - colMeans(prop2_coef)#proposed
bias_boost <- true_coef - colMeans(boost_coef)
bias_full  <- true_coef - colMeans(full_coef)

#RMSE of estimates
rmse_naive <- sqrt(colMeans((naive_coef-rep(true_coef, each=J))^2))
rmse_trad  <- sqrt(colMeans((trad_coef -rep(true_coef, each=J))^2))
rmse_prop2 <- sqrt(colMeans((prop2_coef-rep(true_coef, each=J))^2))#proposed
rmse_boost <- sqrt(colMeans((boost_coef-rep(true_coef, each=J))^2))
rmse_full  <- sqrt(colMeans((full_coef -rep(true_coef, each=J))^2))

#CI of estimator
naive_90CI <- colMeans((naive_coef-1.645*naive_stde<rep(true_coef, each=J))*
                         (naive_coef+1.645*naive_stde>rep(true_coef, each=J))*1)
trad_90CI <- colMeans((trad_coef -1.645*trad_stde<rep(true_coef, each=J))*
                        (trad_coef +1.645*trad_stde>rep(true_coef, each=J))*1)
prop2_90CI <- colMeans((prop2_coef-1.645*prop2_stde<rep(true_coef, each=J))*
                         (prop2_coef+1.645*prop2_stde>rep(true_coef, each=J))*1,
                       na.rm=TRUE)#proposed
boost_90CI <- colMeans((boost_coef-1.645*boost_stde<rep(true_coef, each=J))*
                         (boost_coef+1.645*boost_stde>rep(true_coef, each=J))*1)
full_90CI <- colMeans((full_coef -1.645*full_stde<rep(true_coef, each=J))*
                        (full_coef +1.645*full_stde>rep(true_coef, each=J))*1)
```

Code for Empirical Analysis

```r
# data is available from http://www2.math.uconn.edu/~valdez/data.html
# it has been cleaned by using territory embedding and PCA for telematics variables
########################data
fdata <- read.csv("ffdata.csv")# load rawdata
J = 500 # of total sampling
RMSEs <- matrix(nrow=J, ncol=5)
DEVs  <- matrix(nrow=J, ncol=5)

colnames(RMSEs) <- c("naive-10K", "traditional-90K", "full-90K", "proposed-10K", "boosting-10K")
colnames(DEVs)  <- c("naive-10K", "traditional-90K", "full-90K", "proposed-10K", "boosting-10K")

naive_coef <- matrix(nrow=J, ncol=29)
propd_coef <- matrix(nrow=J, ncol=29)
full_coef  <- matrix(nrow=J, ncol=29)
boost_coef <- matrix(nrow=J, ncol=29)
trad_coef  <- matrix(nrow=J, ncol=13)

naive_stde <- matrix(nrow=J, ncol=29)
propd_stde <- matrix(nrow=J, ncol=29)
full_stde  <- matrix(nrow=J, ncol=29)
boost_stde <- matrix(nrow=J, ncol=29)
trad_stde  <- matrix(nrow=J, ncol=13)

system.time(
  for (j in 1:J) {
    set.seed(j+1000)  #for testing
    test_ind <- sample(1:100000, 100000, replace=T)
    forr.data<-fdata[ test_ind,]
    #sampling-  when using a specific sampling method, comment other two sampling sections.
    ############################################RS
    set.seed(j+2000)
    tele_ind <- sample(1:100000, 100000, replace=T)
    trad_ind <- sample(1:100000, 800000, replace=T)
    ntr <- length(tele_ind)
    ##############################################NIS(advsel)
    #set.seed(j+2000)
    #dz <- 1/(1+exp(1*fdata$NB_Claim))#adjusted
    #dz <- dz/mean(dz)/10
    #tele_ind <- sample(1:100000, 100000, replace=T, prob=  dz)
    #trad_ind <- sample(1:100000, 800000, replace=T, prob=1-dz)
    #ntr <- length(tele_ind)
    ############################################MAR(agesel)
    #set.seed(j+2000)
    #dz <- 1/(1+exp((3*fdata$Insured.age/100)))#adjusted
    #dz <- dz/mean(dz)/10
    #tele_ind <- sample(1:100000, 100000, replace=T, prob=  dz)
    #trad_ind <- sample(1:100000, 800000, replace=T, prob=1-dz)
    #ntr <- length(tele_ind)
    ##########################################################
    S0  <- fdata[ tele_ind,            ]
    # A small dataset that contains both telematics and traditional features
    S1  <- fdata[ trad_ind, c(1:13, 30)]
    # A large dataset that contains only traditional features
    S1s <- fdata[ trad_ind,            ]
    # A large dataset that contains both telematics and traditional features
    S   <- rbind(S0,S1s)#entire train set
```

```
#### Preliminary analysis ####
#naive model
glm.freq.naive <- glm(NB_Claim ~ .-Duration,offset=log(Duration),data=S0, family=poisson())
#full model
glm.freq.full  <- glm(NB_Claim ~ .-Duration,offset=log(Duration),data=S , family=poisson())
#traditional model
glm.freq.trad  <- glm(NB_Claim ~ .-Duration,offset=log(Duration),
                      data=S[, c(1:13, 30)], family=poisson())

#update coefficients and SE from model summaries
naive_coef[j,] <- summary(glm.freq.naive)$coefficients[,1]
naive_stde[j,] <- summary(glm.freq.naive)$coefficients[,2]
full_coef[j,]  <- summary(glm.freq.full)$coefficients[,1]
full_stde[j,]  <- summary(glm.freq.full)$coefficients[,2]
trad_coef[j,] <- summary(glm.freq.trad)$coefficients[,1]
trad_stde[j,] <- summary(glm.freq.trad)$coefficients[,2]

#basis fuctions
b_S0 <- as.matrix(cbind(rep(1,length(S1$Duration)),S0[,c(30,2:13)], S0[,30]*S0[,c(2:13)]))
b_S  <- as.matrix(cbind(rep(1,length(S$Duration)),S[ ,c(30,2:13)], S[ ,30]*S[ ,c(2:13)]))

#function for optimize using nleqslv()
cal_eqn <- function(parm) {
  result <- colSums(as.vector(1+nrow(S1)/nrow(S0)*exp(parm %*%t(b_S0)))*b_S0)-colSums(b_S)
  return(result)  }
#find for parameters of basis functions
fit2 <- nleqslv(rep(0,26), cal_eqn, method = "Newton",
                control = list(allowSingular=TRUE, xtol=1e-15,ftol=1e-8))
#calculate weights from information projection
w3   <- 1+nrow(S1)/nrow(S0)*exp(b_S0 %*% fit2$x)
#############################
SS6<-cbind(S0,w3)#combine weights to S0

#fitted the model with ws
glm.freq.S3 <- glm(NB_Claim ~ .-Duration-w3,offset=log(Duration),
                   weights= w3,  data=SS6, family=poisson())
x_S0 <- model.matrix(glm.freq.S3)

#coef of proposed model
propd_coef[j,] <- summary(glm.freq.S3)$coefficients[,1]

# sandwich formula for variance estimation
Ui <- cbind(c(as.vector(w3)-1, rep(-1, nrow(S1)))* b_S,
  c(w3*(SS6$NB_Claim-fitted(glm.freq.S3)), rep(0, nrow(S1)))*as.matrix(S[,c(1:29)]))

Ui <- Ui - rep(colMeans(Ui), each=nrow(S))
V_U  <- (  t(Ui) %*% Ui)
tau  <- rbind(cbind(t(b_S0) %*% (as.vector(w3-1)*b_S0),
        matrix(0, ncol=ncol(x_S0), nrow=ncol(b_S0))),
  cbind(t(x_S0) %*% (as.vector(w3-1)*(SS6$NB_Claim-fitted(glm.freq.S3))*b_S0),
        -t(x_S0) %*% (as.vector(w3*fitted(glm.freq.S3))*x_S0) ))
invtau <- solve(tau)
propd_stde[j,]  <- sqrt(diag(invtau %*% V_U %*% invtau))[-(1:ncol(b_S0))]
##########################boosting model
glm.freq.boost  <- glm(NB_Claim ~ Annual.pct.driven+Total.miles.driven+Pct.drive.mon
                       +Pct.drive.tue+Pct.drive.wed+Pct.drive.thr+Pct.drive.fri
                       +Pct.drive.sat+Pct.drive.rush.am+Pct.drive.rush.pm+Avgdays.week
                       +Accel.06miles+Brake.06miles+Acbr.others+Left.turns+Right.turns-1,
                         data=S0, offset=log(Duration)+predict(glm.freq.trad, S0)
                         , family=poisson())

  #coefficients and SE
  boost_coef[j,] <- c(summary(glm.freq.trad )$coefficients[,1],
                      summary(glm.freq.boost)$coefficients[,1])
  boost_stde[j,] <- c(summary(glm.freq.trad )$coefficients[,2],
                      summary(glm.freq.boost)$coefficients[,2])
  ################################try forecast
  pred.naive <- predict(glm.freq.naive, newdata = forr.data, type="response")
  pred.full  <- predict(glm.freq.full , newdata = forr.data, type="response")
  pred.S3    <- exp(as.matrix(forr.data[,c(2:29)]) %*%
              propd_coef[j,2:29]+propd_coef[j,1]+ log(forr.data[,1]))
  pred.trad  <- predict(glm.freq.trad , newdata = forr.data, type="response")
  pred.boost <- pred.trad * exp(as.matrix(forr.data[14:29])%*%coef(glm.freq.boost))

  #remove datasets for this split
  rm(tele_ind, test_ind)

  #RMSE
  RMSEs[j,] <- sqrt(c(
    mean((forr.data$NB_Claim-pred.naive)^2),
    mean((forr.data$NB_Claim-pred.trad )^2),
    mean((forr.data$NB_Claim-pred.full )^2),
    mean((forr.data$NB_Claim-pred.S3   )^2),
    mean((forr.data$NB_Claim-pred.boost)^2)))

  #DEV
  DEVs[j,] <- c(
    Poisson.Deviance(pred.naive, forr.data$NB_Claim),
    Poisson.Deviance(pred.trad , forr.data$NB_Claim),
    Poisson.Deviance(pred.full , forr.data$NB_Claim),
    Poisson.Deviance(pred.S3   , forr.data$NB_Claim),
    Poisson.Deviance(pred.boost, forr.data$NB_Claim))
})
```

Appendix C

Basic Setup of Proposed Method

The basic setup of the proposed method is listed in Wang and Kim (2021a). We are interested in estimating β as the unique solution to $\mathrm{E}\{\mathbf{U}(\beta; \mathrm{N}, \mathrm{X}_1, \mathrm{X}_2)\} = 0$ where X_2 is subjected to missingness where $\mathbf{U}$ is a smooth function of β with non singular gradient matrix $\mathrm{E}\{\partial_\beta \mathbf{U}(\beta)\}$.

Now we can denote the actual dataset as $\{(n_i, x_{i1}, \delta_i x_{i2}, \delta_i) : i = 1, ..., M)\}$, where δ_i is the sampling indicator variable defined as

$$\delta_i = \begin{cases} 1, & \text{if } x_{i2} \ \ is \ \ observed, \\ 0, & \text{otherwise} \end{cases}$$

Considering $f(.)$ as the density function, let $f_j(n, x_1, x_2)$ be the conditional density of (N, X_1, X_2) given $\delta = j$ for $j = 0, 1$. Then the density ratio function is defined as $r(n, x_1, x_2) = f_0(n, x_1, x_2)/f_1(n, x_1, x_2)$.

By Bayes theorem, we obtain

$$\frac{\mathrm{P}(\delta = 0|n, x_1, x_2)}{\mathrm{P}(\delta = 1|n, x_1, x_2)} = \frac{\mathrm{P}(\delta = 0)}{\mathrm{P}(\delta = 1)} \times r(n, x_1, x_2).$$

Assuming $c = \mathrm{P}(\delta = 0)/\mathrm{P}(\delta = 1)$ is known, we can express

$$\frac{1}{\mathrm{P}(\delta = 1|N, X_1, X_2)} = 1 + c \times r(N, X_1, X_2). \tag{C.1}$$

and use

$$\omega_i = \frac{1}{\mathcal{P}(\delta_i = 1|n_i, x_{i1}, x_{i2})} = 1 + c \times r(n_i, x_{i1}, x_{i2}).$$

as the propensity score for the i^{th} policyholder with $delta_i = 1$. Thus this problem can be considered as a density ratio function estimation problem. And we can take $\hat{c} = M_1/M_0$, $M_0 = \sum_{i=1}^{N} \delta_i$ and $M_1 = M - M_0$.

Under the assumption of MAR, we consider

$$\delta \perp X_2 | (N, X_1). \tag{C.2}$$

Under (C.2), the density ratio function can be written as $r(n, x_1, x_2) = r(n, x_1)$.

Now a vector of integrable functions of (n,x) are taken as $\mathbf{b}(n, x_1) = (b_1(n, x_1), ..., b_L(n, x_1))$. Then, we assume that

$$\delta \perp X_2 | \mathbf{b}(N, X_1). \tag{C.3}$$

Using $b(n, x_1)$ satisfying (C.3), a smoothed density ratio function is defined as

$$r^*(n, x_1) = \mathrm{E}\{r(n, x_1)|b(n, x_1), \delta = 1\}, \tag{C.4}$$

and it satisfies

$$\frac{1}{M_0}\sum_{i=1}^{M} \delta_i r^*(n_i, x_{i1}) b(n_i, x_{i1}) = \frac{1}{M_1}\sum_{i=1}^{M}(1 - \delta_i) b(n_i, x_{i1}) \tag{C.5}$$

at least approximately.

Then we can use $r^*(n, x_1)$ in (C.4) to construct a smoothed propensity score estimating function of β as follows:

$$\hat{\mathbf{U}}_{SPS}(\beta) = \frac{1}{M}\sum_{i=1}^{M} \delta_i \{1 + \frac{M_1}{M_0} r^*(n_i, x_{i1})\} \mathbf{U}(\beta; n_i, x_{i1}, x_{i2}). \tag{C.6}$$

Furthermore, since $r(n, x_1)$ is a function of (n, x_1) only, we obtain

$$\mathrm{E}\{r(n, x_1)|b(n, x_1), x_{i2}, \delta = 1\} = \mathrm{E}\{r(n, x_1)|b(n, x_1), \delta = 1\} \tag{C.7}$$

Then the results (C.7) implies that

$$\mathrm{E}\{\hat{\mathbf{U}}_{PS}(\beta)|b(n, x_1), x_{i2}, \delta\} = \hat{\mathbf{U}}_{SPS}(\beta),$$

where

$$\hat{\mathbf{U}}_{PS}(\beta) = \frac{1}{N}\sum_{i=1}^{M} \delta_i \{1 + \frac{M_1}{M_0} r(n, x_i 1)\} \mathbf{U}(\beta; n_i, x_{i1}, x_{i2})$$

.

Therefore, the following result is established.

Under (C.2), we obtain

$$\mathrm{E}\{\hat{U}_{PS}(\beta)\} = \mathrm{E}\{\hat{U}_{SPS}(\beta)\} \tag{C.8}$$

and

$$\mathrm{V}\{\hat{U}_{PS}(\beta)\} \geq \mathrm{V}\{\hat{U}_{SPS}(\beta)\}. \tag{C.9}$$

Then the relationships in (C.8) and (C.9) imply that the solution to $\hat(U)_{SPS} = 0$ is more efficient than the solution to $\hat{U}_{PS}(\beta) = 0$.