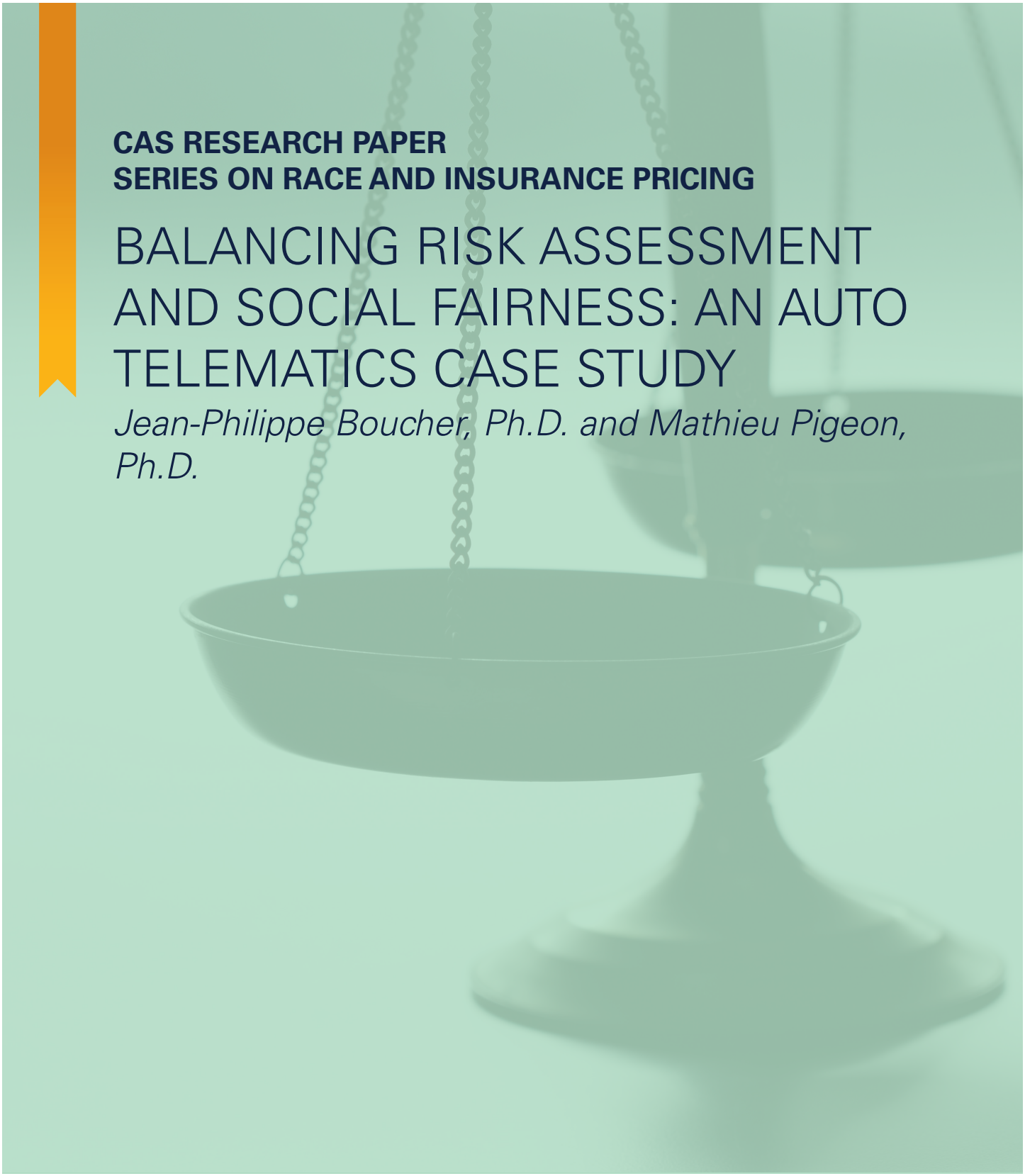


CAS RESEARCH PAPER
SERIES ON RACE AND INSURANCE PRICING

BALANCING RISK ASSESSMENT AND SOCIAL FAIRNESS: AN AUTO TELEMATICS CASE STUDY

Jean-Philippe Boucher, Ph.D. and Mathieu Pigeon, Ph.D.

CASUALTY ACTUARIAL SOCIETY

The CAS Research Paper Series on Race & Insurance Pricing was created to guide the insurance industry toward proactive, quantitative solutions that address potential racial bias in insurance pricing. These reports explore different aspects of unintentional potential bias in insurance pricing, address historical foundations and offer forward-looking solutions to quantify and handle possible bias. Through these reports, the CAS aims to encourage actuaries to discuss this topic with their stakeholders across all areas of insurance pricing and operations. For more information on the series, visit casact.org/raceandinsuranceresearch.

The Casualty Actuarial Society (CAS) is a leading international organization for credentialing, professional education and research. Founded in 1914, the CAS is the world's only actuarial organization focused exclusively on property-casualty risks and serves over 10,000 members worldwide. CAS members are sought after globally for their insights and ability to apply analytics to solve insurance and risk management problems. As the world's premier P&C actuarial research organization, the CAS reaches practicing actuaries across the globe with thought-leading concepts and solutions. The CAS has been conducting research since its inception. Today, the CAS provides thousands of open-source research papers, including its prestigious publication, *Variance* — all of which advance actuarial science and enhance the P&C insurance industry. Learn more at casact.org.

Caveat and Disclaimer

CAS RESEARCH PAPER
SERIES ON RACE AND INSURANCE PRICING

BALANCING RISK ASSESSMENT AND SOCIAL FAIRNESS: AN AUTO TELEMATICS CASE STUDY

Jean-Philippe Boucher, Ph.D.; Mathieu Pigeon, Ph.D.

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, VA 22203
www.casact.org
(703) 276-3100

Contents

Balancing Risk Assessment and Social Fairness: An Auto Telematics Case Study

By Jean-Philippe Boucher, Ph.D., and Mathieu Pigeon, Ph.D.

Executive Summary

The paper explores the possibility of using telematics and usage-based insurance technologies to reduce dependence on sensitive information when pricing insurance. Actuaries commonly rely on demographic factors such as age and gender when deciding insurance premiums. However, some people regard that approach as an unfair use of personal information. Moreover, socioeconomic factors such as marital status, credit score, and geographic location may also be used in insurance pricing, leading to higher or lower premiums for policyholders from similar demographic groups. Furthermore, those factors may also be correlated with other sensitive information, such as income and race or ethnicity, creating concern of proxy discrimination among some consumers and regulators.

Telematics information, which provides detailed insights into driving behavior, may offer a more causal link to the probability of auto accidents than socioeconomic factors. By analyzing factors such as miles driven, driving patterns, and sudden maneuvers, insurers can better understand the underlying risk of accidents. Increasingly, insurers are using telematics technology in their pricing. However, many insurers continue to rely on traditional variables, likely because of the cost to implement telematics technology and consumer data privacy concerns.

Our analysis finds that telematics variables, such as miles driven, hard braking, hard acceleration, and days of the week driven, significantly reduce the need to include age, sex, and marital status in the claim frequency and severity models. Whereas the need for geographic territory and credit score appeared to have been significantly reduced in the model built on synthetic data, the reduction was significantly muted when the approach was validated with the real-world data set.

Although not all of the sensitive variables we tested could be eliminated from the model, the analysis shows there is still value in insurers testing the addition of telematics to their models to potentially reduce reliance on sensitive information that could result in actual or perceived bias. Many insurers still run into roadblocks to adopting telematics, such as the costs and challenges of explaining complex gradient-boosting models to regulators.

Our analyses are based on a synthetic database generated from real-world data from a major Canadian insurer. Thus, all the results obtained can easily be reproduced. The main results obtained are validated on the original insurer data to confirm that the data reproduction process did not introduce significant bias. We built frequency and severity models substituting telematics variables for sensitive variables. Then we ran a comparison to evaluate whether we can eliminate sensitive variables by using telematics variables.

Our analysis considers two main approaches: one based on generalized linear models, a widely used method in insurance pricing, and a gradient-boosting model used in machine learning. The generalized linear model offers a balance between simplicity and accuracy, making it a popular choice. On the other hand, the gradient-boosting model provides significantly more accuracy, but at the cost of transparency and explainability.

The paper proposes a methodology based on synthetic data, and it is essential to note that individual insurers may arrive at different conclusions when using actual data or different rating variables.

1. Introduction

The[1,2] insurance industry often bases its decisions on statistical models, which, in a nutshell, must determine whether an insured person is more or less risky. Therefore, carefully discriminating between risks would seem to be called for to establish the fairest possible premium. However, this practice may contravene principles prohibiting discrimination based on sensitive variables such as age, sex,[3] marital status, and ethnicity. Thus, the line between a *fair* actuarial classification and an *unfair* discriminatory classification is thin and must be studied carefully.

To illustrate the delicate nature of some variables, we can, for example, refer to the practice of *redlining* in the United States to show the racial bias in such segmentation. Redlining refers to delimiting residential neighborhoods according to their level of desirability by assigning them a color (green, blue, yellow, or red). The Federal Housing Administration used redlining to determine eligibility to obtain insurance on a residence's mortgage (Chibanda 2022). The approach was subsequently criticized for discriminating against ethnicity because the neighborhoods identified as the least desirable were mainly those where minorities resided. Redlining became illegal under the Fair Housing Act of 1968.

The use of the sex variable in pricing may also raise concerns regarding fairness. In 1985, Montana was the first state in the United States to ban its use in the insurance industry following the efforts of feminist groups in the fight for unisex pricing (Reid 1985). Several states, such as California, Hawaii, Massachusetts, and Michigan, have followed Montana's lead by excluding this variable from the calculation of automobile insurance premiums. The European Union also banned the use of sex in the estimation of premiums in 2012, and the calculation is now done through variables directly linked to the insured's driving, such as the brand of the car and the mileage traveled (Lichtenstein 2022). Because of these changes, several players in the insurance industry, i.e., companies, regulators, and the scientific community, are developing new methodologies, making it possible to include the notion of fairness in models without sacrificing the models' predictive ability (see, for example, Lindholm et al. [2022] and Embrechts and Wüthrich [2022]).

1.1. TELEcommunications and InforMATICS Technology

Telematics technology is a blend of telecommunications and informatics that allows access to new sources of information through digitization and big data. This data is collected via an onboard diagnostics device or a smartphone application. In the past, auto insurers primarily relied on static attributes related to the vehicle or the insured, which were indirectly related to accident risk. However, with the emergence of telematics technology, insurers

[1] Artificial intelligence (AI) helped generate this project's content, improve grammar, and translate some French expressions.

[2] Computer applications do not claim to be the most efficient or elegant. The goal is mainly to demonstrate how theory can be applied to actual data. If it is possible to optimize the code, one should not hesitate.

[3] As part of this project, we used the term "sex" rather than "gender" in the analysis because we worked with databases containing an Insured.sex variable.

can offer more customized premiums based on the insured's driving habits, style, and distance, which has the potential to more accurately determine their risk.

Usage-based insurance, in which the insured's premium is estimated using their driving data, has become highly popular in the last decade. One piece of GPS-collected information that is directly related to the risk insured is distance driven. The relevance of including this variable in ratemaking has been studied by Ayuso, Guillen, and Pérez-Marín (2014, 2016b), Boucher, Pérez-Marín, and Santolino (2013), and Lemaire, Park, and Wang (2016), among others. Insurers estimate the insured's premium through a pay-as-you-drive or pay-how-you-drive scheme (Tselentis, Yannis, and Vlahogianni 2016). Pay-as-you-drive focuses on driving habits, such as distance, time of day, or road type, while pay-how-you-drive considers driving style, such as aggressive acceleration, sudden lane shifts, or speeding.

Insurers are increasingly promoting usage-based insurance, citing its numerous benefits, such as allowing for more accurate pricing and a better customer experience. Consumers are also increasingly appreciating it. Of 1,005 U.S. insurance consumers surveyed by Willis Towers Watson, 80% were willing to share their recent driving information for a personalized insurance product (Bansag 2017). By accurately assessing people's driving habits, insurers can more accurately determine their risk and offer a fair premium (see, e.g., Lemaire, Park, and Wang [2016] or Verbelen, Antonio, and Claeskens [2018]). Additionally, telematics has the potential to encourage people to drive more safely and drive less overall, which would help reduce traffic congestion, make roads safer, limit greenhouse gas emissions, and make insurance more affordable, among other things.

It is important to note that some obstacles to adopting telematics still exist. Entry costs, such as for technology development, hardware purchases, and advertising, can be a significant obstacle for smaller companies. In addition, the population is increasingly wary of sharing individual data, and a problem of social acceptability could arise. Finally, the costs associated with protecting confidential data and possible legal proceedings related to consumer privacy can weigh negatively in a company's decision to opt for this approach.

1.2. Existing Telematics Models

Over the past 15 years, a wealth of research has been conducted and published in scientific journals spanning actuarial science, statistics, and transport. The extent of this body of work allows us to categorize these contributions based on their success in achieving one or both of the following objectives:

- Demonstrate that a model performs better when one or more elements of telematics are considered. Such elements range from the distance traveled (measured by onboard diagnostics or a smartphone application) to a detailed analysis of position second by second.
- Demonstrate that using one or more elements of telematics can replace use of a sensitive variable such as sex or age.

The models studied range from classic statistical approaches, such as generalized linear models (GLMs), generalized additive models (GAMs), and splines, to machine learning

approaches, such as neural networks and boosting models. Appendix A presents an overview of the scientific literature on telematics models. In recent years, the distance traveled by a motorist has been the most studied measurement provided by a telematics device. Key conclusions from past research include that driver usage significantly influences the expected number of accidents and that some variables, such as sex, may not be necessary if telematics provides enough information on driving habits.

1.3. Objectives

The proposed research project will analyze the importance of several sensitive covariates such as

- territory of residence,
- age,
- sex,
- marital status, and
- credit score.

Some of these covariates, such as territory, marital status, and credit score, are not typically considered protected classes under national or state-insurance related regulations, and they can be used to rate policyholders in current insurance rating plans. However, we often observe a strong link between these covariates and certain protected variables, such as ethnicity (Bender et al. 2021). This raises questions about using such covariates and justifies including them in our studies.

The project's objective is to see how a driver's telematics information can reduce the importance given to sensitive covariates such as the insured's age and sex. Whereas information on policyholder race and ethnicity was not available for this project, one could do a similar analysis to understand whether telematics information could reduce importance given to nonsensitive variables that are perceived as being correlated with race or ethnicity.

The work we present in this paper can be used to extract the main conclusions shown in Section 2.

1.4. Data Used for the Project

1.4.1. Synthetic Data Set

The data that could have been used in the project comes from a major Canadian insurer and is highly confidential. To protect that confidentiality, we generated a synthetic database from the insurer data for use in this analysis. Details can be found in So, Boucher, and Valdez (2021). The database can be accessed at https://emiliano-valdez.scholar.uconn.edu/data/. The synthetic data set generated has 100,000 policies, including observations about driver's claims experience and associated classical risk variables and telematics-related variables. Table 1.1 shows the variables available in the synthetic database.

Table 1.1. Variables in the Synthetic Database

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.noclaim	Number of years without any claims
	Territory	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.xhrs	Percent vehicle driven within x hrs: 2hrs/3hrs/ 4hrs
	Pct.drive.xxx	Percent vehicle driven during xxx: wkday/wkend
	Pct.drive.rushxx	Percent of driving during xx rush hours: am/pm
	Avgdays.week	Mean number of days used per week
	Accel.xxmiles	Number of sudden accelerations 6/8/9/. . ./14 mph per second per 1,000 miles
	Brake.xxmiles	Number of sudden brakes 6/8/9/. . ./14 mph per second per 1,000 miles
	Left.turn.intensityxx	Number of left turns per 1,000 miles with intensity 08/09/10/11/12
	Right.turn.intensityxx	Number of right turns per 1,000 miles with intensity 08/09/10/11/12
Response	NB_Claim	Number of claims during observation
	AMT_Claim	Aggregated amount of claims during observation

We compare the results obtained with synthetic data to those obtained with the original data to ensure the accuracy of the conclusions.

As the synthetic database does not allow the different coverages to be distinguished, it isn't possible to analyze them separately. However, the analysis and the code used could easily be reused to include that information if available. Finally, for the whole study, we've divided the database into a training portion, which contains 80% of the original synthetic data, and a test database, containing the remaining 20% of the synthetic data. All the code and more graphs can be found in the project's GitHub folder (https://github.com/J-PBoucher/CAS_Project2024).

1.4.2. Nonrepresentative Sample

In analyzing telematics data, one must be careful not to jump to general conclusions about the driving behavior of the whole portfolio. Indeed, policyholders who have decided to place a telematics device on their car or to download an application on their phone that tracks all their car trips do not correspond to the general driver population. In our case, approximately 10% to 15% of the insurance company's portfolio chose to use the telematics option for their car insurance. Typically, such insureds correspond to one of two profiles:

1. *Technophile policyholders.* They love new technology in general and want detailed information about their driving habits. Summary driving data is indeed continuously available to policyholders via a website.
2. *Young and/or bad drivers.* To motivate policyholders to buy the telematics option, insurance companies often offer an initial discount, and the renewal discounts range from 0% to 25% depending on the driving experience. Because auto insurance in Ontario is expensive and often unaffordable for some drivers, all discounts are welcome for policyholders with high insurance premiums. As a result, an unusually high proportion of risky insureds use telematics devices or telematics apps. Even if this conclusion contradicts the common belief that the best drivers will choose telematics because they know they are good, it has been proven through the first few years of telematics experience.

1.5. Modeling Approaches

This section presents the modeling approaches we took in the project. We present the more technical details in Appendix B. This paper separately models the frequency and severity components using traditional, telematics, and sensitive explanatory variables. In a regression context, we have a database of size n: $\{z_i; \boldsymbol{x}_i\}_{i=1,2,\ldots,n}$, where $\boldsymbol{x}_i$ is a vector on size q of covariates (continuous and categorical) and z_i is the response variable (e.g., the severity of a claim). The objective is to predict the value of the response variable y^* for a new observation whose covariates are $\boldsymbol{x}^*$.

We have selected two families of models – regression-like and black box. It is important to note that both families have advantages and disadvantages and that choosing to use one or the other depends on multiple factors, such as the need for interpretability, the legal framework, and computing resources. In actuarial practice, complex (but difficult to interpret) models are often used to find insights that can be incorporated into simpler models.

Regression-like models are a family of models for which the interpretation of the parameters is simple. We define a GLM with the following structure:

$$g\left(E\left[Y\right]\right) = \beta_0 + \sum_{j=1}^{q} \beta_j x_j,$$

where $g()$ is the link function. In this paper, we always assume a logarithmic link, i.e.,

$$E\left[Y\right] = e^{\beta_0 + \sum_{j=1}^{q} \beta_j x_j}.$$

We add an elastic net (or GLM-net) regularization to this model to select the covariates and estimate parameters. This method is seen as a combination of lasso and ridge regressions, and we refer to Hastie, Tibshirani, and Friedman (2009) for more details about this approach. One of the advantages of this approach is that it solves the redundancy of variables and the multicollinearity of risk factors. The idea of the procedure is to impose constraints on the coefficients of the model. Excluding the intercept from the procedure, the constraint to be added to the log-likelihood score to be maximized is expressed as follows:

$$\left(\alpha \sum_{j=1}^{q} \left|\beta_j\right| + \frac{1-\alpha}{2} \sum_{j=1}^{q} \beta_j^2\right) \leq \lambda, \lambda > 0, 0 \leq \alpha \leq 1.$$

This penalty constraint depends on the values chosen for the parameters α and λ. If $\alpha = 1$, the elastic net method is equivalent to a lasso regression. In contrast, if $\alpha = 0$, it is equivalent to a ridge regression. In our document, for each model, the optimal values of λ and α were obtained by cross-validation using deviance as a selection criterion.

Black box models are a family of models from the machine learning field, based on boosting, and widely used in the actuarial industry. However, the interpretation of the parameters is complex—hence the term "black box model." Boosting is a meta-algorithm that improves the predictive power of a simpler model (weak learner). It aims to build B models sequentially: the model B depends on the model (B – 1), which depends on the model (B – 2), etc. Each new model is built specifically to improve the predictions made by the previous model. In particular, XGBoost (i.e., eXtreme Gradient Boosting) is an open-source software library that provides a regularizing gradient-boosting framework. In this paper, we use this meta-algorithm with trees as the weak learner.

2. Main Results

Our analysis is focused mainly on the usefulness of telematics variables in replacing, totally or in part, five sensitive variables: credit score, age of the insured, sex of the insured, marital status, and territory. For frequency and severity, the final models consider the following variables (see Table 1.1 for definitions):

- Traditional: *Insured.sex, Marital, Car.use, Region, Credit.score, Insured.age, Car.age, Years.noclaims,* and *Territory*

- Telematics: *Miles.per.day, Avgdays.week, Pct.drive.xhrs, Pct.drive.rushxx, Pct.drive.xxx, Accel.xxmiles, Brake.xxmiles, Left.turn.intensityxx,* and *Right.turn.intensityxx*

For the frequency model, we consider the contract duration and the variable *Total.miles.driven* as exposure measures. For many of these variables, we consider several transformations – we describe those in the next sections.

A first descriptive analysis in which we reviewed the potential impact of each individual rating variable on frequency and severity (see Section 3) showed that among the sensitive variables, two (*Credit.score* and *Insured.age*) have an impact, one (*Insured.sex*) may have a small impact, and two (*Marital* and *Territory*) have some impact on both frequency and severity. In addition, we observe that some sensitive variables present a nonnegligible correlation with each other: *Credit.score*, *Marital,* and *Insured.age* (see Section C.2 in Appendix C).

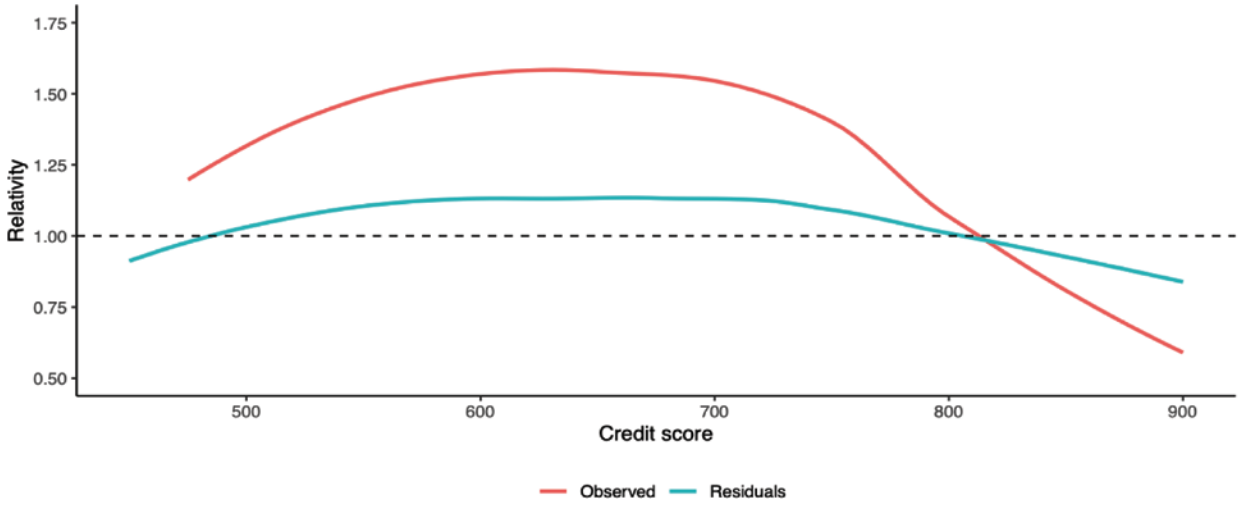
In a basic model with only categorical covariates (see Section 4.1), such as the insured's sex, marital status, vehicle usage, and region, the conclusions from the descriptive analysis are confirmed: the insured's sex has no impact and the marital status has a minimal impact. This suggests that maintaining these two sensitive variables in this particular ratemaking model may not be justified at this stage, but this may not generalize to other insurers.

After adding continuous variables, the different models considered make it possible to improve the scores consistently: XGBoost performs better than GLM-net, which performs better than GLM (trad.). In frequency and severity modeling, the transition from a transparent approach (GLM) to a black box approach (XGBoost) is accompanied by a notable improvement in model performance (see in Sections 4 and 5). In a future project, evaluating the financial gain associated with this transition could allow a more informed decision to be made.

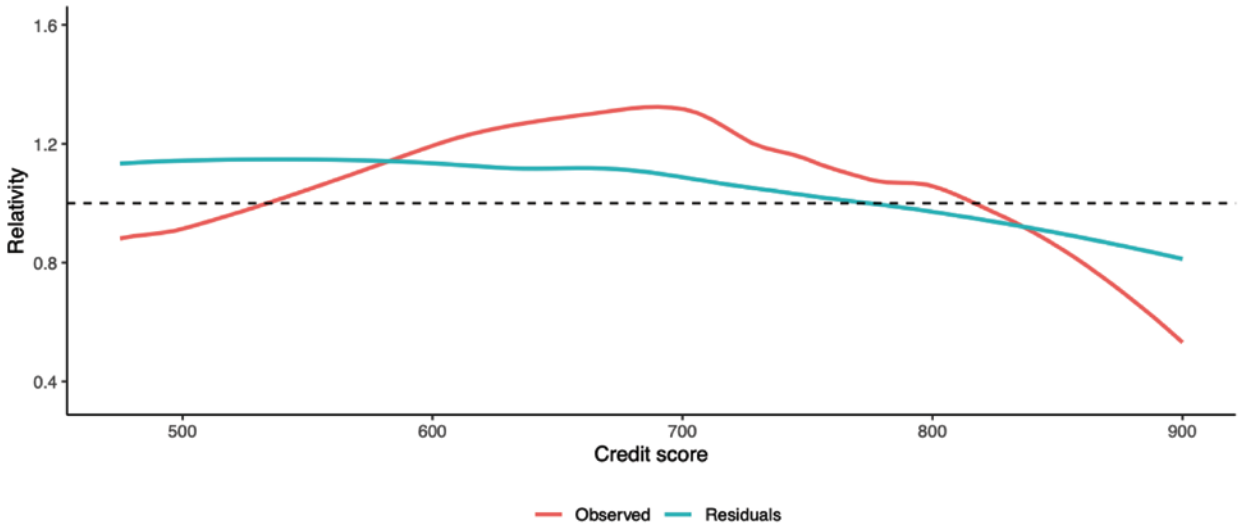
After adding the telematics variables to the analysis on the synthetic database, we observe the following regarding the sensitive variables (see Section 5):

- The credit score, while its impact is diminishing, remains a significant factor. It stands out as the only sensitive variable that telematics data cannot fully negate, affecting both frequency and severity. Figure 2.1 illustrates this conclusion for frequency and the severity.
- Notably, the insured's age, which was once a key factor, now appears to have lost its significance in our analysis for both frequency and severity. Figure 2.2 illustrates this conclusion for frequency and severity.
- Similarly, the territory variable, except for one or two regions, no longer exerts a substantial influence on frequency and severity, as Figure 2.3 shows.
- The effect of marital status on both response variables (frequency and severity), which initially seemed weak in the data, is reduced, or even canceled, by adding telematics variables, thus confirming the lack of interest in this variable in the presence of telematics data (Figure 2.4).
- The effect of the insured's sex on both response variables (frequency and severity) is reduced, even if it does not seem statistically significant in our data, as illustrated in Figure 2.5.

Figure 2.1. Impact of Credit Score Before (Red Line) and After (Blue Line) Adding Telematics Variables in an XGBoost Model



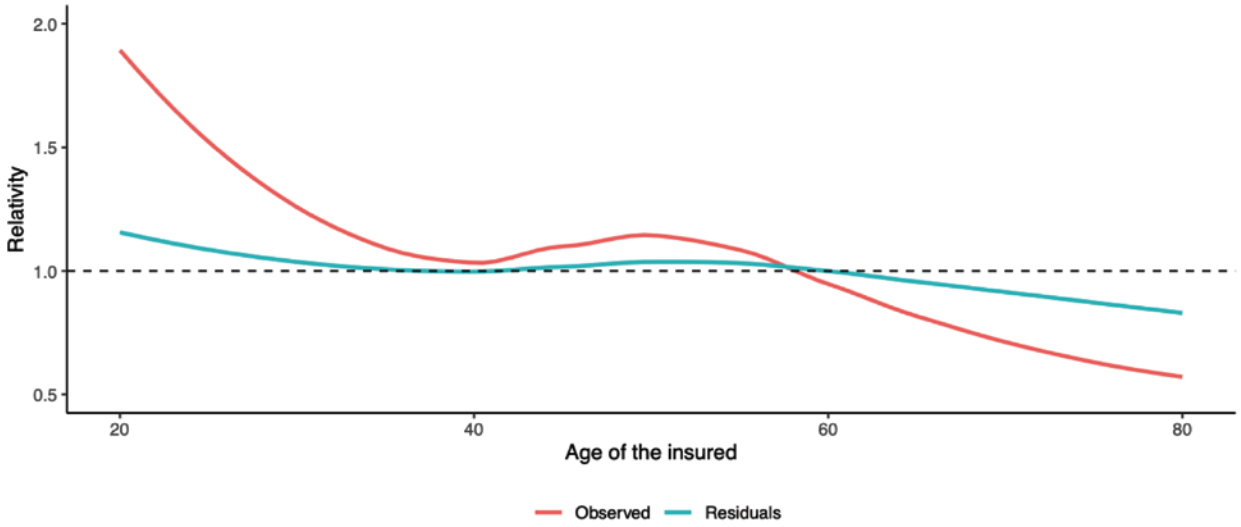
(a) Frequency



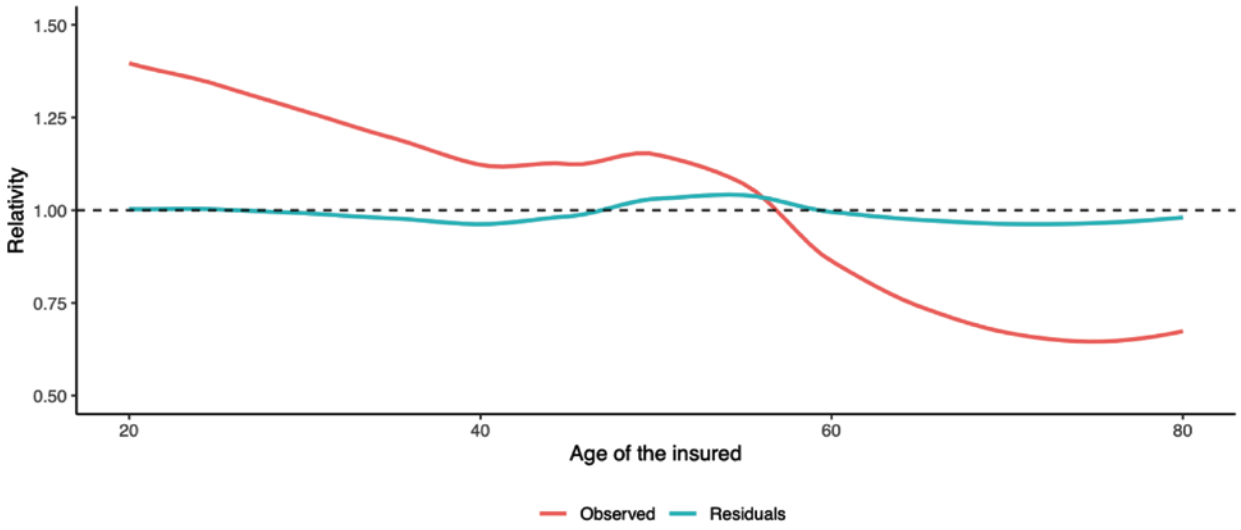
(b) Severity

Note: This type of chart illustrates the predictive power of a covariate (here the credit score) without (red line) and with (blue line) adding telematics variables. A horizontal line indicates that the covariate no longer has predictive power and therefore is no longer useful in the model.

Figure 2.2. Impact of Insured's Age Before (Red Line) and After (Blue Line) Adding Telematics Variables in an XGBoost Model

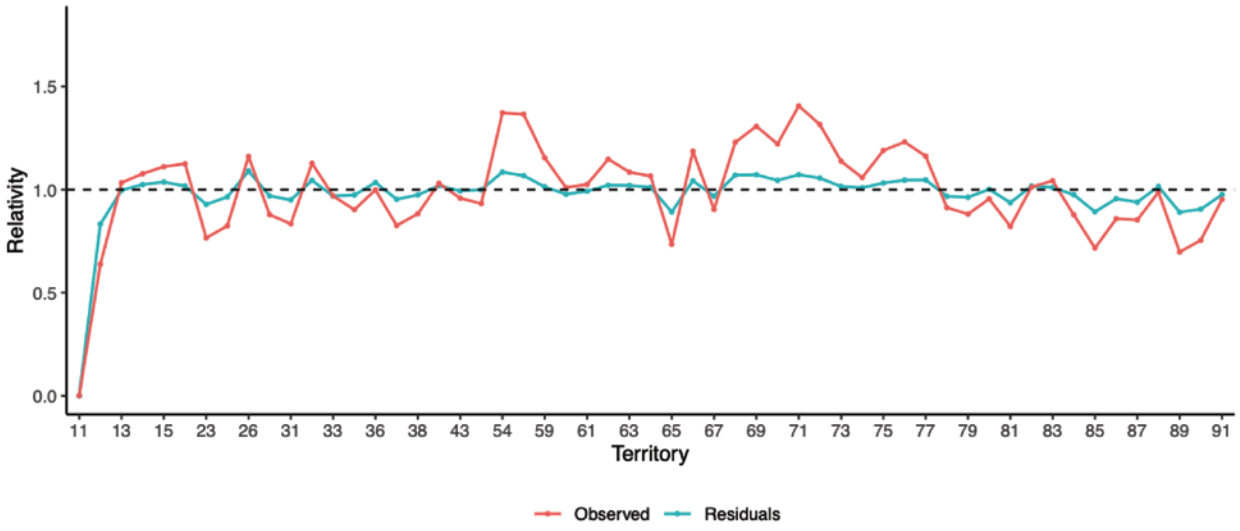


(a) Frequency

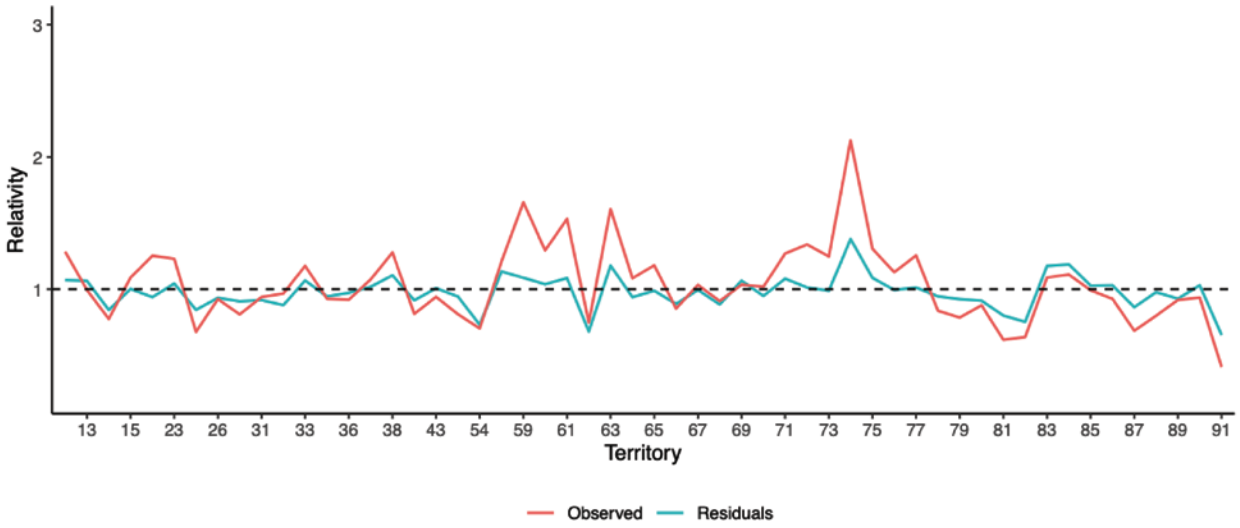


(b) Severity

Figure 2.3. Impact of Territory Before (Red Line) and After (Blue Line) Adding Telematics Variables in an XGBoost Model

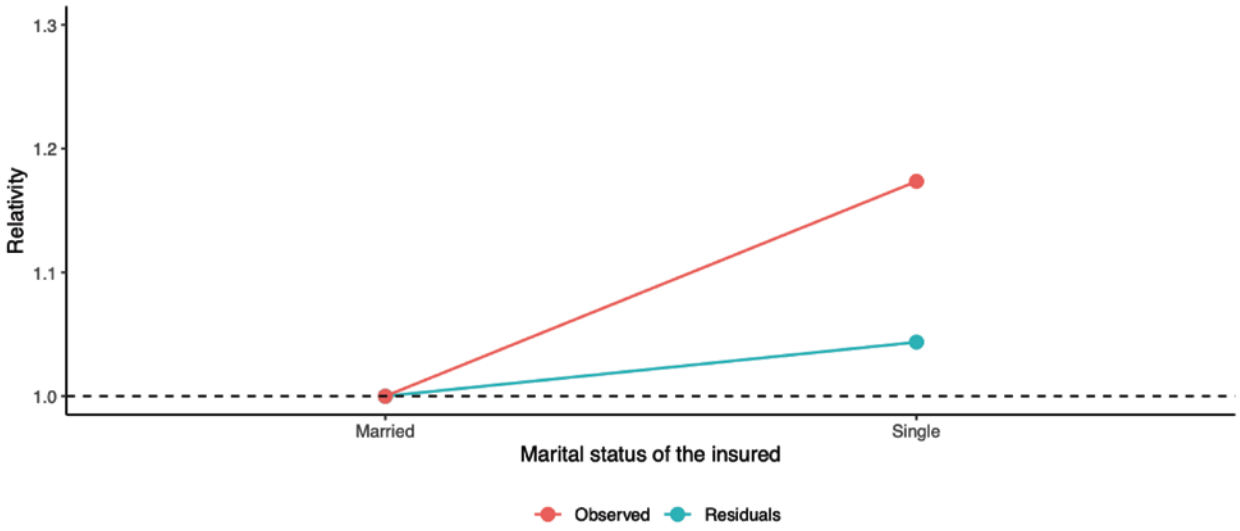


(a) Frequency

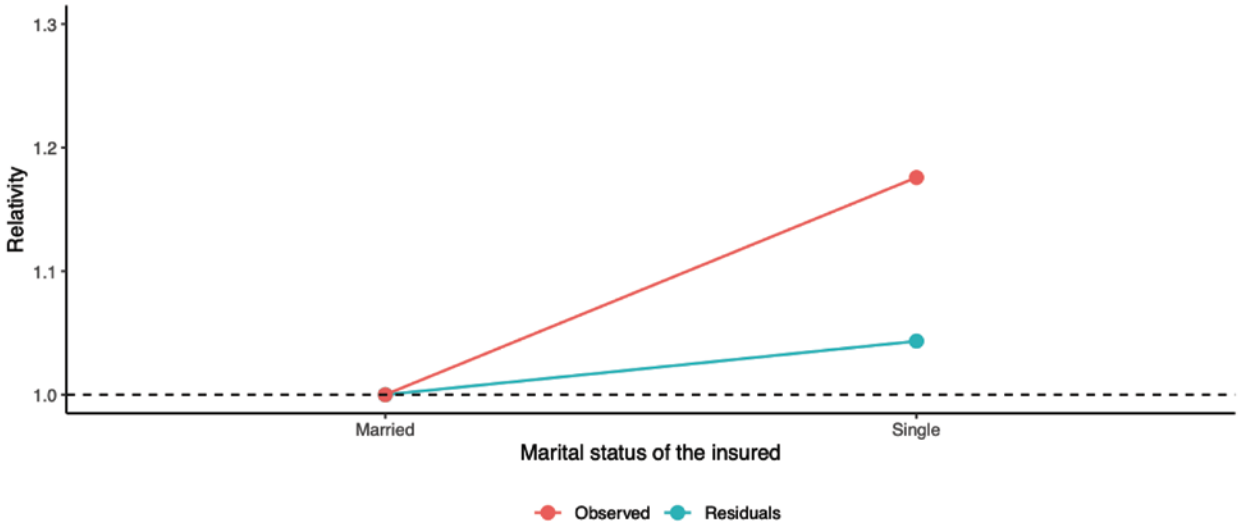


(b) Severity

Figure 2.4. Impact of Marital Status Before (Red Line) and After (Blue Line) Adding Telematics Variables in an XGBoost Model

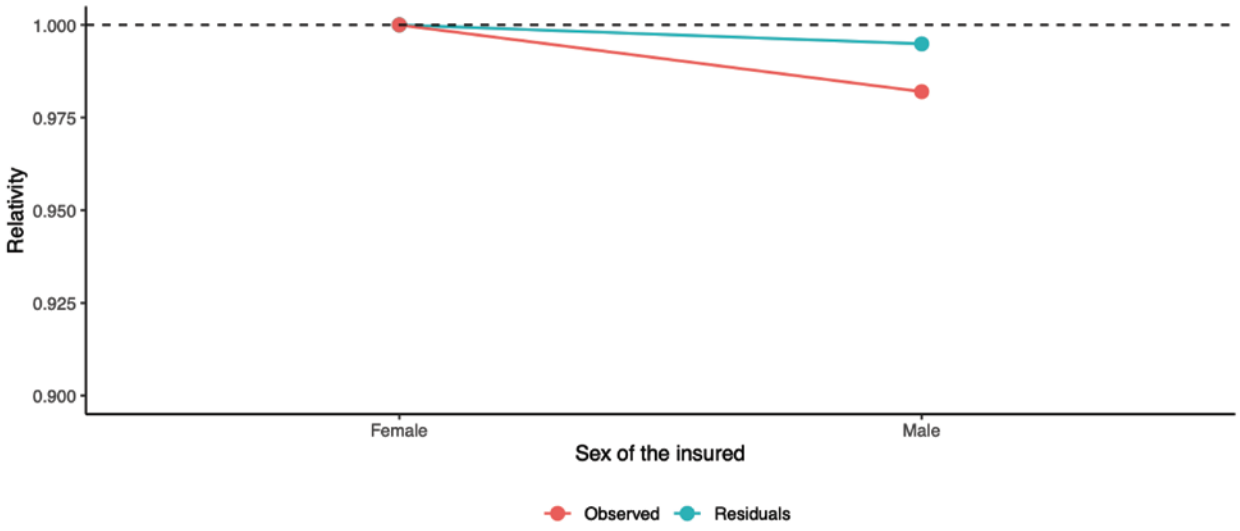


(a) Frequency

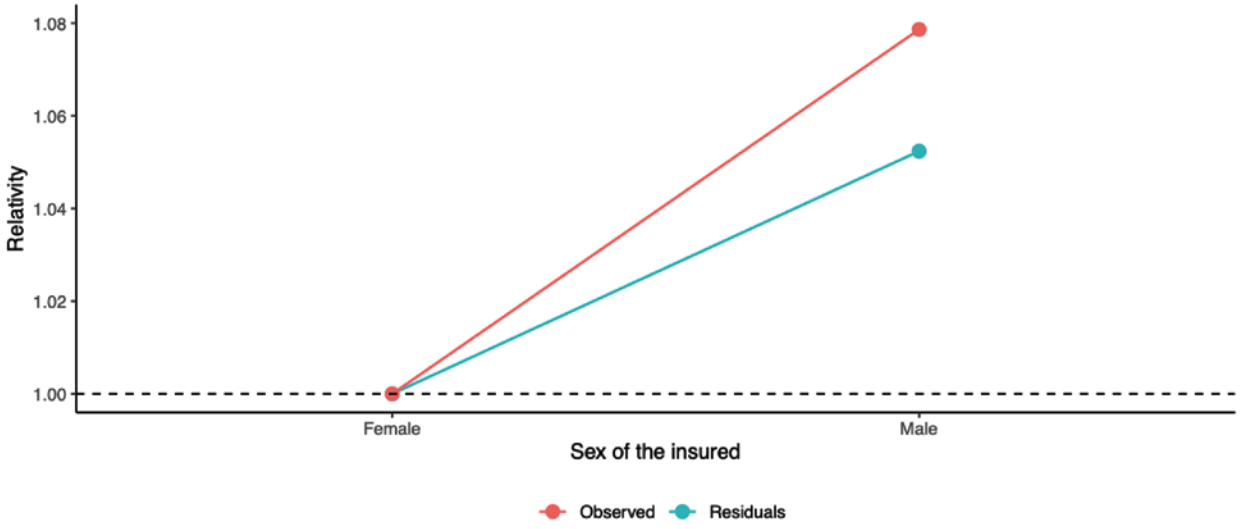


(b) Severity

Figure 2.5. Impact of Insured's Sex Before (Red Line) and After (Blue Line) Adding Telematics Variables in an XGBoost Model



(a) Frequency



(b) Severity

The avoidance of unfair discrimination in the insurance industry is a pertinent issue, as age and marital status are often considered protected information in many jurisdictions and industries. The use of telematics variables to derive the same information could provide a feasible solution to the problem. Such an approach could prove particularly beneficial for certain groups, such as younger groups or those affected by historical systemic barriers that have led to significant differences in marriage rates. By relying on telematics variables, the industry can reduce the risk of unfair discrimination, thereby enabling a fair and just outcome for all insured.

The effectiveness of telematics variables in replacing sensitive variables appears more remarkable when an XGBoost model is used, as compared to the GLM. This

conclusion is not particularly startling, as the XGBoost model has greater flexibility than the GLM-net model. That is primarily because the former incorporates a larger number of parameters and offers the ability to account for interactions between covariates.

Having conducted our analysis on a synthetic database constructed from an actual database from a Canadian insurer, we determined that we should validate the main conclusions on the original insurer database. Most conclusions obtained regarding the impact of telematics on the usefulness of sensitive variables remain valid on the original database, but the effect is sometimes less significant (see Section 6). For example, Figures 2.6, 2.7, 2.8, and 2.9 illustrate the impact of credit score and territory on frequency in the synthetic and in the original data set. In both cases, we observed a similar effect for severity. Although this paper proposes a methodology that is based on synthetic data, it is essential to note that individual insurers may arrive at different conclusions when using actual data or different rating variables.

In the following sections of the paper, we trace our analytical approach, including

- Section 3: evaluation of available covariates, including traditional nonsensitive, traditional sensitive, and telematics covariates;
- Section 3: transformation of covariates where needed;
- Section 4: construction of GLM and XGBoost models with only traditional covariates;
- Section 5: construction of GLM and XGBoost models including telematics variables and excluding traditional sensitive covariates; and
- Section 6: validation of conclusions from the synthetic data set on the original data set.

Figure 2.6. Credit Score – Synthetic Data Set

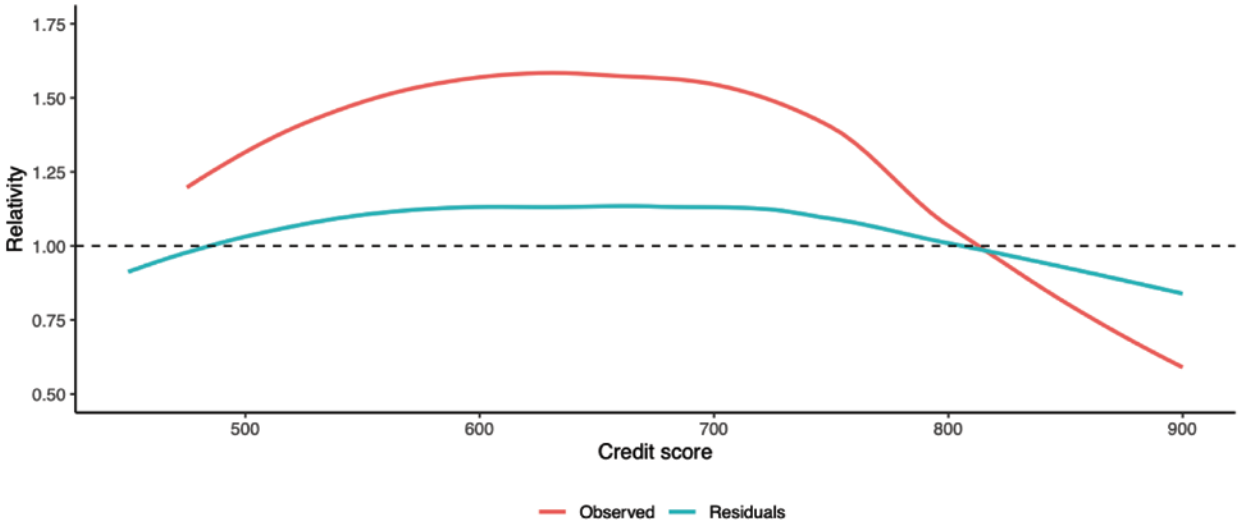


Figure 2.7. Credit Score – Original Data Set

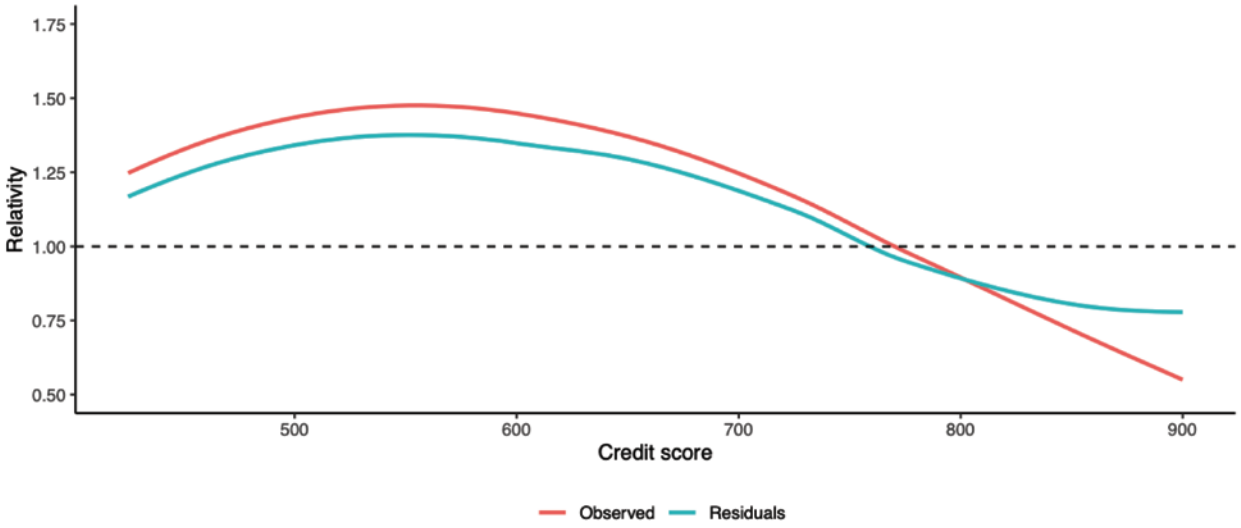


Figure 2.8. Territory – Synthetic Data Set

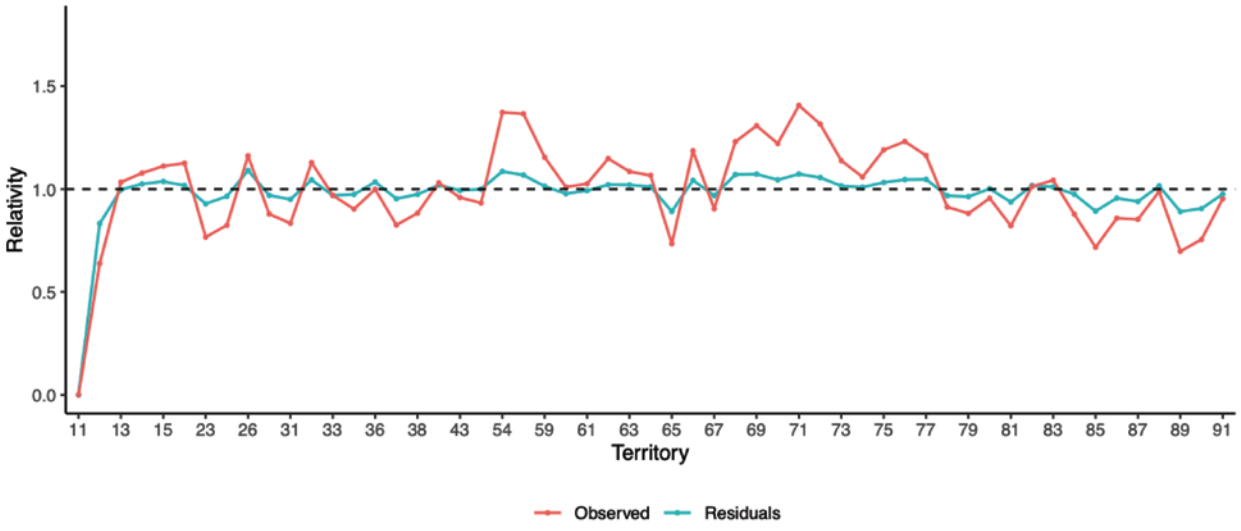
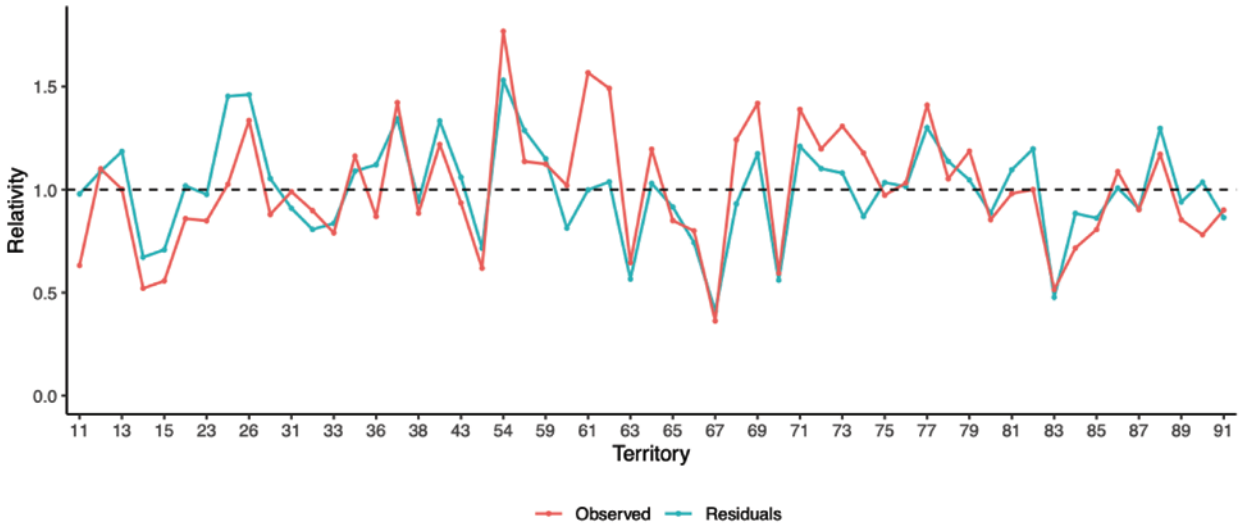


Figure 2.9. Territory – Original Data Set



3. Data Summary

Before delving into the development of more advanced statistical models for variable selection and parameter estimation, it is pertinent to analyze all the covariates available in the data set used for this study. We thus conduct an individual and more in-depth analysis for each of these segmentation variables. This approach will notably allow us to better understand how these covariates could explain the risk of accidents and to propose transformations or groupings of modalities. We cover all segmentation variables available in the database. We have divided variables into different categories:

1. Section 3.1: Traditional covariates
   - Section 3.1.1: Traditional (nonsensitive)
   - Section 3.1.2: Sensitive information
   - Section 3.1.3: Contract duration
2. Section 3.2: Telematics covariates
   - Section 3.2.1: Vehicle usage level
   - Section 3.2.2: Type of vehicle usage
   - Section 3.2.3: Quality of driving

We emphasize that the covariates under study will all exhibit strong correlations among themselves. For instance, younger insured individuals will likely have different credit score distributions than older insured individuals, driving behaviors of men may differ from driving behaviors of women, or the total distance traveled by an insured individual will be linked to the duration of their contract. Thus, the marginal impact of each segmentation variable may partly be explained by the effects induced by other covariates.

3.1. Traditional Covariates

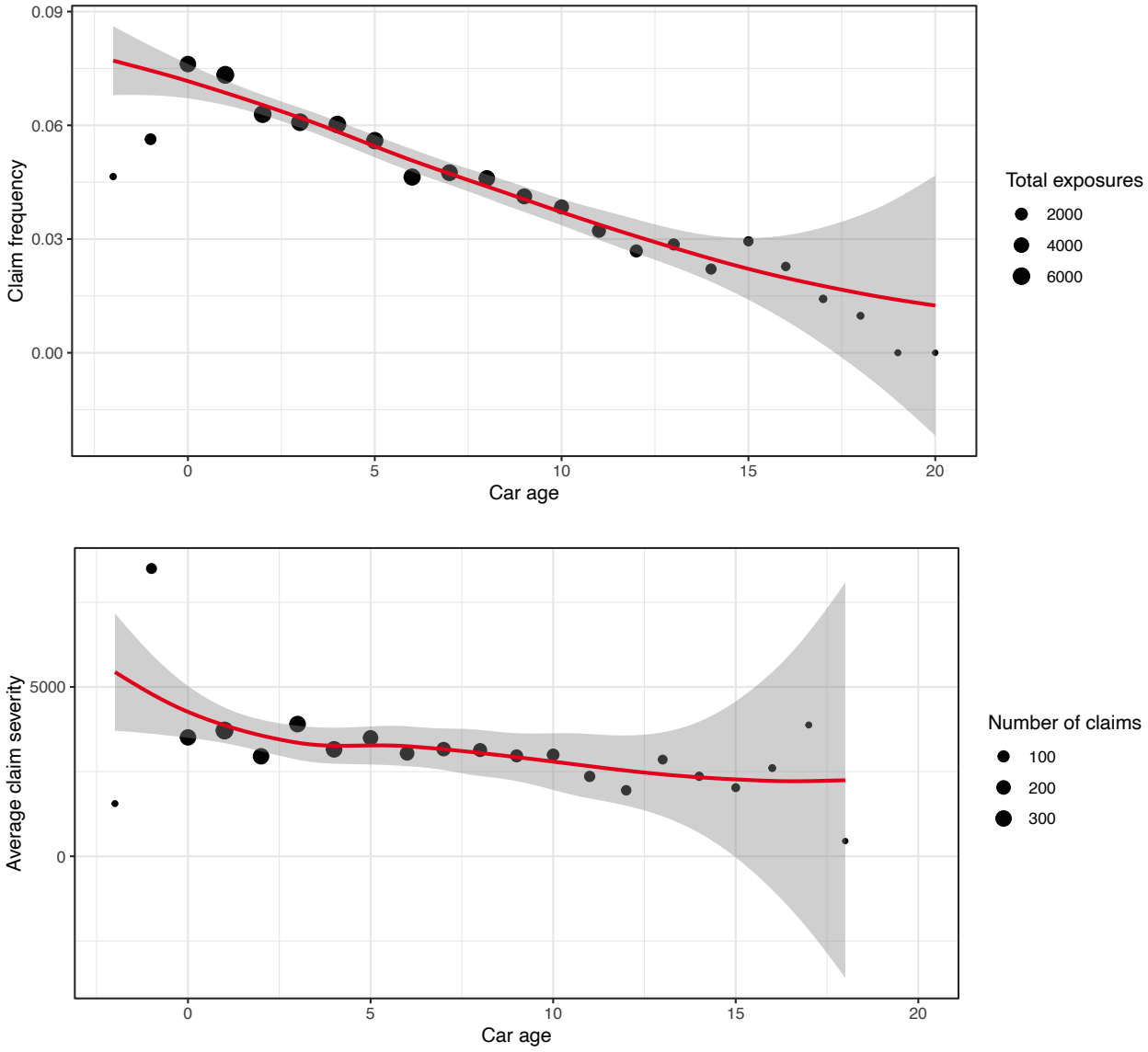
3.1.1. Traditional (Nonsensitive)

We start our analysis by examining nonsensitive traditional segmentation variables, i.e., socially accepted variables in automobile pricing.

3.1.1.1. Age of the Car

Figure 3.1 (top) highlights the relationship between claim frequency and the age of the vehicle. A fairly clear link is observed: the newer the vehicle, the higher its risk

Figure 3.1. Car Age



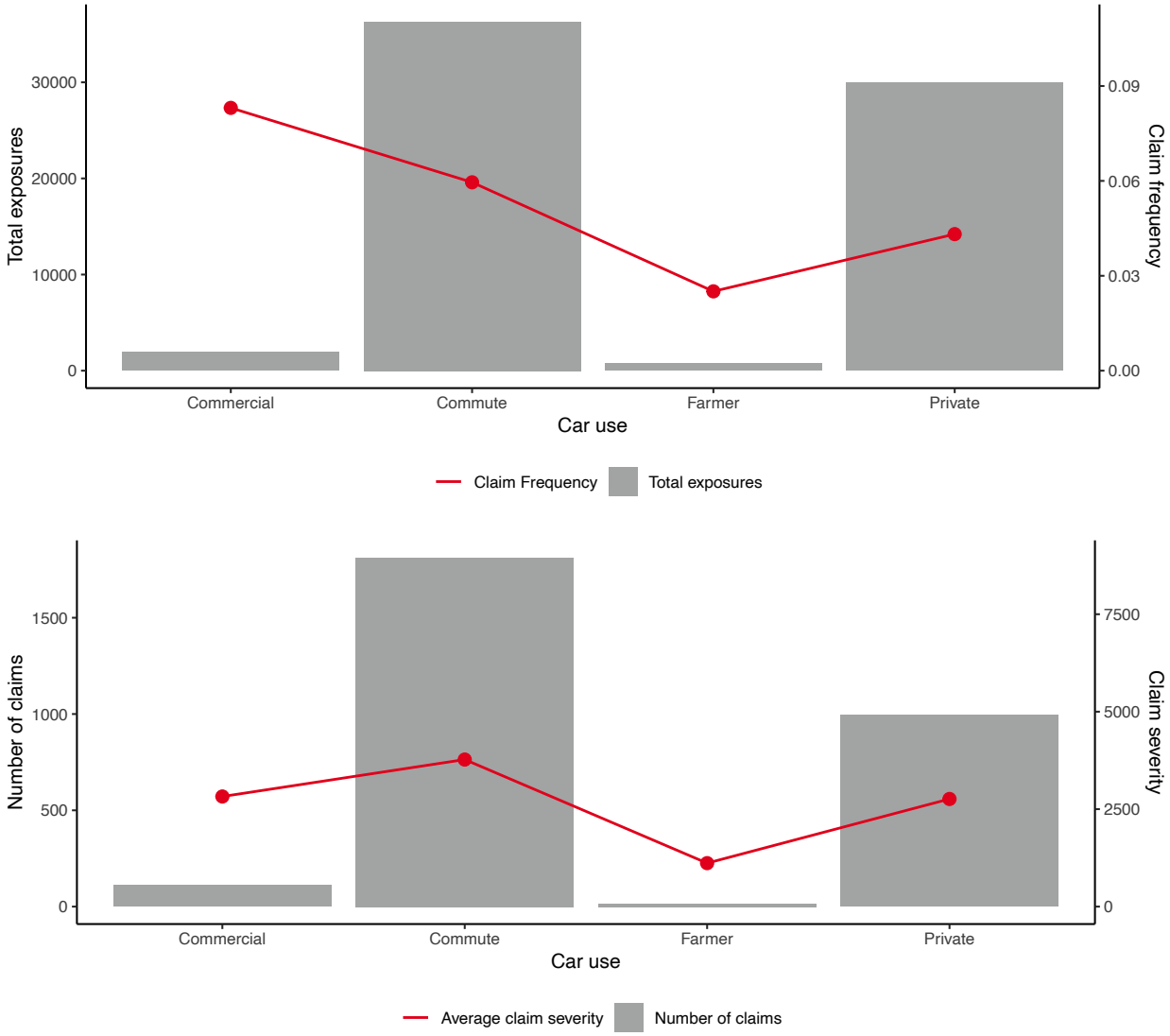
Note: This type of chart illustrates a relationship between a continuous covariate (horizontal axis) and a response variable (vertical axis). The red line represents a smoothing of the data (central trend), and the gray zone is a 95% confidence interval, i.e., an area where we find 95% of the data.

of having an insurance claim. For claim severity (Figure 3.1 [bottom]), we observe the same relationship.

3.1.1.2. Use of the Car

In the data used, we have categorized vehicle usage into four levels: commercial, commute, farm, and private. We observe some difference in the average number of claims between each of the categories, as Figure 3.2 (top) shows. It is also important to note a significant difference in the distribution of this variable: the vast majority of insured individuals are in the commute and private groups. Thus, even if the claim frequency of insured individuals in the commercial and farm groups differs, the small number of contracts in these groups likely limits the predictive capacity of this variable.

Figure 3.2. Use of the Car



Note. This type of chart illustrates a relationship between a categorical covariate (horizontal axis) and a response variable. The red line represents the averages for the different categories (right scale), and the gray area measures volume (left scale).

In Figure 3.2 (bottom), we see that the use of the vehicle has an impact on the average claim severity. We notice a very low exposure for two categories: farm and commercial. If we focus only on the other two categories, the difference remains significant (the *p*-value is very close to 0). However, this effect seems to disappear if we consider the car's age because the test's *p*-value exceeds the generally accepted limit of 5%.

3.1.1.3. Region

In addition to territory, a slightly coarser grouping of the insured individual's residence has been done for insurance contracts: urban and rural. It is important to note that this variable was not constructed from territories. Thus, a territory can contain observations for which the region is "rural" and "urban." The difference in claim frequency between these two types of insured persons is illustrated in Figure 3.3 (top). According to Welch's two-sample t-test, the difference in means is significant (*p*-value very close to 0).

Figure 3.3 (bottom) shows the difference in severity between urban and rural claims. The *p*-value is very close to 5%, and we conclude there is no significant difference between these two region types.

3.1.1.4. Years Without Claim

The variable tracking the number of claim-free years is also of critical importance in modeling risk in automobile insurance. According to various papers (see, for example, Lemaire 1985, ch.7), this variable is considered particularly significant to the extent that if only one segmentation variable were to be used in ratemaking, it would likely be something related to the insured's past claims experience.

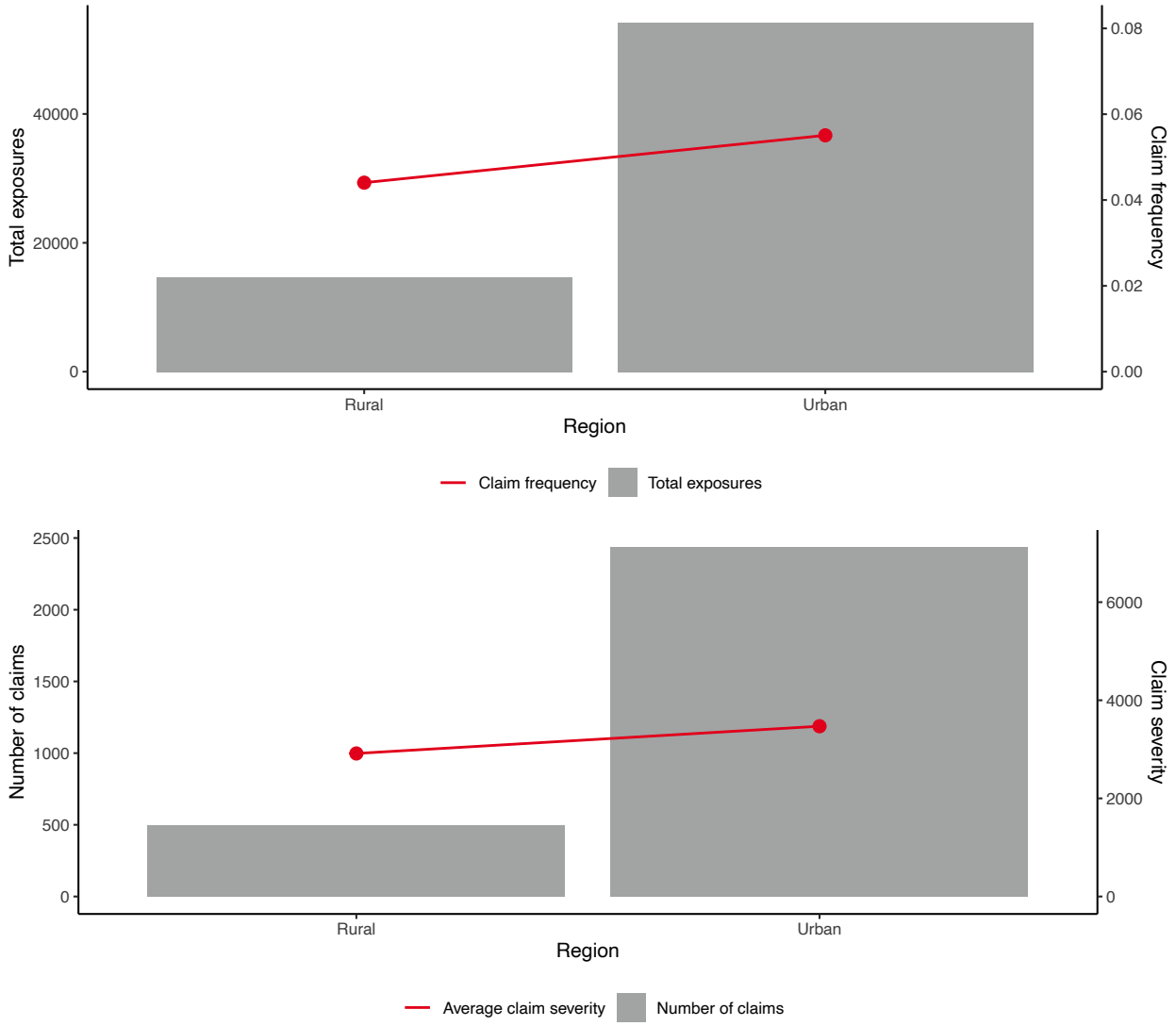
Figure 3.4 (top) illustrates the relationship between the number of claim-free years and the number of reported automobile claims. As indicated by numerous studies, a decreasing relationship can be observed, suggesting that insured individuals who have had few or no claims in the past are also likely to make fewer claims in the future. As many jurisdictions in the United States limit how many years a claim can affect rating, we also analyzed by limiting the history to a maximum of five years, and the conclusions were similar: a decrease in risk with, for example, the frequency going from about 10% to about 5%.

We see a negative trend in Figure 3.4 (bottom), but it seems weaker than in the case of frequency and presents larger variability. The empirical correlation between these two variables is –0.13.

3.1.1.5. Correlation

In Figure C.1 (Appendix C), we can observe the correlation that exists between the traditional segmentation variables. The correlation matrix on the left indicates a strong dependency between vehicle usage and the number of claim-free years. On the right, emphasis is placed on the relationship between the covariates studied in this subsection and the sensitive variables. Thus, we observe a strong dependency between the age of the insured and the number of claim-free years. Vehicle usage is also linked to the age of the insured. We present the same matrices in Figure C.2 for the severity.

Figure 3.3. Region

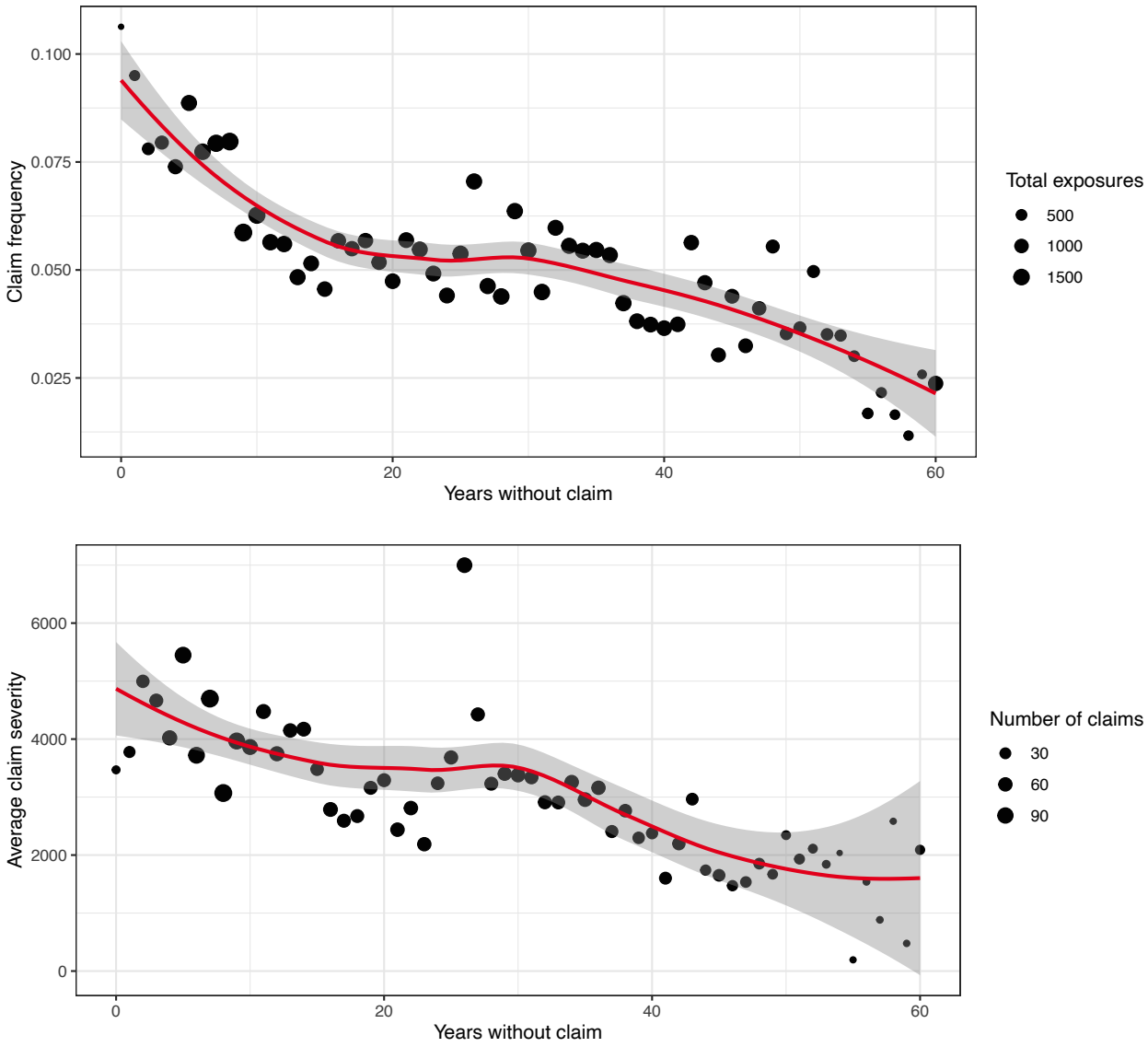


3.1.2. Sensitive Information

3.1.2.1. Credit Score

In Canada, the credit score for auto insurance ratemaking is based on the score the main credit bureaus (Equifax and TransUnion) calculate for mortgages and loans, as opposed to a score constructed specifically for insurance rating. The main factors that may affect the score include how long a person has had credit, how long each credit has been in their report, if they carry a balance on their credit cards, if they regularly miss payments, the amount of their outstanding debts, being close to, at, or above their credit limit, the number of recent credit applications, the type of credit they are using, if their debts have been sent to a collection agency, and any record of insolvency or bankruptcy. In the United States, this is different, because insurers mainly use a credit-based insurance score whose definition can vary greatly from one company to another. Figure 3.5 (top) shows the distribution of the credit score in the insurance portfolio: the points indicate the observed claim frequency, the size of the points

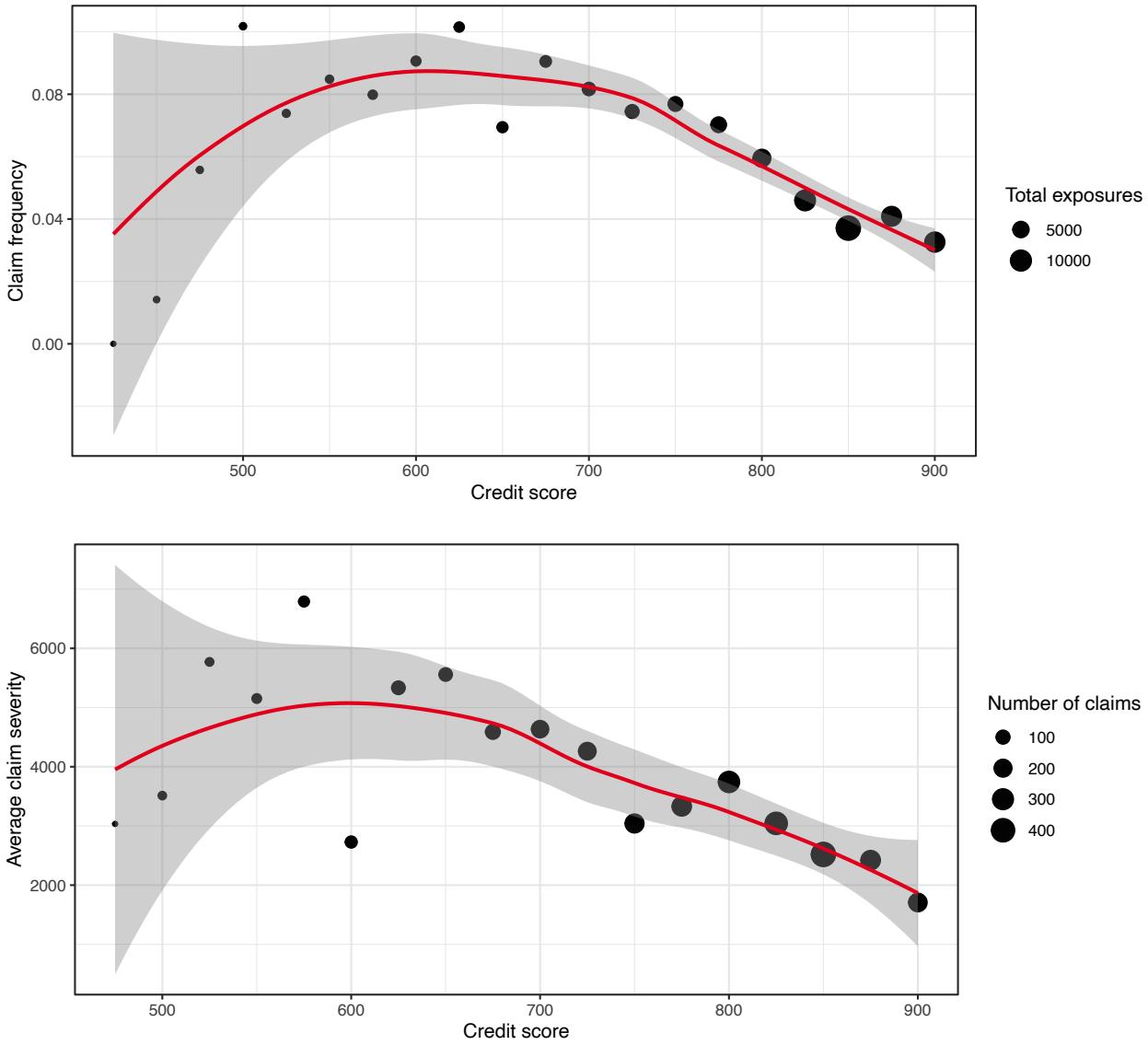
Figure 3.4. Years without Claim



measures the total exposures for each group, and a trend curve (in red) for the frequency has been added. Thus, we can see that the relationship between the number of claims and the credit score is not linear, but generally, a better credit score implies a lower claim frequency.

The link between frequency of auto insurance claims and credit scores has been previously examined. Wu and Guszcza (2003) concluded that this relationship persisted even after controlling for numerous other variables. Despite the statistical association between claims experience and credit score, establishing causality remains elusive. Some argue that a correlation exists between responsible financial behavior and safe driving, or that people with higher credit scores may have easier access to newer vehicles equipped with better safety features, potentially reducing accident risk. Others suggest that the use of credit scoring appears to target young drivers lacking established credit histories, new immigrants, or generally economically disadvantaged populations. Coupled with the opacity of credit

Figure 3.5. Credit Score



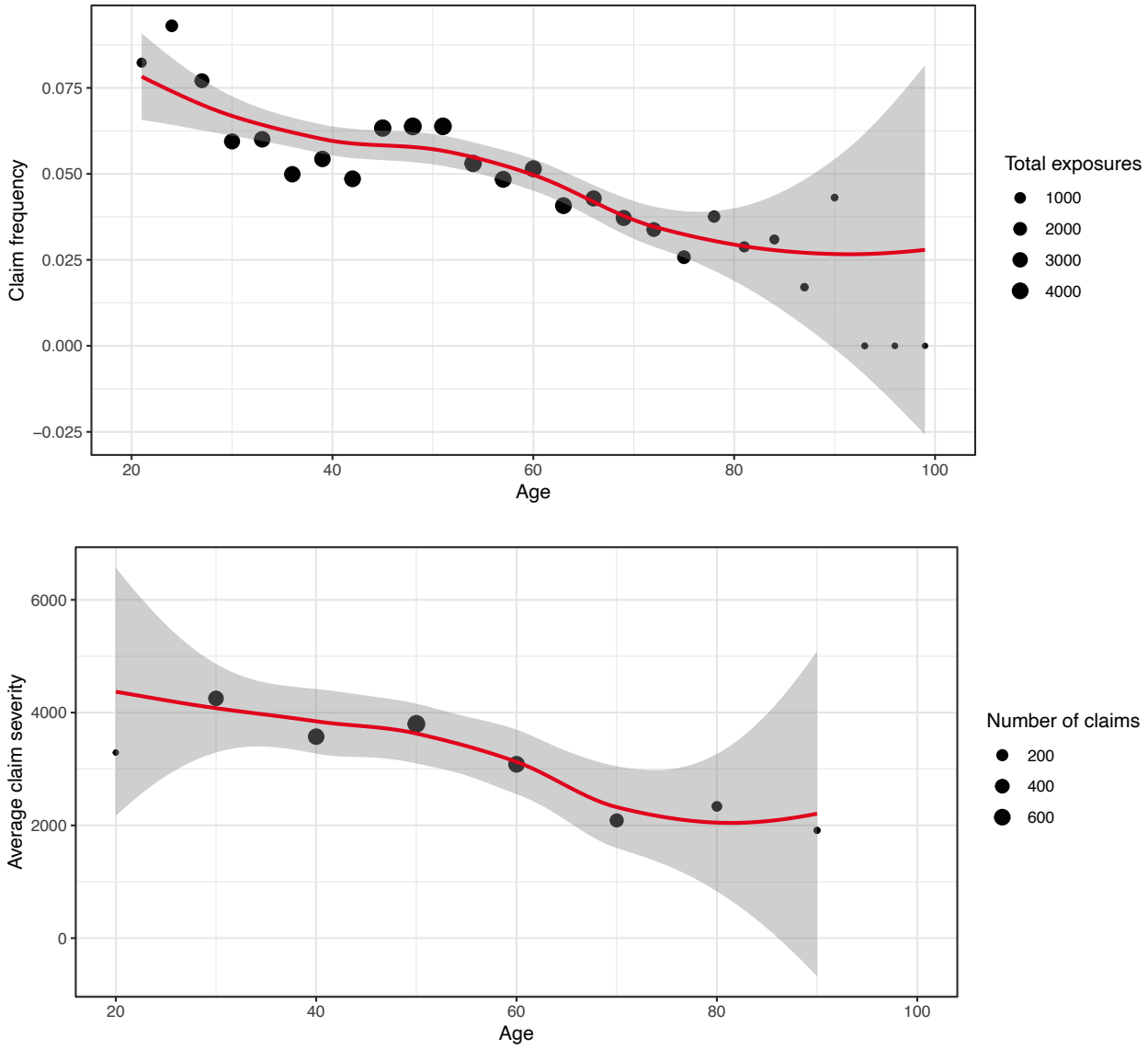
score calculation by private entities, it is reasonable to view credit score as a sensitive variable warranting scrutiny in automobile pricing.

Figure 3.5 (bottom) shows the relation between credit score and claim severity in the portfolio: in this case, the points indicate the observed claims' average severity, the size of the points measures the number of claims for each group, and a trend curve (in red) for the average severity has been added. As we did with frequency, we observe a nonlinear relationship, but overall, a better credit score implies a lower claim severity.

3.1.2.2. Age

Like the credit score, the age of the insured is a sensitive variable in ratemaking. Figure 3.6 (top) illustrates the fairly strong negative relationship between the age of the person insured

Figure 3.6. Age of the Insured

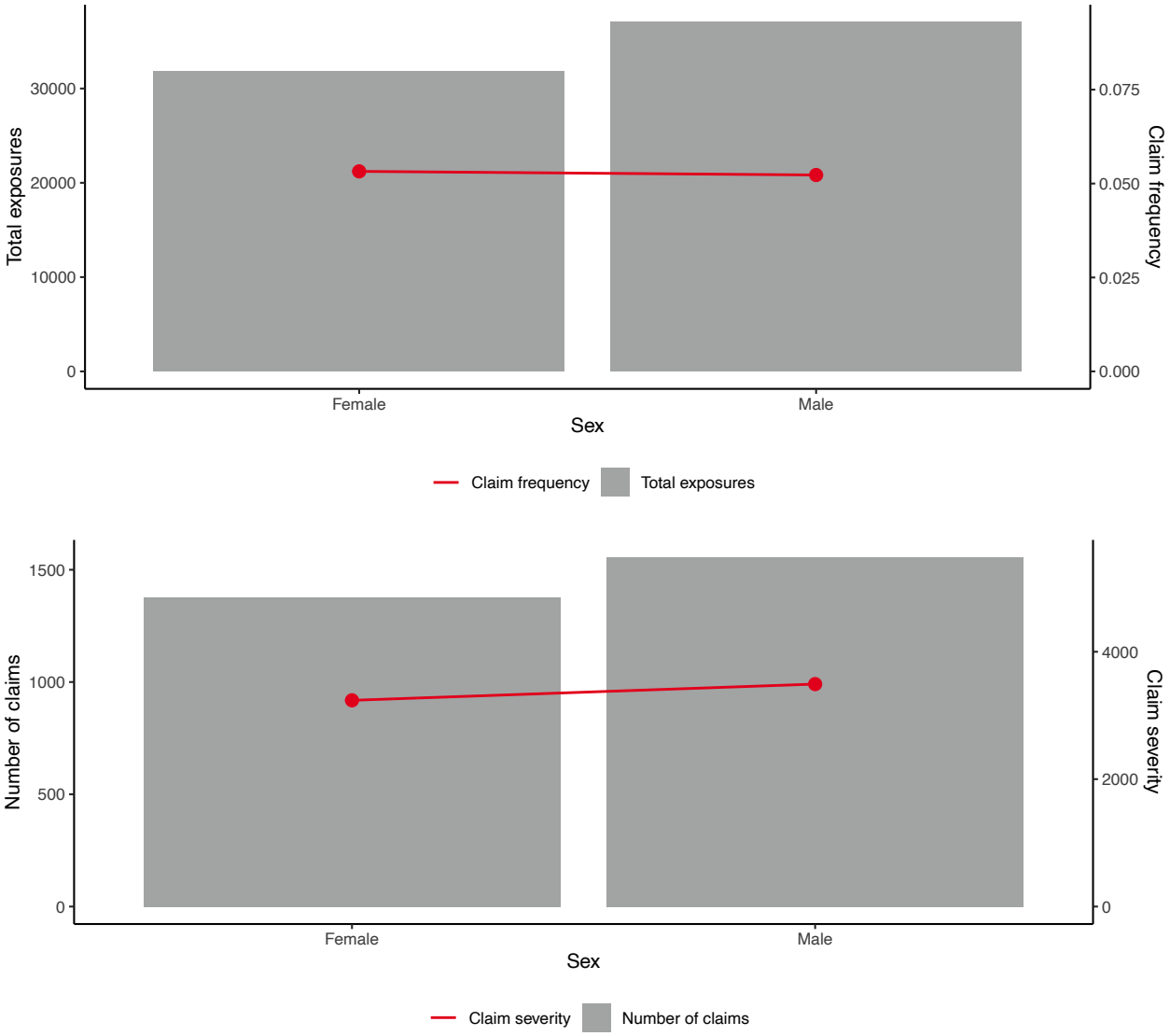


and their claim frequency. As we discussed earlier, age is a controversial segmentation variable. Indeed, one might argue that there is no causal relationship between age and the likelihood of having an accident, and that age is rather used to identify drivers who may lack driving experience, who could be more reckless, and who might be less mature. We can hypothesize that driving behavior measured by telematics devices could better identify such drivers. Figure 3.6 (bottom) illustrates the negative relationship between age and claim severity.

3.1.2.3. Sex

Figure 3.7 (top) shows the distribution of the insured individual's sex in the portfolio (in bars), as well as the observed claim frequency for each of the two groups (in red). It can be seen that there are more men in the portfolio. It can also be observed that the claim frequency

Figure 3.7. Sex of the Insured



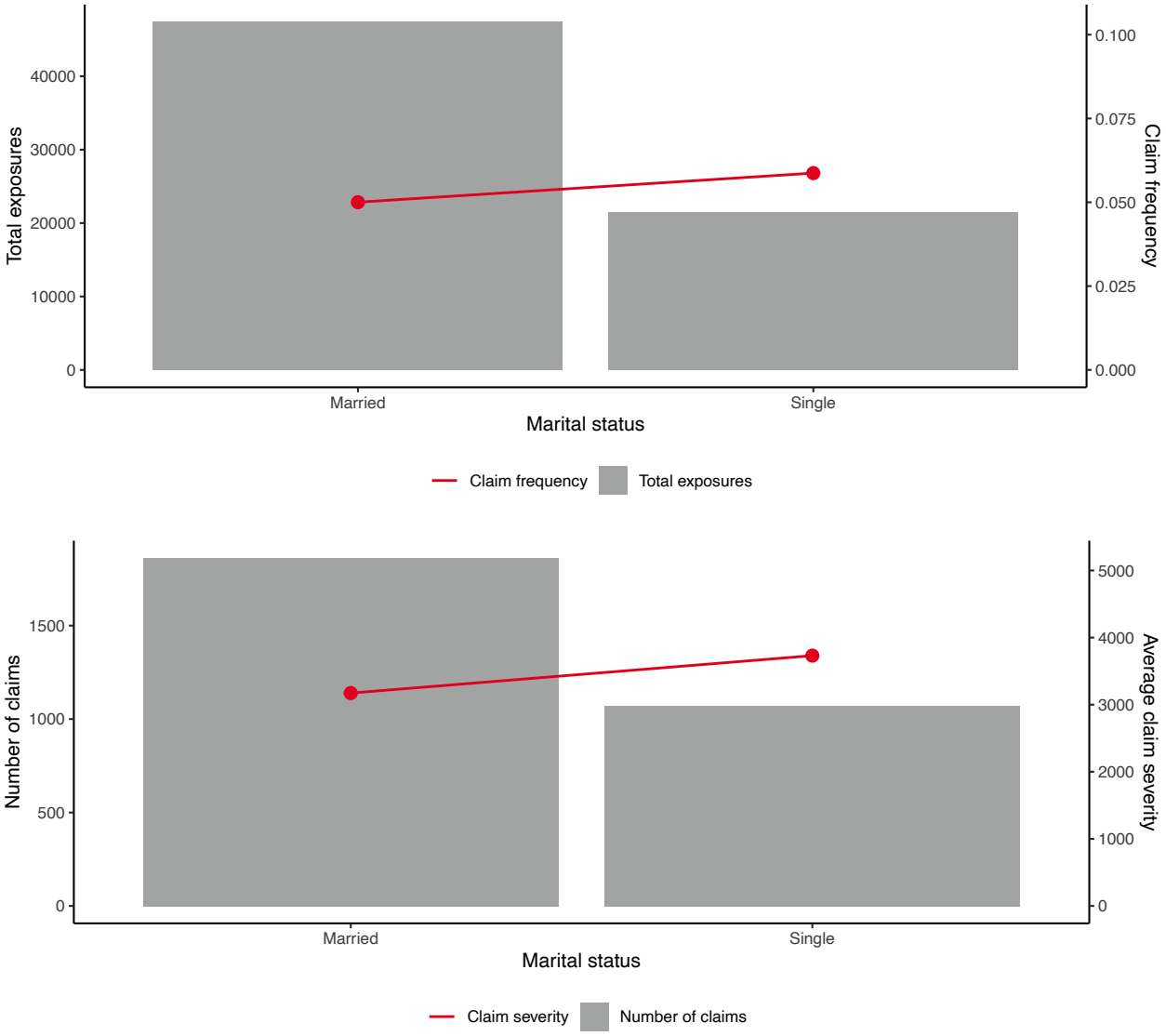
for both men and women is highly similar. Using Welch's two-sample t-test, we obtain a *p*-value of 0.2705, meaning that the null hypothesis – the true difference in means is equal to 0 – cannot be rejected, or, at least, can be rejected with an error probability of 27.05%, which is considerably higher than the generally accepted maximum error probability (5%).

Figure 3.7 (bottom) shows a similar result for the observed average claim severity (in red). The *p*-value of Welch's two-sample t-test is 0.2573. Thus, this variable does not seem to have a significant impact on either of the response variables (frequency and severity) in the synthetic database.

3.1.2.4. Marital Status

The marital status of the insured is not one of the more important segmentation variables (see Figure 3.8, top). However, a statistical test rejects the null hypothesis (the true difference

Figure 3.8. Marital Status of the Insured



in means is equal to 0), so we keep this variable in our analysis. We see in Figure 3.8 (bottom) that marital status has minimal impact on the average claim severity.

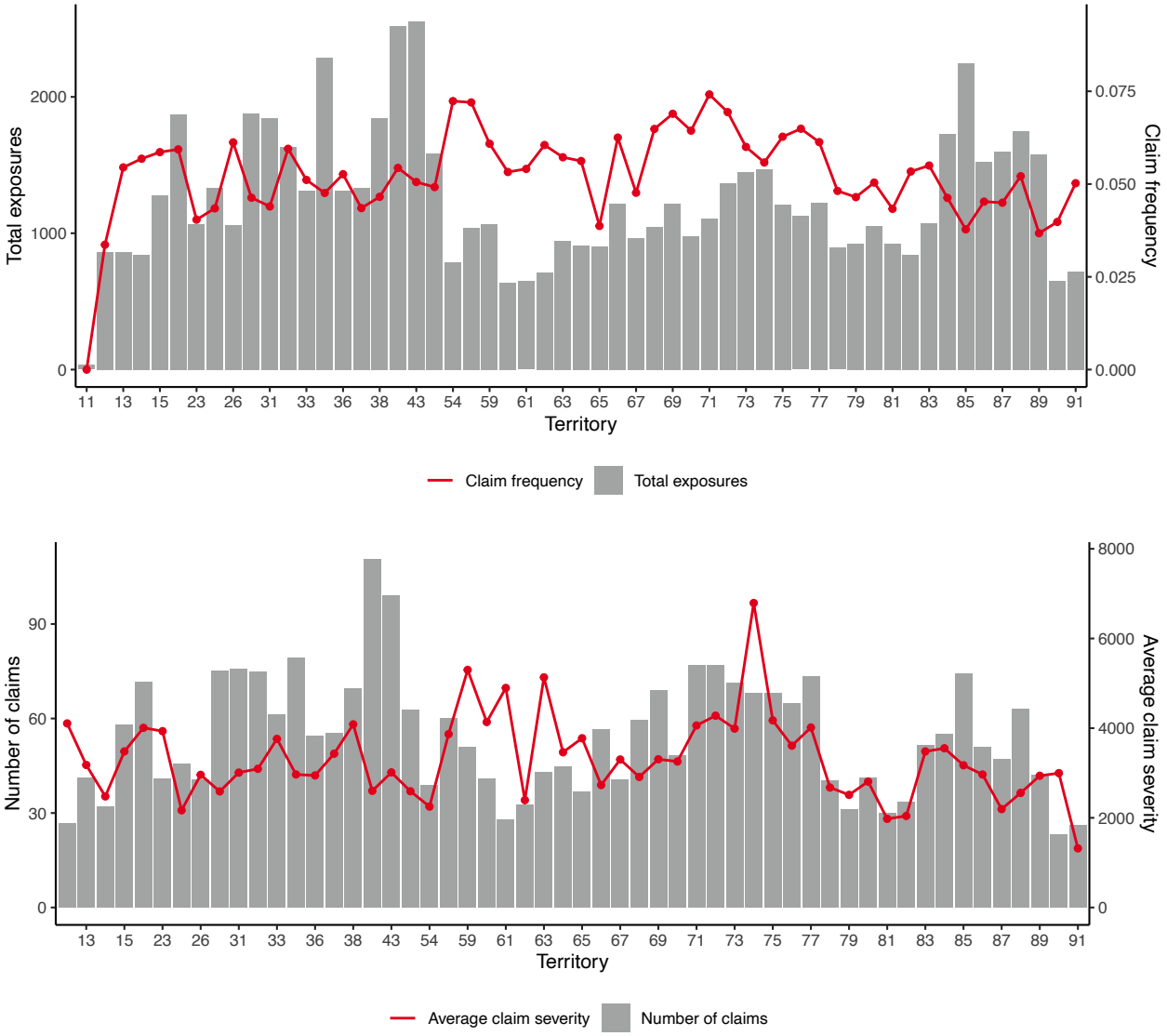
3.1.2.5. Insured's Territory

The pricing of automobile insurance based on the territory of usage is common in many insurance companies worldwide. The approach relies on the principle that the risk incurred by a driver may vary depending on the geographical location where they most frequently drive. Thus, insurers may assess the risks associated with different geographical areas, considering traffic density, crime rates, weather conditions, and road quality. In practice, territories in automobile insurance are often chosen based on practical criteria such as blocks of streets, the same postal code, and divisions created by a river or highway.

Therefore, while using territory in automobile insurance pricing may seem like an objective method of evaluating risks, it can also have significant social implications. Indeed, it has been shown that this approach can contribute to dividing population groups based on criteria such as race, economic level, or even profession. For example, if disadvantaged or ethnically predominant urban neighborhoods are grouped into one territory, while more affluent neighborhoods are grouped into another territory, insurers may inadvertently be building socioeconomic criteria into their pricing models.

To protect the privacy of the insured in the database, their specific addresses and postal codes are not available. Instead, a variable known as *Territory* is used, which is represented by a numerical value ranging from 11 to 91. However, that numerical representation makes it challenging to interpret or correlate with current public data. Figure 3.9 (top) demonstrates the influence of territory on the modeling of claim numbers.

Figure 3.9. Territory



Upon analysis, we find that territory does not have a significant impact on claim frequency, except for category 54. In the absence of this category, a statistical test known as the *F*-test fails to reject the null hypothesis, indicating that territory has no effect on the response variable (with a *p*-value of 0.064). However, due to the abstract nature of the *Territory* variable, it is challenging to interpret its influence or relate it to current public data.

We obtain a similar conclusion (Figure 3.9, bottom) using the average severity as the response variable: territory does not significantly affect claim severity except for categories 74 and 91. This is an unexpected conclusion in both cases as we would assume the insurer's territories were designed based on some observed differences. However, we assume that differences may in fact exist, but the limited data and the large number of territories ensure that they are not statistically significant.

3.1.2.6. Correlation

Finally, we present correlation matrices for frequency (Appendix C, Figure C.3, left) and severity (Fog C, Figure C.3, right). They show the level of dependence between each of the sensitive variables. For instance, it can be observed that the insured's age is correlated with the credit score and marital status.

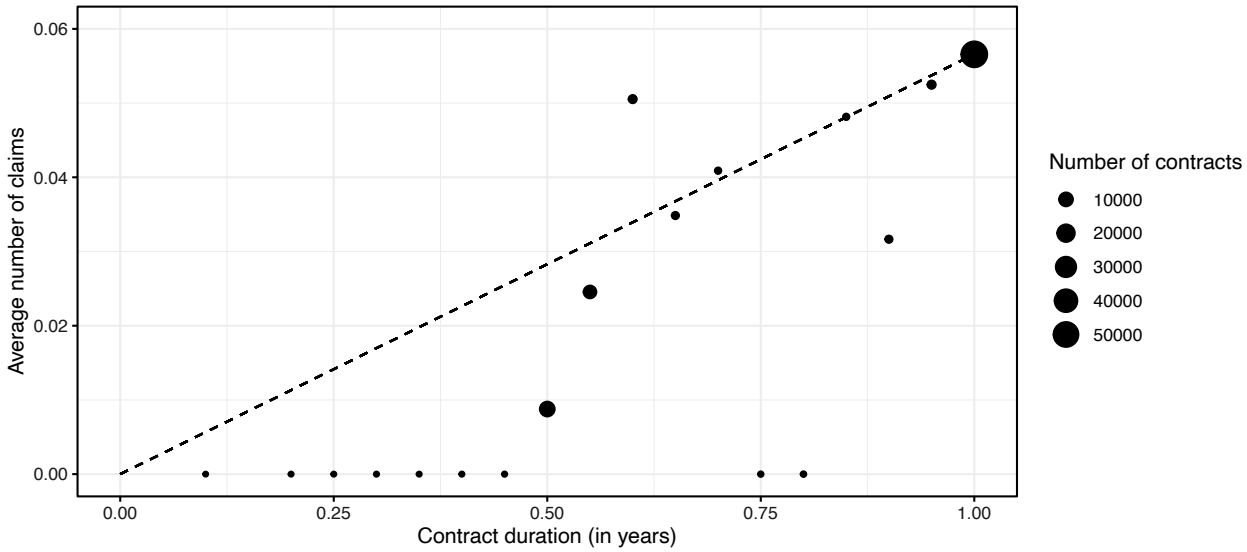
3.1.3. Contract Duration

Special attention should be given to the variable representing the duration of the contract. In traditional ratemaking models based on count distributions, such as the Poisson distribution, it usually is assumed that the duration of the contract is an offset variable rather than a covariate for which a parameter needs to be estimated. For example, when using an offset variable, if an insured individual is covered for only half of the year, the insurer will offer them a premium that is half the size of what it would be for full-year coverage. Some papers question this approach and introduce contract duration as a covariate (Boucher and Denuit 2007; Duval, Boucher, and Pigeon 2022, 2023a, 2023b), suggesting that an insured individual covered for half the year should have a premium that is either larger or smaller than half of that for a full year.

To stimulate further thought, the numerical analysis in Figure 3.10 visualizes the average number of claims observed based on groupings of the duration of the contract. Although we observe an increase in the number of claims based on the contract duration, we also note that the relationship is not perfectly linear compared to the dashed line. But to remain consistent with traditional pricing approaches, we will keep contract duration as a measure of exposure to risk for now. However, with the telematics information available, the distance driven is also available as an alternative for consideration for representing exposure to risk in automobile insurance.

As duration is often not considered in the modeling of severity, we don't use it for the severity analysis.

Figure 3.10. Average Number of Claims vs. Contract Duration



3.2. Telematics Covariates

3.2.1. Vehicle Usage Level

Many scientific articles (see Appendix A) have shown that what appears to be the most relevant telematics information for pricing is not the quality of driving but rather the level of vehicle usage. In this first part of the telematics variables analysis, we therefore focus on vehicle-usage-level variables.

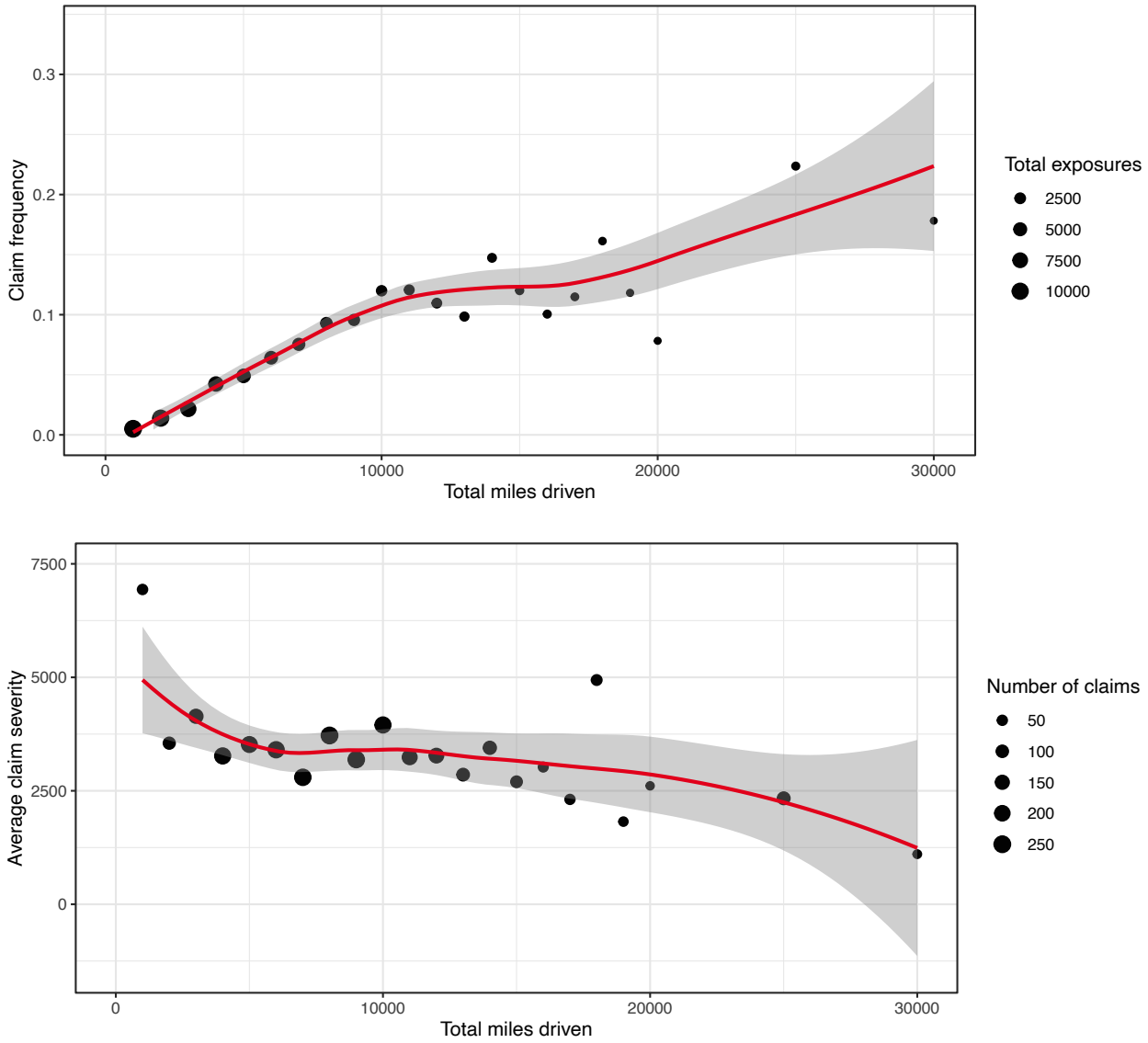
3.2.1.1. Annual Miles Driven

In addition to the declared distance, we also have access to the actual distance traveled by the insured through the telematics device installed in the vehicle. It is increasingly recognized that the distance driven in a car can constitute a more precise measure of risk exposure than simply the duration of the insurance contract. Indeed, the frequency and distance of trips are crucial factors in the likelihood of an accident occurring. For example, one might assume that a driver who regularly covers long distances is statistically more likely to be involved in an accident than a driver who uses their vehicle sporadically, even if both have the same duration of insurance contract.

Figure 3.11 (top) illustrates claim frequency as a function of distance driven. Although the graph appears to show a proportional relationship between distance driven and the number of claims, it also indicates a stabilization in claim frequency for drivers who have driven extensively (between 10,000 and 20,000 miles). That relationship for heavy drivers has been observed in other scientific papers.

Figure 3.11 (bottom) illustrates claim severity as a function of distance driven. We see what appears to be a small negative relationship between these two variables.

Figure 3.11. Total Miles Driven

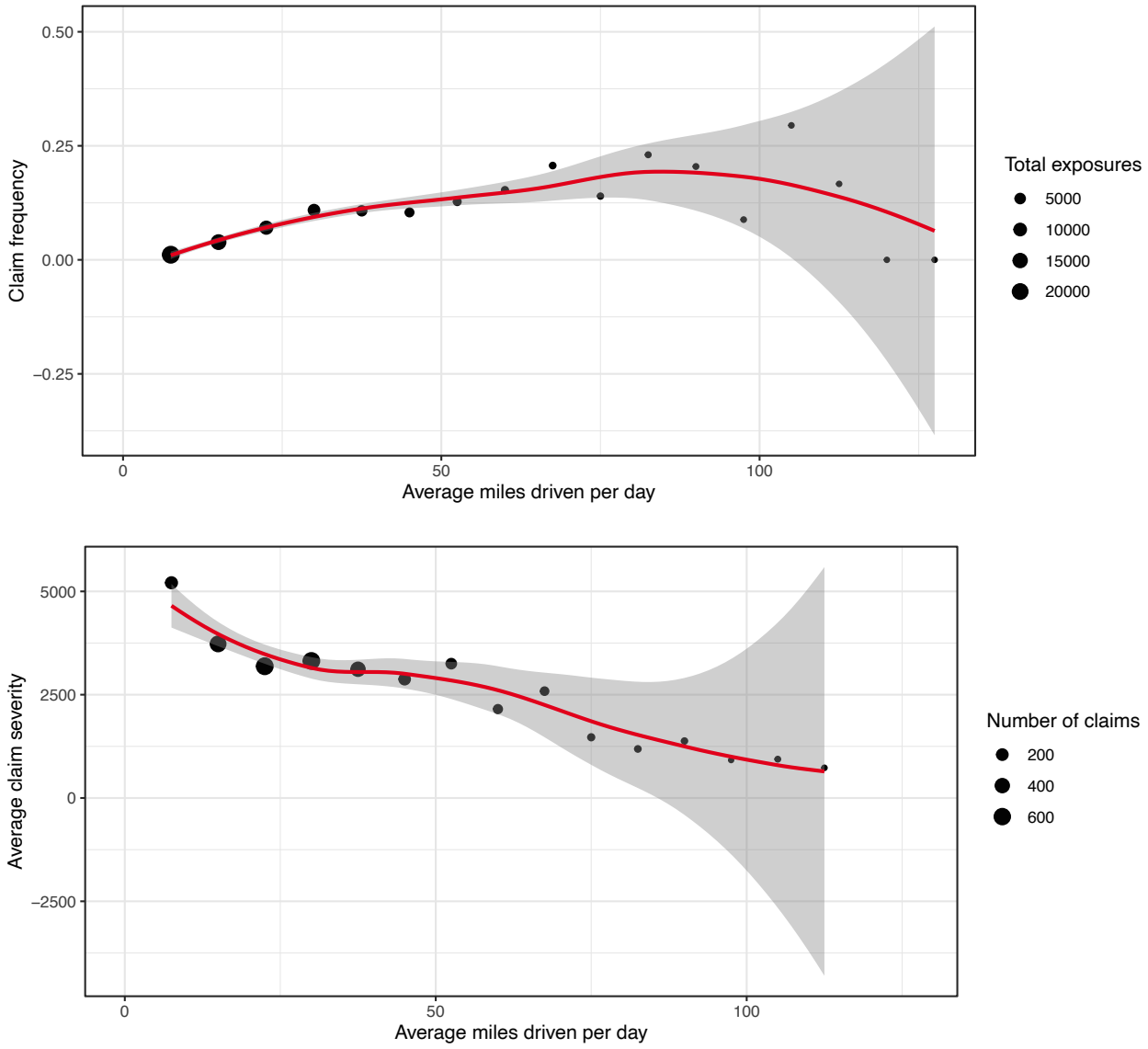


3.2.1.2. Contract Duration Revisited

A sort of competition between contract duration and distance driven (measured) appears to emerge as the adequate measure of risk exposure. Whereas distance driven should be more predictive, contract duration exhibits a more consistent relationship with claim frequency, resembling what risk exposure should entail. We thus propose creating a new covariate for our analysis: the average number of miles traveled per day. This new variable will solely correspond to the duration of the contract.

The new variable can represent a form of driving activity intensity. By replacing the total miles traveled with the average daily miles, we can revert to the duration of the contract as the classical measure of risk exposure. Figure 3.12 illustrates the relationship between average miles driven per day and claim frequency (top) and claim severity (bottom).

Figure 3.12. Average Miles Driven per Day



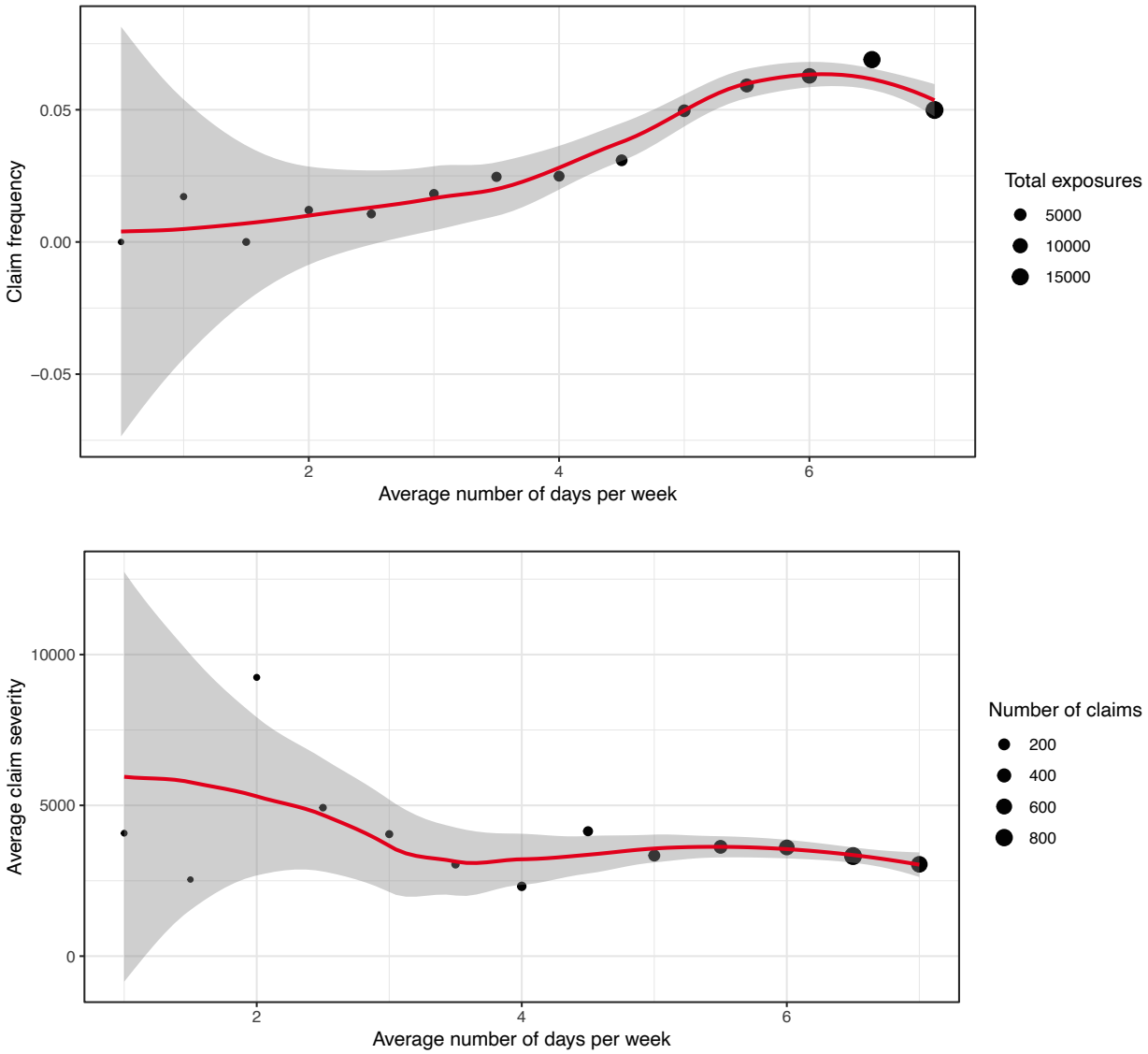
3.2.1.3. Days per Week

Another intensity measure that may come close to the usage percentage is the average number of days the vehicle is used per week. Figure 3.13 (top) shows the relationship between claim frequency and this variable. An increase in claim frequency is observed as the number of days used increases. We note, however, that insured individuals who use their car an average of seven days per week appear to deviate from the general trend. We observe no trend with respect to severity in Figure 3.13 (bottom).

3.2.1.4. Correlation

Once again, we take note of the dependency between the covariates (Figures C.4 and C.5 in Appendix C). It is interesting to note, for example, that the mean number of days used

Figure 3.13. Average Number of Days per Week the Car is Used



is linked to the insured's age. One possible explanation is that drivers in their "middle years" are more likely to drive consistently five days a week for work, whereas younger and older drivers may drive fewer days a week.

3.2.2. Type of Vehicle Usage

Instead of using the telematics device solely to measure levels of vehicle usage, we can also investigate whether certain types of vehicle usage are indicators of a higher risk of claims. In this section, we analyze certain telematics information that we classify as types of usage. This analysis is significantly less interesting when the response variable is severity; we present only the results for frequency to avoid unnecessarily burdening the paper.

3.2.2.1. Days

One pertinent inquiry is to investigate whether vehicle usage on certain days of the week predicts a higher claim frequency. Seven covariates are available in the database, each indicating the percentage of vehicle usage on a particular day of the week. It is worth noting that the sum of the seven percentages for each contract equals 100%. Thus, high vehicle usage on a Saturday corresponds to a high percentage of usage for that day, necessarily implying that the other days will have smaller percentages.

The seven graphs in Figure4 3.14 illustrate claim frequency as a function of vehicle usage for each day of the week. The results obtained for each day are similar and seem to indicate that uniform vehicle usage across all seven days (i.e., 1/7 = 14.2%) is the riskiest situation. In the case of severity, we do not observe the same indication. Thus, vehicle usage for each day appears to signify something, but the information provided by these covariates likely needs transformation.

3.2.2.2. Days (2)

In light of the results obtained from the analysis of vehicle usage for each day of the week, it is appropriate to create new variables that may better represent the risk. We thus create the following variables:

- A variable identifying the maximum value of vehicle usage for each day.
- A variable identifying the minimum value of vehicle usage for each day.
- A variable measuring the difference between the maximum and minimum values that have just been calculated. This variable can thus identify insured individuals who use their vehicle more on specific days or, conversely, insured individuals who typically refrain from using their vehicle on certain days of the week.

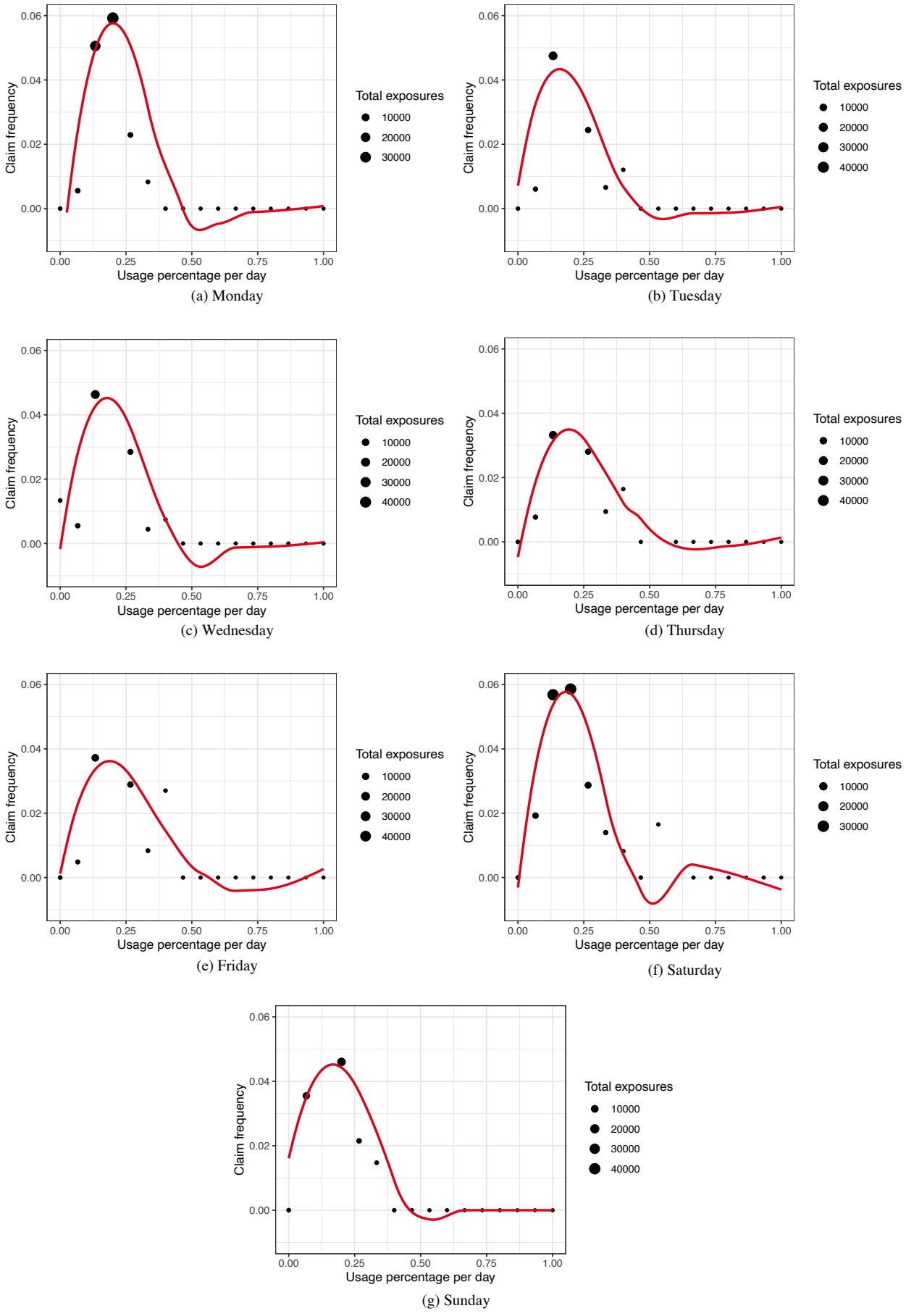
Figure 3.15 illustrates the relationship of each of the three variables with claim frequency. Whereas the graph for maximum use does not seem to point to a significant result in explaining claim frequency, the two other graphs are more interesting: insured individuals who use their vehicle equally on all days of the week display a higher claim frequency than those who use their vehicle more on certain days.

3.2.2.3. Days (3)

Continuing with our days of vehicle usage analysis, we explore two additional variables: (a) a variable identifying the day of the week when the vehicle is most used and (b) a variable identifying the day of the week when the vehicle is least used.

[4] We have removed the confidence interval (gray area) to make the graphs easier to read.

Figure 3.14. Claim Frequency vs. Percentage of Use for Each Day



(a) Monday

(b) Tuesday

(c) Wednesday

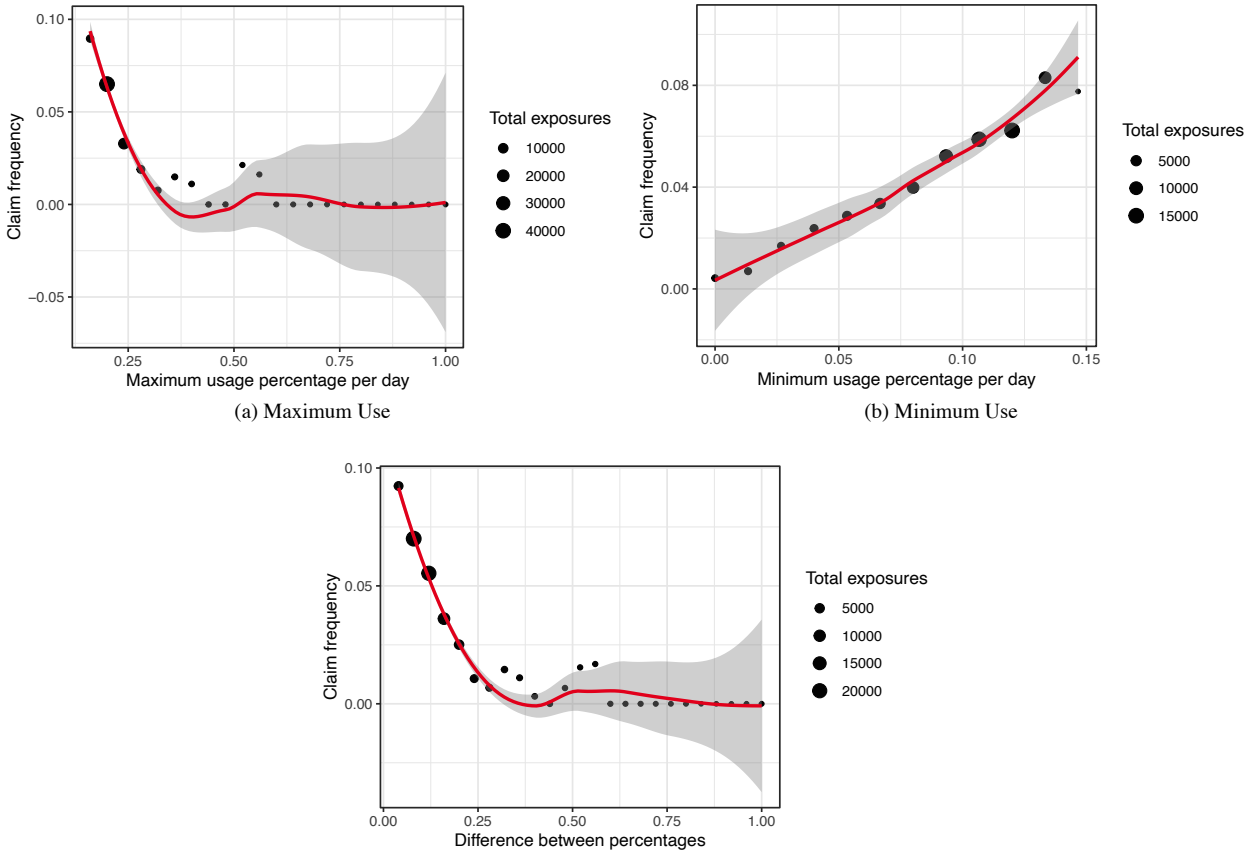
(d) Thursday

(e) Friday

(f) Saturday

(g) Sunday

Figure 3.15. Claim Frequency vs. Use for Each Day



(a) Maximum Use

(b) Minimum Use

(c) Difference between maximum and minimum use

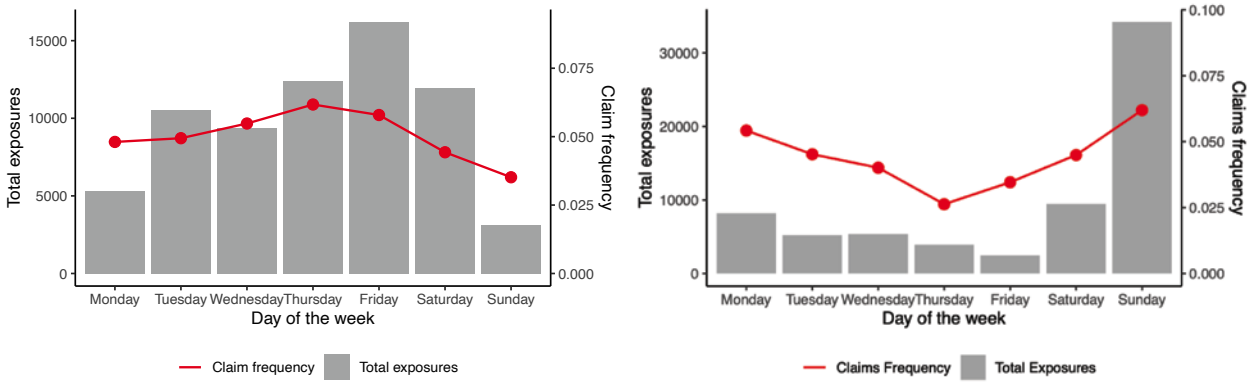
Figure 3.16 attempts to verify whether claim frequency differs for those who use their vehicle more or less on specific days of the week. Friday is evidently the day when insured individuals tend to use their car more frequently. Conversely, Sunday is the day when the car appears to be used the least. For claim frequency, two days appear to be slightly more significant than the others:

- **Thursday:** Insured individuals who use their vehicle most frequently on Thursdays have a higher claim frequency, while those who use their vehicle less on Thursdays have a lower claim frequency.
- **Sunday:** Insured individuals who use their vehicle most frequently on Sundays have a lower claim frequency, whereas those who use their vehicle less on Sundays have a higher claim frequency.

3.2.2.4. Weekend

We performed a grouping of the vehicle usage variables for the days of the week directly in the database. We summed the usage percentages for the days from Monday to Friday in one variable and the usage for Saturday and Sunday in another variable.

Figure 3.16. Claim Frequency vs. Use for Each Day



(a) Day with the maximum use

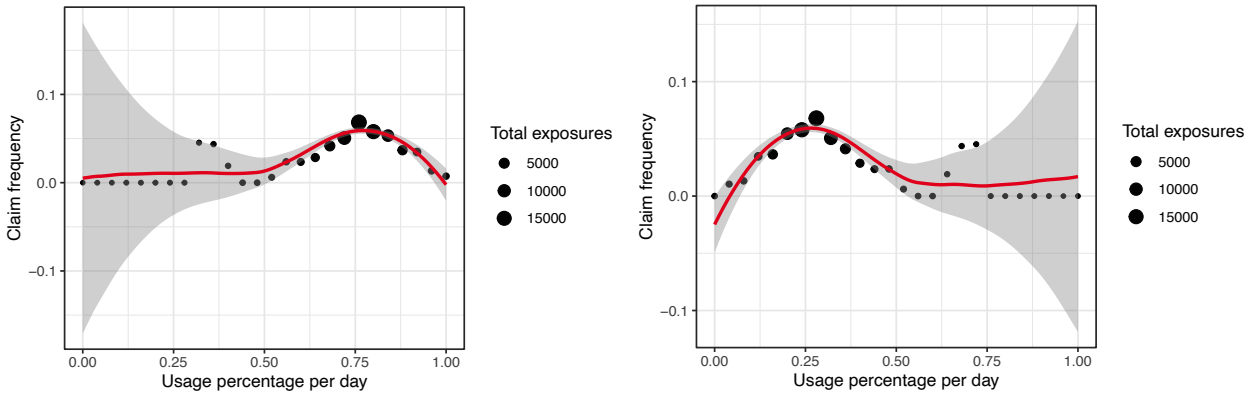
(b) Day with the minimum use

Knowing that the two covariates are complementary (since the sum of both equals 100%), we need to keep only one of them. Figure 3.17 shows the same behavior, but in opposite directions. The final result obtained is similar to what we observed for the individual days of the week, and it is unclear whether these covariates will remain important in the final analysis.

3.2.2.5. Trip Duration

The common explanation for claim probability highlights the use of highways. According to several studies, the risk of an accident per miles traveled is much lower on highways than in urban areas. Thus, the duration of each trip, or the percentage of trips exceeding a certain predetermined duration, could be relevant to analyze. Figure 3.18 shows the graphs of claim frequency as a function of the percentage of trips exceeding two, three, or four hours. Despite the intuition that such information could be relevant, the graphs do not seem to show a particularly strong link between the proportion of long trips and claim frequency.

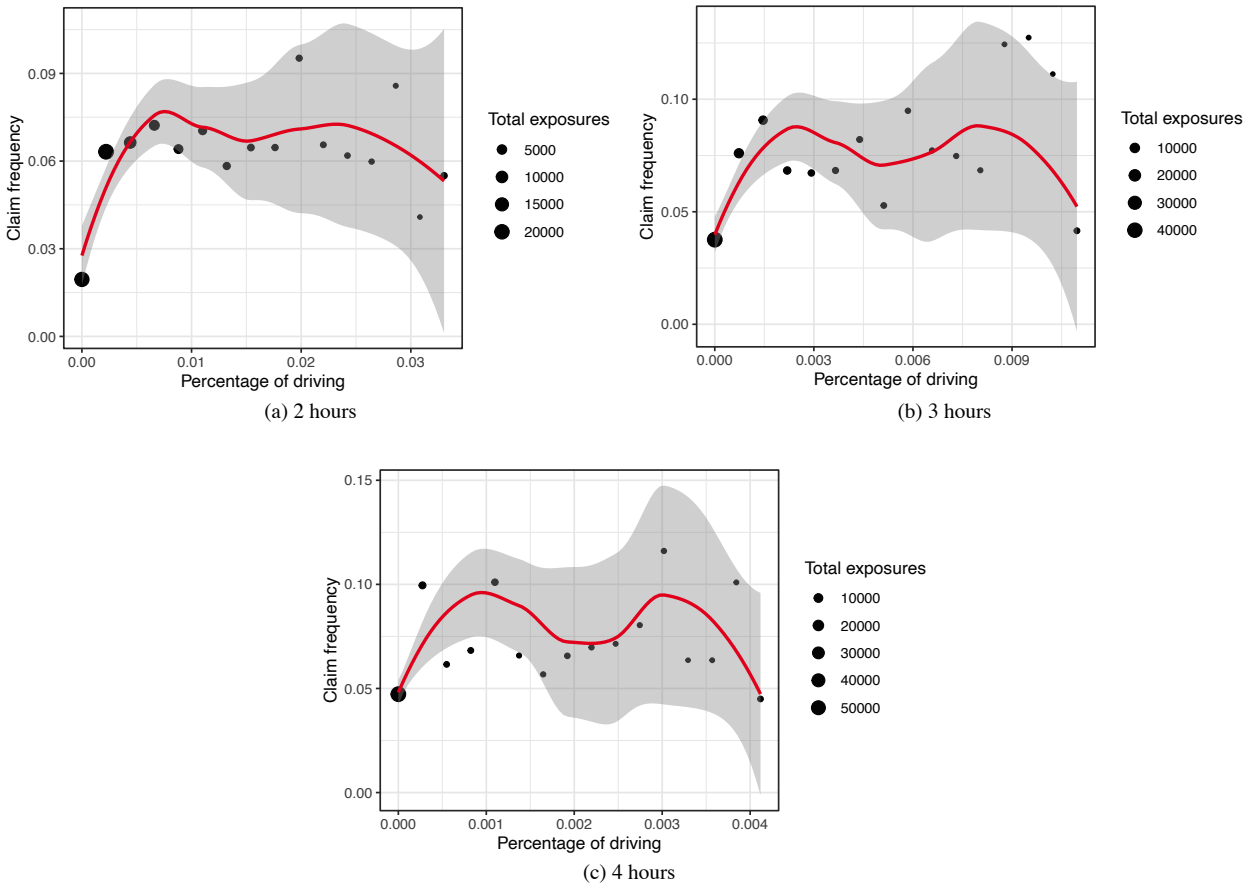
Figure 3.17. Claim Frequency vs. Percentage of Use Weekday and Weekend Day



(a) Week day

(b) Weekend day

Figure 3.18. Claim Frequency vs. Percentage of Vehicle Driven 2, 3 and 4 Hours



(a) 2 hours

(b) 3 hours

(c) 4 hours

3.2.2.6. Rush Hours

Another common hypothesis links the frequency of automobile insurance claims to traffic congestion. Thus, the database makes available the proportion of trips made in times of traffic congestion, whether in the morning or evening. Figure 3.19 shows the graphs of frequency linked to these two variables. Similar to the trip duration theory, the statistical analysis does not validate the hypothesis, and it is not clear whether these variables are relevant for explaining the risk of accidents.

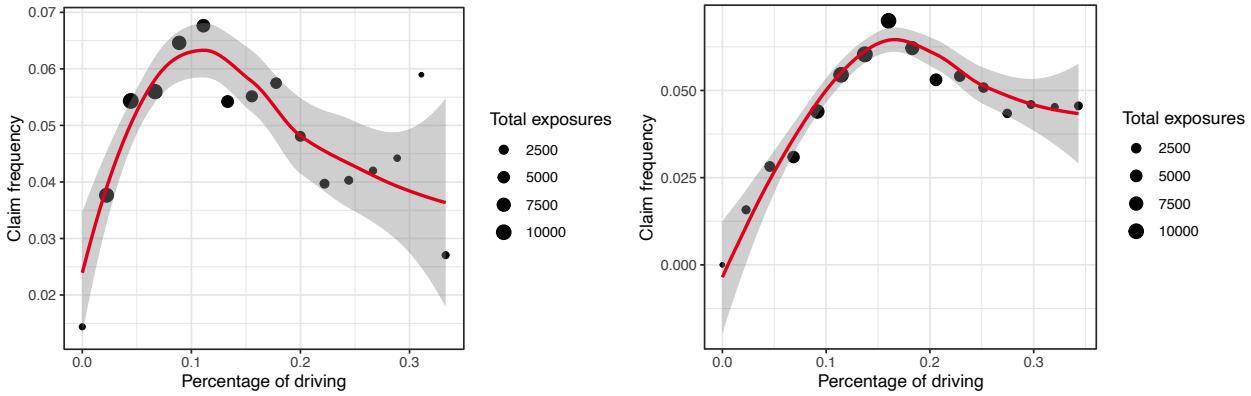
3.2.2.7. Correlation

Figure C.6 in Appendix C illustrates the dependency between each studied covariate. It is interesting to note the apparent link between vehicle usage and the age of the insured.

3.2.3. Driving Behavior

Beyond the intensity of vehicle usage, telematics devices also allow for the compilation of various statistics on driving behavior. These primarily include sudden braking, rapid acceleration, and high-speed turns (both left and right). In this final part of the analysis

Figure 3.19. Claim Frequency vs. Percentage of Vehicle Driven during Rush Hours



(a) AM Rush (b) PM Rush

of segmentation variables available in the database, we therefore work on analyzing and transforming these variables. Again, to make the paper manageable, we present only the results for frequency.

3.2.3.1. Brakes

The database affords us access to a series of variables counting the number of abrupt braking events – with decelerations of 6 miles per hour (mph), 8 mph, 9 mph, 11 mph, 12 mph, and 14 mph – per 1,000 miles traveled. As that description indicates, the number of abrupt braking events is normalized by the distance traveled and not by the number of insured days. Since we choose to use the number of insured days as a measure of exposure to risk, we must transform these variables.

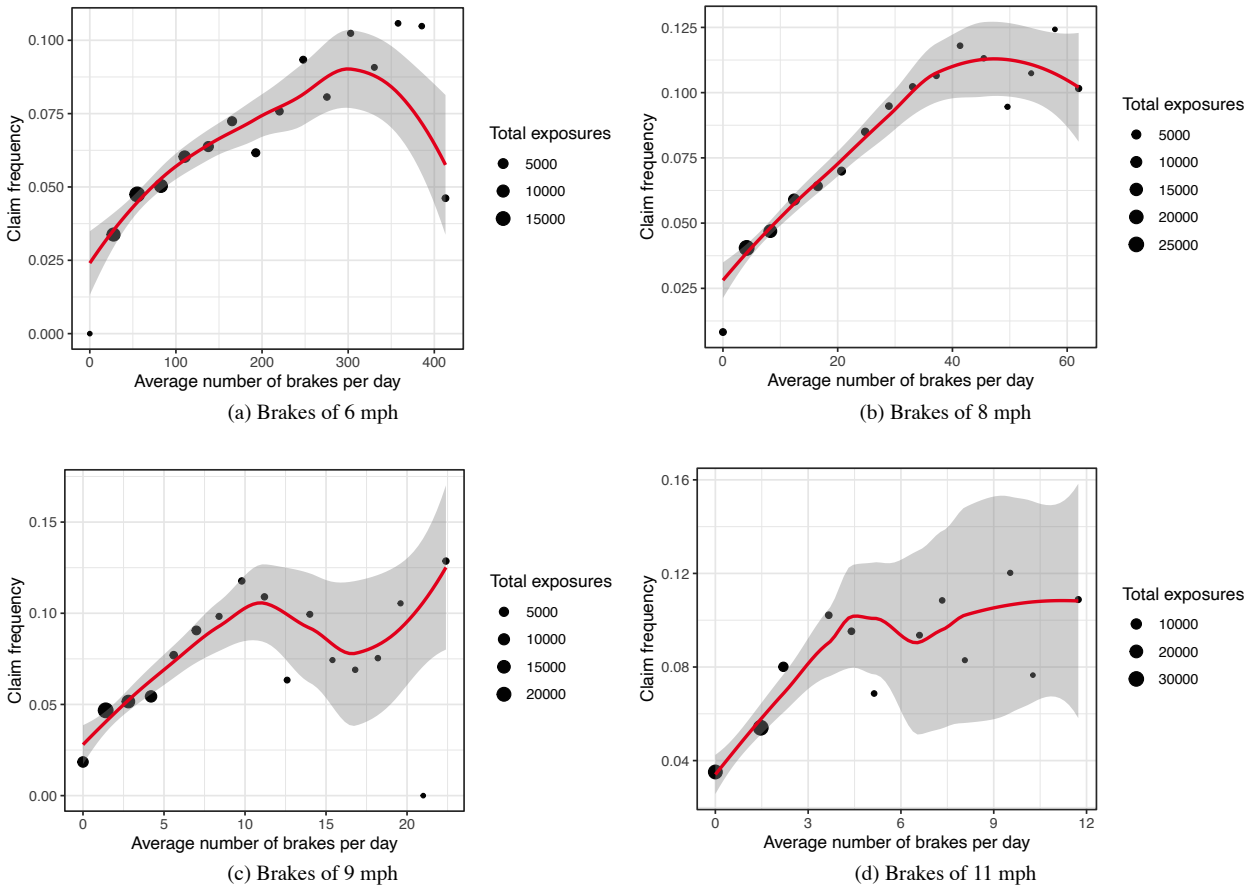
We multiply the number of abrupt braking events per 1,000 miles traveled by each thousand miles traveled, and we obtain the total number of abrupt braking events. We then divide that by the number of insured days to create a new measure of driving quality: the number of daily abrupt braking events. We perform this exercise for the five measures of abrupt braking events. Just as for the average daily distance traveled, we end up with a new variable measuring an intensity, this time the average daily intensity of abrupt braking events.

Furthermore, we add some control to reduce the number of outliers. To do this, for each count of abrupt braking events, we limit the obtained value by the 99th percentile. The four graphs in Figure 3.20 illustrate the relationship between the number of daily abrupt braking events and claim frequency. Since there was very little data for 12 mph and 14 mph decelerations, we do not include those graphs. For all four scenarios, a clear relationship can be observed between an increase in the number of abrupt braking events and the average claim frequency.

3.2.3.2. Accelerations

Similar to our treatment of braking events, we need to convert accelerations, which are normalized by miles driven, into average daily accelerations. We again control the possible

Figure 3.20. Claim Frequency vs. Average Number of Brakes



(a) Brakes of 6 mph

(b) Brakes of 8 mph

(c) Brakes of 9 mph

(d) Brakes of 11 mph

values to not exceed the 99th percentile. The four graphs in Figure 3.21 illustrate the relationship between the average accelerations and claim frequency.

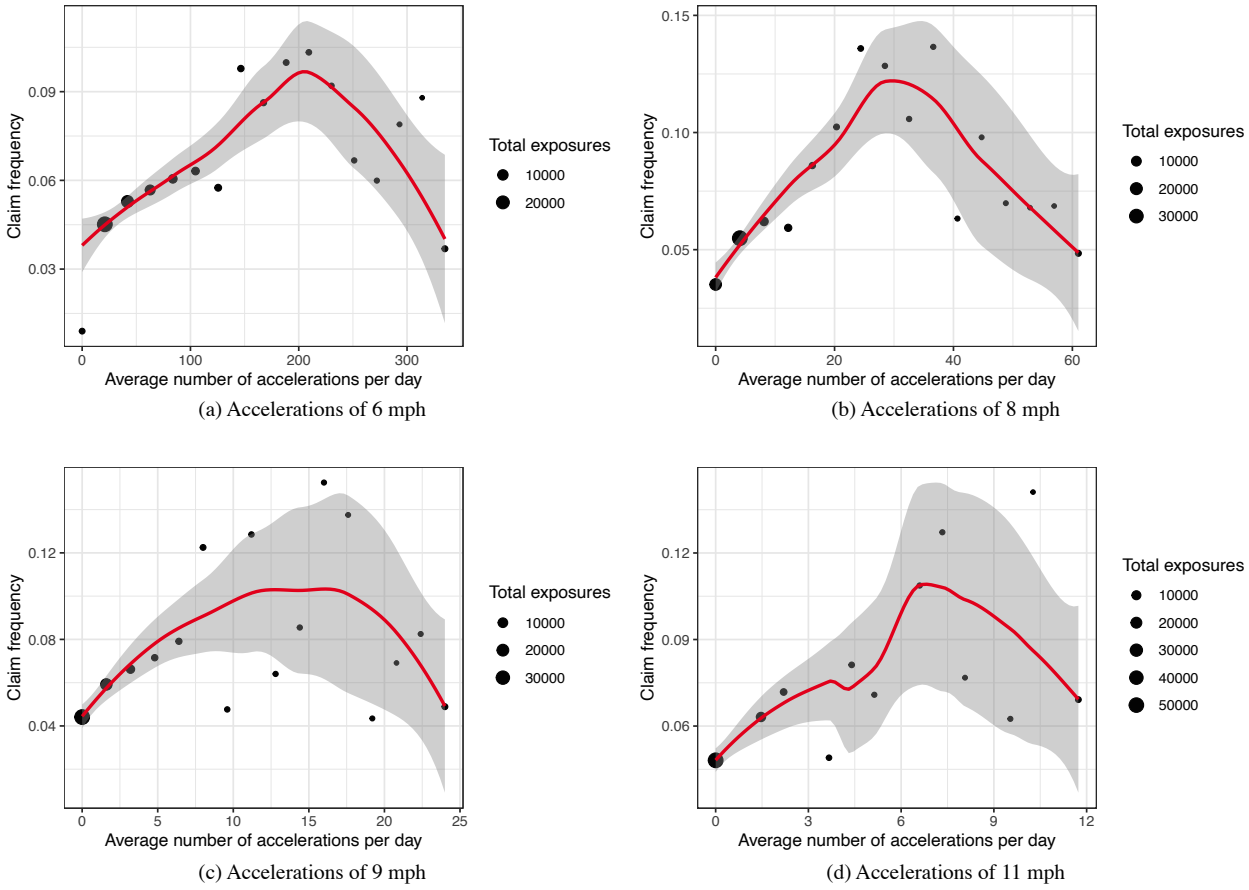
Once again, we observe a clear link between the increase in the number of accelerations and the number of claims.

The telematics device also measures how fast insureds drive when they turn left or right. We observe the same type of results for high-speed left and right turns as we see for decelerations and accelerations (results are not shown but are available if requested).

3.2.3.3. Correlation

We analyze the correlation between all variables measuring driving quality and present that in the tables in Figures C.7 and C.8 in Appendix C. Unsurprisingly, we observe that different accelerations and different braking events are strongly correlated. It's even apparent that insured individuals with high accelerations likely also exhibit strong braking events. The dependency between accelerations, braking events, and sensitive variables is very weak. Thus, the driving quality as measured in the studied database may not be able to replace the predictive capacity of sensitive covariates. The intensity of turns to the left or right also does not seem to explain the sensitive variables.

Figure 3.21. Claim Frequency vs. Average Number of Accelerations



(a) Accelerations of 6 mph

(b) Accelerations of 8 mph

(c) Accelerations of 9 mph

(d) Accelerations of 11 mph

4. Traditional Covariates

Using only traditional covariates, our objective in this section of the paper is to propose various statistical models for estimation and variable selection to predict the number of claims and the average severity. More specifically, we consider basic generalized linear models (GLMs), a larger GLM family including elastic net, and XGBoost models.

As we mention in Appendix B, to compare models and strike a balance between bias and variance while avoiding overfitting, an interesting approach is to assess the prediction quality of models when applied to new data.

The analyses in this section are done using the same data as in Section 3. However, as we concluded at the end of our overview of the data, one does need to transform certain variables. Those transformations are based on splines. We attempt to graphically approximate the results obtained using splines using a simple parametric equation (square, cube, square root, etc.). Again, details are available in the GitHub project folder (https://github.com/J-PBoucher/CAS_Project2024).

4.1. Basic Generalized Linear Models

4.1.1. Single Intercept

A baseline model corresponding to a GLM with an intercept and predicting for each contract only the observed mean multiplied by the exposure is used as a point of comparison. The model is then estimated on the entire *training* set and predicted on the *test* set, which was not used in parameter calibration. Tables 4.1 and Table 4.2 present results on the *test* set for frequency and severity (row "Base"). In these tables, a small score value indicates a better model.

4.1.2. Categorical Covariates

A first regression approach is attempted using only the traditional categorical variables, namely (1) sex, (2) marital status, (3) vehicle use, and (4) region. Even though we should also consider territory since it consists of more than 50 different factors, we do not integrate it into the model immediately. As we saw in the overview of variables in Section 3, the insured's sex did not appear to be a significant variable. This GLM approach confirms that observation. Therefore, that variable is excluded from the model. In Tables 4.1 and 4.2, we can see the impact of adding traditional variables on the prediction quality for frequency and severity – row "GLM (trad.)." We see that adding traditional variables does not substantially enhance prediction on the *test* data set.

Table 4.1. Prediction Scores (Frequency)

Model	Log.	Mean Squared Error (MSE)	Quad.	Spherical	Dawid-Sebastiani
Base	0.1767	0.0455	–0.9215	–0.9598	–2.1988
GLM (trad.)	0.1759	0.0454	–0.9216	–0.9598	–2.2285
LASSO (opt.)	0.1717	0.0449	–0.9223	–0.9600	–2.3406
LASSO (pars.)	0.1733	0.0451	–0.9220	–0.9600	–2.2860
XGBoost	0.1456	0.0365	–0.9315	–0.9637	–2.7278

Table 4.2. Prediction Scores (Severity)

Model	Log.	MSE
Base	9.29504	21.82679
GLM (trad.)	9.27546	21.77177
LASSO (opt.)	9.23357	20.23523
LASSO (pars.)	9.23729	20.21870
XGBoost	9.19011	20.06725

4.2. GLM-Net

As we saw in the previous section, a series of traditional continuous segmentation variables is also available: (1) credit score, (2) age of the insured, (3) age of the vehicle, and (4) number of claim-free years. Furthermore, as we explain later, the territory will also be treated as a continuous variable.

Directly using a continuous variable in a GLM is usually ineffective as it assumes a linear relationship. To avoid overfitting the data, an approach using splines, using the generalized additive model (GAM) theory, is interesting. Such an approach allows the modeler to visualize the general form of the covariate to explain the number of claims. A parametric form can then be proposed to achieve the best possible correspondence with the spline obtained by the GAM. Subsequently, instead of attempting to fit a basic GLM model with all variables, we will work with a GLM-net model that allows for variable selection.

The spline analysis indicates that the following parametric form appears appropriate for capturing the relationship between the number of claims and continuous covariates:

$$s\left(Credit.Score\right) \approx Credit.Score + Credit.Score^2$$

$$s\left(Insured.age\right) \approx Insured.age + log\left(Insured.age\right) + Insured.age^2$$

$$s\left(Car.age\right) \approx Car.age + Car.age^2$$

$$s\left(Years.noclaims\right) \approx Years.noclaims + Years.noclaims^2 + Years.noclaims^3.$$

For the severity, the following parametric form appears appropriate for capturing the relationship:

$$s\left(Credit.Score\right) \approx Credit.Score + Credit.Score^2$$

$$s\left(Insured.age\right) \approx Insured.age + Insured.age^2$$

$$s\left(Car.age\right) \approx Car.age + Car.age^2 + Car.age^3$$

$$s\left(Years.noclaims\right) \approx Years.noclaims + Years.noclaims^2 + Years.noclaims^3.$$

As we saw in the previous section, the insured's territory code corresponds to a categorical variable with a large cardinality. In such a situation, creating a binary variable for each possible territory is not appropriate. Instead, we propose using target encoding based on the territory's rank.

This means that we first calculate the observed frequency for each territory. Then, we rank the frequencies for the fifty-three territories in the database. Next, the rank divided by 53 corresponds to the numerical value of the territory. This form is called rank

encoding. We believe that such a transformation is justified, considering that ranking territories according to risk is an approach insurance companies may take.

With the encoded form of the territory, as we did with the other continuous variables, we propose a parametric form for the spline obtained (for both frequency and severity):

$$s\left(terr.code\right) \approx terr.code + terr.code^2 + terr.code^3.$$

4.2.1. Optimal Value

The parameters of the GLM-net were calibrated using cross-validation to obtain the model's hyperparameters. In particular, we obtain a value of $\alpha = 1$, which corresponds to a lasso model. Using these values, we can calculate the prediction scores of the model based on all covariates. We present these results in Tables 4.1 and 4.2 in the row "LASSO (opt.)."

4.2.2. Parsimonious Model

Instead of using the optimal value of the penalty λ in the elastic net approach, it is often advised to use a penalty value located at one standard error (λ_{1se}). This helps in obtaining a more parsimonious model. These prediction scores are displayed in Tables 4.1 and 4.2 in the row "LASSO (pars.)."

4.2.3. Categorical Covariates

We can now check whether the choice of approach (optimal or parsimonious) modifies the impact of the covariates. For categorical variables, the relativity values obtained for both GLM-net approaches are displayed in Figures 4.1 and 4.2 for frequency and severity, respectively.

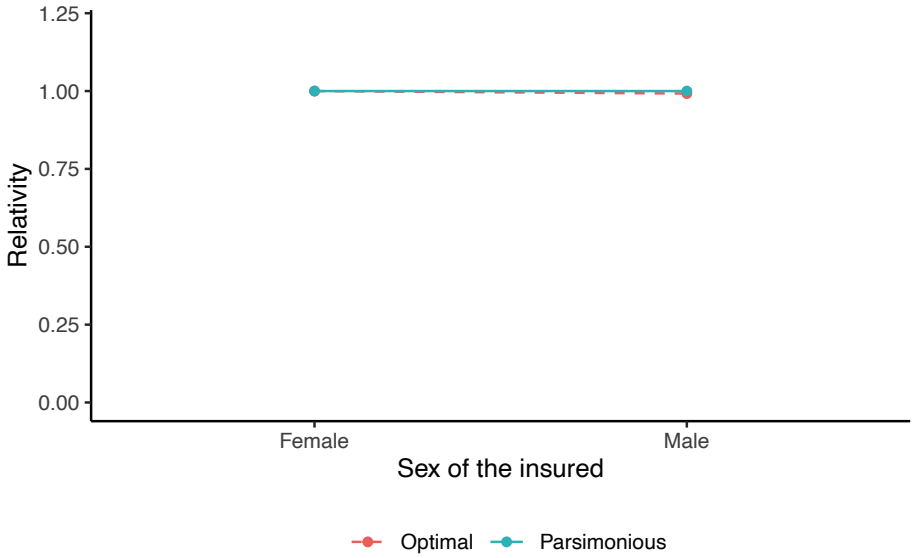
4.2.4. Continuous Covariates

As with the categorical variables, we show the relativities obtained for the continuous variables in Figures 4.3 and 4.4 for frequency and severity, respectively. We observe that the parsimonious approach tends to reduce the impact of segmentation variables on the premium.

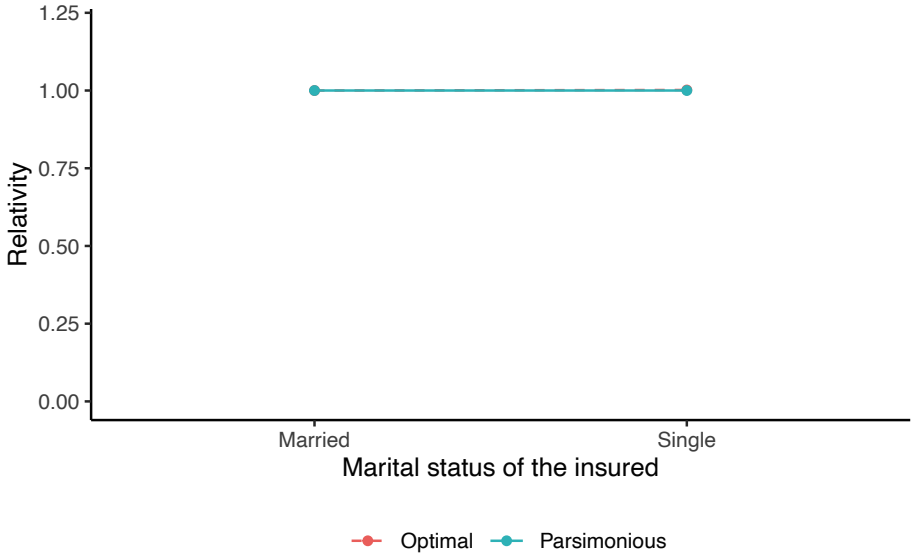
4.3. XGBoost

XGBoost is another approach. However, through cross-validation, the modeler must finely tune the model's hyperparameters, such as the learning rate, the maximum depth of a tree, the minimum sum of instance weight needed in a child, and the subsample ratio of the training instances. We use a grid search approach coupled with Bayesian optimization for the data set used in the project. Then, we compute the model's prediction scores and present results in Tables 4.1 and 4.2 in the row "XGBoost." The scores obtained show a significant improvement compared with the other tested approaches.

Figure 4.1. Interpretation of the Categorical Variables from the GLM-Net Model (Frequency)

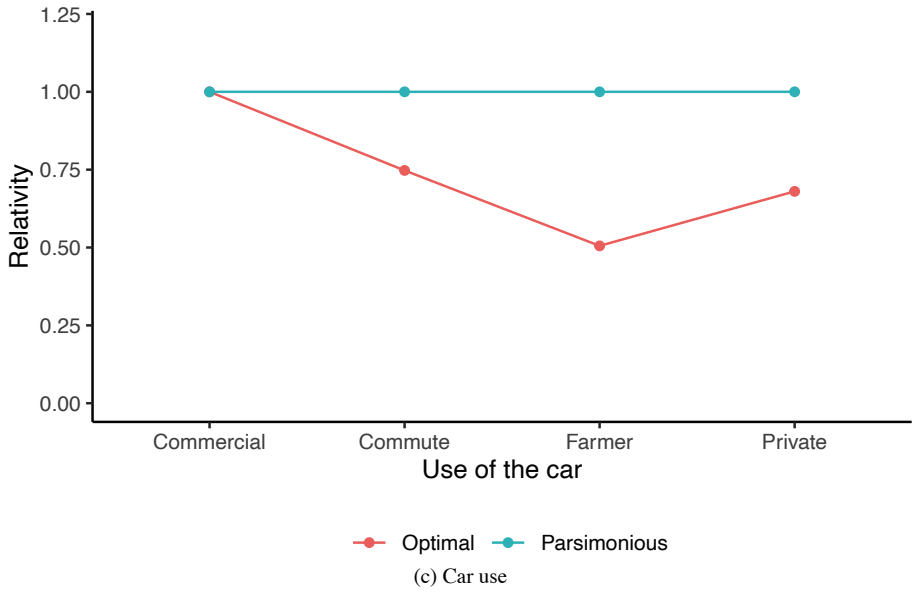


(a) Sex of the insured

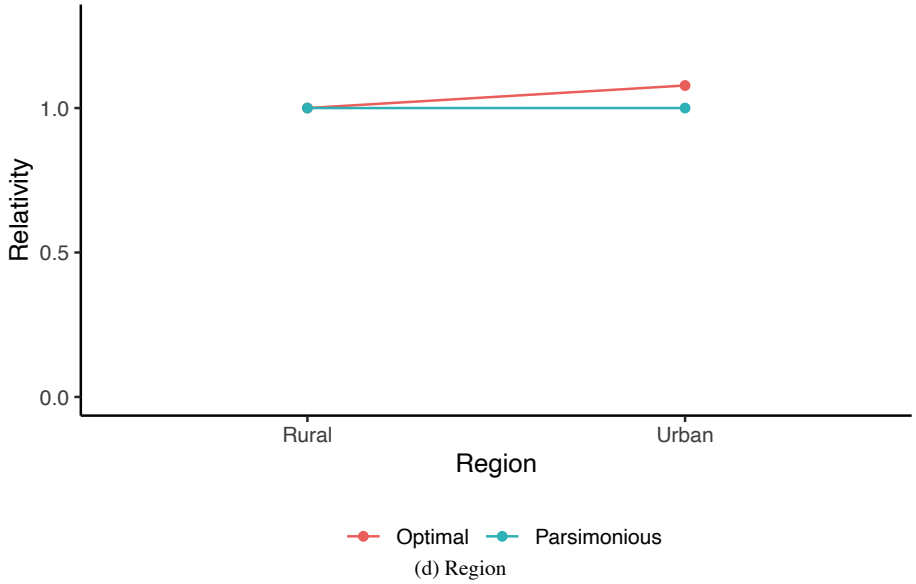


(b) Marital status of the insured

Figure 4.1. Interpretation of the Categorical Variables from the GLM-Net Model (Frequency) *(Continued)*



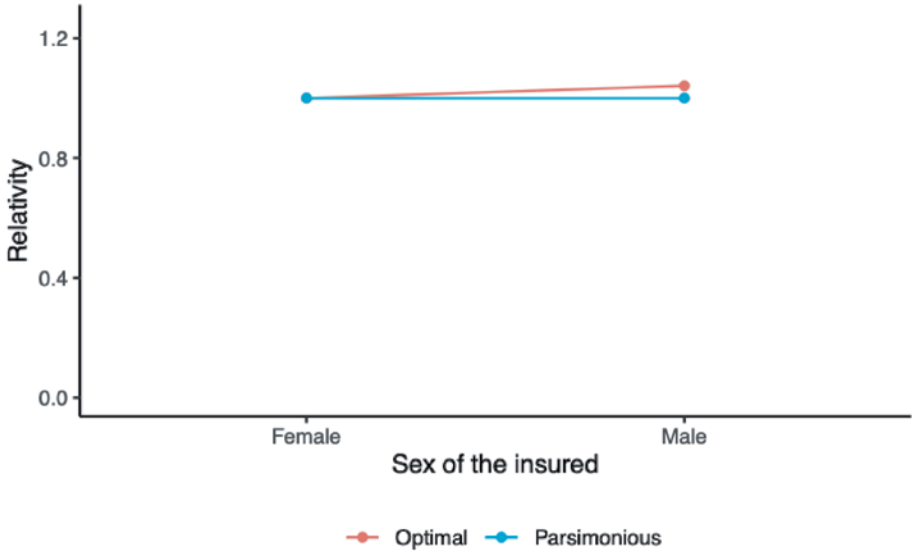
(c) Car use



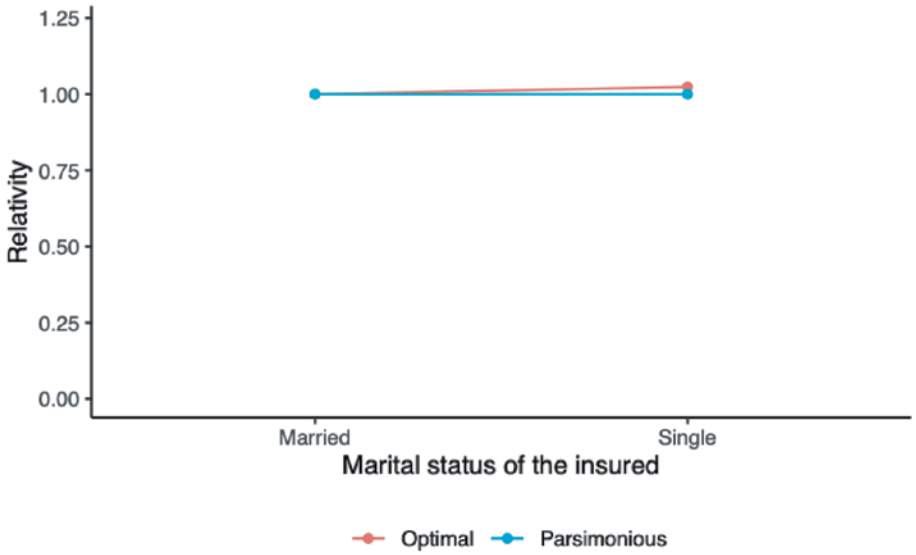
(d) Region

Note: This type of chart illustrates the predictive power of a covariate (here the credit score) without (red line) and with (blue line) adding telematics variables. A horizontal line indicates that the covariate no longer has predictive power and therefore is no longer useful in the model.

Figure 4.2. Interpretation of the Categorical Variables from the GLM-Net Model (Severity)

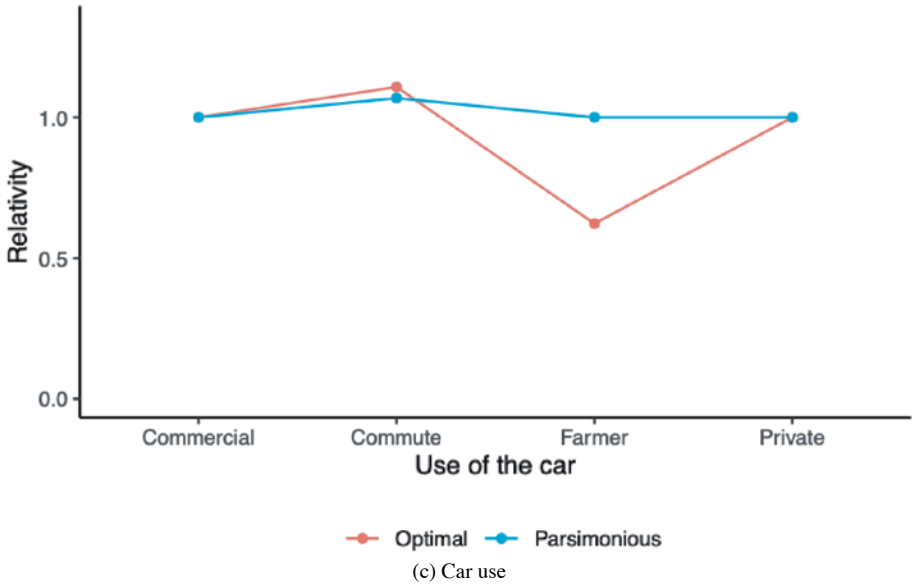


(a) Sex of the insured

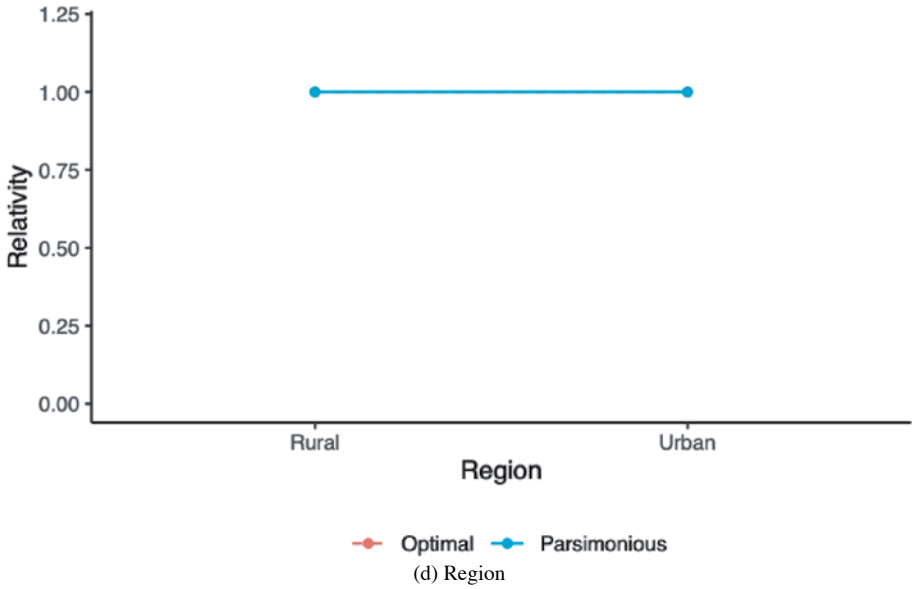


(b) Marital status of the insured

Figure 4.2. Interpretation of the Categorical Variables from the GLM-Net Model (Severity) *(Continued)*



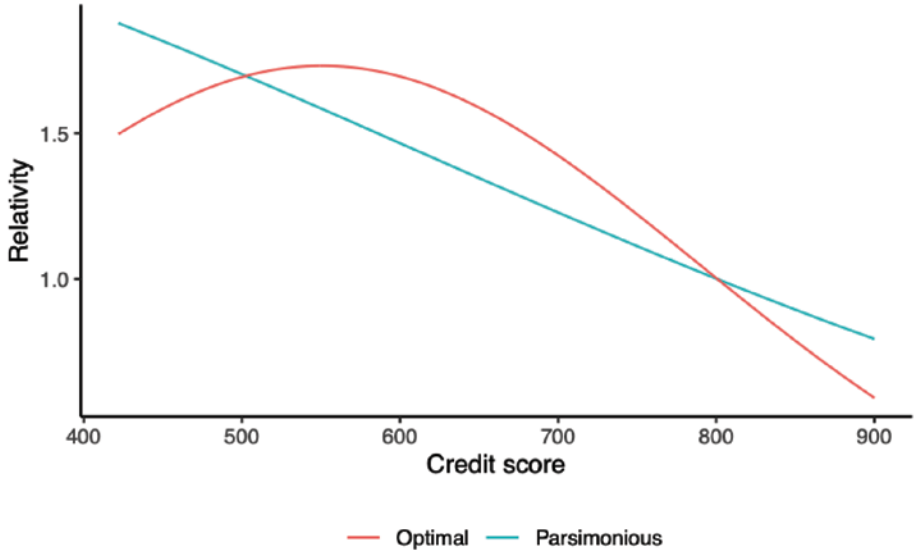
(c) Car use



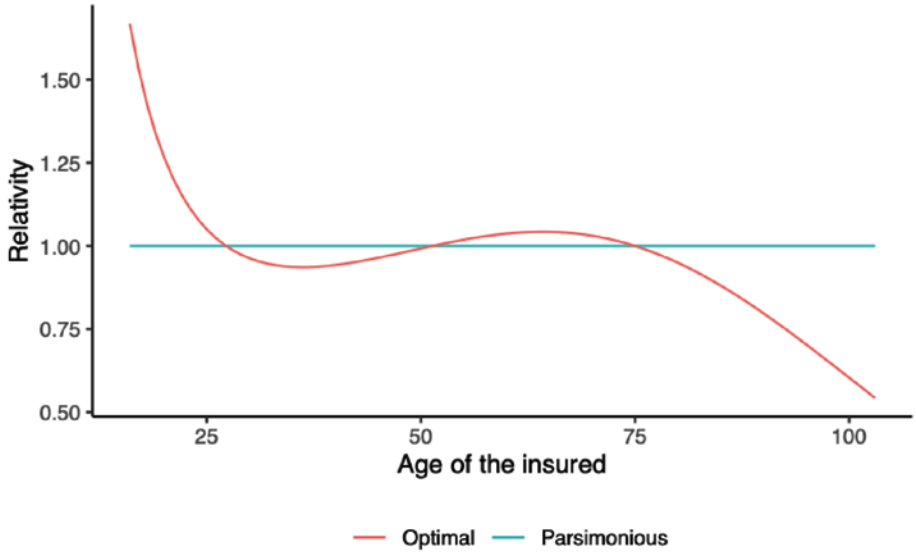
(d) Region

Note: This type of chart illustrates the predictive power of a covariate (here the credit score) without (red line) and with (blue line) adding telematics variables. A horizontal line indicates that the covariate no longer has predictive power and therefore is no longer useful in the model.

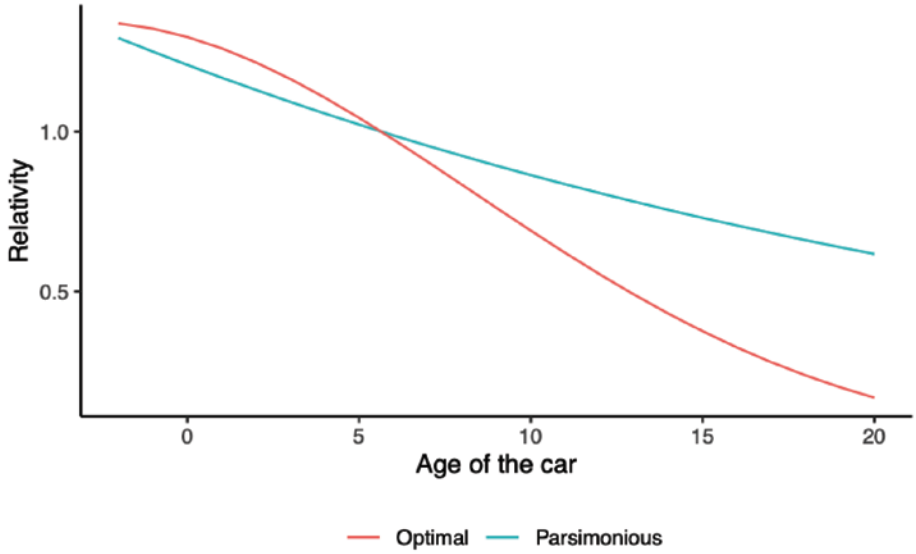
Figure 4.3. Interpretation of Continuous Variables from the GLM-Net Model (Frequency)



(a) Credit score

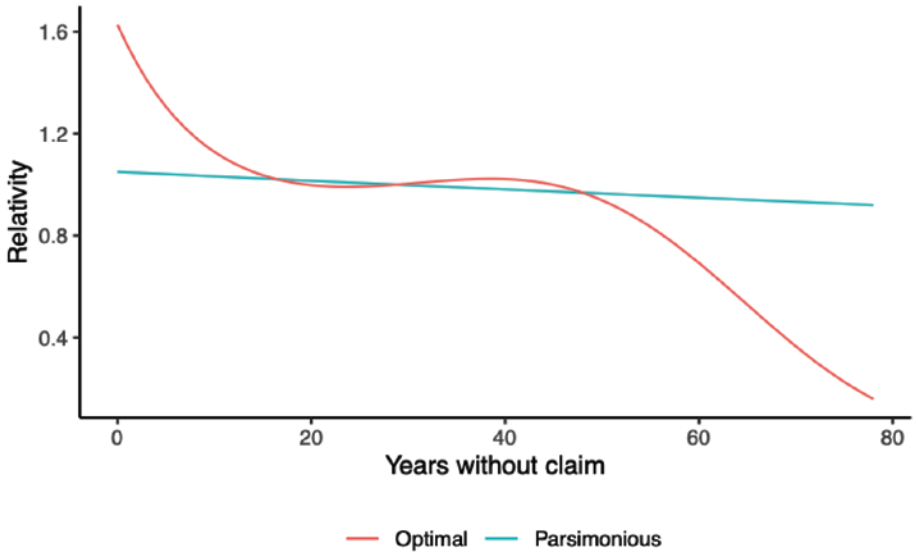


(b) Age of the insured

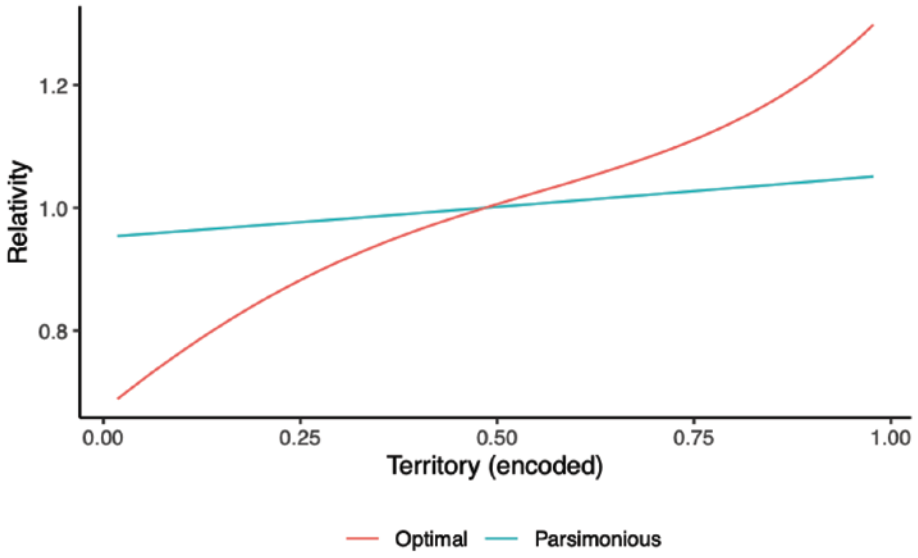


(c) Age of the car

Figure 4.3. Interpretation of Continuous Variables from the GLM-Net Model (Frequency) (*Continued*)

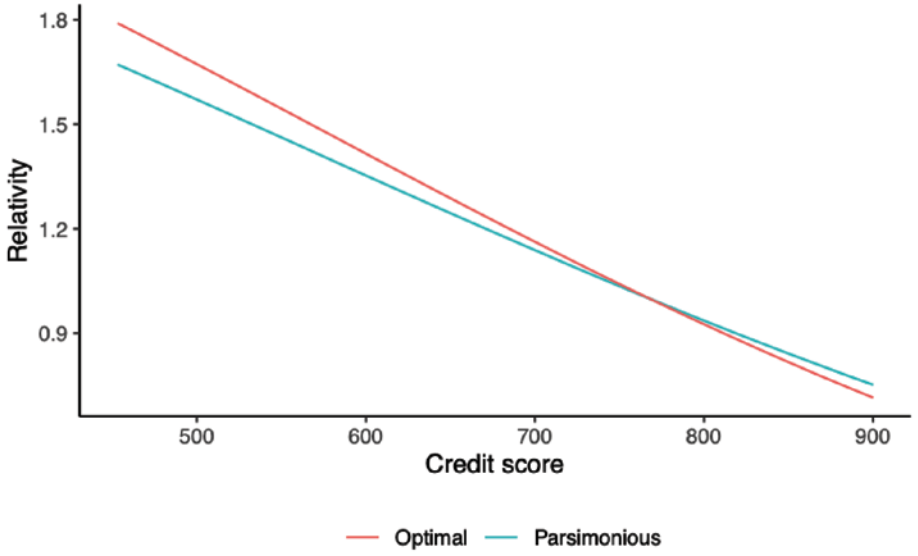


(d) Years without claim

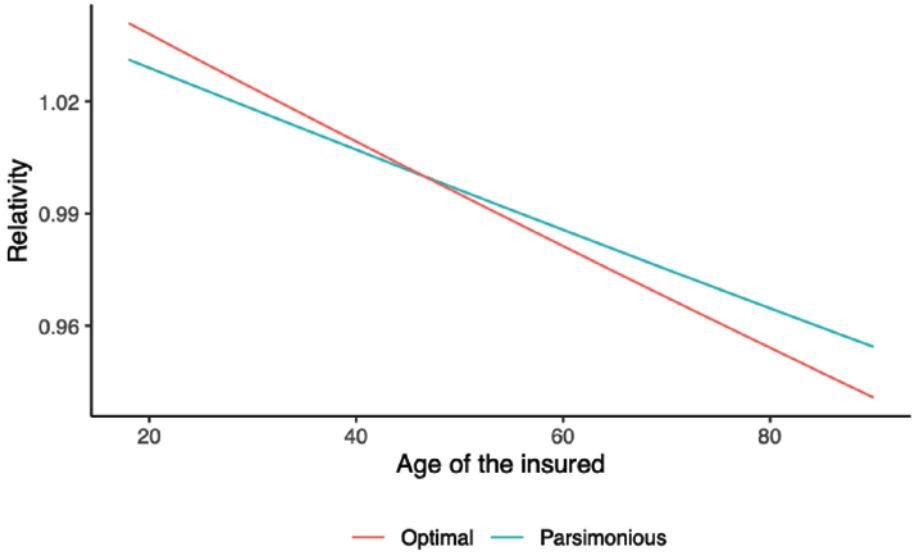


(e) Territory code

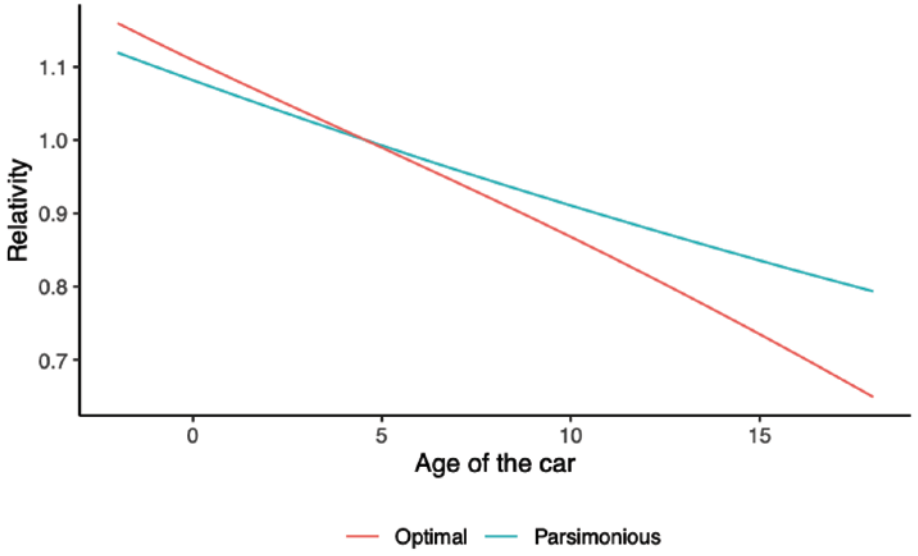
Figure 4.4. Interpretation of Continuous Variables from the GLM-Net Model (Severity)



(a) Credit score

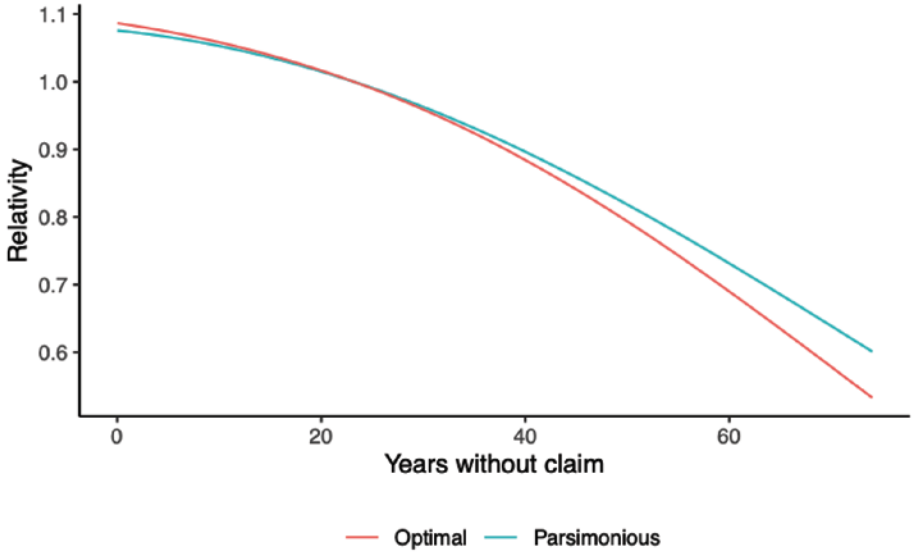


b) Age of the insured

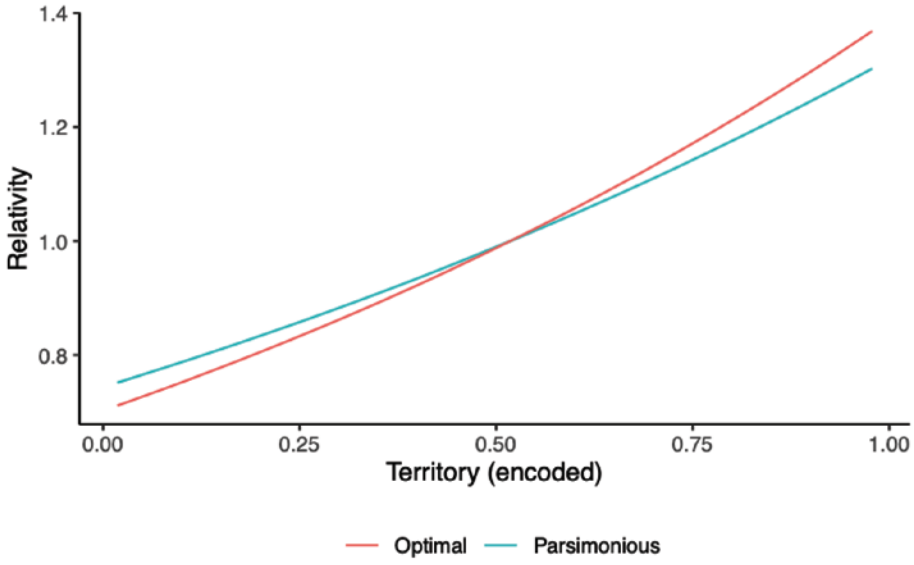


(c) Age of the car

Figure 4.4. Interpretation of Continuous Variables from the GLM-Net Model (Severity) (*Continued*)

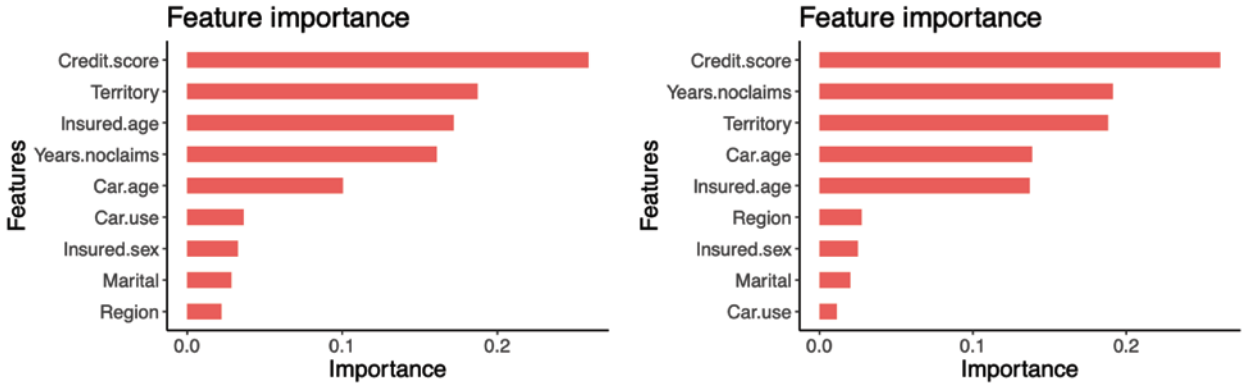


(d) Years without claim



(e) Territory code

Figure 4.5. Variable Importance for Frequency (Left) and Severity (Right)



The XGBoost model has some attractive advantages in practice: (1) fewer assumptions are needed than for the GLM; (2) it can easily handle interactions between variables; and (3) it has a massive number of parameters, which generally increases the predictive potential. On the other hand, (1) an XGBoost model is much more challenging to interpret; (2) training and fine-tuning are more complex to perform; and (3) the model needs to be fully empirically tested.

4.3.1. Variable Importance

A challenge associated with the XGBoost approach is comprehending the full impact of each segmentation variable. Figure 4.5 depicts the most crucial variables in the XGBoost model for frequency (left) and severity (right). We observe that credit score, territory, age of the insured, and years without claim are the most significant covariates in both XGBoost models.

5. Validation

5.1. Validation on the Original Data Set

We conducted our analysis on a synthetic database constructed from an actual database from a Canadian insurer. The use of synthetic databases in actuarial science is slowly developing (see, for example, Gabrielli and Wüthrich [2018] and Avanzi et al. [2021]), but the history is still short. Thus, we consider that we should validate the conclusions on an actual database. Indeed, we want to prevent the mechanics that led to these artificial databases from creating distortions between variables. For example, one could imagine that a database created using a particular technique favors models using that same technique.

Overall, for both frequency and severity, we observe similar links between the average response variable by group and each of the explanatory variables, taken individually. This result is not surprising and confirms that the database creation process was correctly

constructed. The correlation between the variables *Credit.score* and *Insured.age* is also observed in the original database.

The results of the first model containing only the categorical covariates are similar to those obtained with the synthetic database. In particular, we note the weak impact (but which does not seem zero) of the *Marital* variable on both frequency and severity.

Overall, adding continuous variables leads to the same conclusions as those drawn from the synthetic data. Nevertheless, we raise a yellow flag: while the XGBoost model performs significantly better than the other models for the simulated data, that is not the case on the actual data (see tables below). Indeed, there is a slight improvement, but the cost–benefit ratio works against the XGBoost model for actual data. Having not thoroughly analyzed the method for creating the synthetic database, we cannot offer a clear explanation for this phenomenon.

Most conclusions obtained regarding the impact of telematics on the usefulness of sensitive variables remain valid on the original database with the following caveat: the effect is sometimes less significant on the actual data.

5.2. Claim Frequency

Table 5.1 illustrates the various scores achieved for each database, allowing us to delve deeper into the comparison between synthetic data and real data. As before, a small score value indicates a better model. The XGBoost model appears to outperform the other models for both databases; however, as previously mentioned, it exhibits even better performance with synthetic data.

Table 5.1. Prediction Scores (Frequency)

	Log.		MSE	
Model	Synthetic Data	Original Data	Synthetic Data	Original Data
Base	0.17674	0.16794	0.04545	0.04206
GLM (trad.)	0.17359	0.16564	0.04514	0.04186
LASSO (optimal)	0.15536	0.15152	0.04239	0.03992
LASSO (pars.)	0.15704	0.15187	0.04261	0.04014
LASSO* (optimal)	0.15373	0.15072	0.04209	0.03973
LASSO* (pars.)	0.15481	0.15152	0.04226	0.03992
XGBoost	0.12142	0.14995	0.03110	0.03935
XGBoost*	0.12373	0.14893	0.03250	0.03909

5.2.1. Residuals and Sensitive Variables

We can also revisit some of the graphs developed earlier in the paper to better compare the results obtained with the two databases. One of the most critical graphs is the analysis of model residuals concerning sensitive covariates.

As a reminder, we use a model's prediction as an offset variable and assess whether the sensitive covariates still appear to capture a trend. If the resulting curve is horizontal and close to 1 for all possible values of a sensitive covariate, it indicates that telematics variables seem to have captured that covariate's predictive capacity.

Figures 5.1 through 5.10 compare the residual curves for real and synthetic data. The main difference lies in credit score: we observe that reducing its impact through the addition of telematics information is more significant with synthetic data than with real data.

Figure 5.1. Credit Score – Synthetic Data Set

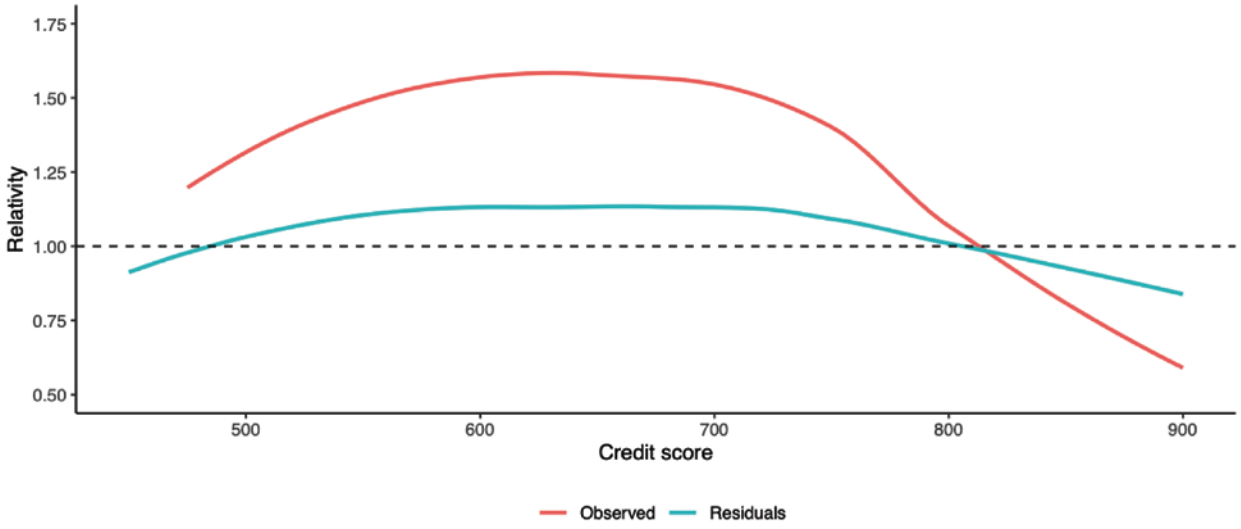


Figure 5.2. Credit Score – Original Data Set

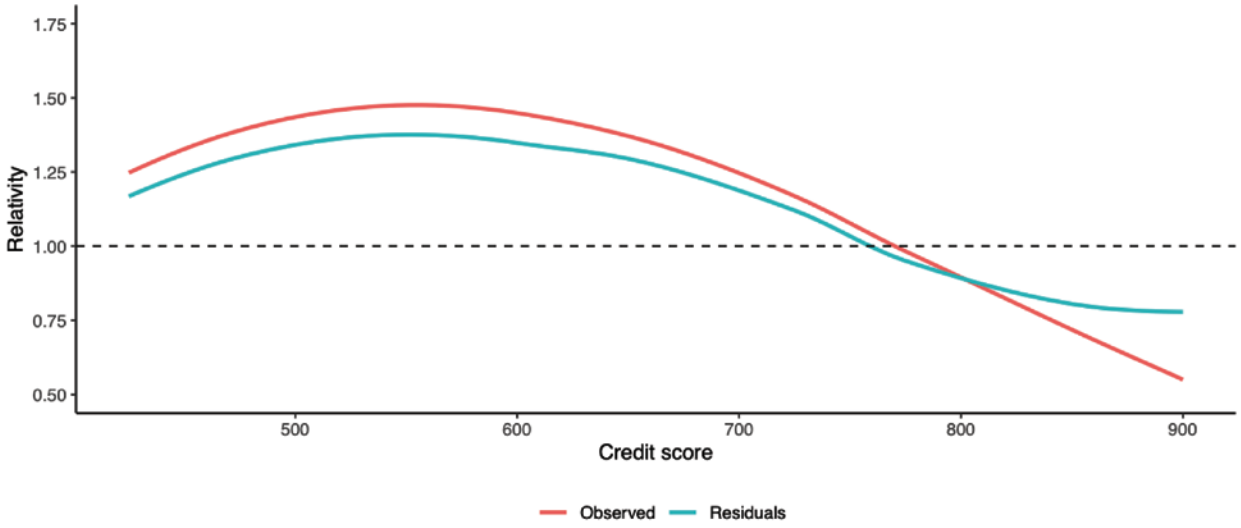


Figure 5.3. Age – Synthetic Data Set

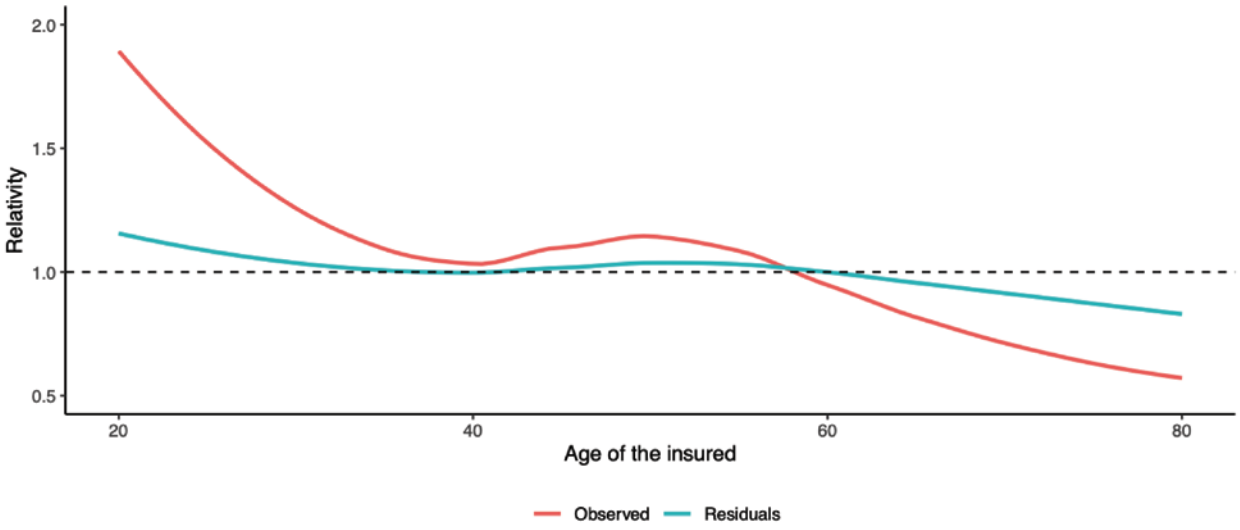


Figure 5.4. Age – Original Data Set

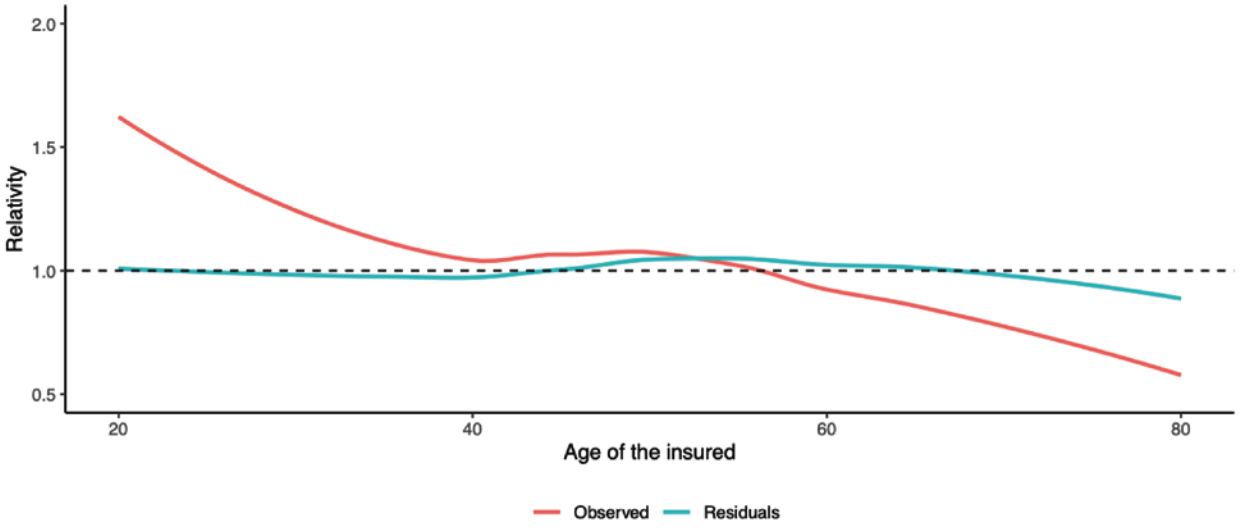


Figure 5.5. Sex – Synthetic Data Set

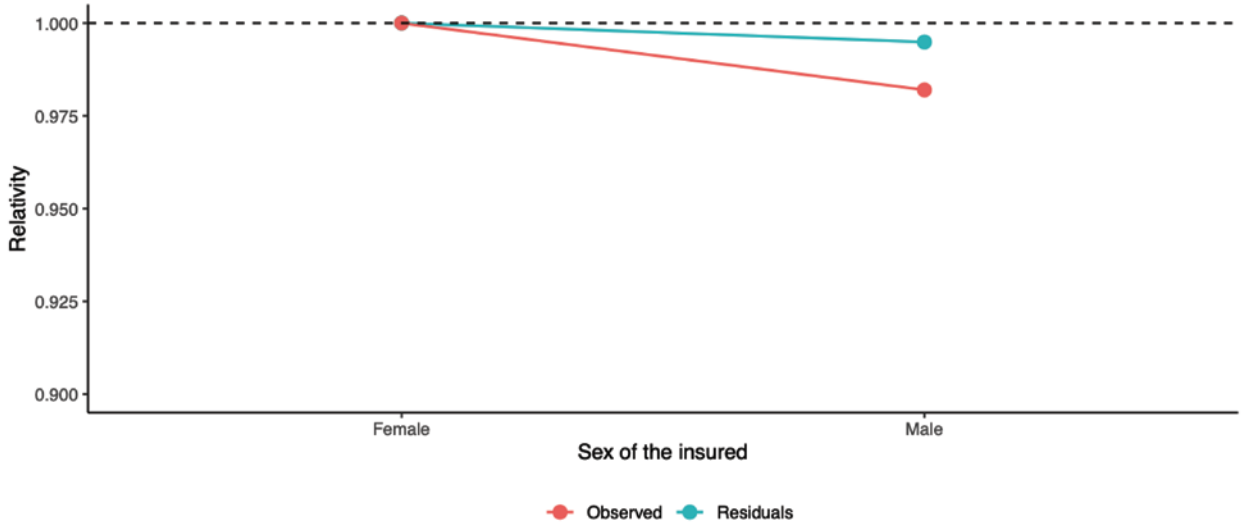


Figure 5.6. Sex – Original Data Set

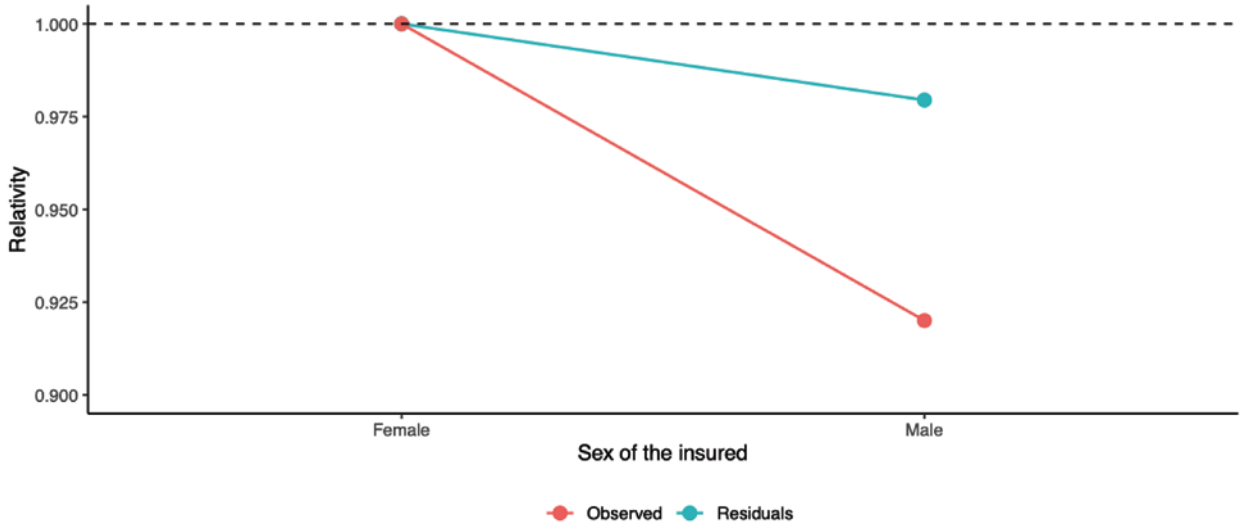


Figure 5.7. Marital – Synthetic Data Set

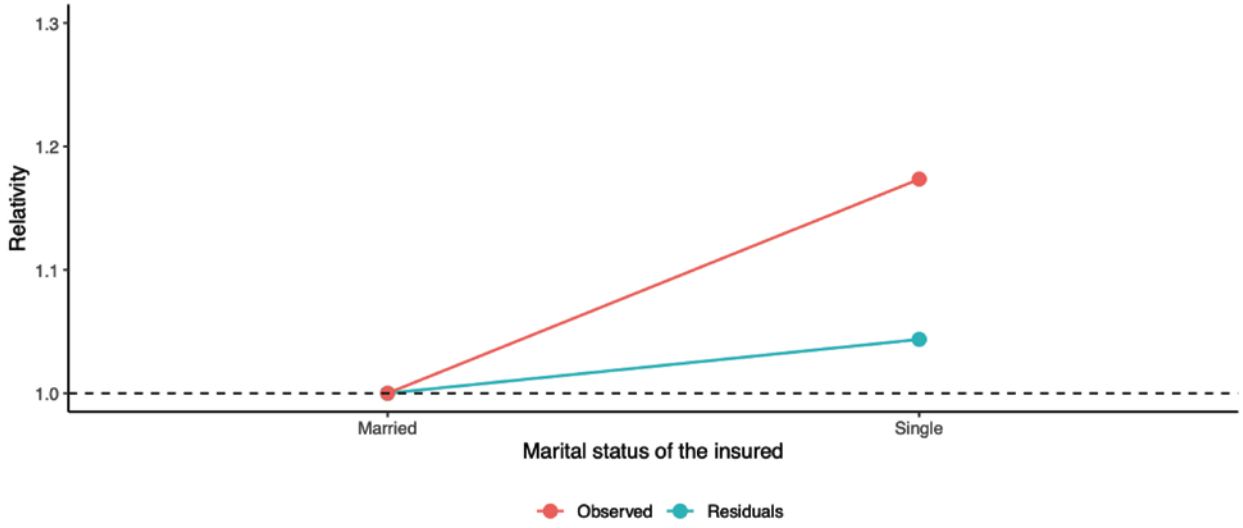


Figure 5.8. Marital – Original Data Set

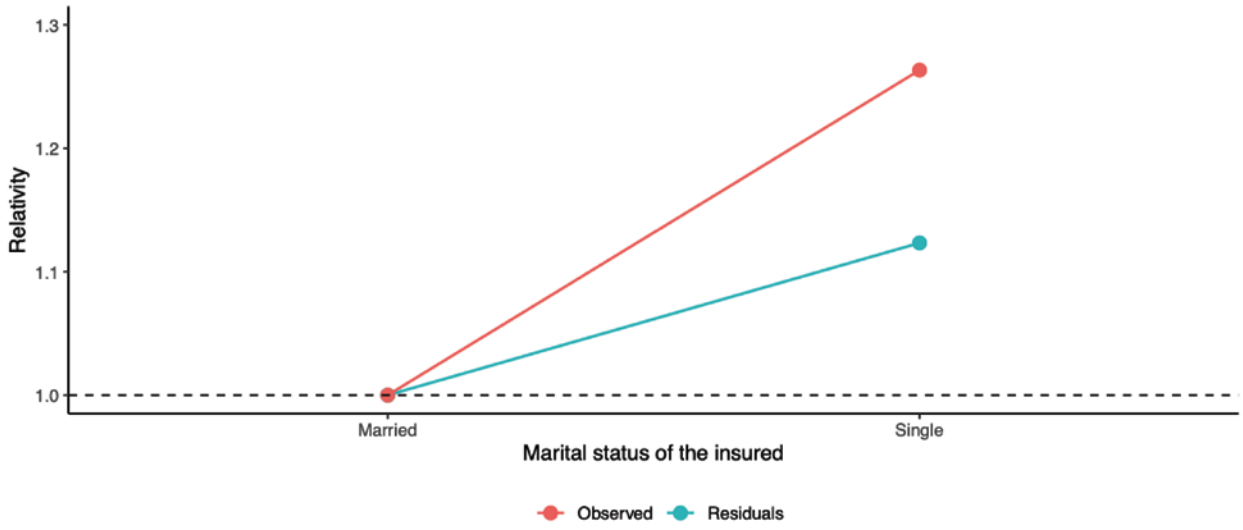


Figure 5.9. Territory – Synthetic Data Set

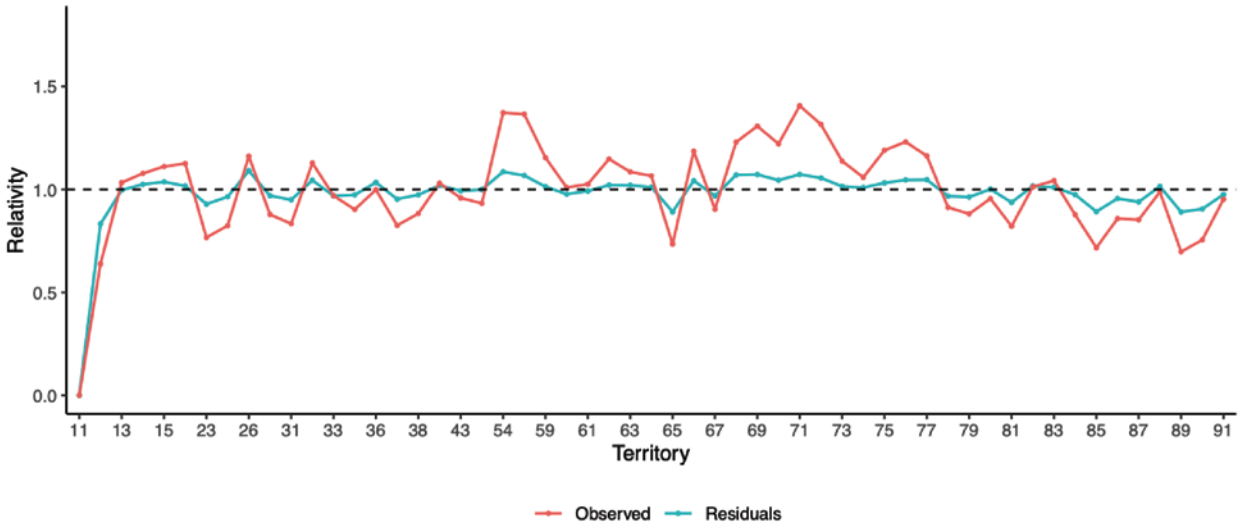
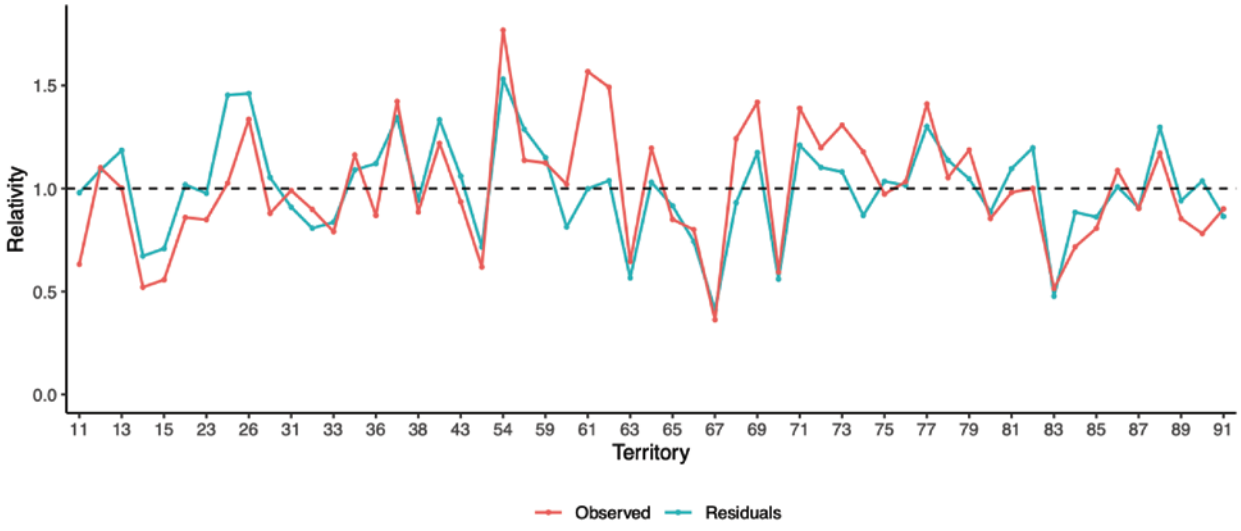


Figure 5.10. Territory – Original Data Set



5.3. Claim Severity

We can also compare the different models for severity. Table 5.2 shows the prediction scores obtained for the various models across both databases. The conclusion aligns with the frequency analysis: the XGBoost model produces the best prediction scores, but the improvement in predictive quality is more significant with synthetic data than with original data. For instance, concerning real data, we observe that the scores obtained for the lasso model (optimal) are pretty close to those achieved by XGBoost. In contrast, the difference between those two models is much more pronounced with synthetic data.

5.3.1. Residuals and Sensitive Variables

Figures 5.11 through 5.20, similarly to the frequency analysis, present trend analysis graphs of residuals based on sensitive covariates. The results are similar except for the credit score, which remains more significant in severity modeling for real data.

Table 5.2. Prediction Scores (Severity)

	Log.		MSE	
Model	Synthetic Data	Original Data	Synthetic Data	Original Data
Base	9.29504	9.51319	21.82679	62.21956
GLM (trad.)	9.23655	9.46042	21.16556	61.55924
LASSO (optimal)	9.22435	9.46367	19.83764	58.13092
LASSO (pars.)	9.24845	9.50317	20.32402	60.99186
LASSO* (optimal)	9.20054	9.42978	19.31021	56.38442
LASSO* (pars.)	9.20843	9.44597	19.59126	57.90176
XGBoost	9.03389	9.41885	15.72516	56.39918
XGBoost*	9.01906	9.24845	15.40489	55.87990

Figure 5.11. Credit score – Synthetic Dataset

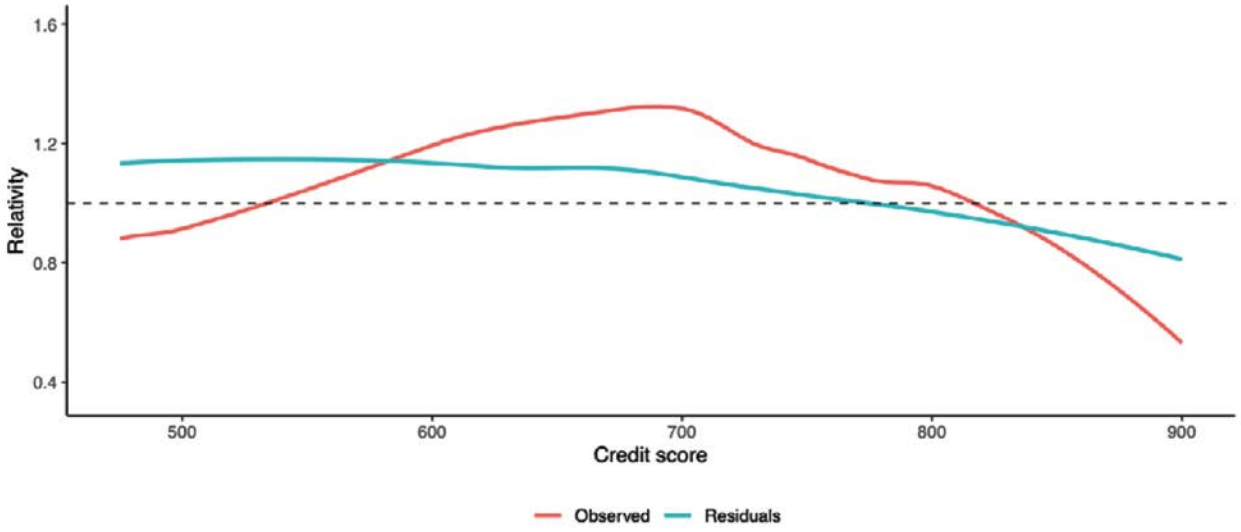


Figure 5.12. Credit score – Original Dataset

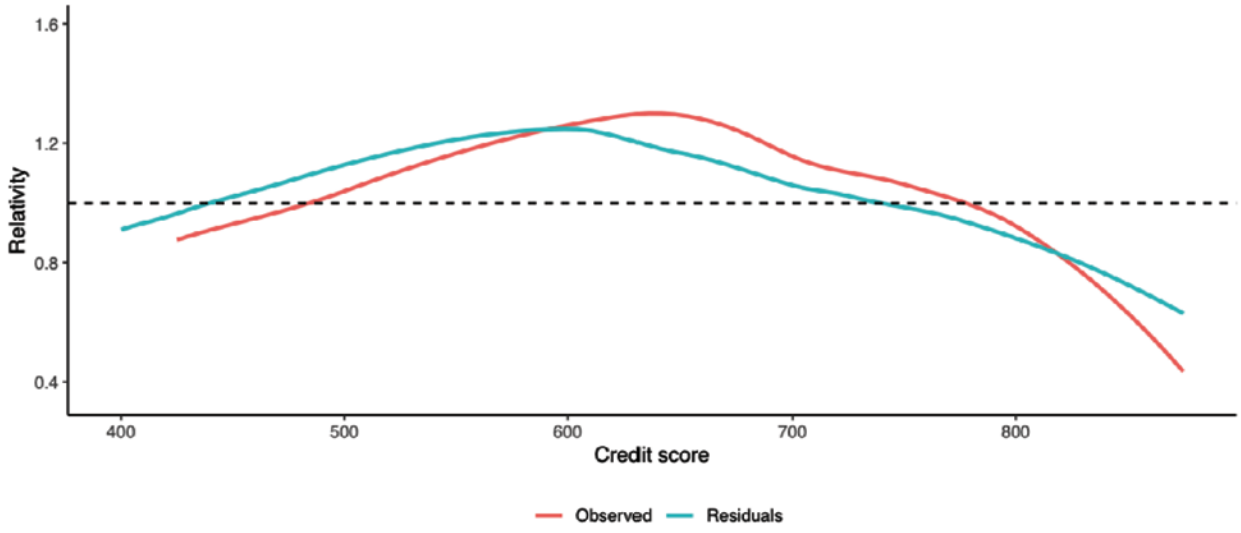


Figure 5.13. Age – Synthetic Dataset

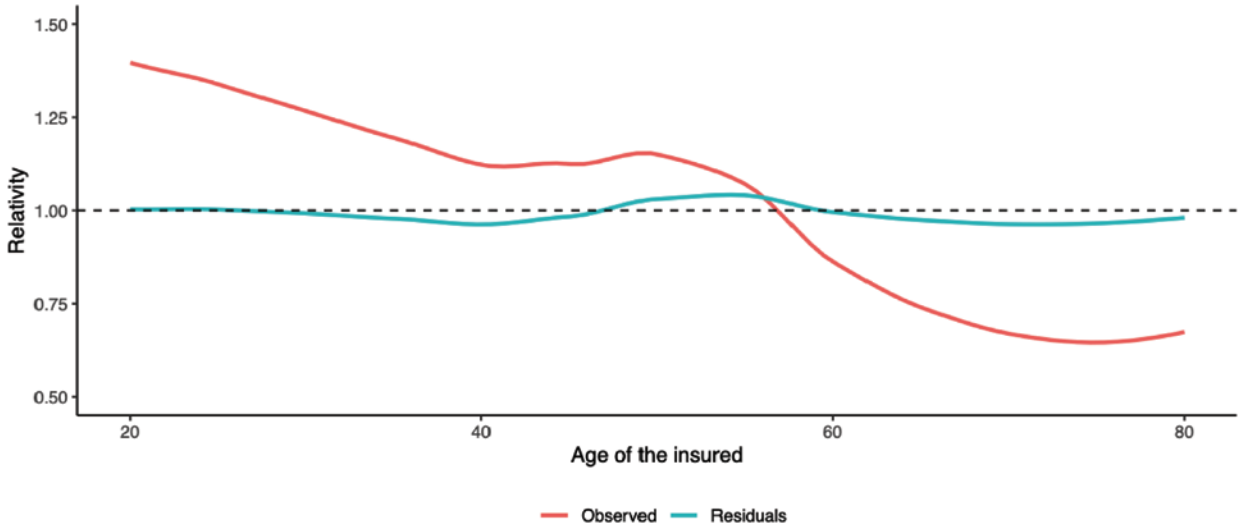


Figure 5.14. Age – Original Dataset

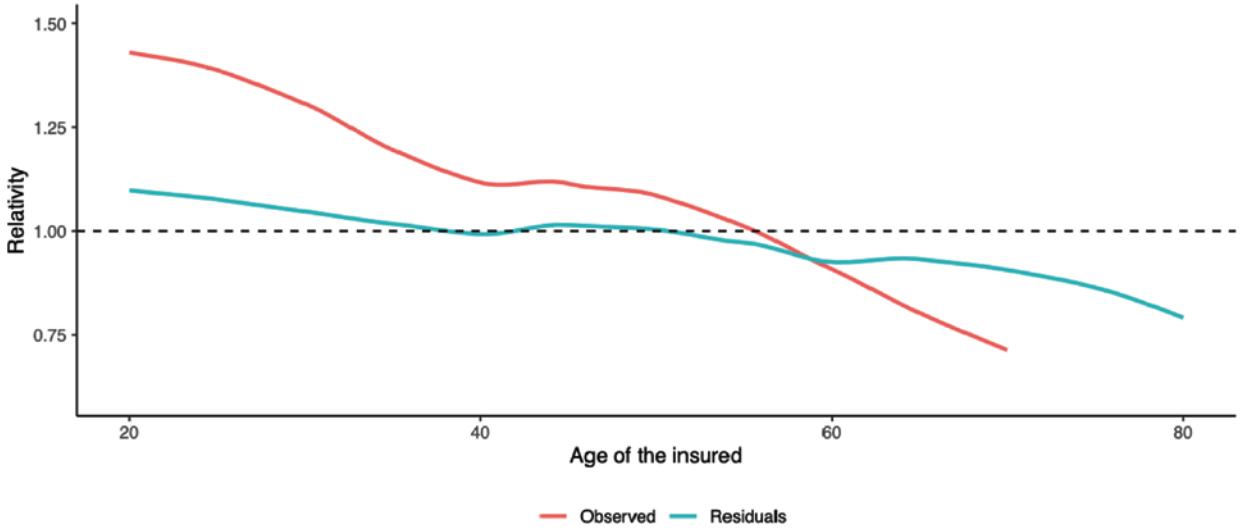


Figure 5.15. Sex – Synthetic Dataset

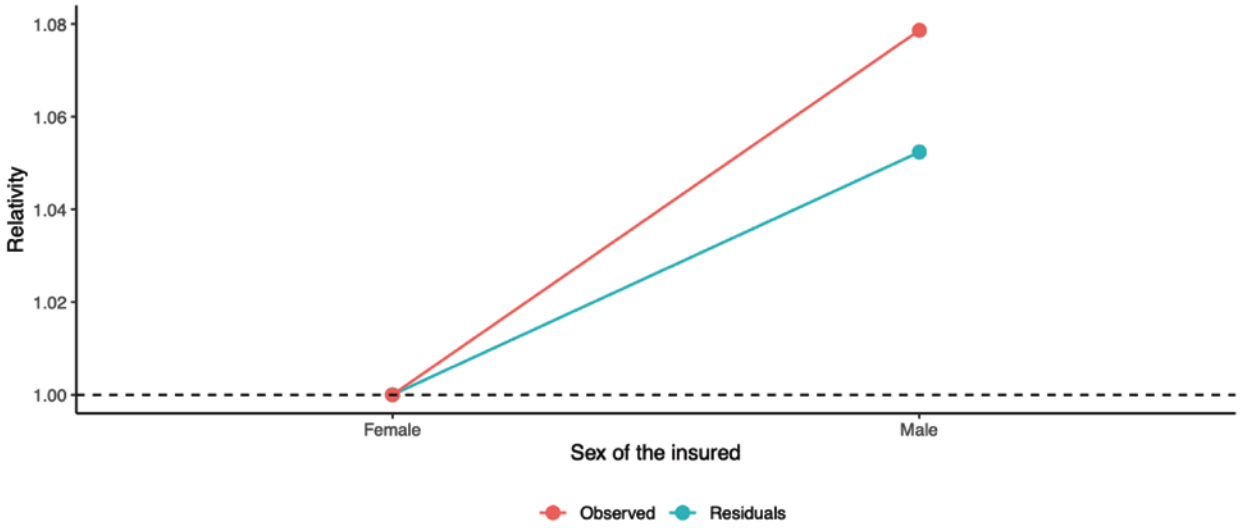


Figure 5.16. Sex – Original Data Set

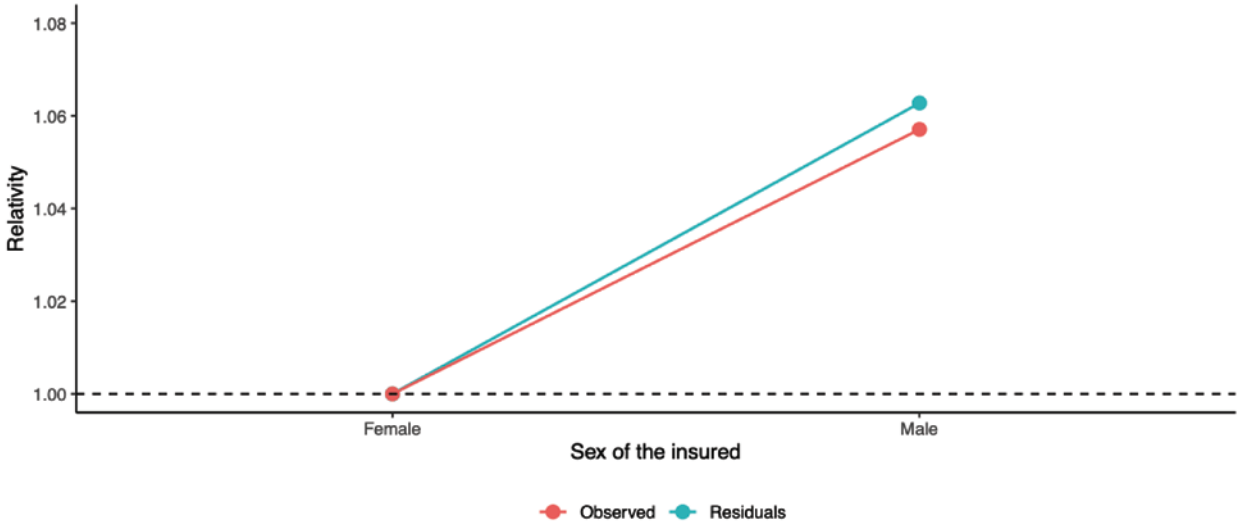


Figure 5.17. Marital – Synthetic Data Set

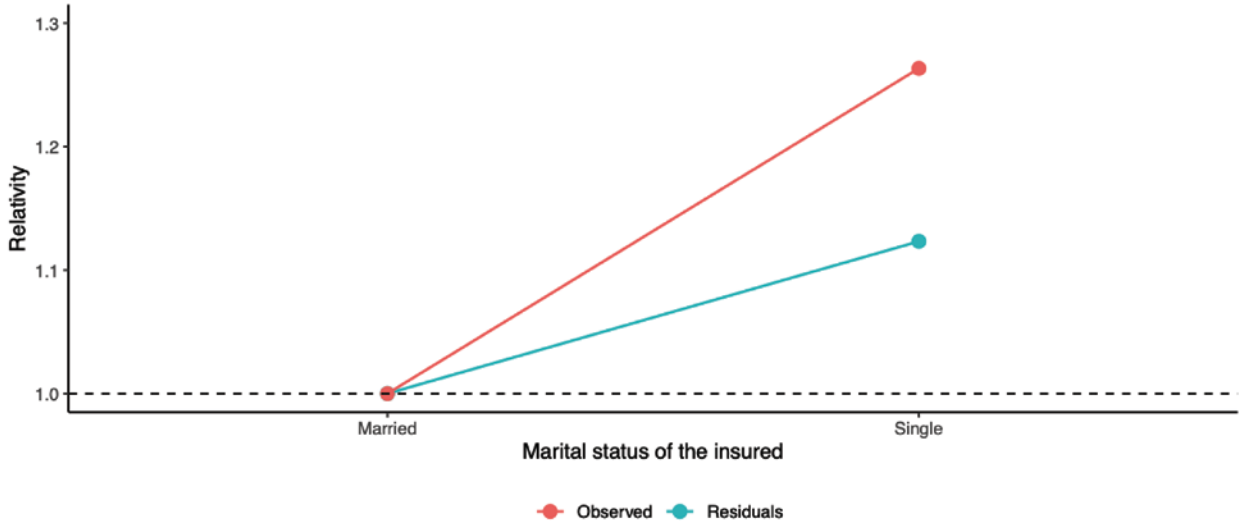


Figure 5.18. Marital – Original Data Set

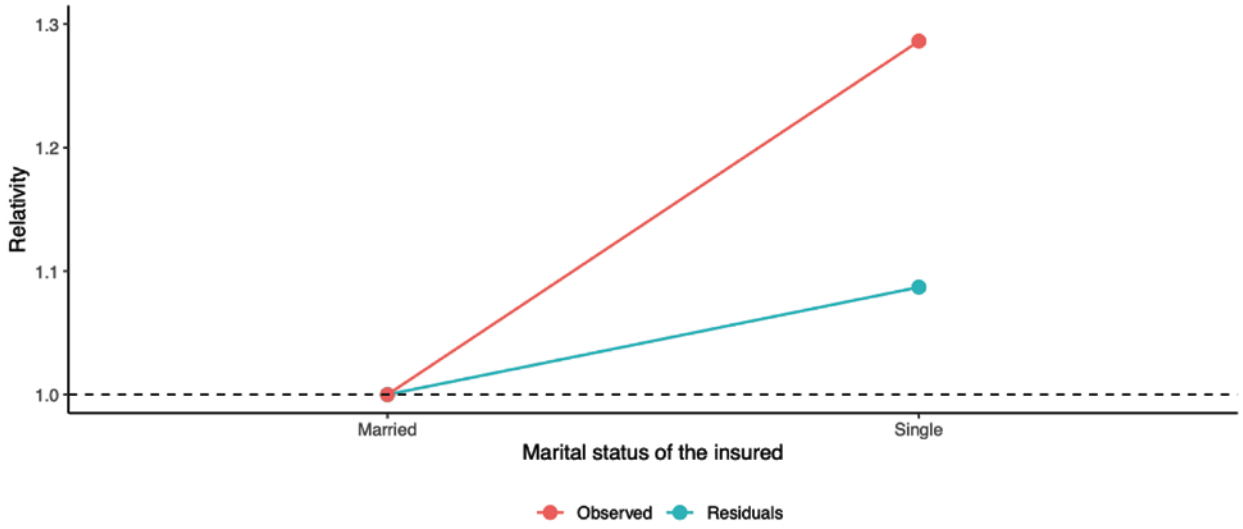


Figure 5.19. Territory – Synthetic Data Set

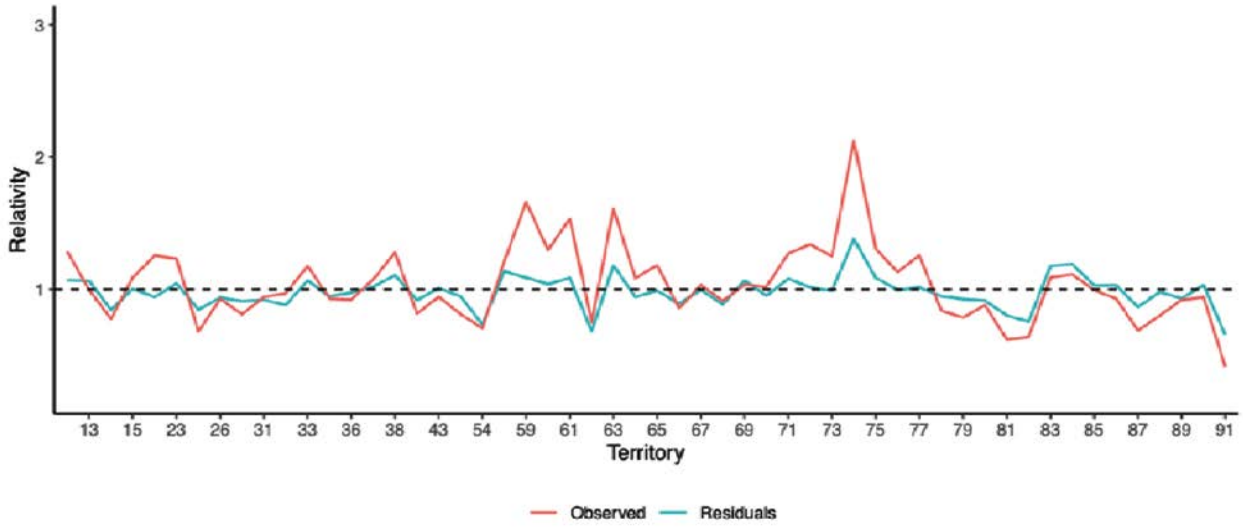
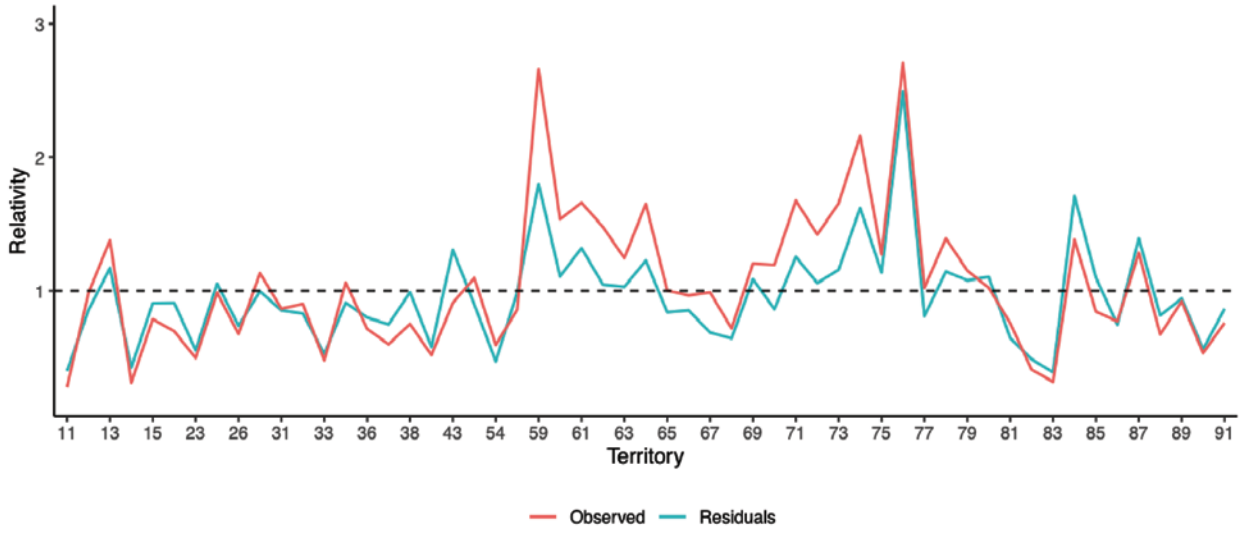


Figure 5.20. Territory – Original Data Set



6. Conclusion

In this paper, we explore the use of telematics and usage-based insurance technologies to reduce the dependence on sensitive information for determining insurance pricing. Our analysis finds that **telematics variables,** such as miles driven, hard braking, hard acceleration, and days of the week driven, significantly **reduce the need to include age, sex, and marital status in the claims frequency and severity models.** Whereas the need for geographic territory and credit score appeared to have been significantly reduced in the model built on synthetic data, the reduction was significantly muted when the approach was validated with the real-world data set.

Although we could not eliminate all of the sensitive variables from the model, this analysis shows there is still value in insurers testing the addition of telematics in their models to potentially reduce reliance on sensitive information that could result in actual or perceived bias. It is a well-established fact that machine learning-based ("black box") approaches have demonstrated higher predictive power compared to regression-based methods. The results of this study affirmed that conclusion. Regression-like models are favored for their balance of simplicity and accuracy. Conversely, gradient-boosting models offer significantly greater accuracy, albeit at the cost of transparency and explainability. The analyses are based on a synthetic database generated from real-world data from a major Canadian insurer, and thus all the results obtained can be easily reproduced. The main results obtained are validated on the original insurer data to confirm that the data reproduction process did not introduce significant bias.

Although the paper proposes a methodology that is based on synthetic data, it is essential to note that individual insurers may arrive at different conclusions when using actual data or different rating variables. In a future project, we intend to improve the synthetic data set further and introduce more advanced training methods to reduce the risk of potential bias.

Appendix A. Overview of the Scientific Literature

Table A.1. Overview of the Sscientific Literature on Telematics Models

Paper	Tools	Main Conclusion
Ayuso, Guillen, and Nielsen (2019)	Count data regression models	Not only the distance traveled by the driver but also driver habits significantly influence the expected number of accidents and, hence, the cost of insurance coverage.
Ayuso, Guillen, and Pérez-Marín (2016a)	Survival models	No gender discrimination is necessary if telematics provides enough information on driving habits.
Boucher, Côte, and Guillen (2017)	Generalized additive model (GAM) for cross-sectional data	Neither distance nor duration is proportional to claim frequency but that frequency tends rather to stabilize once a certain distance or duration has been reached.
Boucher, Pérez-Marín, and Santolino (2013)	Generalization of the offset Poisson regression model	The association between the number of kilometers and claim frequency is not properly captured by a linear relationship.
Boucher and Turcotte (2020)	GAM for location, scale, and shape	The relationship between frequency and distance driven is approximately linear and the apparent nonlinearity is due to residual heterogeneity incorrectly captured by GAMs.
Duval, Boucher, and Pigeon (2022)	Logistic regression with lasso penalty	Telematics data becomes redundant after about three months or 4,000 kilometers of observation from a claim classification perspective.
Duval, Boucher, and Pigeon (2023a)	Anomaly detection algorithm	A routine and a peculiarity anomaly score for each trip can improve classification.
Gao, Meng, and Wüthrich (2019)	*K*-means classification, principal components analysis, neural networks	They recommend the use of speed-acceleration heatmaps for car insurance pricing.
Gao, Wang, and Wüthrich (2022)	Boosting Poisson regression models	Both classical actuarial risk factors and telematics car driving data are necessary to receive the best predictive models.

Table A.1. Overview of the Sscientific Literature on Telematics Models (*Continued*)

Paper	Tools	Main Conclusion
Gao and Wüthrich (2019)	Convolutional neural network	They present a method to appropriately allocate individual car driving trips to selected drivers.
Guillen et al. (2019)	Zero-inflated Poisson model	A learning effect exists for large values of distance traveled; speed limit violations and driving in urban areas increase the expected nu mber of accident claims.
Guillen, Nielsen, and Pérez-Marín (2021)	Poisson regression models	Hard-braking and acceleration events as well as smartphone use while driving increase the cost of insurance.
Huang and Meng (2019)	Classification algorithm	They propose a way to bin continuous telematics variables to create a finite number of risk classes and increase interpretability.
Lemaire, Park, and Wang (2016)	Probit and ordered probit regression models	Annual mileage is an extremely powerful predictor of the number of claims at-fault
Paefgen, Staake, and Fleisch (2014)	Multivariate logistic regression models	There is a nonlinear relationship between mileage and accident risk.
Paefgen, Staake, and Thiesse (2013)	Classification algorithm	They introduce a novel way to aggregate telematics information into what they call an *aggregate risk factor.*
So, Boucher, and Valdez (2021)	Multiclass adaptive boosting algorithm	The proposed algorithm outperforms other learning models designed to handle class imbalances.
Verbelen, Antonio, and Claeskens (2018)	GAM and compositional predictors	Telematics variables increase the predictive power and render the use of gender as a discriminating rating variable redundant.
Wüthrich (2017)	Techniques from pattern recognition and machine learning	Driving styles can be categorized and used for a regression analysis in car insurance pricing.

Appendix B. Theoretical Concepts

Herein we present the general structure underlying our analysis and define the main goodness-of-fit measures we use in the paper. We use formal, mathematical definitions to be precise and to facilitate reproducibility.

B.1. Ratemaking Theory

For an insurance contract t, $t = 1, \ldots, T$, we define the following components:

- The random variable N_t represents the annual claims number.
- For $n_t = n > 0$, let $\boldsymbol{Z}_t = (Z_{t,1}, \ldots, Z_{t,n})$ be the random vector of claims costs associated with this contract. We define this vector only for a positive observed claims number.
- The premium is generally calculated considering specific observable characteristics of each contract. We denote these characteristics by $\boldsymbol{X}_t = (x_{t,0}, \ldots, x_{t,q})$.
- The random variable Y_t represents total cost associated with contract t: $Y_t = \sum_{k=1}^{N_t} Z_{t,k}$, if $N_t > 0$, and 0 if $N_t = 0$.

An insured person exchanges his risk Y_t against a constant π corresponding to an insurance premium. Generally, in actuarial science, we minimize the squared error $min_p E[(Y_t - p)^2 | \boldsymbol{X}_t]$ to obtain the premium: $\pi_t^{(Y)} = E[Y_t | \boldsymbol{X}_t]$. We can consider two strategies to evaluate this value:

1. The frequency–severity approach, where we multiply the frequency component premium and the severity component premium according to some assumptions, i.e.,

$$\underbrace{E\left[Y_t \middle| \boldsymbol{X}_t\right]}_{\pi_t^{(Y)}} = \underbrace{E\left[N_t \middle| \boldsymbol{X}_t\right]}_{\pi_t^{(N)}} \underbrace{E\left[Z_{t,k} \middle| \boldsymbol{X}_t\right]}_{\pi_t^{(Z)}}.$$

   In the frequency–severity approach, for a contract t, we generally assume independence between the frequency and the severity components. Moreover, we assume that $Z_{t,k} | \boldsymbol{X}_t$ are identically distributed.

2. The conditional approach, where we directly model the total loss distribution, i.e.,

$$\pi_t^{(Y)} = E\left[Y_t \middle| \boldsymbol{X}_t\right] = \int y f_{y_t | \boldsymbol{x}_t}(y) dy.$$

   Although the conditional approach is possible, usually using a Tweedie family distribution, it complicates the interpretation of the results. For example, the effect of the same covariate on severity may hide the effect of a covariate on frequency. Thus, this analysis utilizes a frequency-severity approach.

The Poisson distribution is the base model for the number of claims in property and casualty insurance; it has valuable and well-known statistical properties. The probability mass function is

$$Pr\left(N_t = n \middle| \boldsymbol{X}_t\right) = \left(\lambda_{i,t}\right)^n exp\left(-\lambda_{i,t}\right)/n!, n = 0, 1, 2, \ldots,$$

where λ_t is a function. Traditionally, we assume a log-linear relationship between the mean parameter and the policyholder's and claim's characteristics such as sex, age, and marital status, e.g., $\lambda_t = exp(\boldsymbol{X}_t\boldsymbol{\beta})$ and $\boldsymbol{\beta}$ is a column vector containing parameters. The Poisson distribution implies equidispersion, i.e., $E[N_t|\boldsymbol{X}_t] = Var[N_t|\boldsymbol{X}_t]$, which is, usually, a too strong assumption in ratemaking. However, for our project, we restrict ourselves to the overdispersed Poisson, where $Var[N_t|\boldsymbol{X}_t] = \phi E[N_t|\boldsymbol{X}_t] > E[N_t|\boldsymbol{X}_t]$. For our analysis of claim severity, we use the gamma distribution. The probability density function is

$$f_{Z|\boldsymbol{x}_t}(z) = \frac{z^{\alpha-1}e^{-z/\theta}}{\Gamma(\alpha)\theta^{\alpha}}, z > 0,$$

where α is the shape parameter and θ is the scale parameter. The expected value is $\alpha\theta$ and the variance is $\alpha\theta^2$. Usually, severity is less heterogeneous than frequency, and many available covariates have little impact on the prediction. Beyond the gamma distribution, the inverse Gaussian distribution is also a possibility to consider. However, with a cubic variance, the possibility of having statistically significant estimators is even lower.

B.2. Goodness of Fit

The idea of the prediction score is to obtain a numerical value to assess the quality of a model's prediction on new data. By convention, we assume the objective is to minimize the prediction score. More specifically, to evaluate the model *P*, we calculate a penalty *s(P, x)* to determine the prediction error.

For a model *P*, we get the model's prediction score, *S(P)*, by taking the average (or the sum) penalty over *x* observations in a database (which has never been used to estimate any parameters or properties of the model *P*):

$$S(P) = \frac{1}{n}\sum_{i=1}^{n} s(P, x_i).$$

We can list some relevant penalties:

- Logarithmic penalty: $logs(P, x) = -log(Pr(N = x))$, where $Pr()$ is the probability mass function under model *P*, or $logs(P, x) = -log(f_Z(x))$, where $f_Z()$ is the probability density function under model *P*
- Quadratic penalty: $quad(P, x) = -2Pr(N = x) + \left(\sum_{j=0}^{\infty} Pr(N = j)^2\right)$, where $Pr()$ is the probability mass function under model *P*

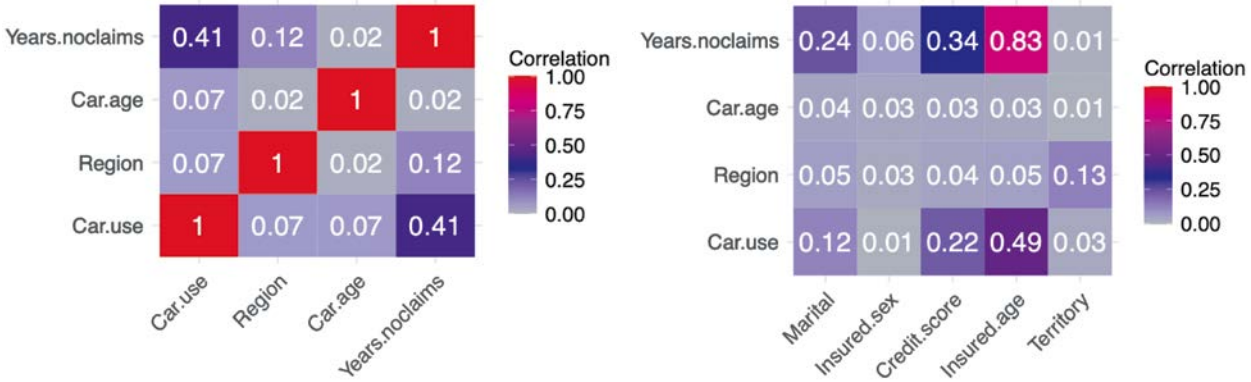
- Squared error penalty: $sq.err(P, x) = (x - \lambda_P)^2$, where λ_P is the predicted value under model P
- Spherical penalty: $sph(P, x) = -\frac{Pr(N = x)}{\sum_{j=0}^{\infty} Pr(N = j)^2}$, where $Pr()$ is the probability mass function under model P
- Dawid-Sebastiani penalty: $DSP(P, x) = \left(\frac{x - \lambda_P}{\sqrt{\lambda_P}}\right)^2 + ln(\lambda_P)$, where $Pr()$ is the probability mass function and λ_P is the predicted value under model P

For more details on the properties of scores and the assessment of counting models, one can consult Czado, Gneiting, and Held (2009).

Appendix C. Correlation Matrices

C.1. Traditional: Nonsensitive Covariates

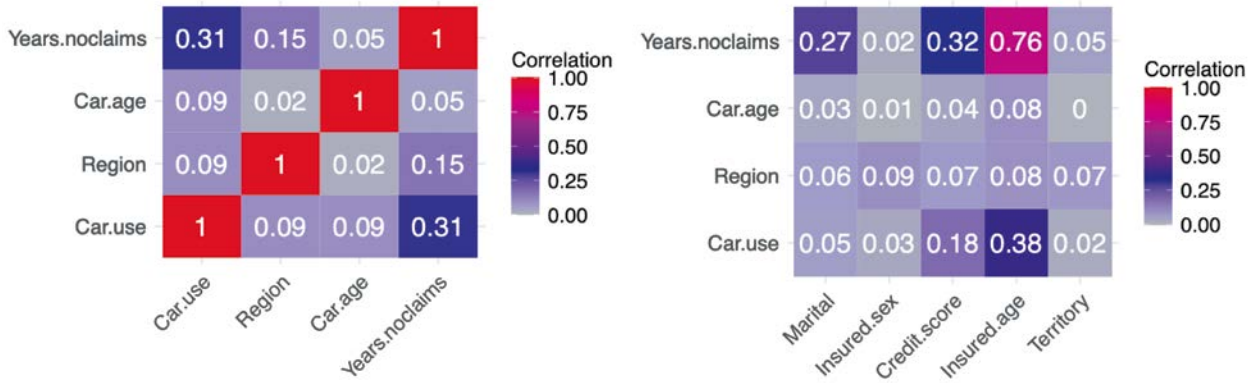
Figure C.1. Correlation between Traditional Covariates (Frequency)



(a) Other traditional covariates

(b) With sensitive covariates

Figure C.2. Correlation between Traditional Covariates (Severity)

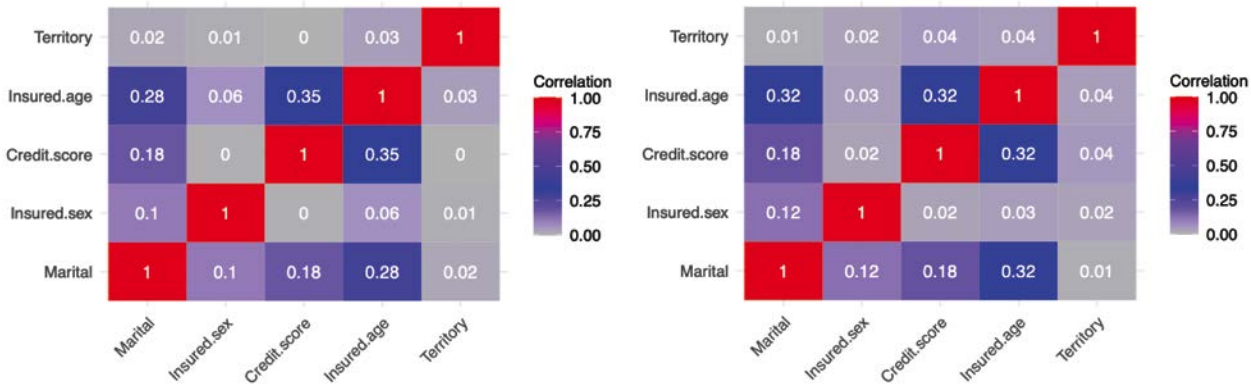


(a) Other traditional covariates

(b) With sensitive covariates

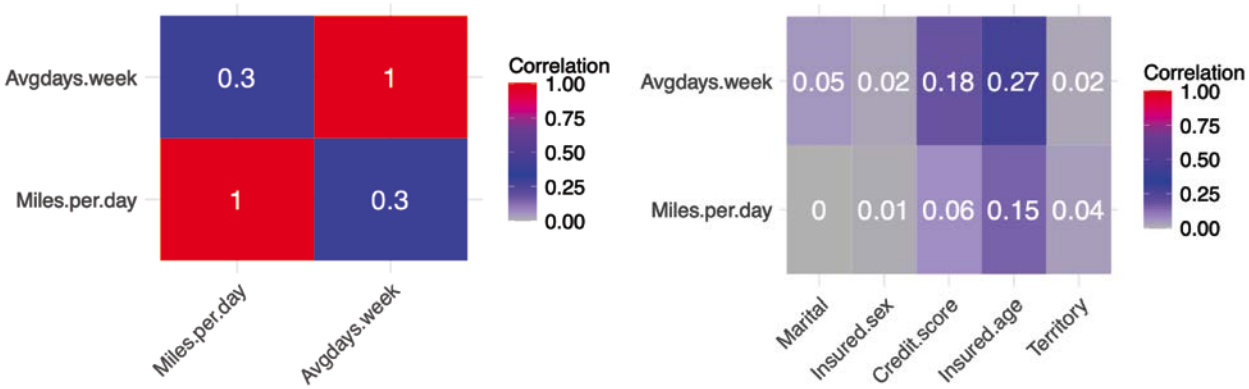
C.2. Traditional: Sensitive Information

Figure C.3. Correlation between Sensitive Covariates for Frequency (Left) and Severity (Right)



C.3. Telematics: Vehicle Usage Level

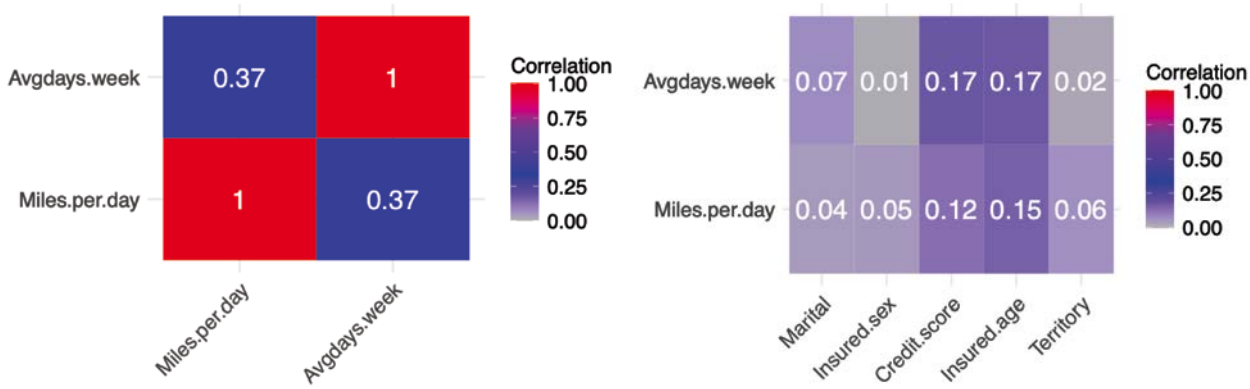
Figure C.4. Correlation between Covariates (Frequency)



(a) Telematics covariates

(b) With sensitive covariates

Figure C.5. Correlation between Covariates (Severity)

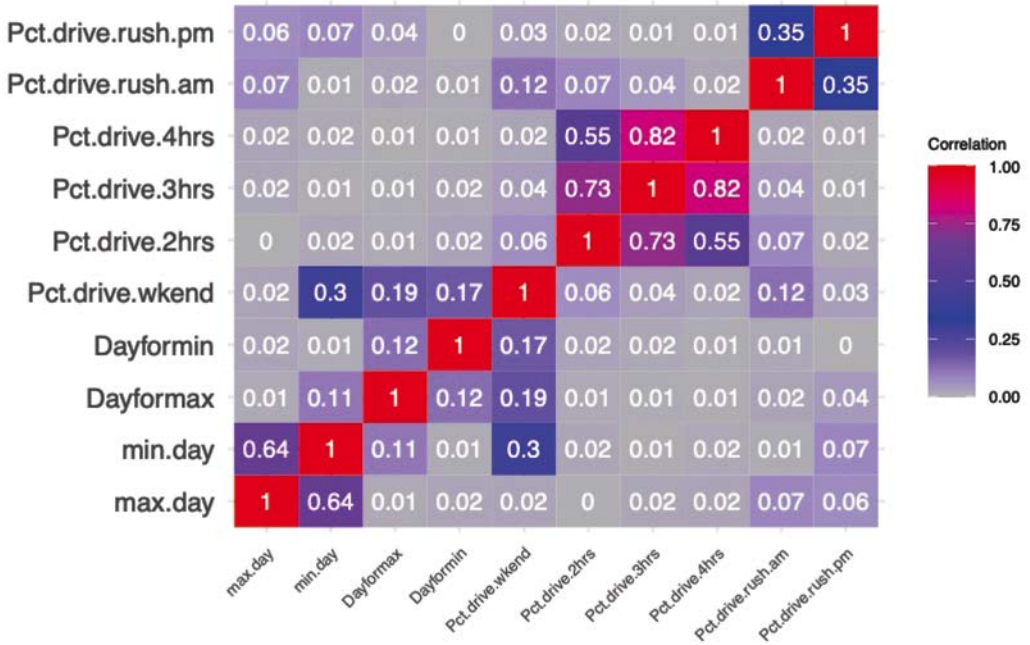


(a) Telematics covariates

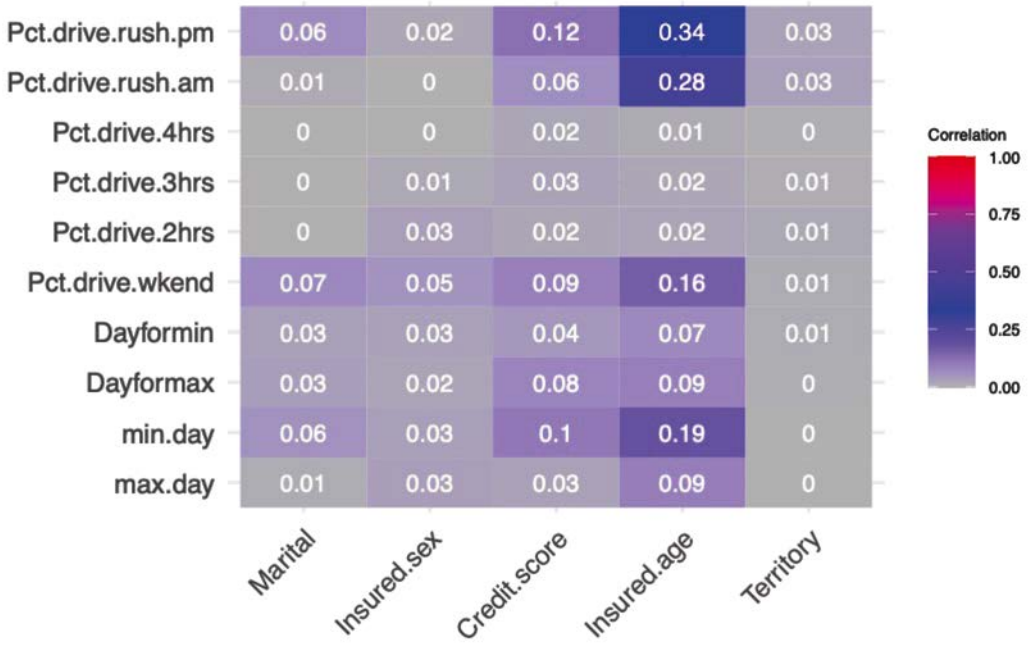
(b) With sensitive covariates

C.4. Telematics: Type of Vehicle Usage

Figure C.6. Correlation between Covariates



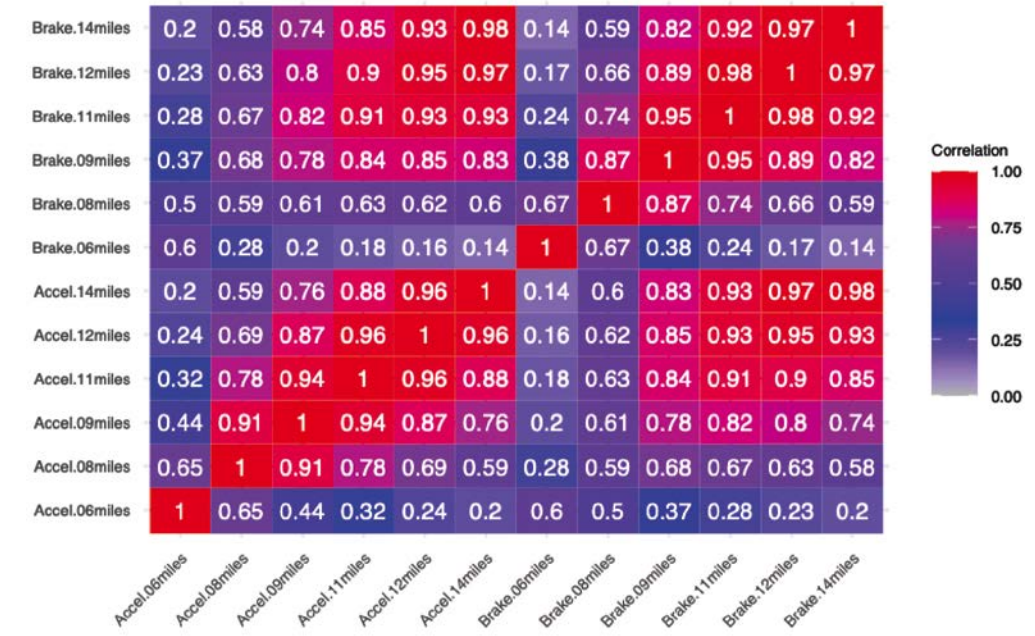
(a) Telematics covariates



(b) With sensitive covariates

C.5. Telematics: Driving Behavior

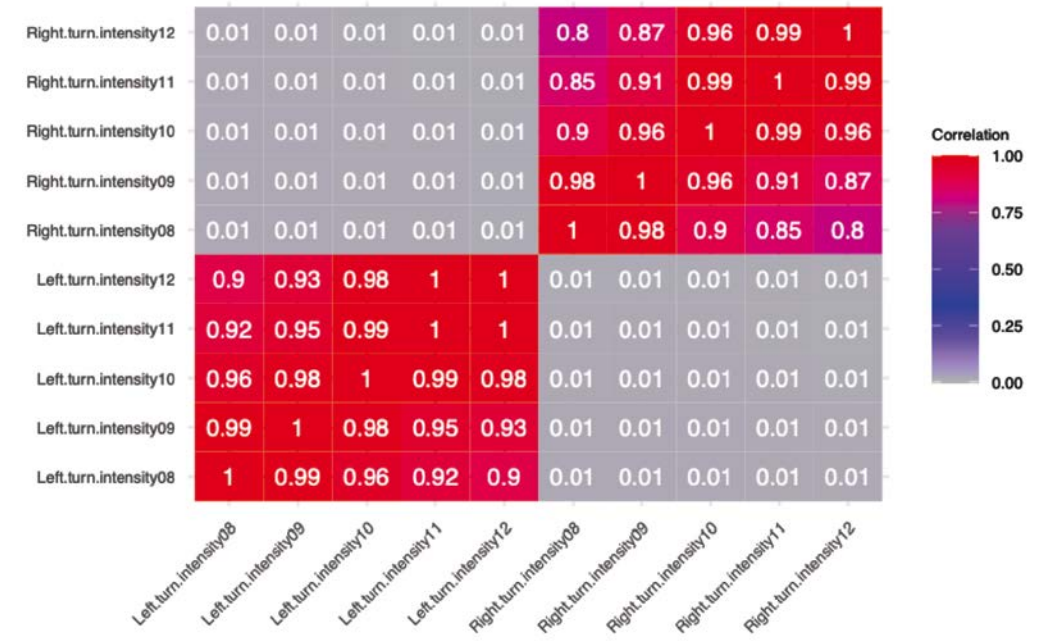
Figure C.7. Correlation between Covariates (Frequency)



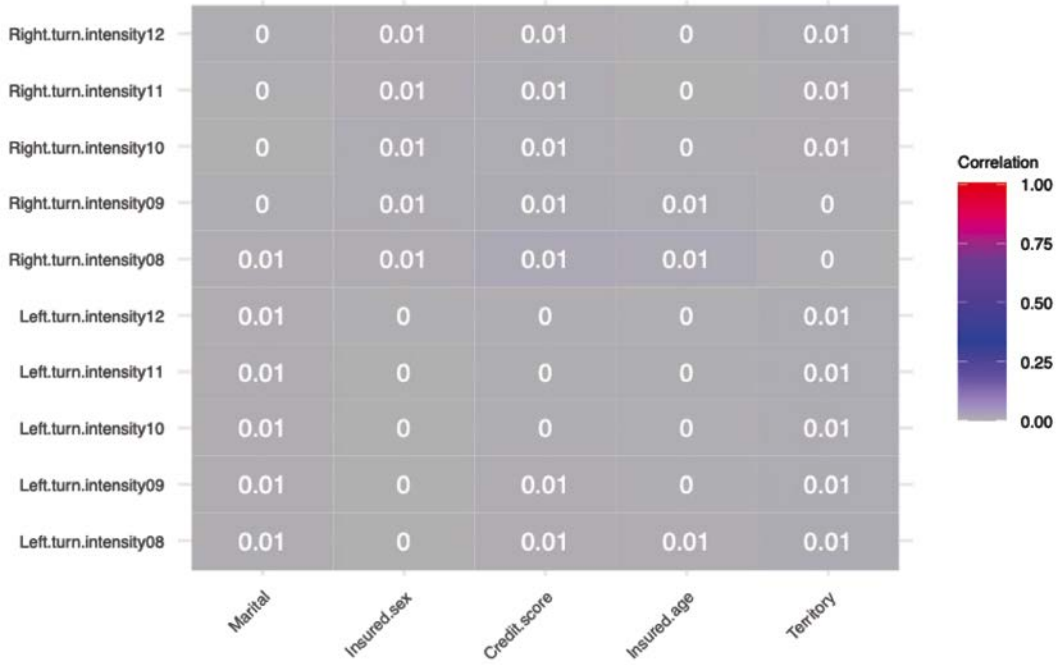
(a) Telematics covariates

(b) With sensitive covariates

Figure C.7. Correlation between Covariates (Frequency) *(Continued)*

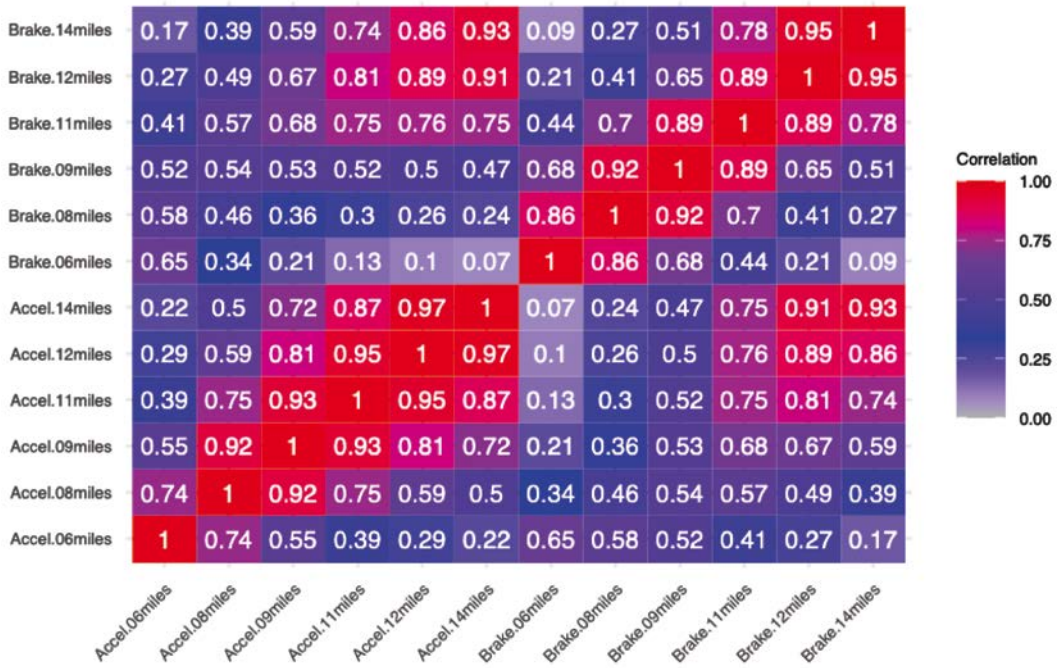


(c) Telematics covariates



(d) With sensitive covariates

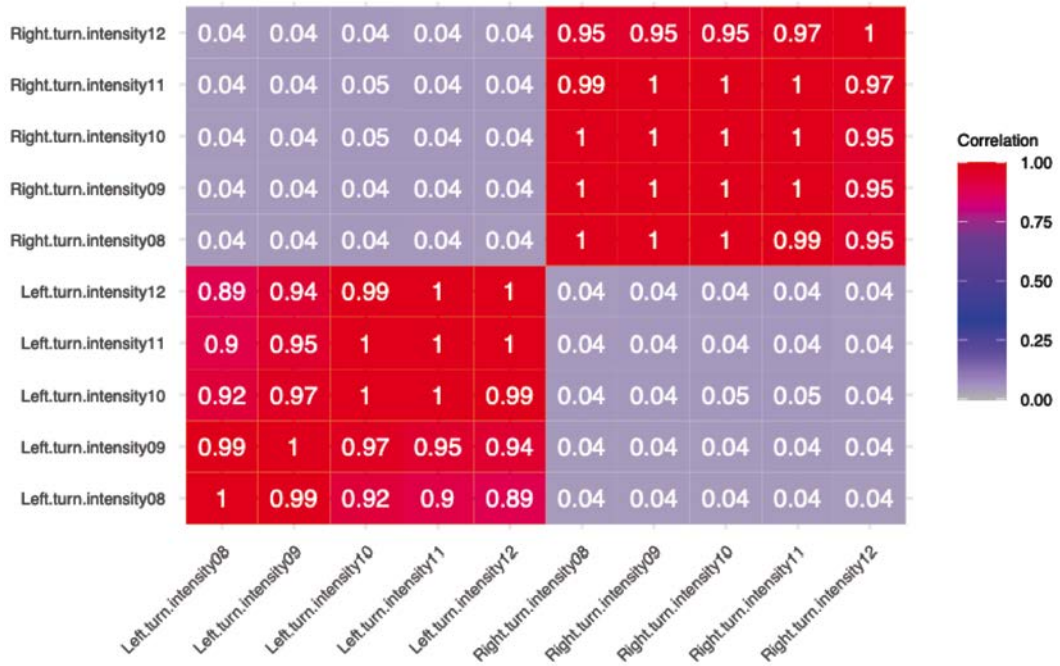
Figure C.8. Correlation between Covariates (Severity)



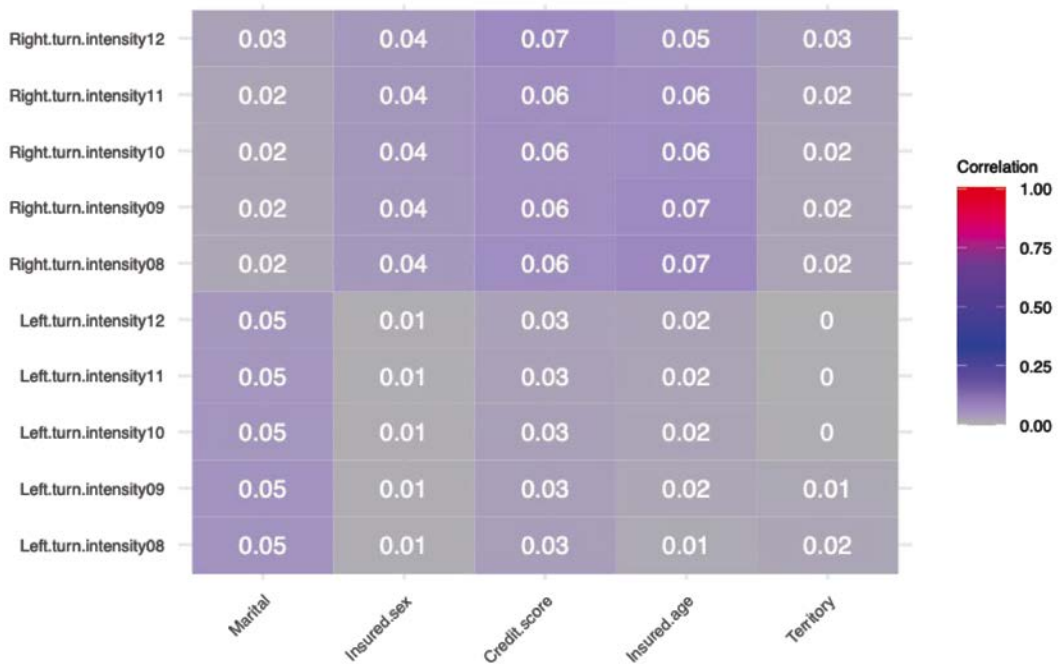
(a) Telematics covariates

(b) With sensitive covariates

Figure C.8. Correlation between Covariates (Severity) *(Continued)*



(c) Telematics covariates



(d) With sensitive covariates

References

Avanzi, B., G. Taylor, M. Wang, and B. Wong. 2021. "SynthETIC: An Individual Insurance Claim Simulator with Feature Control." *Insurance: Mathematics and Economics* 100: 296–308. doi:10.1016/j.insmatheco.2021.06.004.

Ayuso, M., M. Guillen, and J. P. Nielsen. 2019. "Improving Automobile Insurance Ratemaking Using Telematics: Incorporating Mileage and Driver Behaviour Data." *Transportation* 46 (3): 735–52.

Ayuso, M., M. Guillen, and A. M. Pérez-Marín. 2014. "Time and Distance to First Accident and Driving Patterns of Young Drivers with Pay-as-You-Drive Insurance." *Accident Analysis & Prevention* 73: 125–31.

Ayuso, M., M. Guillen, and A. M. Pérez-Marín. 2016a. "Telematics and Gender Discrimination: Some Usage-Based Evidence on Whether Men's Risk of Accidents Differs from Women's." *Risks* 4 (2).

Ayuso, M., M. Guillen, and A. M. Pérez-Marín. 2016b. "Using GPS Data to Analyse the Distance Travelled to the First Accident at Fault in Pay-as-You-Drive Insurance." *Transportation Research Part C: Emerging Technologies* 68: 160–67.

Bansag, R. 2017. "Willis Towers Watson says consumers willing to share driving data" S&P Global Blog, June 7. https://www.spglobal.com/marketintelligence/en/news-insights/trending/iemddg56_yuyoreobpj9wa2.

Bender, M., C. Dill, M. Hurlbert, C. Lindberg, and S. Mott. 2021. *Understanding Potential Influences of Racial Bias on P&C Insurance: Four Rating Factors Explored.* CAS Research Paper Series on Race and Insurance Pricing. Arlington, VA: Casualty Actuarial Society.

Boucher, J.-P., S. Côte, and M. Guillen. 2017. "Exposure as Duration and Distance in Telematics Motor Insurance Using Generalized Additive Models." *Risks* 5 (4).

Boucher, J.-P., and M. Denuit. 2007. "Duration Dependence Models for Claim Counts." *Blätter Der DGVFM* 28 (1): 29–45.

Boucher, J.-P., A. M. Pérez-Marín, and M. Santolino. 2013. "Pay-as-You-Drive Insurance: The Effect of the Kilometers on the Risk of Accident." *Anales del Instituto de Actuarios Españoles* 19: 135–54.

Boucher, J.-P., and R. Turcotte. 2020. "A Longitudinal Analysis of the Impact of Distance Driven on the Probability of Car Accidents." *Risks* 8 (3): 91.

Chibanda, K. F. 2022. *Defining Discrimination in Insurance.* CAS Research Paper Series on Race and Insurance Pricing. Arlington, VA: Casualty Actuarial Society.

Czado, C., T. Gneiting, and L. Held. 2009. "Predictive Model Assessment for Count Data." *Biometrics* 65 (4): 1254–61.

Duval, F., J.-P. Boucher, and M. Pigeon. 2022. "How Much Telematics Information Do Insurers Need for Claim Classification?" *North American Actuarial Journal* 26 (4): 570–90.

Duval, F., J.-P. Boucher, and M. Pigeon. 2023a. "Enhancing Claim Classification with Feature Extraction from Anomaly-Detection-Derived Routine and Peculiarity Profiles." *Journal of Risk and Insurance* 90 (2): 421–58.

Duval, F., J.-P. Boucher, and M. Pigeon. 2023b. "Telematics Combined Actuarial Neural Networks for Cross-Sectional and Longitudinal Claim Count Data." arXiv:2308.01729.

Embrechts, P., and M. V. Wüthrich. 2022. "Recent Challenges in Actuarial Science." *Annual Review of Statistics and Its Application* 9: 119–40.

Gabrielli, A., and M. V. Wüthrich. 2018. "An Individual Claims History Simulation Machine." *Risks* 6 (2): 29.

Gao, G., S. Meng, and M. V. Wüthrich. 2019. "Claims Frequency Modeling Using Telematics Car Driving Data." *Scandinavian Actuarial Journal* 2019 (2): 143–62.

Gao, G., H. Wang, and M. V. Wüthrich. 2022. "Boosting Poisson Regression Models with Telematics Car Driving Data." *Machine Learning* 111 (2): 243–72.

Gao, G., and M. V. Wüthrich. 2019. "Convolutional Neural Network Classification of Telematics Car Driving Data." *Risks* 7 (1).

Guillen, M., J. P. Nielsen, M. Ayuso, and A. M. Pérez-Marín. 2019. "The Use of Telematics Devices to Improve Automobile Insurance Rates." *Risk Analysis* 39 (3): 662–72.
Guillen, M., J. P. Nielsen, and A. M. Pérez-Marín. 2021. "Near-Miss Telematics in Motor Insurance." *Journal of Risk and Insurance* 88 (3): 569–89.
Hastie, T., R. Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition. New York: Springer.
Huang, Y., and S. Meng. 2019. "Automobile Insurance Classification Ratemaking Based on Telematics Driving Data." *Decision Support Systems* 127: 113–56.
Lemaire, J. 1985. *Automobile Insurance: Actuarial Models*. Boston: Kluwer.
Lemaire, J., S. C. Park, and K. C. Wang. 2016. "The Use of Annual Mileage as a Rating Variable." *ASTIN Bulletin* 46 (1): 39–69.
Lichtenstein, E. 2022. "Which States Ban Gender-Rating in Insurance Premiums?" *AgentSync Blog,* March 28. https://agentsync.io/blog/industry-news/which-states-ban-gender-rating-in-insurance-premiums.
Lindholm, M., R. Richman, A. Tsanakas, and M. V. Wüthrich. 2022. "Discrimination-Free Insurance Pricing." *ASTIN Bulletin* 52 (1): 55–89.
Paefgen, J., T. Staake, and E. Fleisch. 2014. "Multivariate Exposure Modeling of Accident Risk: Insights from Pay-as-You-Drive Insurance Data." *Transportation Research Part A: Policy and Practice* 61: 27–40.
Paefgen, J., T. Staake, and F. Thiesse. 2013. "Evaluation and Aggregation of Pay-as-You-Drive Insurance Rate Factors: A Classification Analysis Approach." *Decision Support Systems* 56: 192–201.
Reid, T. R. 1985. "Montana Implements Policy of 'Unisex' Insurance." *Washington Post,* September 30.
So, B., J.-P. Boucher, and E. A. Valdez. 2021. "Cost-Sensitive Multi-Class Adaboost for Understanding Driving Behavior Based on Telematics." *ASTIN Bulletin* 51 (3): 719–51.
Tselentis, D. I., G. Yannis, and E. I. Vlahogianni. 2016. "Innovative Insurance Schemes: Pay as/how You Drive." *Transportation Research Procedia* 14: 362–71.
Verbelen, R., K. Antonio, and G. Claeskens. 2018. "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.
Wu, C.-S. P., and J. C. Guszcza. 2003. "Does Credit Score Really Explain Insurance Losses? Multivariate Analysis from a Data Mining Point of View." *Proceedings of the Casualty Actuarial Society*, Winter, 113–38.
Wüthrich, M. V. 2017. "Covariate Selection from Telematics Car Driving Data." *European Actuarial Journal* 7: 89–108.