Discrimination in Insurance Pricing

by

Carlos Andrés Araiza Iturria

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Actuarial Science

Waterloo, Ontario, Canada, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:
Emiliano Valdez
Professor, Dept. of Mathematics,
University of Connecticut

Supervisors:
Mary Hardy
Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

Paul Marriott
Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

Internal Members:
Ben Feng
Assistant Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

Michael Wallace
Associate Professor, Dept. of Statistics and Actuarial Science,
University of Waterloo

Internal-External Member:
Hilary Bergsieker
Associate Professor, Dept. of Psychology,
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Discrimination is an ongoing problem in the insurance industry that persists, regardless of intent, when the insurer blinds the pricing process from socially controversial or legally prohibited input. In this thesis, we contextualize the problem in property and casualty insurance, considering the prevailing legislation in the United States and the European Union.

In Chapter 1 we introduce the problem of discrimination in insurance, and present contemporary legal cases in the United States, along with recent pricing evidence that supports the hypothesis of discrimination in insurance pricing. We contrast the strengths and weaknesses of some anti-discrimination methodologies for a continuous response variable, from theoretical and practical viewpoints. This introduction opens the door to four research questions, which we contribute an answer to throughout this thesis.

To ensure that the numerical results of our study are realistic, in Chapter 2 we analyze the largest publicly available database of police-reported motor vehicle traffic accidents in the United States. We describe a methodology for extracting a representative sample during the period 2001-2020, and present some results from an analysis of the data. A nationally representative sample of 1,583,520 people involved in 20 years of fatal and non-fatal accidents is analyzed to examine the effects on the injury severity of motor vehicle occupants. We examine the impact of traditional personal automobile insurance rating factors such as gender, age and previous traffic infractions on serious and fatal injuries. An estimated cost of the accidents is used to highlight the rating factors which have the highest influence in prediction accuracy. These results aid in the calibration of a microsimulation model, presented in Chapter 4.

In Chapter 3 we examine the discrimination-free premium in Lindholm et al. (2022a) within a theoretical causal inference framework, and we consider its societal context, to assess when the pricing formula should be used. We consider the insurance pricing problem through the use of directed acyclic graphs. This particular tool allows us to rigorously define an insurance risk factor in a causal framework. We then use this definition in assessing the appropriate application of the discrimination-free premium through three simplified pricing examples, including a health insurance policy and two personal automobile insurance policies with different coverages. From our findings, we suggest criteria for the application of the discrimination-free premium that is dependent on the risk factors and the social context.

In Chapter 4 we describe a microsimulation model which can generate a simulated population of the United States. It is designed to match in aggregate selected characteristics

of the target population. We focus on a 2020 pseudo-population from Wisconsin, which we use to explore personal automobile insurance premium ratings. We contrast four pricing models, in terms of prediction accuracy, and in terms of their discriminatory impact over race, using four different definitions of discrimination proposed in the actuarial and machine learning literature. By adapting definitions for disparate impact and proxy discrimination to a statistical test we show that the traditional assumption of independence between frequency and severity cannot only result in reduced prediction performance, but can also be detrimental to racial minorities.

In Chapter 5 we conclude and present some directions for future research.

Acknowledgements

I would like to profoundly thank my supervisors, Mary Hardy and Paul Marriott, for making this long-term research project possible. Over the last years, they have been the exact definition of a leader for me. In a very professional manner, they always provided timely feedback, motivation and guidance. I'm also grateful for their generous financial support. But more importantly, I'm thankful because their faith in me never wavered, in spite of all the global and personal challenges we faced. This fueled my passion and provided me with the freedom and confidence any researcher needs, which ultimately allowed me to develop new skills and improve as a person in general.

I would also like to thank all the examining committee members, Professor Emiliano Valdez, Hilary Bergsieker, Ben Feng and Michael Wallace for their valuable comments that helped improve my thesis.

Many thanks to the faculty and staff in the Department of Statistics and Actuarial Science at the University of Waterloo for their help during the last years. Special thanks to Mary Lou Dufton, Carla Daniels, Greg Preston and Carlos Mendes.

Dedication

To my family.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

For many United States (U.S.) residents, a car is a necessity in order to access jobs, schools or do essential activities like grocery shopping. In most states having liability insurance is compulsory, making access to affordable car insurance of paramount importance. Therefore, high-priced car insurance can severely impair the lives of those who cannot afford it due to the lack of access to jobs, education or because individuals risk getting a criminal record by driving illegally, without insurance.

Actuaries decide the set of variables – known as rating factors – used in pricing most individual insurance products. They also develop the statistical models designed to represent the expected cost of insurance, and determine loadings for expenses and profit. Those decisions ultimately decide who will get insurance and at what price.

The selection of rating factors has the ultimate goal of reducing the prediction error, but the actuary may also be reflecting predominant social beliefs. For example, in 1896, pricing life insurance using separate mortality tables for White and Black people was deemed normal and reasonable (Wolff, 2006). It was misleadingly shown in Hoffman (1896) that statistically, Black people had higher mortality rates which meant they had to pay higher premiums. The author made the mistake of not stratifying by environmental factors such as social or economic status. More blatantly, Prudential Life Insurance Company charged different premiums based on race to ensure that White policyholders would not view the company as sympathetic or welcoming to Black policyholders (Wolff, 2006).

Over time society changed. With the Civil Rights Act (1964), insurers stopped using race as a rating factor (Scot J. Paltrow, 2000; NAIC, 2020a). Independently of its statistical validity, insurers no longer charge different premiums based on race. You might then think that 59 years after the use of race as a valid rating factor, the difference in insurance coverage or prices across races would be negligible. The problem is that although the use of race as a rating factor was eliminated, indirect discrimination remained.

In mathematical texts, the word discrimination can be used as a synonym for risk classification, which is the essence of insurance. However, throughout this thesis, we restrict 'discrimination' to the definitions from Romei and Ruggieri (2013), which come from a legal perspective:

Definition 1.1. *(Direct discrimination). Arises when rules or practices explicitly treat one person less favorably than another on forbidden grounds. This is also known as disparate treatment.*

Definition 1.2. *(Indirect discrimination). Apparent neutral practice which results in an unfair treatment of a protected group. This is also known as disparate impact.*

A prime example of indirect discrimination is the legal case Griggs v. Duke Power Company (1971). One year after the enacting of the Civil Rights Act (1964), Duke Power started requiring new employees to have a high school diploma, and current employees to take an intelligence test to get promoted. At the time, Duke Power was already racially segregated; for example, the highest paid Black worker made less money than the lowest paid White worker. Additionally, Duke Power was based in the state of North Carolina where one third as many Black people as White people graduated from high school. The relevant question in this case was if these hiring requirements were a valid pre-requisite for coal-workers to do their job. Duke Power's lawyers presented two arguments: first, the intelligence tests are not designed to be discriminatory, they were made by professional psychologists to measure intelligence; second, they were acting in good faith, they were only trying to screen potential new hires to guarantee their employees are capable of doing their everyday tasks. The jury concluded that the tests were not necessary for coal-workers to do their job and thus, the apparently neutral practice resulted in an unfair treatment toward Black employees and applicants. The intent was not relevant.

A comprehensive study of direct and indirect discrimination in insurance should address redlining. Redlining originated as a business practice where residents of certain neighborhoods were discriminated against in availability and costs of loans and mortgage insurance, compared with residents of bordering areas (U.S. Federal Home Loan Bank, 1938b; Rothstein, 2017). Figure 1.1 shows historical redlining in Milwaukee relating to home loan availability. The map was one of over 200 city maps created by the Federal Home Loan Bank and the Home Owners' Loan Corporation as part of a federal initiative to develop underwriting criteria for loans. The most "desirable" neighborhoods were shaded in green; these were typically affluent suburban areas. Blue neighborhoods were less affluent, but "still desirable"; yellow neighborhoods were "declining", and red neighborhoods were labeled "hazardous". Discrimination arose because the red areas had predominantly Black and other minority populations. Insurance companies developed similar maps to deny insurance in areas where the residents were predominantly members of minority groups. For example, an insurer from the U.S. issued a bulletin in 1964 where they encouraged underwriters to "use a red line around questionable areas on territory maps", indicating areas where the insurer wanted no business (Ansfield, 2021).

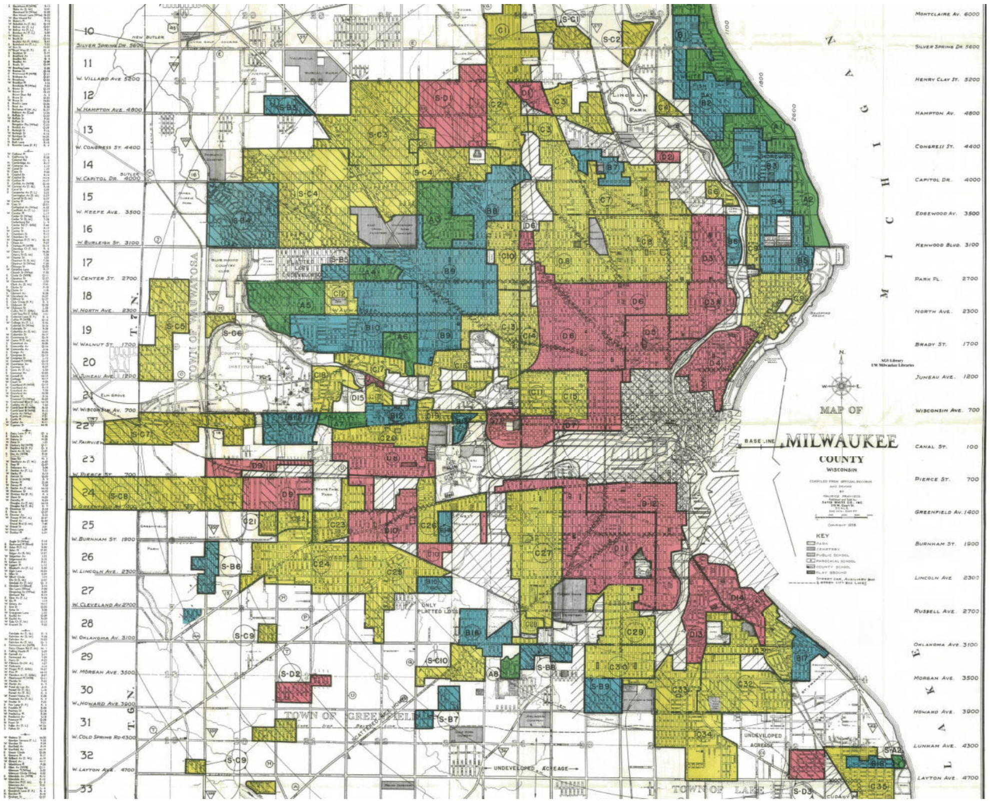


Figure 1.1: Residential Security Map of Milwaukee, Wisconsin. Created by the Federal Home Loan Bank and the Home Owners' Loan Corporation as a guide for mortgage lenders to graphically reflect the desirability of investing in neighborhoods. Retrieved from the University of Wisconsin-Milwaukee Library in U.S. Federal Home Loan Bank (1938a).

Although redlining is often thought of as a problem in the financial industry, it originates from insurance, since mortgage lenders require borrowers to secure property insurance. As mentioned in Stanton (2002), it was not until the landmark case, Dunn v. Midwestern Indem. (1979), that victims of redlining by insurance were allowed to sue insurers under the Fair Housing Act (1968): a U.S. law originally meant to prohibit discrimination when buying or renting a house on the grounds of race, color, religion, sex, familial status, or national origin. As stated in an insurance redlining case, NAACP v. American Family Mut. Ins. Co. (1992), in the city of Milwaukee:

> *"No insurance, no loan; no loan, no house; lack of insurance thus makes housing unavailable."*

1.1 Ongoing Issue

In spite of the legal prohibition, direct discrimination and redlining has prevailed in insurance practice. More recent redlining cases that violate the Fair Housing Act (1968) include:

(1) NAACP v. American Family Mut. Ins. Co. (1992),

(2) United States v. American Family Mut. Ins. Co. (1995),

(3) United States v. Nationwide Mut. Ins. Co. (1997),

(4) United States v. Erie Ins. Co. (2008),

(5) Huskey v. State Farm Fire & Casualty Co. (2022).

Some of the evidence for racial discrimination in these cases is shown in Table 1.1. This shows that racial discrimination is an ongoing issue in the United States. The direct discrimination evidence, which is a very complicated problem in itself, can be attributed to wrongdoing, pervasive racism and ill-intentioned actions from the parties responsible.[1] However, redlining and other forms of indirect discrimination that arise through apparently neutral practices are more difficult to identify and mitigate, regardless of intent. There are many processes and people involved in these practices, that can ultimately lead to indirect discrimination, from those involved in data collection and treatment, to the actuaries responsible for the insurance pricing models.

The identification and mitigation of indirect discrimination in insurance is a topic gaining interest around the world. In 2020 the National Association of Insurance Commissioners

[1] See Glenn (2000) for more evidence of recent underwriting guidelines that classify African Americans and group minorities as undesirable.

Evidence of direct discrimination	Cases
Managers that criticized insurance agents for selling too many policies to African Americans.	(2), (3)
The insurer specifically instructed underwriters and management employees that homeowners from neighborhoods with substantial minority populations were not desirable as clients.	(2), (3)
More rigorous and frequent home inspections for African Americans and minority populations.	(2), (5)
Evidence of redlining	**Cases**
The property insurers used the racial characteristics of neighborhoods as a factor in its underwriting process.	(1), (2), (3), (4), (5)
Statistically significant disparities in coverage between zip codes with predominantly White populations and zip codes with predominantly African American populations; even when controlling for factors related to purchase ability and coverage needs.	(2), (3), (4), (5)
Loss experience data (number of claims per policy and claim amount paid divided by premiums received) does not explain the racial disparities in pricing.	(2)

Table 1.1: Evidence of racial discrimination by legal cases that violate the Fair Housing Act (1968).

(NAIC)[2] created a special committee called Race and Insurance that had among its goals, to determine if current practices exist within the insurance sector that potentially disadvantage minorities (NAIC, 2020b). The Casualty Actuarial Society have commissioned a series of research papers on ‘Race and Insurance Pricing’, including Chibanda (2022) and Mosley and Wenman (2022). In the latter, some recent cases of indirect discrimination are discussed and the study by ProPublica (2017) stands out. This study analyzes over 100,000 premiums and claims paid for mandatory auto insurance by all state insurers in California, Texas, Missouri and Illinois for the most recent five-year period for which the data is available and consistently finds out that, all other things equal, areas with a predominantly minority population pay as much as 30% more than areas with a predominant White population. At a national scale, a similar percentage of overcharge is found for cities and towns with a majority of Black residents, regardless of their driving record, in a 2020 annual report that compares over 25 million rates through their access to a proprietary database (Insurify, 2020). Consumer Federation of America (2015), another study at the national scale, used 293,010 premiums for a single profile from the 5 largest insurers in the United States (which represent more than 50% of the auto insurance market) and found a premium

[2] The NAIC is a regulatory support organization in the United States. Through the NAIC, state insurance regulators establish standards and best practices.

overcharge of 70% in urban zip codes with a predominantly Black population compared to a predominantly White zip code.

To understand the aforementioned analyses we have to understand actuarial practice. Actuaries for years have been using factors like gender, age, marital status, educational attainment and zip code to price insurance products. These variables are used because they correlate with the expected cost of insurance but some of them also correlate with variables like race. For example, in the state of Wisconsin the three largest cities with the most densely populated zip codes are Milwaukee, Madison and Green Bay, as shown in Figure 1.2. But also, in this state, zip code is a strong predictor of race, as can be seen in Figure 1.3, and it is allowed as a rating factor in auto insurance. Zip codes in the city of Milwaukee have the highest percentages of Black, Asian and Hispanic populations in the state. The problem is that as shown in Consumer Federation of America (2015), ProPublica (2017) and Insurify (2020), *ceteris paribus*, people in those zip codes are charged higher premiums than in the neighboring zip codes, and than other zip codes where the majority of the population is White.

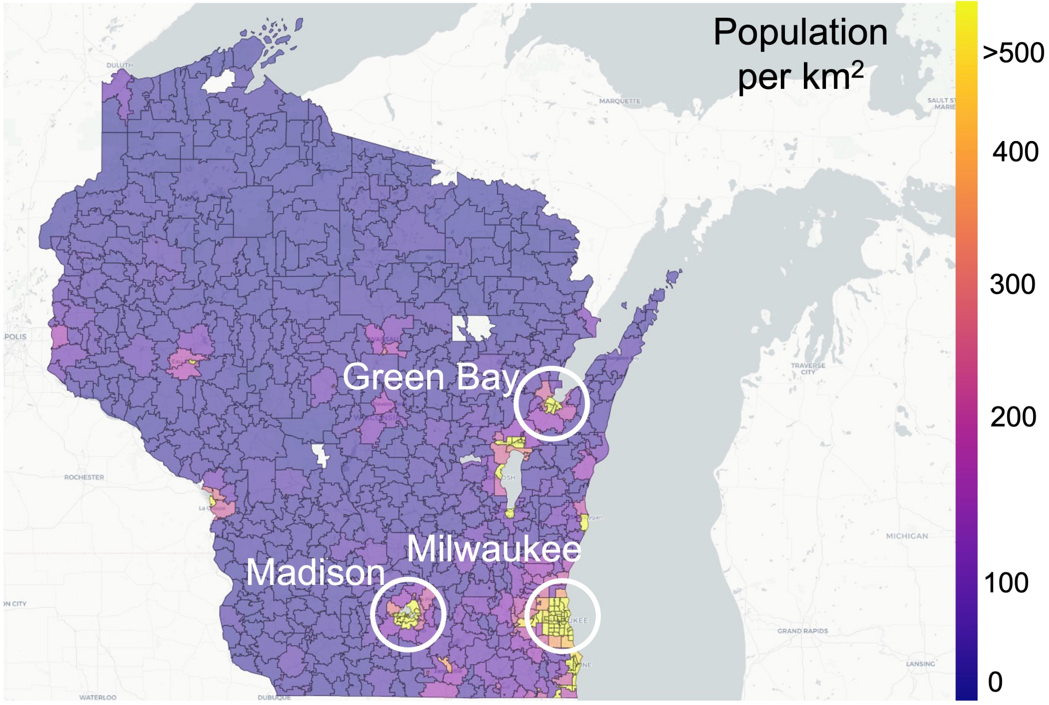


Figure 1.2: Geographical distribution of population density in the state of Wisconsin. See this map for a full interactive display.

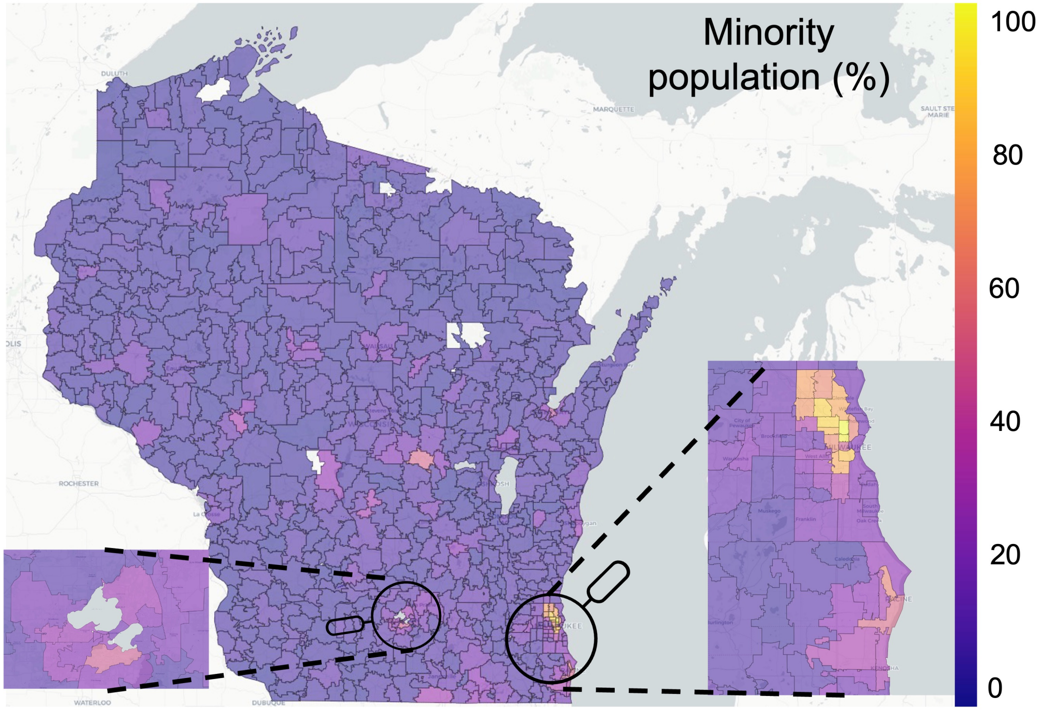


Figure 1.3: Geographical distribution of racial minorities, such as Black, Asian and Hispanic populations in the state of Wisconsin. See this map for a full interactive display.

We make our own exploratory analysis of the current geographical distribution of premiums in Wisconsin to show the particular case for this state and to reinforce the findings of the aforementioned analyses. We retrieve premiums considering as coverage the minimum mandatory liability insurance of the state. This is done for two different driver profiles in 507 out of the 765 zip codes in Wisconsin from a large property and casualty (P&C) insurer (the insurer does not offer automobile insurance in the other 258 zip codes). The first driver profile is a single 28 years old male with no driving record who owns a 2020 Honda Accord and rents his apartment. The second profile is a driver with a higher socioeconomic status: a married 56 years old male who owns a 2021 BMW 5 Series with no driving record, and owns his home. The geographical premium distribution for the first profile can be seen in Figure 1.4, where we find a 70% overcharge, on average, for urban zip codes with a predominantly minority population compared to predominantly White urban zip codes. The premium ordering is exactly the same for the second profile but the overcharge is reduced to 60%, on average. These findings are similar and comparable to those in Consumer Federation of America (2015).

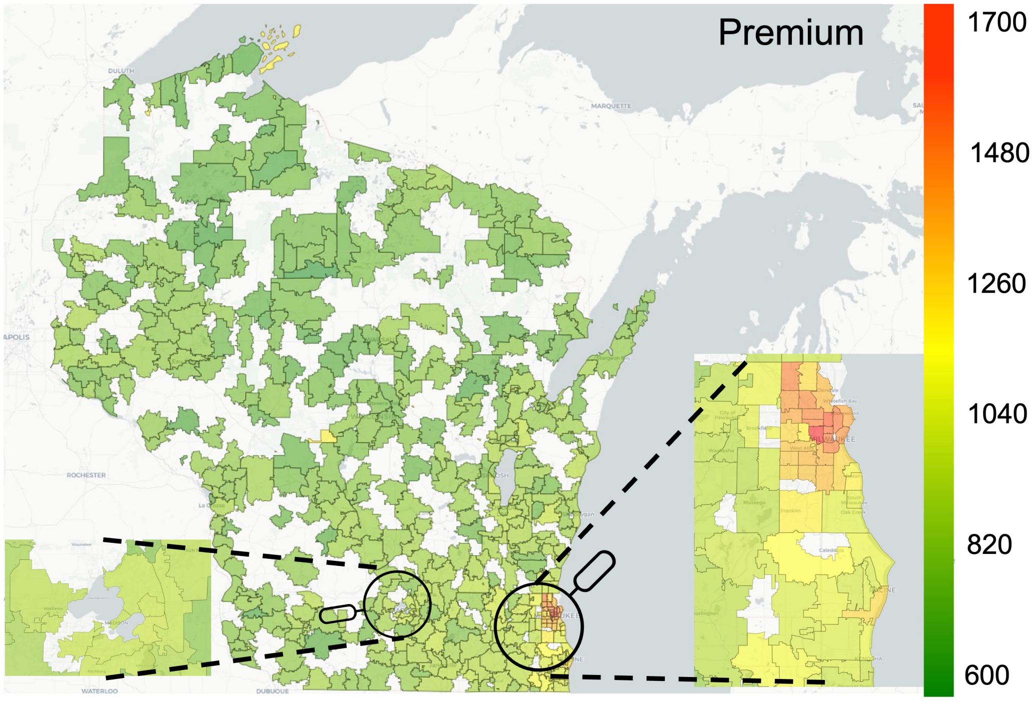


Figure 1.4: Premiums from a large P&C insurer in the state for a 2020 Honda Accord Ex (4 DR 1.5 L Turbo) used only for commuting. Car owned by a 28 years old single male with no driving record, renting his home and a good credit score. Retrieved June 28, 2022.

The premium disparity between zip codes with a widely different racial composition is not an isolated problem of the insurer considered. We also retrieve premiums for the first driver's profile, from two of the largest insurers in the state, for a subset of 41 zip codes where 13 are in Milwaukee, 8 in Madison, 6 in Green Bay and 14 are rural zip codes. We scale the premiums to the unit interval and average with the scaled premiums from the first insurer and show the geographical distribution in Figure 1.5. The average overcharge for minority zip codes in Milwaukee considering these 3 insurers is 52% compared to White urban zip codes and 82% compared to rural zip codes. This systemic evidence supports the hypothesis that indirect discrimination is an ongoing issue in the United States.

Although zip codes may be a valid rating factor, they may also be exploited to increase premiums for minority groups for spurious reasons. Causes of the risk, known as risk factors, need to be analyzed in order to discover spurious correlations. The problem of identifying and mitigating unfair treatment – as part of Definition 1.2 – is becoming more complex with the arrival of machine learning and artificial intelligence (NAIC, 2019). In auto insurance,

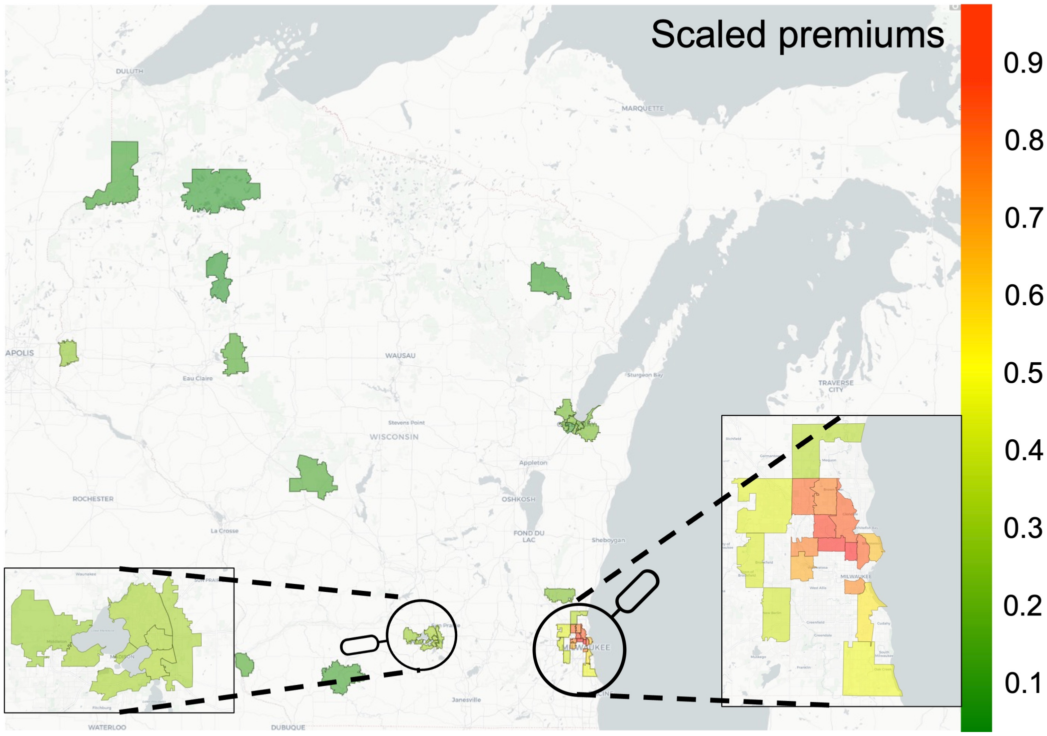


Figure 1.5: Scaled premiums from three of the largest P&C insurers in the state for a subset of 41 zip codes. Retrieved August 22, 2022. See this map for a full interactive display.

telematic systems track where you drive, how fast you drive, at what time of the day you use your car, among many other variables (Verbelen et al., 2017). In health insurance, using a fitness watch or smartphones, an insurer can follow how much the policyholder exercises, the amount of sleep and the time spent sitting down or taking a walk (Scism and Maremont, 2010). With these changes, insurers can now charge premiums directly reflecting policyholder behavior. This might seem like a positive change for the insurance industry; after all, people will be charged according to their behavior. The problem is that, unchecked, this can also generate indirect discrimination. For example, using telematic data, people visiting poor neighborhoods may be charged higher premiums if a lot of auto thefts are reported in that area, if they have longer commutes (O' Neil, 2016) or for directly discriminatory reasons. Then, the apparently neutral behavior variables can become a proxy for income or for race, creating a system where poor people or minority groups are charged more, cannot access insurance and therefore worsening their financial situation. After all, models are mathematical idealizations and as O' Neil states:

> *"Whether or not a model works is also a matter of opinion. After all, a key component of every model, whether formal or informal, is its definition of success. [...] Racism is the most slovenly of predictive models. It is powered by haphazard data gathering and spurious correlations, reinforced by institutional inequities, and polluted by confirmation bias."*

1.2 Literature Review

To identify and mitigate discrimination in insurance pricing first we need to define it in an actuarial context while preserving the spirit of the legal definitions 1.1 and 1.2. In this section we overview the most common definitions of discrimination in actuarial and machine learning texts that can be used when the response variable is continuous. We provide a list of general notions and terminology used in the definitions of direct and indirect discrimination:

- Let Y denote an insurance loss, which is a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathbb{P}$ represents the real world measure.
- Y is dependent on a set of covariates, representing the potential rating factors.
- The set of rating factors can be divided into socially acceptable or nondiscriminatory covariates, denoted by $\boldsymbol{X}$, and protected or discriminatory covariates, denoted by $\boldsymbol{D}$. We use the term '*discriminatory covariate*' to refer to a characteristic whose use in insurance pricing is legally prohibited or socially controversial. In this thesis, we assume, in general, a single discriminatory covariate in each of the models considered.
- Let $\hat{\mu}(\boldsymbol{X})$ denote the estimated pure premium, which corresponds to the expected value of the future loss, given the rating factors $\boldsymbol{X}$. This is the premium assumed to be charged by the insurance company, before loadings for expenses, profit or adverse experience.

1.2.1 Direct discrimination

We present the definition in Xin and Huang (2022) that, adapted to our notation, states:

> *"Direct discrimination occurs when a person is treated less favourably than another person simply because of their membership to a protected group. If the*

> *person's corresponding rating factor is not used by insurers, such discrimination can be completely avoided."*

This definition drives the modeler to completely ignore the discriminatory covariate in the pricing process, also referred to as the 'unawareness' approach. This definition of direct discrimination is consistent with the legal definitions in Romei and Ruggieri (2013) and European Commission (2012). The latter is a European Union (EU) legislation that has banned the use of gender as a rating factor (discussed in detail in Chapter 3). This definition also agrees with the probabilistic definition of direct discrimination in Lindholm et al. (2022a) and the causal definition for unresolved discrimination in Kilbertus et al. (2017).

1.2.2 Indirect discrimination

The identification of indirect discrimination is a more difficult endeavor because it can arise in two ways that are not mutually exclusive:

1. As an indirect consequence of using a subset of variables from $\boldsymbol{X}$ that are correlated with the discriminatory covariate, D.
2. As a disproportionate effect to a group compared to members of another group. This is known as disparate impact.

The first form of indirect discrimination is called proxy discrimination in Frees and Huang (2021) and Kilbertus et al. (2017), and it is called indirect discrimination in Lindholm et al. (2022a). Henceforth, we use proxy discrimination to help us distinguish this form of indirect discrimination from disparate impact. We remark that the legal definitions of indirect discrimination in Romei and Ruggieri (2013) (see Definition 1.2) and in European Commission (2012) do not make a distinction between proxy discrimination and disparate impact.

A fact that increases the complexity of identifying and mitigating discrimination is that a premium obtained through an unawareness approach, which is free of direct discrimination, can coexist with both forms of indirect discrimination. Even if D is not used in the pricing model, proxy discrimination can occur through the predictive power that $\boldsymbol{X}$ has over D. In the same way, disparate impact can occur if the pricing model has a disproportionate impact over a subset of D. This means that statements of *fairness through unawareness*, as coined in Dwork et al. (2012), are incomplete and misleading by not considering the effects of indirect discrimination. In what follows, we state some mathematical definitions and proposed solutions to proxy discrimination and disparate impact in the actuarial and machine learning literature.

Proxy discrimination

The definition in Frees and Huang (2021) states that proxy discrimination occurs when insurers discriminate on a facially neutral characteristic, such as the size of a car engine or geographic area, which correlates with a discriminatory covariate. In consequence, they propose using a modified set of variables that are linearly uncorrelated to the discriminatory covariate. The modified set of variables is $\boldsymbol{X}^* = \boldsymbol{X} - \hat{\boldsymbol{X}}$, where $\hat{\boldsymbol{X}}$ are the predicted values of the linear regression of each nondiscriminatory covariate onto the discriminatory covariate, D.

Under the assumption of linear dependence between the nondiscriminatory covariates and the discriminatory covariate, this methodology satisfies a common conception of non-discrimination called demographic parity (Xin and Huang, 2022). Demographic parity requires $\mathbb{P}(\hat{\mu}|D = d) = \mathbb{P}(\hat{\mu}|D = d')$, and its strengths and disadvantages have been well documented in the machine learning literature (Dwork et al., 2012; Hardt et al., 2016). The main advantage is that members of $D = d$ are equally likely to observe a set of outcomes as members of $D = d'$. However, when the insurance loss and the discriminatory covariate are correlated, the loss in predictive accuracy can be substantial.

An enhanced solution to proxy discrimination that not only removes linear correlation, but also artificially removes the explanatory power that $\boldsymbol{X}$ could have over D, is presented in Lindholm et al. (2022a). Their formula, called a discrimination-free price, is

$$h^*(\boldsymbol{X}) = \int_d \mu(\boldsymbol{X}, d)\, \mathrm{d}\mathbb{P}^*(d), \tag{1.1}$$

where $\mathbb{P}^*$ is an arbitrary probability measure, which guarantees $\mathbb{P}^*(d|\boldsymbol{X}) = \mathbb{P}^*(d)$. This approach is a probabilistic extension of Pope and Sydnor (2011), where the average values of the discriminatory covariate are used in the pricing model. We reproduce their definitions of discrimination and their proposed pricing formula along with examples of its application in Chapter 3. Although formula (1.1) is motivated by statistical independence using observational data, we explore, also in Chapter 3, the discrimination-free price within the theoretical framework of causal inference. This framework helps us rigorously define a risk factor, which yields a criterion for the use of the discrimination-free price in different pricing scenarios.

There has been other work on discrimination discovery and removal using causal inference. For example, Pearl (2009a) addresses discrimination in hiring and argues that the discovery of discrimination is a causal question because "*it requires some knowledge of the data-generating process; it cannot be computed from the data alone, nor from the distributions that govern the data*". Causal inference uses underlying knowledge or judgment about causal connections between variables to extend analysis beyond empirical associations. In our view, extending the application to discrimination in insurance pricing is quite natural. While traditional statistical theory drums into us that correlation does not imply causation, in insurance pricing we are (or should be) specifically interested in causation, leading to the idea that traditional models of dependence may be insufficient for our needs.

An example of discrimination removal using concepts from causal inference is investigated in Kilbertus et al. (2017). In their view, a proxy variable of D should not affect the prediction. Adapted to our notation, we denote a proxy variable $P \in \boldsymbol{X}$, such that P is highly correlated with D. A formal overview of some concepts used from causal inference are presented in Chapter 3, but here we informally describe a causal concept needed for the definition of proxy discrimination in Kilbertus et al. (2017). This concept is known as the *do*-operator, denoted by $do(P = p)$, and it is used to represent an intervention where instead of letting the variable P vary naturally, we force $P = p$ for every observation (Pearl et al., 2016). Intuitively, through the *do*-operator we are removing all external effects over the proxy variable, P, such as those coming from the discriminatory covariate, D. Then, a premium $\hat{\mu}(\boldsymbol{X})$ exhibits no proxy discrimination in expectation, if for all p, p'

$$\mathbb{E}[\hat{\mu}(\boldsymbol{X})|do(P = p)] = \mathbb{E}[\hat{\mu}(\boldsymbol{X})|do(P = p')]. \tag{1.2}$$

We present an example with race as the discriminatory covariate to show the intuition behind this definition of proxy discrimination and some alternatives. Following the explanation for redlining in Section 1.1, we could assume zip code to be a proxy for race. Then, the implementation of (1.2) would result in a premium that charges all policyholders, on average, the same premium independently of zip code. In this particular example, this constraint could be overly restrictive since, geographical location is an important rating factor in insurance losses. However, if premiums are widely different for different groups, then, proxy discrimination persists and corrective action is required. To allow for these cases, we may relax the constraint in Equation (1.2), for example to

$$\frac{\mathbb{E}[\hat{\mu}(\boldsymbol{X})|do(P = p)]}{\mathbb{E}[\hat{\mu}(\boldsymbol{X})|do(P = p')]} < \tau, \tag{1.3}$$

where τ is a threshold that should be sufficient to prohibit wide premium differences between different zip codes. This approach is similar to what is done to remove disparate impact (Feldman et al., 2015), which we introduce below.

Disparate impact

The legal background of disparate impact started with the federal response to Griggs v. Duke Power Company (1971) with the Equal Employment Opportunity Act (EEOA, 1972), which amended the Civil Rights Act (1964). This law was passed by the Supreme Court with the goal of equal employment opportunity without discrimination on the grounds of race, color, sex, religion or national origin. Five federal agencies are responsible for the enforcement of this law and in 1978 they issued guidelines to establish a uniform federal government position (Equal Employment Opportunity Commission, 1979). In this guideline, they define disparate impact as a substantially different rate of selection in hiring, promotion or other employment decision which works to the disadvantage of members of a race, sex or ethnic group. Mathematically, we can express this definition of disparate impact on employment as

$$\frac{\mathbb{P}(\text{Hiring} = \text{Yes} \mid D = d)}{\mathbb{P}(\text{Hiring} = \text{Yes} \mid D = d')} < \tau, \tag{1.4}$$

where τ is the threshold that would not be considered as a *substantially different rate*. The enforcement agencies have adopted a rule of thumb known as the 'four-fifths rule' which sets the threshold τ to be 4/5 or 0.8. This rule gives a numerical basis to draw initial inference, and in the case of non-compliance, request additional information of the overall selection process from the employer in order to determine disparate impact.

In the case where a large number of employment selections are made, disparate impact needs to be determined statistically. However, note that (1.4) is a classification problem since the response variable is binary (see Feldman et al. (2015) for insights on the statistical test and the removal of disparate impact). In Xin and Huang (2022) it is pointed out that (1.4) is a relaxation of demographic parity, which means the problem can be adapted to insurance pricing as

$$\frac{\mathbb{E}[\hat{\mu}(\boldsymbol{X})|D = d]}{\mathbb{E}[\hat{\mu}(\boldsymbol{X})|D = d']} < \tau. \tag{1.5}$$

Analogously to employment, we can determine disparate impact as a substantially different average premium which works to the disadvantage of a protected group. We list some considerations of this definition of disparate impact in insurance pricing:

- The main advantage of complying with (1.5) is that, independently of which rating factors were used, on average, premiums differ by less than τ for different protected groups, as compared to the unprotected group. Compared to demographic parity, the constraints over the pricing model are relaxed since it only needs to be tuned to avoid substantial differences across groups.

- It provides a mathematical way of determining disparate impact, which in consequence results in an unambiguous definition of a discriminatory premium. From a legal standpoint, this is an advantage since current regulations are ambiguous and provide inadequate guidance to actuaries in insurance companies. For example, insurers in the U.S. are required to comply with the P&C Model Rating Law from the NAIC that establishes that premiums should not be unfairly discriminatory (NAIC, 2010). The law states its own definition of unfair discrimination: *"Unfair discrimination exists if, after allowing for practical limitations, price differentials fail to reflect equitably the differences in expected losses and expenses."* The word 'equitable' is not defined in the document, providing leeway for interpretation which can be troublesome.

- The main disadvantage is the predictive accuracy trade-off that occurs as a result of the model tuning to remove disparate impact. This is discussed in Miller (2009), but the author concludes that the solution for disparate impact is to ban certain rating factors, but that is likely to lead to inaccurate risk assessments. Rating factors do not necessarily need to be banned in order to eliminate discrimination, as we see in the solutions to proxy discrimination from Frees and Huang (2021), Lindholm et al. (2022a) and Kilbertus et al. (2017). In Chapter 3 we discuss this accuracy trade-off, since we also believe a fair premium should reflect the risk transferred to the insurer, independently of the predictive power that $\boldsymbol{X}$ has over D.

1.3 Research Objectives and Contributions

1. **Data availability.** One of the main and recurring challenges in actuarial science problems is the lack of public insurance data (So et al., 2021). We are able to see gross premiums, but policyholder data and their respective insurance losses are proprietary information. We tackle this issue in Chapter 2 and 4. In the former, we provide the largest public database of motor vehicle traffic accidents in the United States, along with the methodology and code that made its construction possible. In the latter chapter, we use microsimulation to obtain a population from Wisconsin that is calibrated to statistics from the U.S. Census Bureau. Both methodologies can be replicated and adapted to solve other problems in insurance.

2. **How to identify discrimination?** In Section 1.1 we saw a geographical distribution of premiums that raises the question: is there clear evidence of indirect discrimination,

considering that premiums are, on average, 52% more expensive in predominantly minority zip codes? The definitions of direct and indirect discrimination presented in Section 1.2 can help answer this question. Specifically, in Chapter 3 we rigorously explore the discrimination-free premium proposed in Lindholm et al. (2022a). This analysis leads to a criterion we propose to identify cases where there is discrimination and in consequence, it is justified to use their discrimination-free formula. In Chapter 4 we exemplify with simulated insurance data the identification of discrimination (using the definitions presented in Section 1.2). Further, we adapt the definition of disparate impact when the response variable is continuous and propose a suitable nonparametric statistical test for its identification.

3. **How can discrimination arise?** In Chapter 3 we show that a relatively small measurement error in a rating factor can result in indirect discrimination. In Chapter 4 we show that the traditional actuarial assumption of independence between frequency and severity can have a significant discriminatory impact. These are two examples that can make discrimination occur in insurance pricing. However, it is important to note that this question is broad and ambitious since discrimination can arise in insurance pricing even before any losses are observed. For example, in the redlining cases in Section 1.1 it is mentioned that agents were discouraged from selling policies to African Americans. With our numerical examples we seek to clarify some ways in which unintentional discrimination can arise. Intentional acts will substantially worsen disparity, but they are not the focus of this work.

4. **How can we mitigate discrimination?** Mathematical definitions of direct and indirect discrimination lead to ways in which we can minimize or mitigate discrimination. But as we state in Chapter 3, the calculation of premiums cannot be considered as a purely scientific endeavor. Legislation and the social context need to be considered by the actuary to make sure the rating process is discrimination-free and avoids exacerbating social inequalities.

1.4 Thesis Outline

This thesis is organized as follows. In Chapter 2 we describe a methodology for extracting a representative sample of motor vehicle traffic accidents from the United States during the period 2001-2020, and present some results from an analysis of the data. A nationally representative sample of 1,583,520 people involved in 20 years of fatal and non-fatal accidents is analyzed to examine the effects on the injury severity of motor vehicle occupants. We examine the impact of traditional personal automobile insurance rating factors such as gender, age and previous traffic infractions on serious and fatal injuries. An estimated cost of the accidents is used to highlight the rating factors which have the highest influence in prediction accuracy. These results aid in the calibration of the microsimulation in Chapter 4.

In Chapter 3 we examine the discrimination-free premium in Lindholm et al. (2022a) within a theoretical causal inference framework, and we consider its societal context, to assess when the pricing formula should be used. We consider the insurance pricing problem through the use of directed acyclic graphs. This particular tool allows us to rigorously define an insurance risk factor in a causal framework. We then use this definition in assessing the appropriate application of the discrimination-free premium through three simplified pricing examples, including a health insurance policy and two personal automobile insurance policies with different coverages. From our findings, we suggest criteria for the application of the discrimination-free premium that is dependent on the risk factors and the social context.

In Chapter 4 we describe a microsimulation model which can generate a simulated population of the United States. It is designed to match in aggregate selected characteristics of the target population. We focus on a 2020 pseudo-population from Wisconsin, which we use to explore personal automobile insurance premium ratings. We contrast four pricing models, in terms of prediction accuracy, and in terms of their discriminatory impact over race, using four different definitions of discrimination proposed in the actuarial and machine learning literature. By adapting definitions for disparate impact and proxy discrimination to a statistical test we show that the traditional assumption of independence between frequency and severity cannot only result in reduced prediction performance, but can also be detrimental to racial minorities.

In Chapter 5 we conclude and present some directions for future research.

Chapter 2

A Consolidated Database of Police-Reported Motor Vehicle Traffic Accidents in the United States for Actuarial Applications

We describe a methodology for extracting a representative sample of motor vehicle traffic accidents in the United States during the period 2001-2020, and present some results from an analysis of the data. The source of the publicly available data is the US National Highway Traffic Safety Administration. A nationally representative sample of 1,583,520 people involved in 20 years of fatal and non-fatal accidents is analyzed to examine the effects on the injury severity of motor vehicle occupants. We examine the impact of traditional personal automobile insurance rating factors such as gender, age and previous traffic infractions on serious and fatal injuries. An estimated cost of the accidents is used to highlight the rating factors which have the highest influence in prediction accuracy.

2.1 Introduction

Practitioners and researchers in the insurance industry need relevant and representative data in order to assess and improve pricing and underwriting processes. However, due to proprietary issues, obtaining insurance data from major insurers can be a challenge for researchers (Gabrielli and Wüthrich, 2018).

In this chapter, we make two contributions. First, we provide a database, and the code for its creation, which will allow researchers to explore ratemaking using the largest publicly available database on motor vehicle traffic accidents (MVTAs) in the United States. The database and code are available with open access through the digital repository in Araiza Iturria et al. (2021a,c). The code can be used to update the database annually. We filter, standardize and pool 20 years of data from two public sources of MVTAs that include both fatal and non-fatal injuries. The pooling of the databases involves using statistical methods to ensure a representative dataset. Further, due to the richness of data on fatal accidents – which tend to be the most costly – we are able to provide a detailed picture of the tail of the cost distribution.

Second, we discuss examples of inferential analysis using the database. Specifically, we analyze annual trends of traditional personal automobile insurance rating factors that effect the injury severity of vehicle occupants involved in MVTAs. Rating factors such as age, gender, socio-economic group and zip code are explored. The motor vehicle occupant's race is imputed to the database, because in this thesis we seek to mitigate the detrimental consequences caused by racial and gender discrimination in insurance. The rating factor trends are shown for the latest 20 years of U.S. public crash data, from 2001 to 2020. Then, we transform the maximum injury severity of each MVTA, given its zip code, into an estimated dollar amount. With this, our goal is both a direct analysis of the data from the point of view of insurance but also to have data which allows calibration of the microsimulation model in Chapter 4, which is used to explore pricing discrimination.

The data is collected by an agency of the U.S. Department of Transportation called the National Highway Traffic Safety Administration (NHTSA). In particular, we use two NHTSA databases, the Crash Report Sampling System (CRSS) and the Fatality Analysis Reporting System (FARS).

The chapter is organized as follows. In Section 2.2 we describe the structure of the non-fatal and fatal datasets, along with a detailed description of the process of filtering, standardizing and pooling the databases in order to obtain a nationally representative sample of MVTA. In Section 2.3 the multinomial logistic regression model used in this chapter to understand the effects on injury severity is described along with its assumptions. In Section 2.4 we present the results from our analysis of the data highlighting both strengths and limitations of this observational approach. In Section 2.5 we highlight rating factors that, due to their predictive ability in the severity distribution, are used in the microsimulation of Chapter 4. In Section 2.6 we offer some concluding comments.

2.2 Data

Since 2016, the NHTSA has prepared the CRSS, which is a nationally representative sample of around 50,000 police-reported MVTAs in the U.S., taken from the estimated 5-6 million police-reported MVTAs that occur each year. The accidents include those that result in a fatality, injury or property damage. Before 2016, the information was provided in the General Estimates System (GES), which had a different sampling design than CRSS (Zhang et al., 2019b), but with the same target population. Like the CRSS, the GES used information collected solely from police crash reports. It is important to note that, as mentioned in National Center for Statistics and Analysis (2020), injury and property damage estimates from the GES and the CRSS are not directly comparable because the sample designs are different. Therefore, annual trends are presented separately for the periods 2001-2015 and 2016-2020.

The FARS database, published annually, is a national census of all MVTAs that result in a fatality within 30 days of the event. The FARS data is significantly more detailed than the CRSS/GES data, as it is obtained from a range of sources including police crash reports, death certificates, state driver licensing files, and emergency medical services reports (National Center for Statistics and Analysis, 2021).

Although the FARS data has a comprehensive set of variables, including detailed information on the 30,000-40,000 fatal accidents that occur annually, from 1988 to 2018 it only represents around 0.5% to 0.6% of the estimated total police-reported MVTAs in the U.S. (National Center for Statistics and Analysis, 2020). Thus, even though fatal accidents tend to be the most expensive insurance losses, analyzing rating factors from the FARS database alone would likely result in biased estimates of the predictive relationship between rating factors and insurance costs, as the fatal accidents may not be representative of the vast majority of accidents that result in less severe injuries.

In order to take advantage of the individual strengths of each database, we pool the FARS and GES/CRSS data, using a framework presented and validated in Yasmin et al. (2015), who obtained nationally representative estimates for driver injury severity by pooling data from GES and FARS for a single year (2010). Their study is more limited in scope than ours, in that they focus exclusively on accidents involving two or fewer cars, they only consider drivers (not passengers), and their goal is to analyze survival times of injured drivers, in order to influence trauma triage policies. In this chapter, we extend the analysis to multiple years and to multiple-vehicle accidents, and we include all the motor vehicle occupants involved.

The sampling design of the NHTSA uses geographically contiguous areas for operational efficiency, assigning weights based on variables such as the population size, the number of fatal accidents, the total number of miles driven by motorcycles, among others (Zhang et al., 2019a). Therefore, the pooling methodology combines a weighted sample and a census in a two-step approach.

The first step uses the weights included in the GES/CRSS database, to obtain a sample that is nationally representative. In the second step, we remove all accidents from GES/CRSS that include a fatal injury and replace them with a sample from the FARS database. To exploit the richer information available from the FARS dataset, we oversample the fatal accidents and apply weightings that result in unbiased estimates when considering all type of accidents together. This method utilises the richer FARS data, which is associated with the right tail of the loss distribution. This is done because we expect it to observe heavy tails which are strongly influential in insurance pricing.

2.2.1 Filtering

Both the FARS and GES/CRSS data for each year consist of three main files:

1. A file containing details of each accident;
2. A file with details of each vehicle involved in each accident;
3. A file with details of each person involved in each accident, including pedestrians and other non-motorists.

For insurance research purposes, we are mainly interested in the drivers and vehicle occupants. Hence, our database uses the file with details of each individual vehicle occupant as a base file, to which we add the vehicle and accident information from the other files.

The FARS data has been published since 1975 but we only use records since 2001 because that is when the NHTSA started to record the latitude and longitude of the crash. Using the crash coordinates, we can obtain the zip code where the accident happened. We use 'zip code' interchangeably with Zip Code Tabulation Area (ZCTA), as is common practice. This is due to the fact that the U.S. Postal Service zip codes are not geographical areas, they are a collection of mail delivery routes. ZCTAs are generalized areal representations of U.S. Postal Service zip codes created by the U.S. Census Bureau. The algorithm for the creation of ZCTAs can be consulted in U.S. Census Bureau (2015).

The algorithm to retrieve the zip code from each crash latitude and longitude requires coordinates for all the ZCTA regions in the United States. We used the shape files corresponding to the 2010 Zip Code Tabulation Area (ZCTA), which can be downloaded from the U.S. Census Bureau website.

To obtain the zip code of each MVTA we identify if the MVTA's coordinates are inside or outside the corresponding ZCTA region. To do this, we used the R package sp (Pebesma et al., 2020), designed for spatial location information. Using the function point.in.polygon we can test for each MVTA if it is located inside or outside each ZCTA.

Before the filtering process, the GES/CRSS (FARS) database for the period 2001-2020 contains 1,069,601 (690,087) MVTAs, involving 2,678,388 (1,740,268) people. Since we are interested in personal auto insurance, we exclude MVTAs where all vehicles involved are 'non-standard', including public utility vehicles, limousines, trucks, buses, motorcycles, or where the vehicle is unknown. We also exclude accidents where all drivers involved have a commercial license (this variable is only available for the FARS database) and/or where the vehicle has a special use, including taxis, military vehicles or emergency response vehicles. Because injury severity is an important variable in our analysis, MVTAs where all people involved have injuries with missing values, such as injury severity not reported or unknown are removed from the database. Finally, we remove a small number of MVTAs where at least one occupant died prior to the crash. The filtered GES/CRSS (FARS) database for the period 2001-2020 consists of 1,958,922 (993,818) rows, one for each person involved in any of the 728,945 (351,717) MVTA of interest. The filtering procedure with record counts at each step of the process can be seen in Table 2.1.

2.2.2 Injury severity

The injury severity for each individual involved in an MVTA is classified using the KABCO scale, which was developed by the National Safety Council (NSC) (National Safety Council, 2017). In decreasing order of severity, the KABCO categories are: fatal injury (K), serious injury (A), minor injury (B), possible injury (C) or no apparent injury (O). The KABCO scale injury classification is made at the scene of the accident by a police officer, with the exception of fatal injuries. The NSC manual indicates that 98% of the deaths from MVTA occur within 30 days (National Safety Council, 2017). Therefore, if a death occurs within 30 days, the injury severity is recorded as K (fatal injury). The specific guidelines for the other categories may vary by state, details can be found in FHWA (2016).

The KABCO scale implemented in both GES/CRSS and FARS is useful for insurance purposes as it allows us to estimate, broadly, potential insurance losses, following Council et al. (2005). Their paper analyses costs as a function of maximum injury severity in the MVTA, speed limits (used as a proxy for urban/rural location), single or multiple vehicle

	Filtering criterion	GES/CRSS		FARS	
		No. MVTA	No. people	No. MVTA	No. people
	Records in unfiltered database	1,069,601	2,678,388	690,087	1,740,268
1	All vehicles involved are non-standard.	331,424	706,405	326,949	722,649
2	All vehicles involved have a special use.	1,754	3,353	838	1713
3	All drivers involved have a commercial license.	-	-	10,542	21,949
4	Injury severity of all occupants is unknown or not reported.	7,436	9,633	0	0
5	Accidents where at least one occupant died prior to crash.	42	75	41	139
	Records in filtered database	728,945 (68.1%)	1,958,922 (73.1%)	351,717 (50.9%)	993,818 (57.1%)

Table 2.1: Filtering procedure with the number of records removed by each filter for the 2001-2020 GES/CRSS and FARS databases. Percentages in parenthesis are the number of records in the filtered database divided by the number of records in the unfiltered database.

event, whether the crash occurred at an intersection, and whether the crash involved a pedestrian, animal or object. This approach is used and explained with more details in Section 2.5.

2.2.3 Imputation and feature engineering

Variables such as state, county and driving record (including previous accidents, suspensions of license, speeding or driving under the influence of alcohol or drugs) are essential for insurance pricing and underwriting. Other variables, including race and Hispanic ethnicity are helpful to explore and mitigate unintended pricing discrimination, as discussed in Lindholm et al. (2022a) and in Chapter 3. Since state, county, driving record, race and Hispanic ethnicity are missing from the GES and CRSS databases, we impute them under assumptions described below. Other variables are created to simplify the pooling procedure in Section 2.2.4 or to improve the classification needed in the estimation process. The full list of variables deemed relevant for insurance research is given in Table A.1 in Appendix

A, which describes how the variables from the databases are standardized.

The maximum severity variable, denoted MAX_SEV, is the highest injury severity of all the people involved in the same MVTA. This variable is used in Section 2.2.4 to facilitate the sampling process. The relative frequency of MVTAs in each category of maximum injury severity is shown in Table 2.2. Further, for each individual record we add a variable NUM_VEH denoting the number of vehicles involved in the MVTA.

MAX_SEV	Label	Frequency (%)
K	Fatal injury	1.6
A	Serious injury	13.6
B	Minor injury	19.4
C	Possible injury	19.7
O	No apparent injury	45.6

Table 2.2: Relative frequency of the maximum injury severity per MVTA of the filtered and standardized GES/CRSS database for the period 2001-2020.

The individual's home state, county and zip code are important for insurance pricing. For both the GES and CRSS databases, the variables STATE and COUNTY are not directly available but as a proxy, we use the state and county of the driver's zip code (denoted by DR_ZIP in Table A.1). In the FARS data, the variables STATE and COUNTY reflect the location of the accident.

Using the 2018 American Community Survey (ACS) we add two further variables to each record in the database. The first is the population count for the zip code, denoted by POP2018. This allows us to distinguish between rural and urban areas following the classifications in U.S. Census Bureau (2021a), in a similar way to Council et al. (2005), who used speed limits. The second is marital status, which is a common rating factor used in insurance underwriting. The marital status, denoted by MARITAL, is randomly assigned using probabilities based on age, gender and zip code. For parsimony, we use only two mutually exclusive categories for marital status: Married or Single.

Race and Hispanic origin are recorded only for people with a fatal injury in the FARS database, as it is recorded on the death certificate. People who were involved in fatal accidents but do not have a fatal injury in the FARS dataset, and people involved in accidents in the GES/CRSS dataset, do not have race or Hispanic origin recorded. Therefore, we have imputed race and Hispanic origin using the 2018 ACS data, based on each individual's zip code, gender and age category. More information on the categories for 'race' can be found in Appendix A.

The MODEL of the vehicle is not relevant without the MAKE variable. Thus, a variable called MAKEMODEL is created as a concatenation of the former two variables. There are 658 and 622 unique combinations of MAKEMODEL for the filtered and standardized GES/CRSS and FARS databases, respectively. If we include this variable in a statistical model, the high number of make-model vehicle combinations creates a data sparsity problem, which would be exacerbated by the inclusion of the vehicle year, denoted MOD_YEAR in Table A.1. To simplify the thousands of possible make-model-year combinations, we used data from CarGurus (2022a) to transform the large number of combinations into an estimate of each vehicles' price. For proprietary reasons, we are unable to store, disclose or share these vehicle prices. Instead, we convert them into a categorical variable which we then assign to a proxy of socio-economic categories. In the ACS of 2018 there are 10 different income groups for households in the United States, its distribution can be observed in Figure 2.1. We used the percentage of households that belong to each income group nationally to create percentiles for the vehicle prices and have 10 socio-economic groups (SEG) for the FARS and GES/CRSS database that match the income distribution of the United States. We recognize that the allocation of socio-economic grouping using the value of the vehicle is quite rough.

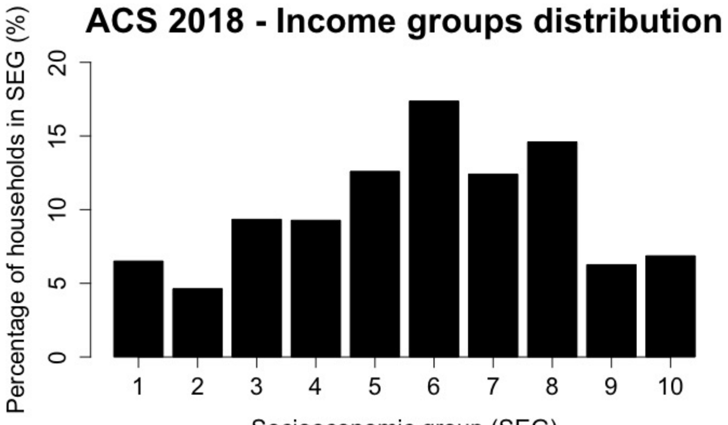


Figure 2.1: Distribution of households in each income group for the United States, data from the 2018 American Community Survey. The socioeconomic groups are in ascending order, such that SEG 1 corresponds to the lowest income group and SEG 10 to the highest income group as defined in the 2018 ACS.

Auto insurance premiums are typically adjusted if the policyholder has had an accident or a driving offence in the previous few years. Variables in Table A.1 that provide this information in the FARS dataset are `PREV_ACC`, `PREV_SUS`, `PREV_DWI` and `PREV_SPD`. These four variables – from now on, collectively called 'the driving record'– are not collected by the NHTSA for GES/CRSS. However, due to their importance in insurance, we impute them in the final database. First, for the FARS data, we summarize the four variables into one new binary variable, denoted by `PREV`, which is set to 1 if the person has had one or more accidents or driving offences in the last 5 years, and set to 0 otherwise. We use a binary logistic regression from the filtered and standardized FARS dataset to impute the variable `PREV` for the GES/CRSS dataset. This implicitly assumes that the distribution of driving records is similar between fatal and non-fatal accidents, when conditioned on age, gender, SEG, state, rural/urban zip code, and time of the accident. To account for the sampling differences between GES and CRSS, we had to additionally include a categorical covariate that is set to 1 if the accident's year of occurrence is before 2016 and 0 otherwise. The parameter coefficients estimates for each state can be seen in Figure 2.2.

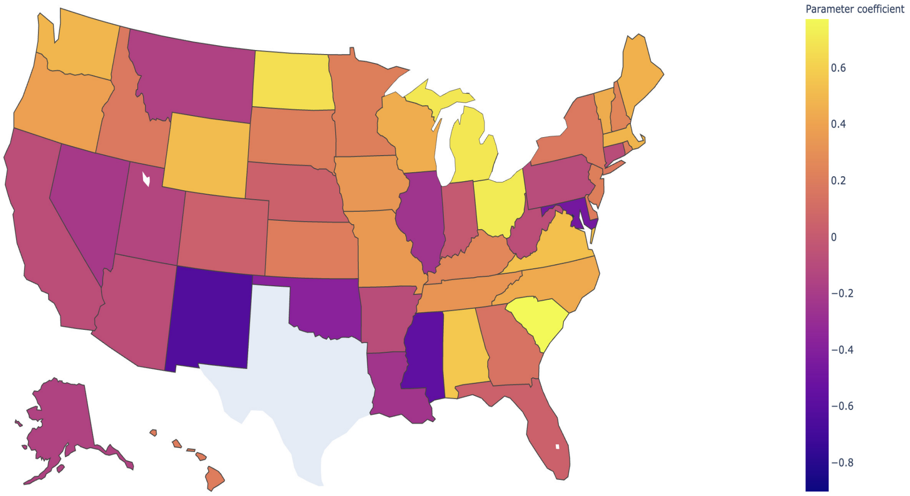


Figure 2.2: State parameter coefficient estimates obtained in the `PREV` imputation process. Coefficients are estimated from a binary logistic regression when conditioning on age, gender, SEG, state, rural/urban zip code, time of the accident and year group. The state of Texas is used as a baseline because of its large representation in the FARS dataset.

2.2.4 Pooling procedure

To create the pooled sample of GES/CRSS and FARS data, we follow the two-step approach used in Yasmin et al. (2015). As mentioned in Section 2.2, the first step is to obtain a weighted sample from the filtered GES/CRSS database. We seek a nationally representative sample of MVTAs conditioned on their maximum severity category. We therefore use a proportional sampling method such that the number of MVTAs in each KABCO category is in proportion to their representation nationally (Schutt, 2011).

The NHTSA under-samples MVTAs where all the people involved had no apparent injuries, and then adds large weights to these accidents in order to keep the sample representative of the national distribution (Zhang et al., 2019a). We sample without replacement which means we need to use a sample size that is smaller than the number of MVTAs in each maximum severity category[1]. For year j, let N_j denote the total number of MVTAs in the filtered GES/CRSS database; $N_j^{(i)}$ and $W_j^{(i)}$ are the total number of MVTAs and the sum of weights for all MVTAs in category i, respectively, where $i \in \mathcal{I} := \{K, A, B, C, O\}$. Then, for year j, the largest sample size possible, expressed as a proportion of the total, is,

$$r_j^* = \underset{r \in (0,1)}{\arg\max} \left\{ N_j^{(i)} > r \cdot N_j \cdot \frac{W_j^{(i)}}{\sum\limits_{k \in \mathcal{I}} W_j^{(k)}}, \forall i \in \mathcal{I} \right\}.$$

For parsimony, we keep the proportion of the sample size constant for all years in the period of study. This means the largest constant sample size possible, expressed as a proportion, has to be smaller than $\min\left(r_{2001}^*, r_{2002}^*, \ldots, r_{2020}^*\right) = 0.58$. We set all sample sizes to be 55% of the total number of accidents in each year, times the weighted proportion each maximum severity category represents for that year, or,

$$n_j^{(i)} = 0.55 \cdot N_j \cdot \frac{W_j^{(i)}}{\sum\limits_{k \in \mathcal{I}} W_j^{(k)}},$$

rounded to the nearest integer. To keep the notation consistent, the lowercase letters correspond to the GES/CRSS sample and denote the same thing as the uppercase letters for the filtered GES/CRSS database.

[1] We sample without replacement to avoid introducing additional bias. A sample with replacement could over-represent MVTA with large weights that tend to be from heavily populated states such as California or Texas.

For each year j and $i \in \mathcal{I}$, we sample $n_j^{(i)}$ entries from the filtered GES/CRSS database, without replacement, using as a probability distribution the weight of each MVTA divided by $W_j^{(i)}$. For the year 2020, Table 2.3 shows the distribution of weights and number of MVTAs by maximum severity category. As can be observed in the gray-shaded columns, the sum of weights, as a percentage, for the CRSS database matches the percentage of MVTAs in the sample obtained through the proportionate sampling method. This allows us to ignore the weight variable created by the NHTSA in subsequent steps since the number of MVTAs in our sample now reflects accidents at a national scale.

	MAX_SEV	$W_{2020}^{(i)}$	$\dfrac{W_{2020}^{(i)}}{\sum\limits_{k\in\mathcal{I}} W_{2020}^{(k)}}$ (%)	$N_{2020}^{(i)}$	$\dfrac{N_{2020}^{(i)}}{N_{2020}}$ (%)
Filtered CRSS	K	16,443	0.5	763	2.2
	A	79,218	2.3	3,799	10.9
	B	350,066	10.2	5,732	16.5
	C	563,547	16.5	8,356	24.0
	O	2,411,629	70.5	16,110	46.3
	MAX_SEV	$w_{2020}^{(i)}$	$\dfrac{w_{2020}^{(i)}}{\sum\limits_{k\in\mathcal{I}} w_{2020}^{(k)}}$ (%)	$n_{2020}^{(i)}$	$\dfrac{n_{2020}^{(i)}}{n_{2020}}$ (%)
CRSS Sample	K	2,333	0.1	92	0.5
	A	12,547	0.5	443	2.3
	B	163,133	6.4	1,956	10.2
	C	275,869	10.9	3,149	16.5
	O	2,084,004	82.1	13,478	70.5

Table 2.3: Distribution of weights and number of MVTA by maximum severity for the filtered CRSS (top) and for the CRSS sample (bottom). Values shown solely for the database corresponding to 2020.

The second step consists in removing from the GES/CRSS sample all MVTAs that include a fatal injury, that is, the $n_j^{(K)}$ cases for each year j, and replacing them with a larger sample size from the FARS database for year j, denoted m_j. We remark that the intention is to oversample fatal MVTAs by setting $m_j >> n_j^{(K)}$, as the FARS data is more comprehensive than the CRSS/GES data.

After the fatal accidents from the GES/CRSS sample have been replaced with a larger sample from the FARS database, a different weight than that of the NHTSA has to be added, to keep the sample nationally representative. Each case from the FARS database includes a weight equal to the number of fatal cases removed from the GES/CRSS sample divided by the total number of FARS cases added, that is, $n_j^{(K)}/m_j$. For example, in Table 2.3, the CRSS sample of 2020 includes 92 fatal MVTAs. If we replace them with a sample of 8,000 MVTAs from the FARS database, then each includes a weight of 92/8000. All non-fatal MVTAs from the GES/CRSS sample are assigned a weight of 1.

To determine m_j, we decided to use the same constant proportion of 55% for each year. Let M_j denote all year j MVTAs from the FARS database. Then the sample size of fatal MVTAs is,

$$m_j = 0.55 \cdot M_j,$$

rounded to the nearest integer. These are sampled, without replacement, through an equal weights probability distribution, and then merged with the GES/CRSS sample. The oversampled fatal MVTAs in the resulting sample allows us to explore the most severe events without sparsity problems. The pooling procedure described here and validated in Yasmin et al. (2015) resulted in a nationally representative sample for the period 2001-2020 of 1,583,520 individuals, and 592,976 different MVTAs.

2.3 Multinomial Logistic Regression Model

One goal of this work is to explore the relationship between insurance rating factors and injury severity in the United States. In this section we use regression models to do this. Regression models rank high in interpretability and are commonly used inference tools in insurance practice.

Injury severity has multiple categories, making a multinomial logistic regression model a feasible choice. In Yasmin et al. (2015) a generalized ordinal logistic (GOL) regression is used with a modified injury severity response variable. For their objective – analyzing survival time of victims to influence more efficient emergency and triage procedures – a GOL provides useful information. However, the GOL regression, which consists of fitting several logistic regression models where the response variable is binary collapsed by all possible partitions, is less suitable for insurance applications. For example, knowing that females have a higher probability of minor injuries compared to severe injuries does not provide a basis for ratemaking. Further, even though injury severity is a variable that, by construction, can be ordered, using other ordered logistic regressions, such as proportional odds or partial proportional odds, can be too restrictive. Instead, it is more practical for the insurer to have a model such as a multinomial logistic regression that allows comparison of parameter estimates for all the different injury categories, which, as shown in Council et al. (2005), has different monetary consequences.

A multinomial distribution is a general form of the binomial distribution where instead of a binary allocation of events (eg 'pass/fail'), events are assigned to more than two categories. Similarly, while logistic regression models a binary response variable, a multinomial logistic regression models a response variable that may fall into one of $n > 2$ exclusive and exhaustive categories.

A random vector follows a multinomial distribution with k categories and n trials, denoted $\mathbf{Y} = (Y_1, \ldots, Y_k) \sim \mathrm{MN}(n, p_1, p_2, \ldots, p_k)$, if the random variables Y_j for $j \in \{1, \ldots, k\}$ indicate the number of times category j is observed over the n trials. The probability mass function is,

$$\mathbb{P}\left(Y_1 = n_1, Y_2 = n_2, \ldots, Y_k = n_k\right) = n! \prod_{j=1}^{k} \frac{p_j^{n_j}}{n_j!}, \tag{2.1}$$

$$\text{where} \quad \sum_{j=1}^{k} n_j = n \quad \text{and} \quad \sum_{j=1}^{k} p_j = 1.$$

In a regression setting, the parameters p_j, for $j \in \{1, \ldots, k\}$, in equation (2.1) can be expressed as a conditional probability of a parameter matrix $\beta_{\ell \times k} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_k\}$ and a data matrix $X_{n \times \ell}$, where ℓ is the number of independent variables including a constant vector for the intercept. The conditional probability for observation $i \in \{1, \ldots, n\}$, denoted by y_i, with independent covariates denoted by $\boldsymbol{x}_i$ is,

$$\mathbb{P}(y_i = j|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_1) + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_2) + \ldots + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}, \quad j = 1, 2, \ldots, k. \tag{2.2}$$

To avoid non-identifiability issues and without loss of generality, we set the additional constraint $\boldsymbol{\beta}_1 = \mathbf{0}$ (Tutz et al., 2015). This means we have selected $j = 1$ as the reference category. To visualize how all categories are compared to a reference category, using equation (2.2) we can obtain,

$$\boldsymbol{x}_i^\top \boldsymbol{\beta}_j = \ln\left(\frac{\mathbb{P}(y_i = j|\boldsymbol{x}_i)}{\mathbb{P}(y_i = 1|\boldsymbol{x}_i)}\right), \quad j = 1, 2, \ldots, k. \tag{2.3}$$

In (Osborne, 2015, p. 399-401) it is pointed out that a multinomial logistic regression with k categories is similar to $k-1$ binary logistic regressions, with the reference category as the contrast for each of the $k-1$ regressions. Equation (2.3) helps us visualize that fact, but in addition, for the binary regressions to be similar, we need the independence of irrelevant alternatives (IIA) assumption to be satisfied. The IIA is discussed further in the following Section 2.3.1.

In our case, using the KABCO scale, the reference category is 'no apparent injury' (O) as it is the most frequent category in our sample. Therefore, the parameter estimates obtained for each category let us draw conclusions relative to the group of people involved in MVTAs who had no apparent injuries.

2.3.1 Modelling assumptions

In order to use this modelling approach we require some model inference assumptions to be met. Specifically, using standard first order methods for the regression, we require: (i) independence of observations, (ii) no inappropriately high collinearity between predictors and (iii) a fully represented (not sparse) data matrix (Osborne, 2015). Because our statistical analysis of MVTA data, presented in Section 2.4, includes both the driver and vehicle occupants of all the vehicles involved in the same accident as separate records, there is some cluster-type dependence in the data. However, we assume that dependence is relatively weak, given that our sample size consists of 1,583,520 individuals and that, on average, there are 2.7 people involved in each MVTA. Nevertheless, the length of the confidence intervals reported in this chapter might be underestimated. In Section 2.4, the stable parameter estimates and confidence intervals obtained for the period of study provides evidence that the estimation procedure does not suffer from data groups that are not fully-represented.

The independence of irrelevant alternatives (IIA) is a property of the multinomial logit regression model, which states that the ratio in Equation (2.3) for category j is independent of the ratio for category j' with $j \neq j'$ and $j, j' \in \{2, \ldots, k\}$. This property can be checked using the Hausman-McFadden test (Hausman and McFadden, 1984). The test consists of estimating the unknown parameter matrix β, using all the categories (called the unrestricted model) and compare them with parameter estimates from a restricted model, constructed by removing one category (and respective data) at a time. The null hypothesis states that the multinomial logistic specification is correct. If the parameter estimates from the restricted and unrestricted model are approximately the same, the null hypothesis is not rejected. For details on the computation of the test statistic we refer the reader to Hausman and McFadden (1984) and Cheng and Long (2007).

In Table 2.4, we show the test statistic, denoted HM, with associated p-values for the Hausman-McFadden test (Hausman and McFadden, 1984). The 20 multinomial regressions fitted for the period from 2001 to 2020 satisfy the IIA assumption for the four different injury severities. Under the null hypothesis, the asymptotic distribution of HM is χ^2 with degrees of freedom equal to the rank of the restricted covariance matrix. We remark that it is reported in Hausman and McFadden (1984) that negative statistics can be obtained in finite sample applications when the difference of covariance matrices does not result in a positive definite matrix. Hausman and McFadden (1984) conclude that negative statistics are evidence that the IIA assumption holds. This decision rule is also used in Cheng and Long (2007). Nevertheless, we observe in Table 2.4 the distribution of the p-value under the null hypothesis is mostly a center of mass at 1. We explore further this issue using an alternative test statistic proposed in Vijverberg (2011) that ensures mathematically that the difference in covariance matrices is positive definite. The asymptotic properties of their proposed test statistic are the same as those of HM, but in finite samples the former is preferred since it is more closely χ^2-square distributed. The test statistic from Vijverberg (2011) also confirm the non-rejection of the IIA assumption for the 20 multinomial regressions. These alternative test results are not included because they do not provide any more relevant insights.

2.3.2 Evaluation and validation

The parameter matrix β from the multinomial logistic regression is obtained using weighted maximum likelihood estimation. For a weight vector $\boldsymbol{w} \in \mathbb{R}^n$, the independence of observations assumption allows us to multiply all conditional probabilities in Equation (2.2) to obtain the weighted log-likelihood

$$\ln\left(\mathbb{P}(\beta|\boldsymbol{y}, X, \boldsymbol{w})\right) = \sum_{i=1}^{n}\sum_{j=1}^{k} \mathbb{I}_{(y_i=j)} w_i \ln\left[\frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j)}{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_1) + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_2) + \ldots + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)}\right],$$

where $\mathbb{I}_{(y_i=j)}$ denotes the indicator function of observation y_i having the value j (Hu and Zidek, 2002).

To assess the goodness-of-fit of the multinomial logistic regression model we use McFadden's pseudo-R^2 (McFadden, 1977), denoted by ρ^2;

$$\rho^2 = 1 - \frac{\text{Log-likelihood of fitted model}}{\text{Log-likelihood of null model}}.$$

Year	Possible Injury		Minor Injury		Serious Injury		Fatal Injury	
	HM	p-value	HM	p-value	HM	p-value	HM	p-value
2001	5.77	1.00	5.78	1.00	Neg	-	Neg	-
2002	3.14	1.00	0.59	1.00	Neg	-	Neg	-
2003	Neg	-	6.83	1.00	Neg	-	0.65	1.00
2004	10.29	1.00	Neg	-	0.55	1.00	Neg	-
2005	6.95	1.00	7.98	1.00	1.49	1.00	0.28	1.00
2006	4.25	1.00	Neg	-	Neg	-	0.47	1.00
2007	73.39	0.43	2.11	1.00	Neg	-	Neg	-
2008	7.81	1.00	Neg	-	Neg	-	Neg	-
2009	11.12	1.00	4.17	1.00	Neg	-	Neg	-
2010	Neg	-	2.53	1.00	Neg	-	Neg	-
2011	12.8	1.00	Neg	-	Neg	-	Neg	-
2012	Neg	-	9.27	1.00	Neg	-	Neg	-
2013	Neg	-	1.39	1.00	Neg	-	4.69	1.00
2014	37.78	1.00	1.33	1.00	Neg	-	0.68	1.00
2015	Neg	-	Neg	-	29.65	1.00	Neg	-
2016	22.36	1.00	Neg	-	Neg	-	Neg	-
2017	9.65	1.00	10.72	1.00	1.14	1.00	0.26	1.00
2018	Neg	-	27.85	1.00	2.48	1.00	Neg	-
2019	1.05	1.00	Neg	-	Neg	-	1.18	1.00
2020	32.21	1.00	Neg	-	Neg	-	Neg	-

Table 2.4: Hasuman-McFadden test statistic and p-values for the 20 multinomial regressions of Section 2.4. A negative test statistic is replaced by 'Neg' and we use this as evidence that the IIA assumption holds.

McFadden's ρ^2 can be interpreted as a measure of the explanatory power of the fitted model over the null model. Parallel to the traditional R^2 from ordinary least squares, a better fit means a higher value of ρ^2, but conclusions drawn from this goodness-of-fit measure should be made with care; as mentioned in McFadden (1977), values for ρ^2 tend to be considerably lower than those of R^2. Also Osborne (2015) warns users to be cautious as the log-likelihood is highly influenced by sample size, unlike R^2, which is standardized and more easily interpretable.

For the 20 fitted multinomial regressions, an average $\hat{\rho}^2$ of 0.344 is obtained, with a minimum $\hat{\rho}^2$ of 0.318 for the year 2011 and maximum value of 0.389 for 2020. To have a frame of reference for these values, we conducted a simulation experiment to compare the performance of the estimated $\hat{\rho}^2$ obtained using the MVTA data. In the simulation study, the data generation process (DGP) is known for a multinomial regression model with 5 different categories. We ran 100 simulations using 46,000 observations (the average number of observations for the 20 regressions fitted) and 24 binary predictors (the number of covariates in each regression as described in Section 2.4). With these parameters, we show the results obtained from the simulation study are comparable to those from the estimation procedure due to the high dependence of McFadden's ρ^2 on the sample size and number of predictors.

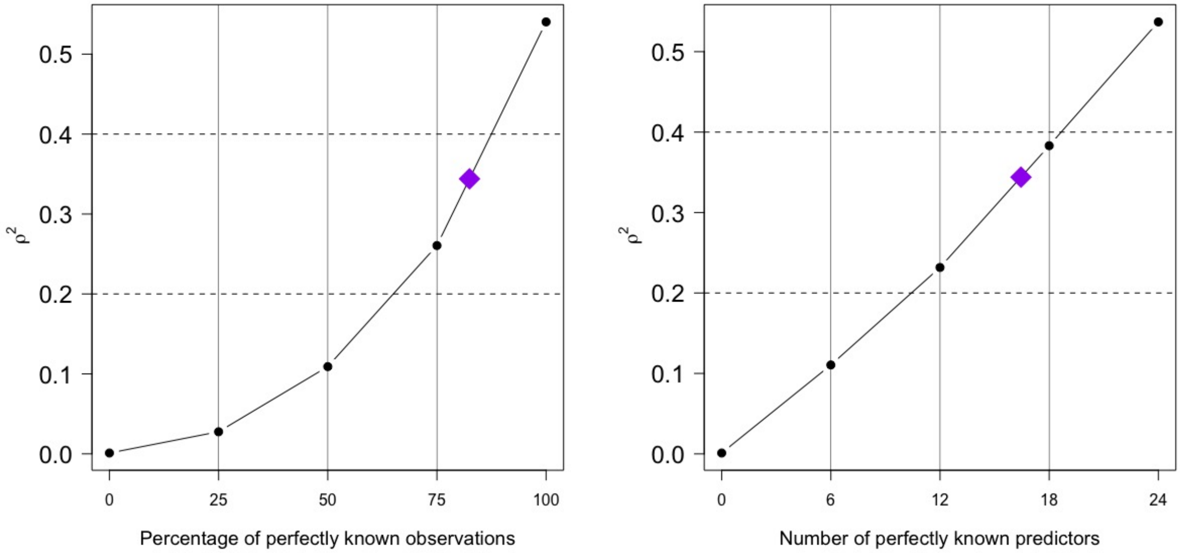


Figure 2.3: The circles represent the estimated median of McFadden's ρ^2 from 100 simulations with full knowledge of the DGP. Each simulation from the multinomial logistic regression has 46,000 observations and 24 binary predictors (including the intercept). The diamond is the average $\hat{\rho}^2$ of the 20 fitted multinomial logistic regressions of MVTA data for the period 2001-2020. The horizontal dotted lines at 0.2 and 0.4 are the range of values of ρ^2 that represent an excellent fit (McFadden, 1977).

For the 100 simulations, the estimated median of the McFadden's ρ^2 was calculated using subsets of the true data that generated the response variable of the simulation experiment. The subsets were obtained through two different approaches: horizontal replacement and vertical replacement. The horizontal replacement (on the left-side plot of Figure 2.3) consists in replacing a percentage of observations with random covariates, while other observations remain intact and, therefore, part of the true DGP. The vertical replacement (on the right-side plot of Figure 2.3) consists in replacing a predetermined number of predictors for all of the observations with random values, while other predictors remain intact. Under both scenarios, the average McFadden's $\hat{\rho}^2$ obtained for the 20 multinomial regressions fitted in this chapter are in the range of (0.2, 0.4) proposed by McFadden (1977) as representing an excellent fit. We are not over-fitting the data and on average, we capture the effects over injury severity on 82.5% of the people involved in police-reported MVTA in the period 2001-2020.

2.4 Injury Severity Results

Now that we have described the data and validated the modelling approach, we present some results illustrating the methodology, and discuss what can and cannot be inferred from these results. We show graphically some of the results obtained from fitting independent annual multinomial logistic regressions for the period 2001-2020 to the sample obtained in Section 2.2. Our objective in this section is to study national trends of the relationship between injury severity and traditional insurance rating factors. We remark that we are not doing inference in a causal sense since this is a purely observational study. Nevertheless, we can generate causal hypotheses from the data and explore there consequences in insurance practice, as done in Chapter 4.

The final selection of covariates (which is a subset of the covariates in Table A.1) and their respective categories are made by balancing the fit, as measured by McFadden's ρ^2, and including covariates that are traditionally used for personal automobile insurance pricing and underwriting. The covariates, shown in Table 2.5, can be divided into three general categories that describe characteristics of the person, the vehicle and the accident. We remark that race – imputed for most observations through the procedure described in Section 2.2.3 – is deemed not significant by the multinomial logistic regression model. This evidence is used in the design of the microsimulation in Chapter 4.

Variables describing the vehicle occupant's characteristics and circumstances are the age category, gender, driving record, if the occupant was under the influence of alcohol or illegal drugs, the position of the occupant in the vehicle (either passenger or driver) and an interaction between gender and the position of the occupant. The vehicle characteristics are summarized in the covariate socio-economic group (see the discussion in Section 2.2.3) and if the vehicle driver's speed was related to the crash. Covariates describing the accident are weather, hour of the accident, if it occurred in a rural or urban area (defined by the population size in the occupant's zip code) and the collision type. Records of the sample with 1,583,520 people that have missing values in any of the previous covariates are not used in the estimation procedure. This decision can over or under-represent a specific type

Person		Vehicle		Accident	
Covariate	Categories	Covariate	Categories	Covariate	Categories
Age	0-17, 18-29*, 30-39, 40-49, 50-59, +60	SEG	Low (1-3)*, Middle (4-7), High (8-10)	Hour	Rush hour (17-22), Late-night drive (23-5), Day drive (6-16)*
Gender	Male*, Female	Speed related	Yes, No*	Weather	Clear*, Atmospheric condition
Position in vehicle	Driver*, Passenger			Urbanity	Rural (<2,500), Urban cluster (2,500-50,000)*, Urbanized area (>50,000)
Gender and position in vehicle	Male driver*, Female driver, Male passenger, Female passenger			Type of collision	MVT*, Non-collision, Object not fixed, Fixed object
Driving record	Yes, No*				
Alcohol or illegal drugs	Yes, No*				

Table 2.5: Final selection of covariates and their respective categories for the 20 multinomial logistic regressions. Categories with an * are used as a baseline to avoid non-identifiability problems. Information on the creation of covariates can be seen in Section 2.2.3 and details on their categories in Appendix A.

of MVTAs, which in consequence means we tell a specific story about the data that can be different from that where there are no missing values.

Considering all the aforementioned covariates with their respective categories, the statistical challenge is to obtain 24 parameter estimates (including the intercept) for each of the four injury severity categories (K, A, B and C). Thus, for each year of MVTA data, we estimate 96 parameters in total. A table and figures by year and covariate of the total $96 \times 20 = 1,920$ parameter estimates with their 90% confidence intervals can be found in Araiza Iturria et al. (2021b). The estimation procedure is done in `R` using the `multinom` function from the `nnet` package (Ripley and Venables, 2021).

2.4.1 Gender

In the U.S., gender is still a permissible rating factor, whereas the European Union (EU) banned the use of gender as a direct rating factor with the objective of equal treatment between men and women in relation to insurance premiums and benefits (European Commission, 2012), see Chapter 3. This EU policy has generated a lot of debate and research in actuarial fairness because observed gender differences in the distribution of insurance costs now has to be ignored in ratemaking. This may result in either males or females paying more than they would if gender were still permitted as a rating factor. We note that Ayuso et al. (2016) use an example from a Spanish insurer, that indicates that men have riskier driving patterns than women. On the other hand, using U.S. drivers' accident data of 2010, Yasmin et al. (2015) found evidence of higher injury risk propensity for females than males.

These results led us to explore gender differences in injury severity in our database, focusing particularly on fatal injuries, which tend to be the most expensive in terms of insurance. Our first illustration of the parameter estimates obtained for a single covariate in the period 2001-2020 can be seen in Figure 2.4. These parameters determine the Relative Risk Ratio (RRR), which for a parameter β is e^{β}. If $\beta > 0$, then RRR$>$ 1, indicating an increased risk relative to the baseline group, which in this case is male drivers. Similarly, $\beta < 0$ gives RRR$<$ 1, indicating a decreased risk relative to the baseline.

Overall for non-fatal injuries, female passengers have more severe injuries relative to male drivers. For fatal injuries, the parameter estimates are not significant in any individual year, but become significant if we treat the parameter as a constant for all years. Note that we separate the regression assuming constant β into two periods; the first uses the sample that includes GES data (2001-2015) and the second uses the sample that includes CRSS data (2016-2020). By assuming a constant parameter we effectively increase the sample size, such that the confidence intervals for the two periods are reduced by a factor of $\sqrt{15}$ and $\sqrt{5}$, respectively. This effect can be observed in the bottom-right corner of Figure 2.4. Based on the aggregated 2001-2015 and 2016-2020 data, *ceteris paribus*, there is a significant positive RRR of 123% and 120%, respectively, in fatalities for female passengers, relative to male drivers, conditional on an accident having occurred.

It is interesting to see the difference in the relative risk for female drivers and passengers. In Figure 2.5 we show the fatal injury parameter estimates for all female drivers and passengers (aggregated) on the left side, for female drivers only in the center, and for female passengers only on the right. We see that combining drivers and passengers gives significant estimates for 2001-2015 and not significant for 2016-2020, but by considering drivers and passengers separately we can see a difference. Female passengers appear to have

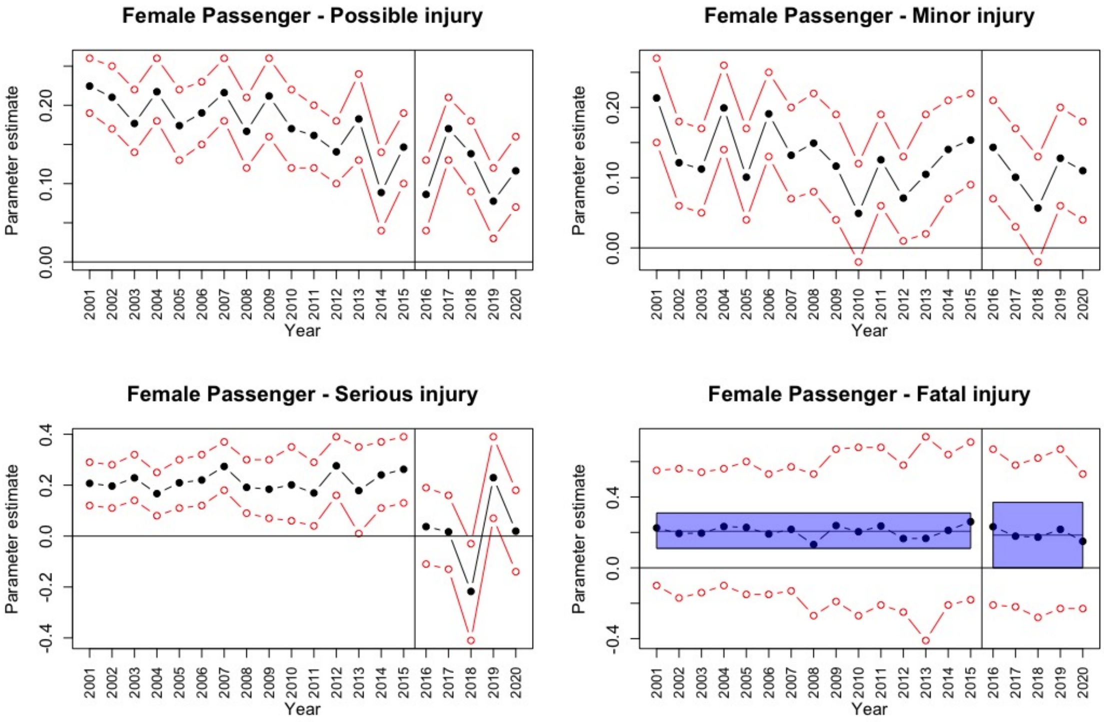


Figure 2.4: Parameter estimates of the interaction between gender and vehicle position for the category: Female Passenger (or non-driver). Red hollow points are the 90% point-wise confidence intervals for the period 2001-2020. The vertical line between 2015 and 2016 is to separate the data coming from GES and the data from CRSS which are not statistically comparable due to their sample design differences. The rectangle for fatal injuries (bottom-right) denotes the parameter estimates and confidence intervals obtained using data without yearly splits.

a significant positive fatal injury risk, for both the 2001-2015 and 2016-2020 data, whereas female drivers have a significant negative fatal injury risk (at least for the 2001-2015 data), in both cases relative to the baseline group, which is male drivers. The MVTAs were not studied by the gender of the driver, which could have changed the results and inference of this analysis.

We remark that these results are observational, not causal. We do not speculate on the underlying reasons, which cannot be inferred from the data. Statements such as *'[...] female truck drivers are also found vulnerable to severe injury. Perhaps a combination of physiological and behavioral factors significantly affects the injury severity of truck driver and caused the observed differences between male and female drivers.'* (Chen and Chen, 2011), or *'The result perhaps is indicative of the lower physiological strength of female drivers'*

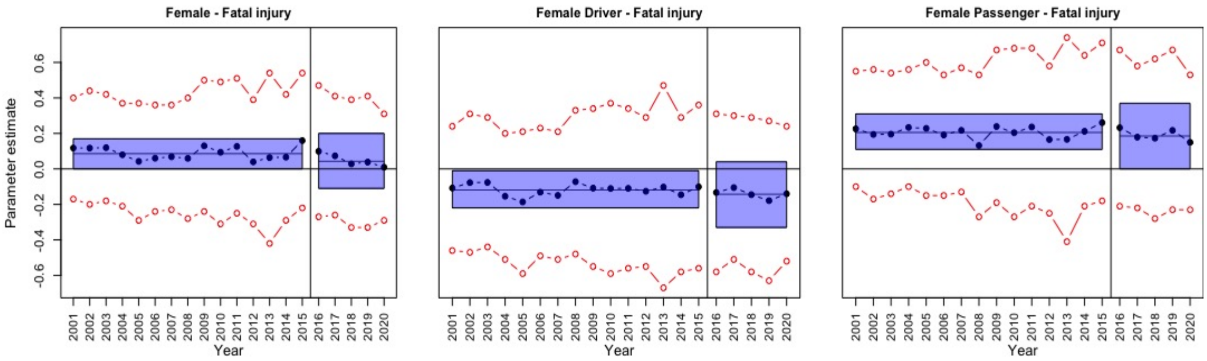


Figure 2.5: Parameter estimates of fatal injuries for the covariates female (left-side), female driver (center) and female passenger (right-side). Red hollow points are the 90% point-wise confidence intervals for the period 2001-2020. The vertical line between 2015 and 2016 is to separate the data coming from GES and the data from CRSS which are not statistically comparable due to their sample design differences. The rectangle for fatal injuries denotes the parameter estimates and confidence intervals obtained using collective data for GES and CRSS.

(Yasmin et al., 2015) cannot be justified or substantiated by the data alone, and can be very misleading, especially if they are picked up and further disseminated by individuals with little or no training in statistical inference.

2.4.2 Alcohol and/or illegal drugs

Our second illustration involves the effect of alcohol and illegal drugs. Parameter estimates are shown in Figure 2.6. Conditionally on being involved in a MVTA, and all other things constant, those that were under the influence of alcohol or illegal drugs have a significant RRR in the range of 170-283% and 179-415% (depending on the accident year) of having a minor injury and serious injury, respectively, relative to the baseline group. Moreover, fatal accidents involving people under the influence of alcohol or illegal drugs have a significant linear increasing trend in the relative risk ratio. For GES data, the estimated RRR for 2001 is 311% with an average increase of 2% per year. For CRSS data, the estimated RRR for 2016 is 489% with an average yearly increase of 6%. These linear trends are selected over the constant model because a likelihood-ratio test deems the gradient coefficient significant. The confidence intervals in the bottom-right corner of Figure 2.6 widen near the initial and last years of accident data as a natural consequence of assuming a linear relationship.

We might infer from these results that there is an increasing trend increase in the risk of fatal injuries for people involved in an MVTA who are under the influence of alcohol or

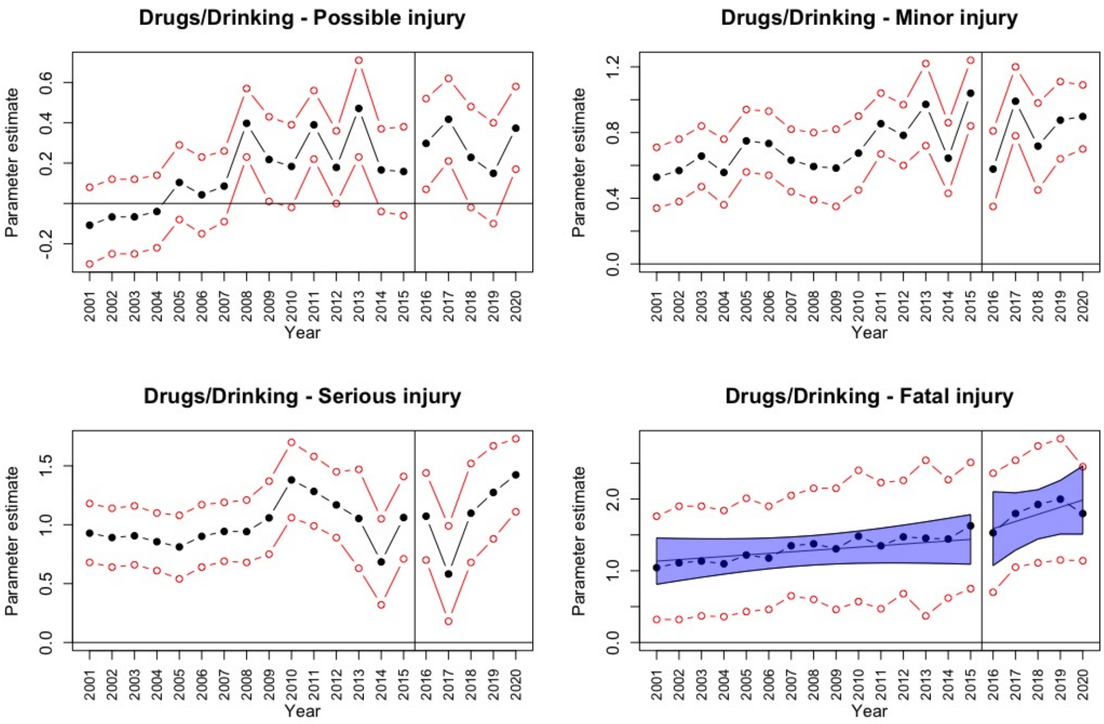


Figure 2.6: Parameter estimates for the covariate that is set to one if the vehicle occupant was under the influence of alcohol or illegal drugs. Red hollow points are the 90% point-wise confidence intervals for the period 2001-2020. The vertical line between 2015 and 2016 is to separate the data coming from GES and the data from CRSS which are not statistically comparable due to their sample design differences. The rectangle for fatal injuries (bottom-right) denotes the parameter estimates and confidence intervals obtained using data without yearly splits.

illegal drugs, compared to people who are not. However, our observational results must be interpreted with caution. For our sample, the variable DRINKING is missing for 30% of the data and 37% missing for DRUGS. In this chapter, the missing data for these two covariates is managed by recoding missing values as zero values – that is, recoding the missing values as involving people not under the influence of alcohol or drugs, so that the effect of drinking and/or drugs is underrepresented in our sample. Furthermore, we are susceptible to other data recording limitations out of our control which might bias our results. An example is that during the period 2001-2008 alcohol and drugs information in non-fatal accidents was only collected for drivers and non-motorists, not for passengers (National Center for Statistics and Analysis, 2019).

2.4.3 Magnitude of covariates

Figure 2.4 and 2.6 allow us to visualize possible trends in the effects that individual covariates have had on the injury severity of motor vehicle occupants that were involved in police-reported MVTAs. On the other hand, in Figure 2.7 we compare the magnitude of all the covariates considered for a single year. Specifically, we show the covariates for 2019 for the serious and fatal injury categories.[2] We make the following observations.

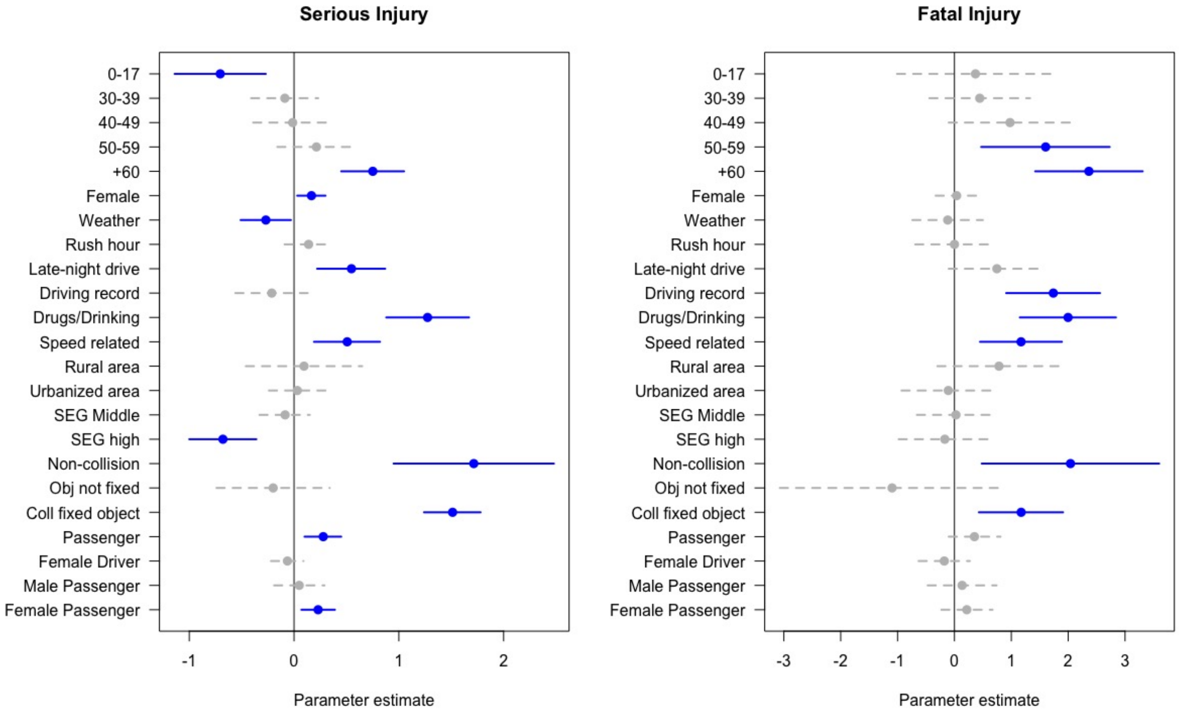


Figure 2.7: Parameter estimates and their 90% confidence intervals for all the covariates describing 2019 injury severity: serious injury (left-side) and fatal injury (right-side). Blue solid-lined (gray dotted-lined) estimates are (non) significant at a 90% level for 2019. The estimated intercept for serious injury is -4.93 and -8.44 for fatal injuries.

(1) For both serious and fatal injuries, the three largest parameter estimates correspond to the type of collision covariate, specifically, with the non-collision category, the

[2] In the analysis, the last year of available data is 2020 but we do not show it here because it was an atypical year in driving patterns and behaviors due to the COVID-19 pandemic (NHTSA, 2021).

drugs/drinking covariate, and the age ≥ 60 covariate. For fatal accidents, the driving record is also important.

(2) The most common type of non-collision for the period 2001-2020 is vehicle rollover. Among the people involved in non-collisions, around 78% and 86% were involved in rollovers in the GES/CRSS and FARS filtered datasets, respectively.

(3) We created age categories by splitting the population into representative groups. We made the split to have a non-adult group (ages 0-17) and age groups for each decade until reaching age 60 and older. The baseline group is ages 18-29. We observe in Figure 2.7 a monotonic increasing effect of age on the severity of the injury.

(4) The driver's record also has a high relative risk with respect to fatal injuries – an effect observed for each of the 20 years of data, see Figure 2.8. In some ratemaking systems, for example using bonus-malus scales, premiums are calculated as a function of the frequency of accidents and not of severity. This result indicates that the driver's record may relate to both frequency and severity risk.

(5) The parameter estimates obtained in this chapter are consistent with the results obtained in Yasmin et al. (2015), who consider a single year, 2010. When we consider the 2010 data, (see Araiza Iturria et al. (2021b) for the parameter estimates), consistent with Yasmin et al. (2015), we find evidence in 2010 of lower injury risk propensity when there are not clear atmospheric conditions and higher injury risk propensity for females (when we combine both drivers and passengers), for older individuals, for people under the influence of alcohol (including illegal drugs here), for late-night driving (coded here as 11 p.m. until 5 a.m.) and for collisions involving fixed objects instead of against a moving vehicle.

(6) As in every real world data modeling challenge, we are exposed to model misspecification and parameter estimation errors as a result of insufficient information or measurement errors. For instance, in the U.S. traffic regulations, definitions and laws vary by state but in our numerical results the data is aggregated such that state-level differences are not accounted for. An example of definition discrepancies throughout the states is mentioned for the KABCO scale in Section 2.2.2. The decision of not including state-level differences is used to avoid a data sparsity problem.

(7) A documented source of bias in the data is the observed evidence of fatal crash underestimation in GES data (Zhang et al., 2019b). This is one of the reasons that the sampling design changed from GES to CRSS.

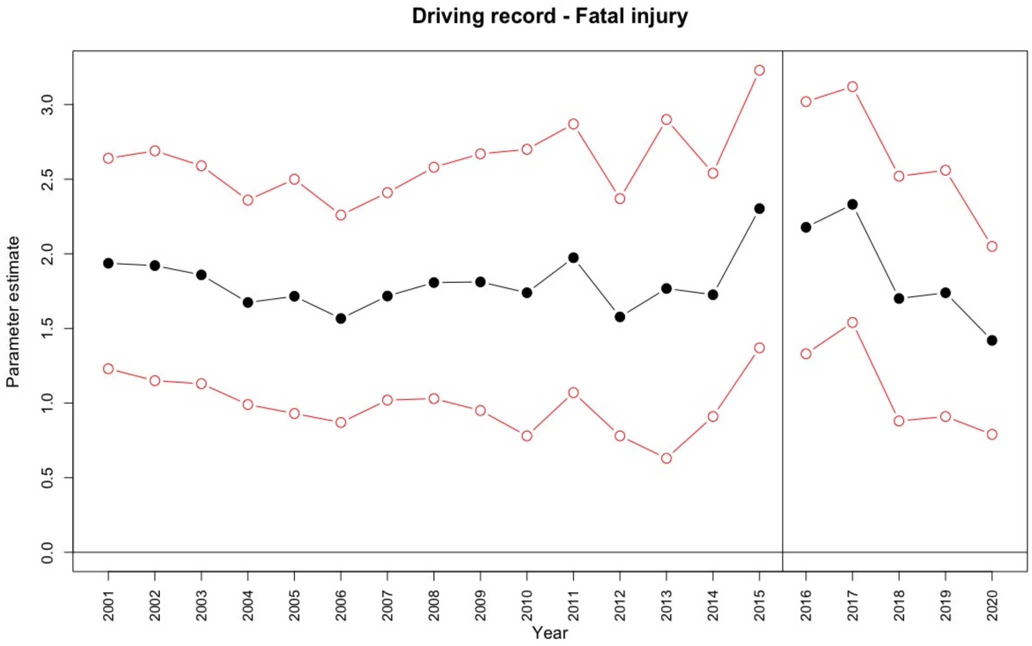


Figure 2.8: Parameter estimates for the covariate that is set to one if the vehicle occupant had previous accidents or driving offences in the last 5 years. Red hollow points are the 90% point-wise confidence intervals for the period 2001-2020. The vertical line between 2015 and 2016 is to separate the data coming from GES and the data from CRSS which are not statistically comparable due to their sample design differences.

2.5 Variable Importance on Severity

For the calibration of the microsimulation in Chapter 4 we are interested in quantifying the cost associated with the MVTAs available in the database presented here. Through an estimated cost we can perform an analysis on the severity of nationally representative police-reported MVTA in the United States. In Section 2.4, we explore as a response variable the injury severity of the drivers and passengers involved in the MVTAs for the period 2001-2020. As mentioned in 2.2.2, we can estimate the cost of the MVTA through the maximum injury severity. This transformation from injury severity to a monetary loss has been explored by the Federal Highway Administration (FHWA) in Council et al. (2005)[3] and Blincoe et al. (2010). The latter is the most recent report of an estimate of

[3] We cite the report from the FHWA because its peer-reviewed version, available in Zaloshnja et al. (2006), removed the tables that allow us to transform from KABCO to an economic cost.

MVTA costs, but its objective is to provide a monetary value of the economic and societal impacts (such as lost quality-of-life) in the U.S. for 2010. While the objective in Council et al. (2005) is to attach estimated costs to the KABCO scale by analysing different types of crashes – an approach that suits better our goal.

The monetary translation in Council et al. (2005) is available by type of accident (rear-ended, sideswipe, opposite direction, among others), if the accident occurred at a signalized intersection, among other considerations that in total make up 22 crash types. We do not use all these crash types since each of them is split by the 5 injury severities in KABCO, meaning that some estimates are unreliable given that they were calibrated with only 3 years of data. For our goal, Table 2.6 (extract of Table 12 in Council et al. (2005)) is more reliable and a better fit to transform the KABCO scale into a monetary loss because it is aggregated by maximum injury severity and rural/urban area. We focus on the economic cost of the MVTA which is a function of: medically-related costs, emergency services, property damage and lost productivity. The severity is inflation-adjusted from 2001 to June 2022 dollars by multiplying by 1.6731, this factor is obtained using the Consumer Price Index (as instructed by the authors) (USBLS, 2022).

Maximum injury severity in MVTA	Rural/Urban	Average economic cost per MVTA	Std. error	95% Conf. Interval	
K	Rural	$ 2,137,655	28,877	2,080,486	2,194,825
A		$ 191,429	17,292	157,195	225,661
B		$ 77,740	11,342	55,285	100,197
C		$ 49,192	4,201	40,875	57,509
O		$ 10,870	1,233	8,431	13,311
K	Urban	$ 1,869,164	50,900	1,768,394	1,969,934
A		$ 169,195	17,872	133,813	204,577
B		$ 58,750	4,509	49,824	67,676
C		$ 45,832	9,637	26,755	64,911
O		$ 10,526	708	9,125	11,926

Table 2.6: Severity per MVTA by maximum injury severity and rural/urban area in 2022 dollars.

A reliable severity analysis is based on insured drivers involved in a MVTA. Since the database comprises a unique severity value for the MVTA and possibly more than one driver, we consider a subset of the database which includes one driver per MVTA. We assume a driver is financially responsible for the total cost of the MVTA, as is done in jurisdictions regulated by at fault accidents. Consequently, the first step is to assign a driver – for those aged 18 or older – who is at fault for the MVTA. We assume if the MVTA involved only one driver, then that person is at fault. For MVTAs with more than one driver, we assign randomly the driver who is at fault. From the sample of 1,583,520 people, this procedure generated a new database with 543,211 drivers who are at fault for the period 2001-2020; we removed 596 MVTA with no registered driver, 20,288 where the age of the driver is unknown and 28,881 where the age is 17 years old or younger. As shown in what follows, our results are robust to the random assignment of financial responsibility.

For the drivers database, we simulated a severity distribution conditional on the maximum injury severity perceived in each MVTA and by its rural/urban designation – a zip code is rural if its population size is less than 2,500 habitants as defined in U.S. Census Bureau (2021a), and urban otherwise. To account for the dispersion of the average economic cost per MVTA, we assume the severity, denoted by S, has a conditional Gamma$(\alpha, 1/\beta)$ distribution with parameters as shown in Table 2.7. As compared to the confidence intervals in Table 2.6, this assumption, on average, increases the lower limit of the confidence intervals by 2% and decreases the size of the interval by 1%. This occurs because the assumption of a gamma distribution has a heavier tail than the distribution of costs in Council et al. (2005).

Maximum injury severity in MVTA	Rural/Urban	α	$1/\beta$	95% Conf. Interval		
K	Rural	5,480	390	$	2,081,429	2,194,621
A		123	1,562	$	159,041	226,775
B		47	1,655	$	57,116	101,494
C		137	359	$	41,302	57,760
O		78	140	$	8,589	13,417
K	Urban	1,349	1,386	$	1,770,721	1,970,232
A		90	1,888	$	135,987	205,977
B		170	346	$	50,245	67,911
O		221	48	$	9,184	11,957
C		23	2,026	$	28,929	66,561

Table 2.7: Shape and rate parameters for the gamma distribution of the average severity per MVTA by the KABCO scale and rural/urban area, along with their corresponding 95% confidence intervals in 2022 dollars.

The log of the simulated severity distribution for the driver database can be seen in Figure 2.9, where the fatal accidents are sampled according to their weight to depict insurance losses of MVTAs that are nationally representative. The distribution is right-skewed and the severity of approximately 80% of the MVTAs is under $47,000. In Section 2.4 we compare passenger and driver characteristics that had different injuries with people that had no injuries through a multinomial regression. In this section, the response variable is continuous and at least four different local maxima can be observed in the histogram[4], which would have to be modeled through a highly parametrized GLM to obtain valid results. Instead, we study the severity through a random forest; this supervised learning algorithm is highly flexible, but at the expense of interpretability.

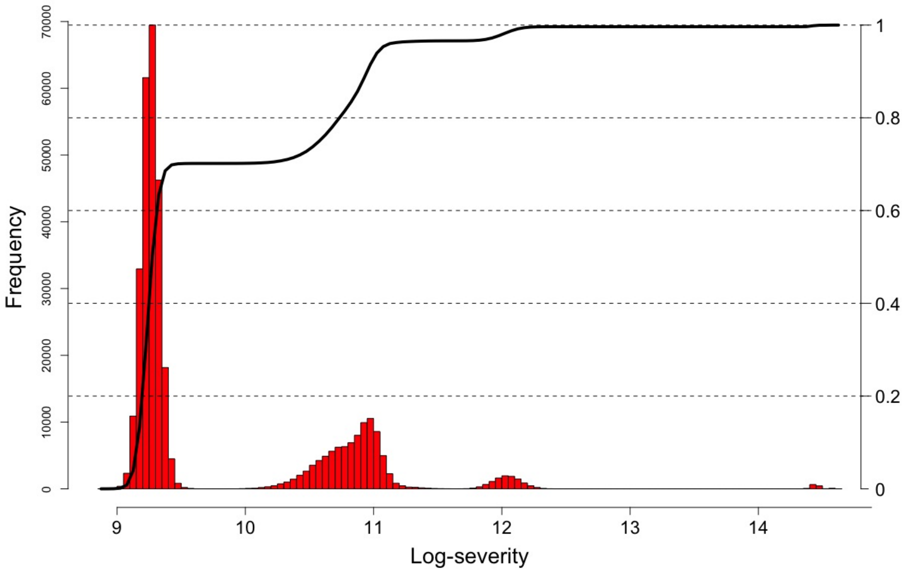


Figure 2.9: Histogram and cumulative density function of the log-severity for the driver database.

Using a random forest we provide the importance – formally defined in what follows – that each variable has in predicting the severity of a MVTA, done separately for the sample of 2001-2015 and 2016-2020. For the estimation we use the `ranger` package (Wright and Ziegler, 2017); a fast and efficient implementation of random forests suited for high-

[4] This multimodality is a consequence of the combination of maximum injury categories and rural/urban areas in the simulation, whereas real severity data might have a distribution with less pronounced modes and more observations in between modes.

dimensional data[5]. As a response variable we use the log-severity and include the weights to account for the oversampling of fatal accidents. This weight is used by the algorithm before growing the trees, to select observations with higher weights at the bootstrap stage. For more details on random forests, see Hastie et al. (2017).

For the feature space, we consider 9 variables: age, gender, weather, hour, drugs/driving, speed related, urbanity, SEG and type of collision with the same categories as presented in Table 2.5. Compared to the analysis in Section 2.4, imputed variables, along with the position in the car, are removed from the variable importance analysis. The former group of variables is removed because they encode our assumptions, and the latter due to fact that it is a driver-only database.

Coding categorical variables as categories, versus using a dummy variable for each category, has an important impact in the final results when using random forests. To select the type of variable coding, along with the optimal number of trees and number of features to sample at each node, we cross validate with 5 samples that are split 80/20 between training and test set. Coding the features as categorical variables has the smallest test error; the optimal number of trees is 500 (1000) and we use 5 (9) variables for the period 2001-2015 (2016-2020). For the drivers database with non-fatal accidents from GES, the optimal number of variables to sample from at each split is less than the dimension of the feature space, which means that there is a benefit by applying a random forest, because it de-correlates predictions through the average of trees constructed from independent bootstrap samples. The fact that the random forest samples from all features at each split, as in the 2016-2020 sample, means that there is low correlation but also low strength if a smaller number of features is sampled. This is mentioned in Breiman (2001) for the case where all input variables are categorical.

We specifically use the permutation variable importance as proposed by Breiman (2001) and recommended in Strobl et al. (2007) for when the cardinality of categorical variables differ, as is our case. Intuitively, the permutation variable importance is a measure of the increase in prediction error for a given random forest after permuting the value of a variable for all observations (breaking the association between the variable of interest and the response variable) compared to the unperturbed forest. In variable selection problems, this measure has the advantage of considering individual variable predictive power along with multivariate interactions (Strobl et al., 2007). Its drawback is that it does not consider the problem of multicollinearity, as more sophisticated and computationally expensive measures do, such as the conditional variable importance (Strobl et al., 2008).

[5] For classification and regression trees, ranger is a C++ implementation of the commonly used `randomForest` package.

For variable $\boldsymbol{X}_j$ in a p-dimensional feature space $X = \{\boldsymbol{X}_j\}_{j=1}^{p}$, we denote its permuted importance by $VI(\boldsymbol{X}_j)$. In a regression setting, this measure for tree t is given by,

$$VI^{(t)}(\boldsymbol{X}_j) = \frac{\sum_{i \in \overline{B}^{(t)}} \left(s_i - \hat{s}_i^{(\pi_j)}\right)^2}{|\overline{B}^{(t)}|} - \frac{\sum_{i \in \overline{B}^{(t)}} \left(s_i - \hat{s}_i\right)^2}{|\overline{B}^{(t)}|} \qquad t \in \{1, 2, \dots, B\}, \tag{2.4}$$

where $\overline{B}^{(t)}$ is the out-of-bag sample, B is the total number of trees, s_i and $\hat{s}_i$ are the observed and predicted severity for observation i, respectively, and $\hat{s}_i^{(\pi_j)}$ denotes the prediction after randomly permuting the value of $\boldsymbol{X}_j$ in observation i. The value $VI(\boldsymbol{X}_j)$ is obtained by averaging $VI^{(t)}(\boldsymbol{X}_j)$ across all trees. Larger (smaller) values of $VI(\boldsymbol{X}_j)$ indicate a variable that has a higher (lower) importance as measured by predictive accuracy.

A comparison of the permutation variable importance, measured as the increase in prediction error of the severity of MVTAs using tuned hyper-parameters for the periods 2001-2015 and 2016-2020, can be observed in Figure 2.10. For both periods, the most important variable is if the driver was under the influence of alcohol or illegal drugs, with an increase in the prediction error of 17% and 11% when the variable is noised up. This result is consistent with the findings in Section 2.4.2 since most expensive MVTAs are fatal accidents and their RRR is significantly high (higher for the period 2016-2020). Interestingly, people aged 60 years and older, and non-collisions were also two of the covariates with parameter estimates that had the highest magnitude for fatal injuries in Section 2.4.3, but given their other categories in the random forest, their variable importance is not as high. In particular, age increases the prediction error of severity by 4% (2%) for the period 2001-2015 (2016-2020) and type of collision (which includes non-collisions) is not important. For the period of 2016-2020, speed related is highly important to predict the severity of the MVTA (the RRR is significant for minor, serious and fatal injuries). Although the two samples are not directly comparable, as explained in Section 2.2, the difference in prediction error using the variable speed related for the period 2001-2015 is influenced by the fact that this variable was underreported for fatal MVTAs before 2008, see Table A.2.

An average over the permutation variable importance of 50 different randomly assigned at fault driver samples did not change the current ordering of variables, nor does the small discrepancy in values have any significant visual impact. This indicates that the values in Figure 2.10 are robust to the assumption of randomly assigning drivers who are at-fault for MVTAs that involve two or more vehicles (72% of the drivers database). This robustness points out the strength in prediction accuracy that driving under the influence, age, speeding, and urbanity have over MVTA severity and these variables therefore, have to be influential in the design of the microsimulation in Chapter 4. We remark that from this

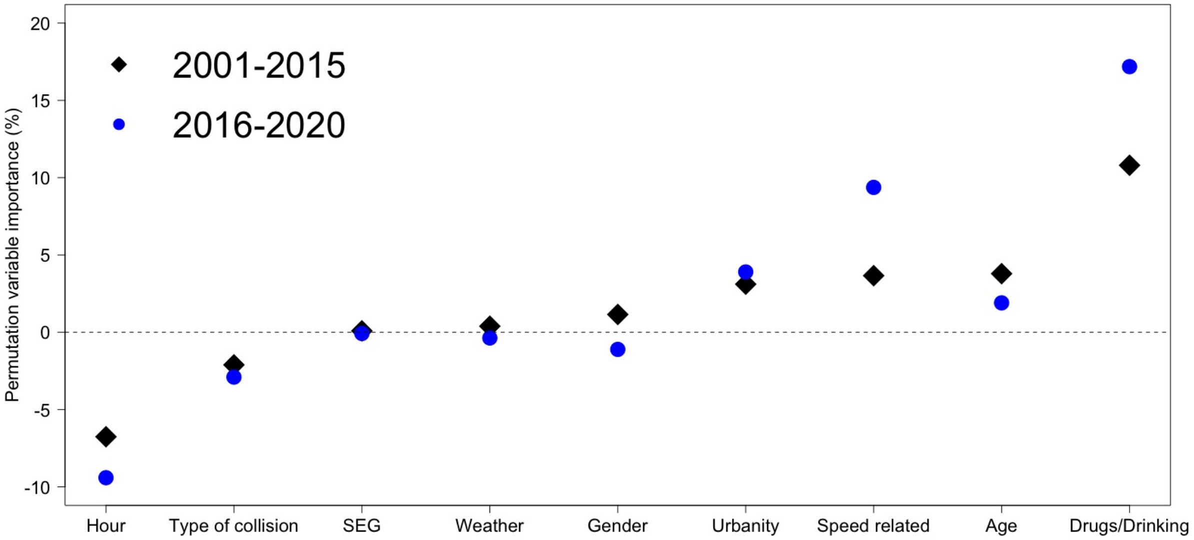


Figure 2.10: Permutation variable importance on the severity of MVTAs for the periods 2001-2015 and 2016-2020. Variables are sorted by the observed importance for the period of 2001-2015.

robustness result no arguments can be made to compare at fault with no fault insurance because on the latter system, the MVTA severity amount would have to be split between the drivers instead of assigned entirely to the driver held responsible.

2.6 Conclusions

In this chapter we have described the construction and validation of a representative database, which we have also made publicly available, of people involved in police-reported MVTAs in the United States during the period 2001-2020. Code that can be used for extracting the data is also provided to allow researchers to modify the underlying filtering processes and additional assumptions, or to extend the database as future years' data become available. We have described the known limitations of the dataset, including inconsistency between states in the allocation of injuries and recording of variables, missing data, especially with respect to the potential influence of alcohol and/or drugs, speed-related MVTAs, and the under-recording of fatal accidents in the GES data.

We have illustrated how analysing the multinomial logistic regression parameters can give informative insights into the key variables related to injury severity; a variable that we transformed into an estimate of severity to explore which variables are relevant when predicting insurance losses. In particular, we have shown that when considering the interaction between gender and the occupant's seat position, and all other things constant, female drivers have a lower relative risk of fatal injuries than male drivers, while female passengers have a higher relative risk. But interestingly, using the random forest, gender has shown no importance in predicting the MVTA severity. Also, our results support the monotonicity of the age effect on the most serious injuries while having an important role in prediction. Finally, we find an association between injury severity and previous driving record, which indicates that it could be valuable to explore the association between the past driving record and the claim severity for auto insurance policyholders.

By making the database (and associated code) publicly available, we hope to facilitate further investigation of factors associated with different levels of MVTA injury severity, particularly as they relate to auto insurance underwriting and ratemaking.

Chapter 3

A Discrimination-Free Premium Under a Causal Framework

We examine the discrimination-free premium in Lindholm et al. (2022a) within a theoretical causal inference framework, and we consider its societal context, to assess when the pricing formula should be used. We consider the insurance pricing problem through the use of directed acyclic graphs. This particular tool allows us to rigorously define an insurance risk factor in a causal framework. We then use this definition in assessing the appropriate application of the discrimination-free premium through three simplified pricing examples, including a health insurance policy and two personal automobile insurance policies with different coverages. From our findings, we suggest criteria for the application of the discrimination-free premium that is dependent on the risk factors and the social context.

3.1 Introduction

In an insurance contract, a policyholder transfers an economic risk to the insurer. The insurer is obligated to compensate the customer for certain unpredictable losses during a time period in exchange for a premium. An *actuarially fair* risk premium (as coined in Arrow (1963)), before loadings for expenses, profit, or adverse experience, corresponds to the expected value of the future loss transferred to the insurance company; this is also known as the equivalence premium principle (Ohlsson, 2010; Dickson et al., 2020). The premium for a risk is dependent on a number of variables, called rating factors, that characterize the individual policyholder and the insured object. A rating factor need not have a direct influence on the loss, but all rating factors are assumed to be correlated to the loss.

Rating factors may be classified as nondiscriminatory or discriminatory, as mentioned in Section 1.2. We use the term '*discriminatory covariates*' to mean characteristics whose use in insurance pricing is legally prohibited or socially controversial. For example, most countries, and most (but not all) states in the United States, have banned the use of race as a rating factor in insurance. The European Union (EU) has also banned the use of gender (European Commission, 2012), although the legislation (described in more detail in Section 3.6) does allow the use of a rating factor which is correlated with gender when it is a 'true risk factor'. The term 'true risk factor' is left undefined, but the implication is that it must have a direct and unarguable connection to the insured risk. In this chapter, we suggest a more rigorous definition of a risk factor using the framework of causal inference, and we consider the potential impacts of associations between discriminatory (or banned) rating factors and nondiscriminatory rating factors on premium calculations.

There is increasing interest in the development of discrimination-free insurance pricing structures. Frees and Huang (2021) present an overview of discrimination in insurance pricing from actuarial, economic and social perspectives. The Casualty Actuarial Society have commissioned a series of research papers on 'Race and Insurance Pricing', including Mosley and Wenman (2022) and Chibanda (2022), and the Australian Institute of Actuaries offer guidelines on anti-discrimination compliance in Dolman et al. (2020). Avraham (2017) proposes a theoretical legal framework, to apply a fair anti-discrimination policy in insurance, arguing that in order to select relevant factors, we must determine whether a feature constitutes a cause of the risk, rather than merely correlating with it. Lindholm et al. (2022a) introduce a discrimination-free pricing formula (previously shown in Section 1.2) based on probabilistic interpretations of direct and indirect discrimination-free prices. Our work builds on theirs, and so their definitions of direct and indirect discrimination, along with their proposed pricing formula are reproduced in the following section.

In this chapter, we use the framework of causal inference to develop a more rigorous definition of a risk factor, and to identify when it would be appropriate to use the discrimination-free pricing formula of (3.2) for a specific insurance product. While traditional statistical theory deals with associations between variables, causal inference uses underlying knowledge or judgment (or experimental evidence) about causal connections between variables, to extend the analysis beyond empirical correlations. Pearl (2009a), considering discrimination in hiring, argues that the discovery of discrimination must be a causal question because "*it requires some knowledge of the data-generating process; it cannot be computed from the data alone, nor from the distributions that govern the data*". This reasoning is directly applicable in this chapter and therefore, explicit statements of our causal assumptions are necessary. Furthermore, to impact processes where only observational data is available, Qureshi et al. (2020) use propensity score analysis in two datasets with the objective of causal discrimination discovery. Although the datasets are not related to insurance pricing, their methodology is applicable in an insurance context. They warn of the biased results that stem from variables that can correlate with the response variable but are also associated with other covariates.

This chapter is organized as follows. In Section 3.2 we recall some notation, and review the definitions of discrimination, and the formula for the discrimination-free premium, from Lindholm et al. (2022a). In Section 3.3 we describe some concepts from causal inference that are applied in subsequent sections. In Section 3.4 we use a causal structure to define an insurance risk factor. In Section 3.5 we provide three simplified insurance examples, and evaluate the application of the discrimination-free premium given the specific causal frameworks. From these results, we suggest criteria for its use. Section 3.6 highlights the dependence that the social context has on the application of the discrimination-free premium. In Section 3.7 we discuss some issues that can arise as a consequence of using causally based inference tools in insurance applications. We conclude in Section 3.8.

3.2 Direct and Indirect Discrimination

We assume an insurance loss, denoted by Y, is a random variable in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathbb{P}$ represents the real world measure. Y is dependent on a set of covariates, representing the potential rating factors. The set of rating factors can be divided into nondiscriminatory covariates, denoted by $\boldsymbol{X}$, and discriminatory covariates, denoted by $\boldsymbol{D}$. We denote the marginal and conditional distribution of the discriminatory covariates under $\mathbb{P}$ by $\boldsymbol{D} \sim \mathbb{P}(\boldsymbol{d})$ and $\boldsymbol{D}|\boldsymbol{X} = \boldsymbol{x} \sim \mathbb{P}(\boldsymbol{d}|\boldsymbol{x})$.

A key component in Lindholm et al. (2022a) is that there are multiple ways to achieve a discrimination-free price. In their probabilistic formulation, it is not necessary to work with the real world distribution of the discriminatory covariates $\mathbb{P}(\boldsymbol{D})$. Consequently, we work with arbitrary measures, $\mathbb{P}^*$, which can be decomposed as

$$\mathbb{P}^*(Y, \boldsymbol{X}, \boldsymbol{D}) = \mathbb{P}(Y|\boldsymbol{X}, \boldsymbol{D})\mathbb{P}(\boldsymbol{X})\mathbb{P}^*(\boldsymbol{D}).$$

The predictive part of the model, $\mathbb{P}(Y|\boldsymbol{X}, \boldsymbol{D})$, uses the real world measure and can hence be consistently estimated from observational data. The component $\mathbb{P}^*(\boldsymbol{D})$ allows for an arbitrary reweighting of costs across the discriminating covariates, in ways that are independent of $\boldsymbol{X}$. In other words, it eliminates one form of indirect discrimination (described as proxy discrimination in Section 1.2) since the predictive ability that the nondiscriminatory covariates could have over the discriminatory covariates is artificially removed.

Lindholm et al. (2022a) makes the following definitions of direct and indirect discrimination-free pricing, based on an arbitrary probability measure $\mathbb{P}^*$ on $(\Omega, \mathcal{F})$, and on $\boldsymbol{Z}$, which is a sub-vector of $(\boldsymbol{X}, \boldsymbol{D})$.

Definition 3.1 (Direct discrimination)**.** *A price avoids direct discrimination if it can be written as*

$$\mu^*(\boldsymbol{Z}) := \mathbb{E}^*[Y|\boldsymbol{Z}], \tag{3.1}$$

where $\boldsymbol{Z}$ is $\sigma(\boldsymbol{X})$-measurable, and the expectation is taken w.r.t. $\mathbb{P}^*$ where $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^*)$.

Definition 3.2 (Indirect discrimination)**.** *The price $\mu^*(\boldsymbol{Z})$ of Definition 3.1 is said to also avoid indirect discrimination if $\boldsymbol{Z}$ and $\boldsymbol{D}$ are independent under $\mathbb{P}^*$.*

The discrimination-free premium introduced in Lindholm et al. (2022a), which satisfies the requirements of Definitions 3.1 and 3.2 is,

$$h^*(\boldsymbol{X}) := \int_{\boldsymbol{d}} \mu(\boldsymbol{X}, \boldsymbol{d}) \, \mathrm{d}\mathbb{P}^*(\boldsymbol{d}), \tag{3.2}$$

where $\mathbb{P}^*$ is dominated by $\mathbb{P}$, and $\mu(\boldsymbol{X}, \boldsymbol{d}) = \mathbb{E}[Y|\boldsymbol{X}, \boldsymbol{d}]$ is a price computed under the real world measure $\mathbb{P}$. This price can be estimated by any of the common statistical methods used in insurance, such as a generalized linear model (GLM), including as rating factors both the discriminatory and nondiscriminatory covariates. A simple walk-through example of the application of the discrimination-free premium, contrasted with an unawareness price, is given in Appendix C.

While $\mathbb{P}^*$ is formally arbitrary, some choices are more sensible than others. Lindholm et al. (2022a) use $\mathbb{P}^*(\boldsymbol{d}) = \mathbb{P}(\boldsymbol{d})$, where the marginal distribution $\mathbb{P}(\boldsymbol{d})$ is calculated empirically as the observed relative frequency of $\boldsymbol{d}$ in the portfolio. Note however, that the discrimination-free premium $h^*(\boldsymbol{X})$ introduces a portfolio bias, because $\mathbb{E}[Y] \neq \mathbb{E}[h^*(\boldsymbol{X})]$ (except for the special case where $\boldsymbol{X}$ and $\boldsymbol{D}$ are $\mathbb{P}$-independent and $\mathbb{P}^*(\boldsymbol{d}) = \mathbb{P}(\boldsymbol{d})$), which suggests the possibility that $\mathbb{P}^*(\boldsymbol{d})$ could be selected to correct for the bias. This is one of three methods proposed by Lindholm et al. (2022a) to correct for bias; the other two are a level additive adjustment and a level multiplicative adjustment, where either the additive or multiplicative adjustment is selected such that the total premium income is equal to the total expected loss. In this chapter, we use the multiplicative adjustment to the discrimination-free premium. The first method – altering the marginal distribution of the nondiscriminatory covariates $\mathbb{P}^*(\boldsymbol{d})$ to remove the portfolio bias – although mathematically appealing, is non-intuitive, and would be difficult to justify to regulators and other stakeholders. The level additive method approach is discarded because it can result in negative prices for some policyholders, and can actually exacerbate discrimination.

3.3 Causal Inference

In this section, we introduce some important definitions from causal inference, which we then use to relate the discrimination-free premium in (3.2) with the adjustment formula, a common method used in causal studies to generate unbiased estimates of causal quantities.

3.3.1 Markovian directed acyclic graphs (DAGs) and structural causal models (SCMs)

A causal structure is composed of a set of nodes, and a collection of arrows, or directed edges. The nodes represent (possibly random) variables, denoted by $\boldsymbol{V}$ and $\boldsymbol{U}$. The former is the set of observed or measured variables, and it is also called the set of endogenous variables. The variable set $\boldsymbol{U}$ is composed of unobserved or unmeasured variables. A directed edge from one node to another represents a direct causal effect. Some causal structures can be visually expressed through graphs such as those in Figure 3.1. The starting node for each edge is called a 'parent' of the end point, known as a 'child' node. If two nodes can be traced along the directed edges, the first node is the ancestor, and every node on the path is a descendant of the first node. In Figure 3.1, labeled $\mathcal{G}$, we have for the endogenous variables that

(i) Y is a child of the parent nodes D and X, denoted by $\mathbf{Pa}(Y) = \{D, X\}$, and it is also a descendant of D since it can be traced along the indirect path $D \to X \to Y$;

(ii) X is a child of D.

For $\mathcal{G}$ in Figure 3.1, no path exists from any node back to itself; directed graphs with this property are called directed acyclic graphs. In what follows, we require the causal structures to be acyclic to obtain valid results. A DAG can allow us to visually express the assumed causal relationship between variables used to price an insurance product. In order to capture the randomness involved, we endow the DAG with a probability measure over sets $\boldsymbol{V}$ and $\boldsymbol{U}$, denoted by $\mathbb{P}$ for $\mathcal{G}$.

The variables in $\boldsymbol{U}$ are also called exogenous, or error terms because they act in a DAG solely as parent nodes which means that for every $U_i \in \boldsymbol{U}$ we have $\mathbf{Pa}(U_i) = \emptyset$. In other words, every variable in the endogenous set $\boldsymbol{V}$ is a child or descendant of a variable in $\boldsymbol{U}$. In what follows, we assume the exogenous variables are jointly independent. This is a strong assumption which, coupled with the acyclic structure, allows the model to be

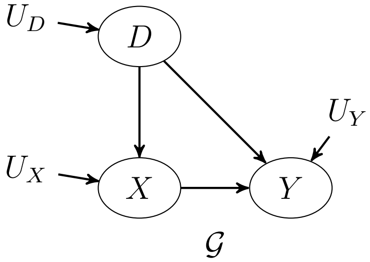


Figure 3.1: Example of a DAG, denoted by $\mathcal{G}$, where the endogenous set $\boldsymbol{V}$ is composed of $\{D, X, Y\}$ and $\boldsymbol{U} = \{U_D, U_X, U_Y\}$.

Markovian (Pearl, 2009b). If the unobserved variables are dependent, that is a signal that the causal structure is incomplete, as there is some mechanism not represented in the DAG responsible for the dependence (Spirtes et al., 2000). For example, $\mathcal{G}$ in Figure 3.1 is called Markovian iff

$$\mathbb{P}(U_D, U_X, U_Y) = \mathbb{P}(U_D)\mathbb{P}(U_X)\mathbb{P}(U_Y).$$

In this chapter, $\mathcal{G}$ is the only DAG that explicitly includes both the endogenous and exogenous sets. Henceforth, we assume the existence of the set $\boldsymbol{U}$ is implicit to simplify the notation of causal structures and to avoid cluttering the DAGs that are used. This simplification does not undermine or weaken the examples included.

A Markovian model implies the parental Markov condition[1] which states that each variable is independent of all its nondescendants, given its parents (also called the local Markov condition in Lauritzen (1996)). This means that an important benefit from using a Markovian DAG is that it visually encodes independence relations in a probability distribution. For example, if we focus solely on the chain $D \to X \to Y$, this DAG entails that: (1) D and X are dependent, (2) X and Y are dependent, (3) D and Y are dependent, and (4) D and Y are independent, conditional on X.[2]

A DAG provides a qualitative understanding of causal relationships (Pearl et al., 2016). To quantify these relationships, we use a structural causal model (SCM), which consists of a causal structure and a set of functions, denoted by $\boldsymbol{F}$, that assigns values to each variable in the endogenous set $\boldsymbol{V}$. For example, a structural causal model for $\mathcal{G}$ in Figure 3.1 can be

[1] Through the causal Markov condition, which can be found as Theorem 1.4.1 in Pearl (2009b).

[2] Formally, to read dependencies in distribution from the DAG, we also need the minimality condition to hold, which says that the probability distribution does not satisfy any additional independencies to those imposed by the associated DAG (Neal, 2020).

expressed as $\boldsymbol{V} = \{D, X, Y\}$, $\boldsymbol{U} = \{U_D, U_X, U_Y\}$ and $\boldsymbol{F} = \{f_D, f_X, f_Y\}$. Then, the values of the variables in $\boldsymbol{V}$ can be assigned through the functions in $\boldsymbol{F}$ as:

$$\begin{aligned} D &= f_D(U_D) \\ X &= f_X(D, U_X) \\ Y &= f_Y(D, X, U_Y) \end{aligned}$$

3.3.2 The adjustment formula

Suppose $\mathcal{G}$ in Figure 3.1 describes a causal structure for an insurance loss and let $\mathbb{P}(y) = \mathbb{P}(Y \leq y)$. Our objective is to discover the causal effect that the nondiscriminatory covariate X has over our response variable Y, in order to compute the premium. Further, we might be legally required to understand the role of the discriminatory covariate D in the pricing problem at hand. Simply calculating the conditional probability $\mathbb{P}(y|X = x)$ does not solve the problem because in $\mathcal{G}$, the variable D is associated with both X and Y, which means the causal effect of X on Y is confounded by the influence of D. A variable is a confounder when it is associated with an explanatory variable and the response variable.

In order to discover the causal effect of X on Y in an unbiased way, we would need to intervene and remove the influence of D over X. This would guarantee – if the causal model is a valid representation of reality – that the change in the outcome variable must be due solely to the change in the nondiscriminatory covariate, X. This intervention in causal inference is done through the *do*-operator, denoted by $do(X = x)$ (Pearl et al., 2016). This operator reads as an intervention where instead of letting the variable X vary naturally, we fix $X = x$ for every observation, which amounts to removing all external effects over X. Graphically, this intervention is equivalent to eliminating all directed edges that point into X. For example, we display the endogenous variables of $\mathcal{G}$ in Figure 3.1 in the left-hand DAG of Figure 3.2, and after the intervention $do(X = x)$, the directed edge $D \to X$ from $\mathcal{G}$ is removed, resulting in the right-hand DAG, labeled by $\mathcal{G}^*$. We endow this modified DAG with a probability measure denoted by $\mathbb{P}_m$.

Thus, to quantify a causal effect we need to obtain $\mathbb{P}(y|do(X = x))$. To do this, in addition to the assumption of a Markovian model, we require what is known as the modularity assumption in Neal (2020)[3], which requires that, if we intervene on a set of nodes $\boldsymbol{X}$, setting them to constants, then for all $V \in \boldsymbol{V}$, we have

[3] This assumption can be found in indices (ii) and (iii) in Definition 1.3.1 of Pearl (2009b).

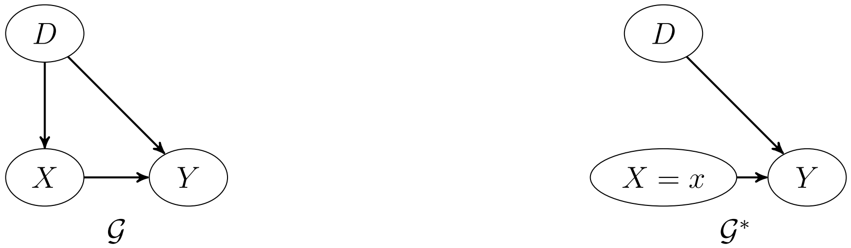


Figure 3.2: Example of a Markovian DAG (left-hand side), denoted by $\mathcal{G}$, where the causal effect of X on Y is confounded by D. In the right-hand side DAG, denoted by $\mathcal{G}^*$, the variable $X = x$ is fixed, removing the effect that D could have over X.

(i) If $V \in \boldsymbol{X}$, then

$$\mathbb{P}(v|\mathbf{pa}(V)) = \begin{cases} 1 & \text{if } v \text{ is the value that } V \text{ was set to by the intervention;} \\ 0 & \text{otherwise.} \end{cases}$$

(ii) If $V \notin \boldsymbol{X}$, then $\mathbb{P}(v|\mathbf{pa}(V))$ remains invariant,

where $\mathbf{pa}(V)$ denotes values of the parents of variable V. If V has no parent variable in $\boldsymbol{V}$, then $\mathbb{P}(v|\mathbf{pa}(V)) = \mathbb{P}(v)$. When the modularity assumption holds, we say that v being consistent with the intervention (Pearl, 2009b). The modularity assumption in Figure 3.2 means that after the intervention in $\mathcal{G}^*$, the nondiscriminatory covariate fixed to $X = x$ is independent of the value of D, which results in the equality $\mathbb{P}(y|do(X = x)) = \mathbb{P}_m(y|X = x)$. Furthermore, the process by which the response variable Y responds to X and D, and the marginal distribution of D are unchanged after the intervention in Figure 3.2. This results in the following invariance equations (Pearl et al., 2016):

$$\begin{aligned} \forall\, y,\, x,\, d: \qquad & \mathbb{P}(y|X = x, D = d) = \mathbb{P}_m(y|X = x, D = d), \\ \text{and} \qquad & \mathbb{P}(d) = \mathbb{P}_m(d). \end{aligned} \tag{3.3}$$

The invariance results are computed using the particular structure encoded in $\mathcal{G}$ and $\mathcal{G}^*$. These equations are the building blocks for computing causal effects from the combination of data and the causal assumptions encoded in the DAG. They allow the so-called 'post-intervention' quantity $\mathbb{P}(y|do(X = x))$ to be computed in observational studies (that is, where a physical intervention is impossible, such as the case of insurance pricing) through the use of pre-intervention conditional probabilities, in a formula known as the adjustment formula. For $\mathcal{G}$, the adjustment formula to uncover the causal effect of X on Y is,

$$\mathbb{P}(y|do(X=x)) = \int_d \mathbb{P}(y|X=x, D=d)\, \mathrm{d}\mathbb{P}(d),$$

consequently, the expected value after removing external effects over X is,

$$\mathbb{E}[Y|do(X=x)] = \int_d \mathbb{E}[Y|X=x, D=d]\, \mathrm{d}\mathbb{P}(d). \tag{3.4}$$

3.3.3 The adjustment formula and the discrimination-free premium

We note that if X and D are univariate, the discrimination-free premium in (3.2) is identical to equation (3.4) when we use the invariance equation in (3.3), with $\mathbb{P}^* = \mathbb{P}_m$. The invariance equation sets the arbitrary probability measure to be equal to the measure associated with the modified DAG $\mathcal{G}^*$, and by transitivity of the probability measures, corresponds to using $\mathbb{P}^*(d) = \mathbb{P}(d)$ in the discrimination-free premium. This means that for the causal structure in Figure 3.2, the discrimination-free premium is equivalent to the expected value of Y, when we remove the external effects over the nondiscriminatory covariates, $h(x) = \mathbb{E}[Y|do(X=x)]$.

The connection between the adjustment formula and the discrimination-free premium in Lindholm et al. (2022a) is a foundation for this chapter. Consequently, we include a step-by-step derivation of equation (3.4). The derivation provides some insight into the role of the marginal distribution $\mathbb{P}^*(d) = \mathbb{P}(d)$ in the discrimination-free premium. We refer back to our Markovian model in Figure 3.2 that satisfies the modularity assumption, where the initial question for the development of the adjustment formula is the desire to obtain the causal effect $\mathbb{P}(y|do(X=x))$ in $\mathcal{G}$. By definition of the *do*-operator, we remove all external effects over X, leading to the modified probability measure $\mathbb{P}_m$ in $\mathcal{G}^*$, where

$$\begin{aligned}
\mathbb{P}(y|do(X=x)) &= \mathbb{P}_m(y|X=x) \\
&= \int_d \mathbb{P}_m(y|X=x, D=d)\, \mathrm{d}\mathbb{P}_m(d|x) \\
&= \int_d \mathbb{P}_m(y|X=x, D=d)\, \mathrm{d}\mathbb{P}_m(d) && \text{(by independence in } \mathcal{G}^*\text{)} \\
\mathbb{P}(y|do(X=x)) &= \int_d \mathbb{P}(y|X=x, D=d)\, \mathrm{d}\mathbb{P}(d) && \text{(by the equalities of invariance)}
\end{aligned}$$

This result is known as the adjustment formula in Pearl et al. (2016). Since premiums are obtained through the future expected loss and by assuming an insurance loss Y is $\mathcal{L}^1$-measurable, we can obtain the post-intervened expected value using the adjustment formula as follows,

$$\begin{aligned}
\mathbb{E}[Y|do(X=x)] &= \int_y y \, \mathrm{d}\mathbb{P}(y|do(X=x)) \\
&= \int_{y,d} y \, \mathrm{d}\mathbb{P}(y|X=x, D=d) \, \mathrm{d}\mathbb{P}(d) \qquad \text{(by the adjustment formula)} \\
&= \int_d \mathbb{E}[Y|X=x, D=d] \, \mathrm{d}\mathbb{P}(d) \\
&= \int_d \mu(x,d) \, \mathrm{d}\mathbb{P}(d) \\
\mathbb{E}[Y|do(X=x)] &= h(x)
\end{aligned}$$

It is important to note that the same formula has come from two quite different arguments. The first was based on a particular, if natural, choice of reweighting a discriminatory price to remove the direct and indirect effects of the discriminatory variable. The second is a computation of the causal effect of X on Y in the case when the parental Markov condition and the modularity assumption hold.

We can compute equation (3.4) in a more general DAG, by obtaining a factorization of $\mathbb{P}(\boldsymbol{v}|do(\boldsymbol{X}=\boldsymbol{x}))$ that only requires pre-intervention conditional probabilities. This can be done with the truncated factorization formula in equation (3.5) below, which holds under the parental Markov condition and the modularity assumption (Pearl, 2009a).

$$\mathbb{P}(\boldsymbol{v}|do(\boldsymbol{X}=\boldsymbol{x})) = \prod_{V \notin \boldsymbol{X}} \mathbb{P}(v|\mathbf{pa}(V)) \qquad \text{for all } \boldsymbol{v} \text{ consistent with } \boldsymbol{x}. \tag{3.5}$$

In order to apply these formulas, we need the confounding set $\boldsymbol{D}$ to satisfy the backdoor criterion (Pearl et al., 2016). This requires that any node in the confounding set $\boldsymbol{D}$ should not be a descendant of X, and should block every path between X and Y that contains a directed edge into X. In a multivariate setting, the set $\boldsymbol{D}$ has to satisfy the backdoor criterion relative to any pair (W, Z) such that $W \in \boldsymbol{X}$ and $Z \in \boldsymbol{Y}$ (Pearl, 2009b). The backdoor criterion is satisfied with respect to variable D in all of the DAGs considered in this chapter. When the backdoor criterion is not satisfied, and if the specific causal structure allows it, we would have to rely on a tool called *do*-calculus to uncover the causal effect. See Pearl (2009b) for details.

3.4 Risk Factors

The causal definitions described in Section 3.3 allow us to define an insurance risk factor more formally. First, we note the characteristics that differentiate a rating factor from a risk factor. A rating factor is a variable or characteristic used in pricing insurance contracts. The selection of rating factors is restricted both by availability and, in some cases, by law. On the other hand, a risk factor is an attribute, characteristic or exposure of an individual or event that increases the likelihood of a future loss (SOA, 2022). In the causal inference framework, given a DAG, we interpret this direct connection as follows.

Definition 3.3. *(Insurance risk factor) Given a directed acyclic graph, $\mathcal{G}$, over $\boldsymbol{V}$, where $Y \in \boldsymbol{V}$ represents the insurance loss, a variable $X \in \boldsymbol{V}$ is called a risk factor of Y iff X is a parent of Y.*

The set of risk factors of Y is therefore the set of parent nodes of Y. Definition 3.3 implies that if we change the value of a risk factor, all other things equal, it will have a direct causal effect on the conditional distribution of the child node. Therefore, to achieve an actuarially fair premium we seek to identify the risk factors of an specific insurance product and measure their causal effects in an unbiased way in order to, as accurately as possible, estimate the expected future loss.

For example, consider a personal automobile insurance policy with a third-party liability (TPL) coverage. In the United States, TPL is compulsory in most states. It provides cover for expenses of a third party for which the policyholder is liable, and which result from an insured risk. It could reasonably be hypothesized that distance driven has a causal effect on the claims distribution, since it is a measure of risk exposure. Given this hypothesis, we can draw the simple DAG shown in Figure 3.3.

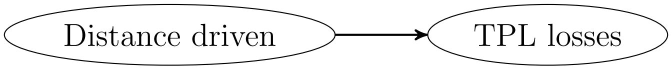


Figure 3.3: DAG for a TPL coverage example where the variable distance driven is hypothesized to be a risk factor of the claims distribution.

A DAG, such as the one in Figure 3.3, must be associated with a SCM in order to obtain quantitative results. In the example in Figure 3.3 with a single risk factor, as a basis for a pricing model, it is in the interest of the insurer to quantify the effect that distance driven has on the risk of an automobile accident. We note that a DAG carries our assumptions on the direction of the causal structure which implies the SCM is asymmetric with respect to its arguments.

Defining a risk factor relative to a choice of DAG provides a rigorous tool to differentiate between rating factors and risk factors in a causal framework. Using Definition 3.3 we can explore insurance pricing problems and determine the role of each variable in the causal structure in order to determine if the discrimination-free premium should be used.

In practice, the design of the DAG will rely on the modeler's expert judgment of the underlying processes. In some cases, such as the causal effect of distance driven on TPL losses, the directed arrow representing a causal process, seems quite natural. In other cases, the connection may be more questionable. It is important to bear in mind that expert judgment can be flawed and consequently, lead to misleading or spurious results.

3.5 The Discrimination-free Premium Under a Causal Framework

In this section, we explore three different examples of insurance pricing problems, to analyze the applicability of the discrimination-free premium from equation (3.2) in different, stylized contexts.

3.5.1 Example A: Health insurance pricing

In this section, we reframe the numerical illustration from Lindholm et al. (2022a) as a causal model. Lindholm et al. (2022a) explore their discrimination-free pricing method using a model of health insurance costs. The cost (Y) is assumed to be a function of three covariates: gender (D), age (X_1), and smoking status (X_2), so the set of variables is $\boldsymbol{V} = \{Y, D, X_1, X_2\}$. The set of nondiscriminatory covariates is comprised of age and smoking habits, denoted by $\boldsymbol{X} = (X_1, X_2)^\top$ and the univariate discriminatory covariate is gender. In their numerical illustration, gender has both a direct and indirect impact on costs; direct through gender specific health costs, and indirect through the impact of gender on smoking habits. Smoking habits and age have a direct effect on costs, but no effect on each other. While in Lindholm et al. (2022a) this model is expressed as a GLM, we have interpreted it through the DAG $\mathcal{G}_A$, shown on the left-hand side of Figure 3.4.

We assume the model is Markovian and the modularity assumption holds. Before the intervention, the DAG $\mathcal{G}_A$ shows that gender is acting as a confounding variable by being

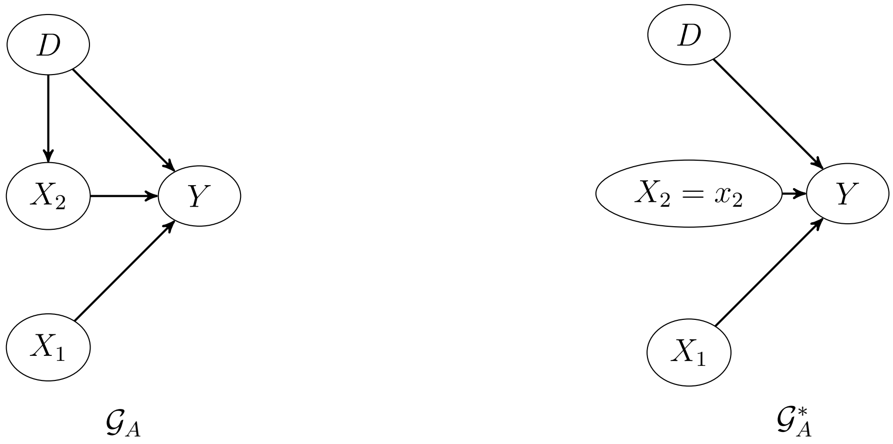


Figure 3.4: $\mathcal{G}_A$, is a causal framing of an example in Lindholm et al. (2022a); $\mathcal{G}_A^*$, is the modified DAG, created by artificially removing the effect of the discriminatory covariate over the nondiscriminatory covariates.

associated with with health insurance costs both directly and indirectly. In the modified DAG, $\mathcal{G}_A^*$, we have removed the indirect effect. This allows us to compute a premium for each policyholder that avoids direct and indirect discrimination (as defined in Section 3.2) by quantifying the causal effect of the nondiscriminatory covariates on health insurance costs, represented by $\mathbb{E}[Y|do(\boldsymbol{X} = \boldsymbol{x})]$. In $\mathcal{G}_A$, the only directed edge that points into $\boldsymbol{X}$ is $D \to X_2$, which means the *do*-operator can be simplified to $do(\boldsymbol{X} = \boldsymbol{x}) = do(X_2 = x_2)$.

In this illustration, using the discrimination-free premium is justified because of the confounding bias present in $\mathcal{G}_A$, coupled with the fact that the causal assumptions required for the equivalence between the discrimination-free premium (3.2) and the adjustment formula with an expected value (3.4), hold. In particular, for a policyholder with characteristics $\boldsymbol{X} = \boldsymbol{x}$, the discrimination-free premium is calculated in its discrete form as,

$$h(\boldsymbol{x}) = \sum_{d \in \{\text{woman, man}\}} \mu(\boldsymbol{x}, d)\, \mathbb{p}(d),$$

where $\mathbb{p}(d)$ denotes the probability mass function of D. The equality of marginal distributions $\mathbb{P}^*(d) = \mathbb{P}(d)$ can be chosen because D satisfies the backdoor criterion relative to X_2.

In this example, the application of the discrimination-free premium has three important advantages:

1. The confounding bias introduced by gender is removed, allowing the premium to reflect specifically the causal effect of smoking habits over expected health insurance costs.
2. The premium is gender-neutral, and, furthermore, does not allow inference of gender based on smoking habits.
3. The premium fulfills the solidarity mechanism established in European Commission (2012) where (for example) birthing-related costs do not result in differences between premiums charged for men and women.

There could be a case in this example where the conditional probabilities underlying the population are such that the application of the discrimination-free premium could result in a lower premium for smokers, as compared to the unawareness premium. The correct interpretation for this scenario is that we are subsidizing either females or males by removing the influence of $\mathbb{p}(d|x_2)$, and it should not be seen as a subsidy for smokers.

This analysis utilizes the particular choice of causal assumptions characterized by $\mathcal{G}_A$ in Figure 3.4. The choice of a DAG represents a story about how we believe the world operates, and different assumptions are possible which would be consistent with the GLM equations in Lindholm et al. (2022a). Different DAGs would result in different valuations of direct causal effects. Pearl (2009a) argues that assessments of fairness are causal statements and so only exist relative to explicit statements of our causal story.

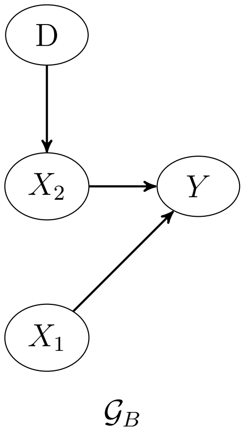
3.5.2 Example B: Third-party liability pricing

We now consider an example relating to TPL within an auto insurance policy. As usual, we let Y denote the insurance loss. We assume three other nodes in the causal structure: gender (D), age (X_1) and driving habits (X_2). As before, gender is a discriminatory covariate, while age and driving habits are nondiscriminatory covariates. We assume the causal structure, denoted by $\mathcal{G}_B$ in Figure 3.5, which is similar to graph $\mathcal{G}_A$ above, but without the direct causal link between D and Y.

Note that X_1 and X_2 are risk factors; D is not a risk factor in this case. This means that conditional on X_1 and X_2, gender is independent of the TPL costs, that is

$$\mathbb{E}[Y|\boldsymbol{X}, D] = \mathbb{E}[Y|\boldsymbol{X}]. \tag{3.6}$$

In this example, because of the conditional independence (3.6) and regardless of the choice of probability measure $\mathbb{P}^*$, the discrimination-free premium is equal to the premium



$\mathcal{G}_B$

Figure 3.5: DAG, denoted by $\mathcal{G}_B$, for a personal automobile insurance example showing the causal structure over third-party liability costs (Y), gender (D), age (X_1) and driving habits (X_2).

obtained by ignoring gender completely (the unawareness premium), because $Y|\boldsymbol{X}$ is independent of D, so that

$$h^*(\boldsymbol{X}) = \int_d \mathbb{E}[Y|\boldsymbol{X}, d]\mathrm{d}\mathbb{P}^*(d) = \int_d \mathbb{E}[Y|\boldsymbol{X}]\mathrm{d}\mathbb{P}^*(d) = \mathbb{E}[Y|\boldsymbol{X}].$$

Note that there may still be a statistically significant difference in TPL costs by gender at the population level. The path $D \to X_2 \to Y$ in $\mathcal{G}_B$ implies there is a quantifiable indirect effect on the distribution of losses by gender.

However, in practice, we may find a difference between the premium calculated with gender as a rating factor and the unawareness premium. This can arise due to sampling bias, model misspecification, incorrect variable selection, measurement error, discretization, or overfitting. Each of these can create spurious dependence such that $\hat{\mathbb{E}}[Y|\boldsymbol{X}, D] \neq \hat{\mathbb{E}}[Y|\boldsymbol{X}]$. In this case, the discrimination-free premium would incorrectly average prices by gender.

To illustrate this problem, we present a simulation study where the variables have the causal structure implied by $\mathcal{G}_B$ in Figure 3.5. The simulation assumptions and parameters are as follows:

Number of policyholders:	$n = 58,000$
Gender distribution:	$D \sim \text{Bin}(1, 0.4)$, where $D = 0$ for females and $D = 1$ for males.
Age distribution:	$X_1 \sim \text{Unif}(18, 80)$
Driving habit distribution:	$X_2 = 0$ for good drivers, $X_2 = 1$ for bad drivers;

$$\text{If } D = 0,\ X_2 \sim \text{Bin}(1, 0.0393)$$
$$\text{If } D = 1,\ X_2 \sim \text{Bin}(1, 0.9241).$$

The selection of parameters for the Bernoulli distributions is made such that the conditional probabilities $\mathbb{P}(D = 1|X_2 = 1) = 0.94$ and $\mathbb{P}(D = 1|X_2 = 0) = 0.05$ are satisfied. A justification for these conditional probabilities and the number of policyholders is made in Example 3.5.3 which follows. The TPL losses are generated by

$$Y \sim \text{Gamma}(\exp(5.15 + 0.10X_1 + 0.22X_2)).$$

We also assume that our information about X_2 is subject to measurement error. This can be interpreted as modeling X_2 with a proxy, denoted $X_2^{(p)}$. We set the proxy variable to have a 1% probability of being misclassified if the policyholder has good driving habits ($X_2 = 0$) and a 2% probability of misclassification for policyholders with bad driving habits ($X_2 = 1$). We assume the rate of misclassification is larger for drivers with bad driving habits because they are slightly under-represented in the portfolio ($\mathbb{P}(X_2 = 1) = 0.39$), so that we have less data from this group to accurately approximate the covariate X_2.

To compute the discrimination-free premium, first we specify the model with the covariates that we have access to, denoted by M_1, which is a Gamma GLM with a log-link, denoted by g, to estimate the conditional expectation of Y as,

$$g\left(\mathbb{E}[Y|\boldsymbol{X}, D]\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{(p)} + \beta_3 D. \tag{3.7}$$

We also denote by M_2 the model M_1 without the discriminatory covariate. Because we have perfect knowledge of the underlying model, we know that the true parameters are $\{\beta_0, \beta_1, \beta_2, \beta_3\} = \{5.15, 0.10, 0.22, 0.00\}$. Parameter estimates for both models can be seen in Table 3.1. Due to the causal dependence between D and X_2, and by transitivity to $X_2^{(p)}$, the estimate $\hat{\beta}_2$ changes significantly when the discriminatory covariate is included.

Model		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$\hat{M}_1$	$\hat{\mathbb{E}}[Y\|\boldsymbol{X}, D]$	5.1509 (0.0003)	0.10 (10^{-5})	0.1756 (0.0004)	0.0441 (0.0004)
$\hat{M}_2$	$\hat{\mathbb{E}}[Y\|\boldsymbol{X}]$	5.1534 (0.0004)	0.10 (10^{-5})	0.2137 (0.0002)	-
True values		5.15	0.10	0.22	0.00

Table 3.1: Parameter estimates for the model in equation (3.7), obtained using simulated data where X_2 is subject to measurement error. Standard errors of the parameter estimates are included in parenthesis.

Because the *estimated* conditional loss $\hat{\mathbb{E}}[Y|\boldsymbol{X}, D]$ is not equal to $\hat{\mathbb{E}}[Y|\boldsymbol{X}]$, the choice of probability measure $\mathbb{P}^*$ now has an impact in the computation of the discrimination-free premium. Here, we choose to set the arbitrary probability measure equal to the empirical measure, that is, we compute $\hat{\mathbb{P}}(D = 1)$ as the empirical proportion of male policyholders in the portfolio, which is 0.4037. As mentioned in Section 3.2, the discrimination-free premium has a bias at the portfolio level, since $\boldsymbol{X}$ and D are not independent. We use the log-link in the GLM to remove the portfolio bias by proportionally adjusting the discrimination-free premium through $\gamma = e^c$, where $\hat{c} = \log\left(\hat{\mathbb{E}}[Y]/\hat{\mathbb{E}}[\hat{h}(\boldsymbol{X})]\right) = 0.18\%$. This way, the insurer can compare the impact of $\hat{c}$, an estimated constant of proportional adjustment, with the parameter estimates obtained from the GLM.

In Figure 3.6, we compare for each value of X_2, the following types of premiums: (1) a premium for males, given by $\hat{\mathbb{E}}[Y|\boldsymbol{X}, D = 1]$, (2) a premium for females, given by $\hat{\mathbb{E}}[Y|\boldsymbol{X}, D = 0]$, (3) an unawareness premium, given by $\hat{\mathbb{E}}[Y|\boldsymbol{X}]$ and (4) the discrimination-free premium (labeled as DFP), given by $\hat{h}(\boldsymbol{X})e^{\hat{c}}$, after proportionally adjusting for the portfolio bias. This figure shows the predictive ability that the unawareness premiums can have on the discriminatory covariate, independently of the value that the other nondiscriminatory covariates can take, such as age in Figure 3.6. The numeric values are shown in Table 3.2.

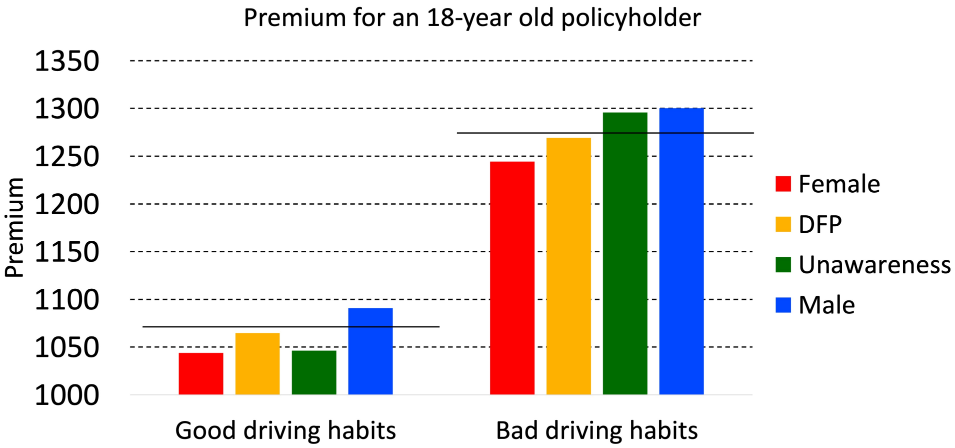


Figure 3.6: Estimated gender-specific, unawareness, and discrimination-free premiums for an 18-year old policyholder. The solid line represents the average between the lowest and highest premiums for each value of X_2.

In this scenario, because of the causal structure implied by $\mathcal{G}_B$, the discriminatory covariate should not have had a significant parameter estimate in M_1. In Figure 3.6 all the

Premium	$X_2 = 0$	$X_2 = 1$
Male	1,091	1,300
Female	1,044	1,244
Unawareness	1,046	1,296
DFP	1,065	1,269

Table 3.2: Estimated premiums for 18-year old policyholders by driving habits, based on gender-specific, unawareness, and discrimination-free pricing formulas.

premiums on the left-hand side should be equal, and all the premiums on the right-hand side should be equal. It is our attempt to differentiate policyholders by good and bad driving habits through a proxy variable that created a spurious direct relationship between D and Y. Note that the proxy variable had, on average, a 1.4% misclassification rate with respect to X_2. The correct premium, in this case, is the unawareness premium. If we use the discrimination-free premium, then part of the price that should be borne by high-risk policyholders – those with undesirable driving habits – would be unfairly transferred to the low-risk group, for each age category. Therefore, in this example, ignoring gender completely is better than including it in the discrimination-free pricing formula. Identifying whether gender is truly a risk factor, or simply an ancestor of the loss variable, could improve non-discrimination pricing accuracy. Hence, this is an insurance pricing example where, if the social context allows it, we suggest the discrimination-free premium should not be applied.

We suggest that the discrimination-free premium should not be applied, regardless of the fact that driving habits are a good predictor of gender, or in other words, where the conditional probability $\mathbb{P}(D = d|X_2 = x_2)$, for some values of d and x_2, is sufficiently high, so that the unawareness premium $\mathbb{E}[Y|\boldsymbol{X}]$ is highly correlated with gender. This suggestion is consistent with the EU regulation explored further in Section 3.6. Because of our causal assumptions implied by $\mathcal{G}_B$, there are no causal justifications in this insurance product to use the discrimination-free premium. However, causal arguments may in some cases be superseded by social arguments. We illustrate such a case in the following section.

3.5.3 Example C: Automobile theft coverage

There are social contexts where, regardless of the causal structure underlying the insurance product, we believe there is a case for using the discrimination-free premium. To illustrate, we again consider auto insurance, but instead of TPL, we look at costs arising from theft or vandalism of the policyholder's vehicle. The variables in the set $\boldsymbol{V}$ are: losses from theft and vandalism of the policyholder's vehicle (Y), race (D), age (X_1) and zip code (X_2). Age and zip code comprise the set of nondiscriminatory covariates, race is a discriminatory covariate. We assume the DAG, represented by $\mathcal{G}_C$, in Figure 3.7, which is identical to $\mathcal{G}_B$ in Figure 3.5.

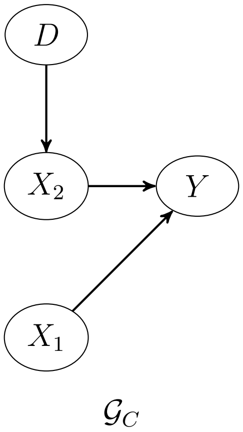


Figure 3.7: DAG, denoted by $\mathcal{G}_C$, for a personal automobile insurance example showing the causal structure over automobile theft and vandalism costs (Y), race (D), age (X_1) and zip code (X_2).

The explanation of the DAG is that certain zip codes have a higher automobile theft rate, resulting in the directed edge $X_2 \to Y$, and people of different ages may be more cautious about the security of their car which creates a causal effect between age and losses. We also assume the directed edge $D \to X_2$, which reflects the different racial profile of different zip codes.

To illustrate a social context, we use data from Milwaukee, Wisconsin. Zip codes are commonly used as rating factors in auto insurance. As shown in Section 1.1 for the city of Milwaukee, zip codes are strong predictors of race because 61% of the Black population in the state lives in a cluster of 14 zip codes out of the 775 in Wisconsin (U.S. Census Bureau, 2018). This racial distribution of the population by zip codes for the city of Milwaukee can be seen in Figure 3.8. Additionally, keeping all other rating factors constant, those zip codes which have a majority of Black population have higher premiums than the neighboring zip codes, and than other zip codes where the majority of the population is White. An example of this pricing practice retrieved from *The General Insurance* for one U.S. insurer can be seen in Table 3.3.

The DAG $\mathcal{G}_C$, as in example B, satisfies the conditional independence condition (3.6) which means that once our causal assumptions in the pricing model consider age and zip

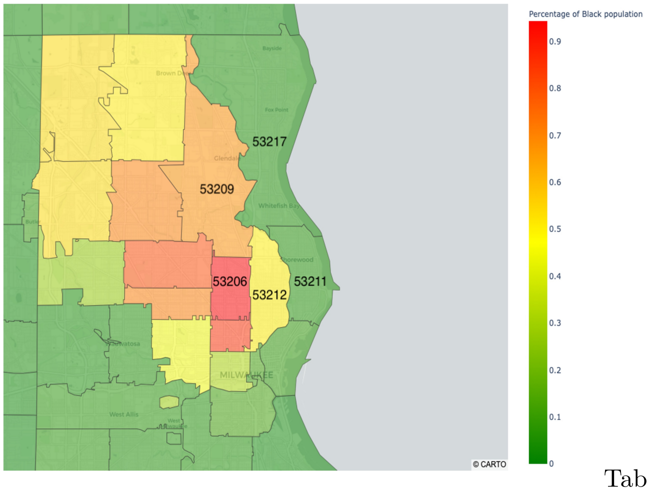


Figure 3.8: Percentage of Black population by zip code in the city of Milwaukee, Wisconsin. Data from the 2018 American Community Survey. A full interactive map of Wisconsin can be found in WI.

Zip code	Premium
53211	1,035
53217	1,062
53209	1,141
53212	1,286
53206	1,286

Table 3.3: Annual premiums from *The General Insurance* for a 2017 Honda Accord EX 4Dr and MT owned by a single 27 year old male, with no previous insurance or accidents, good credit and renting his residence. Retrieved on 04-03-2021.

code, race is independent of the automobile theft prices. We saw in the previous example, using the same DAG, that many plausible scenarios can lead to $\hat{\mathbb{E}}[Y|\boldsymbol{X}, D] \neq \hat{\mathbb{E}}[Y|\boldsymbol{X}]$. In fact, we can repurpose the simulation from the previous example, since both examples have equivalent causal structures. For simplicity, we assume the portfolio only includes policyholders with two possible race values and two zip codes. We represent a Black policyholder by $D = 1$, a White policyholder by $D = 0$, zip code 53206 by $X_2 = 1$, and zip code 53211 by $X_2 = 0$ (both zip codes are included in the pricing practice shown in Table 3.3). With these new labels, the number of policyholders n, the conditional and marginal probabilities used for the selection of parameters in example B is consistent with the population data of zip codes 53206 and 53211 in Milwaukee from the 2018 American Community Survey (U.S. Census Bureau, 2018). The average 1.4% measurement error of X_2 in this example can be thought of as a small minority of policyholders that park their car in a zip code different than the address used in their policy application. Using these labels for the covariates, we can compare again in Figure 3.9 and Table 3.4 the premiums for the automobile theft coverage example.

In Table 3.4, the unawareness premium $\hat{\mathbb{E}}[Y|\boldsymbol{X}]$ is intentionally set to approximately replicate the premiums in Table 3.3. To be consistent with this pricing practice, all other things equal, premiums are higher for zip code 53206 than for 53211. Similarly to the previous example, the unawareness premium is highly correlated with race, due to the indirect path $D \to X_2 \to Y$. The discrimination-free premium (after proportionally

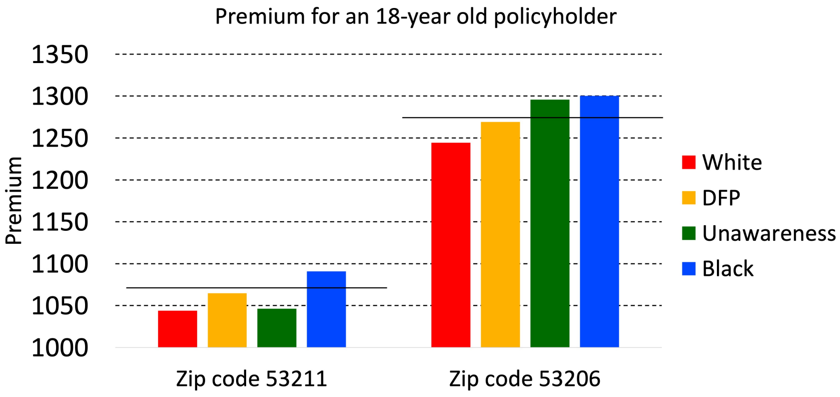


Figure 3.9: Estimated race-specific, unawareness, and discrimination-free premiums for an 18-year old policyholder. The solid line represents the average between the lowest and highest premiums for each value of X_2.

Premium	53211	53206
Black	1,091	1,300
White	1,044	1,244
Unawareness	1,046	1,296
DFP	1,065	1,269

Table 3.4: Estimated premiums for an 18-year old policyholder by zip code.

adjusting for the portfolio bias) provides a financial relief, compared with the unawareness premium, for policyholders living in zip code 53206 – which is predominantly Black since $\mathbb{P}(D = 1|X_2 = 1) = 0.94$. This results holds in this framework because $\hat{\beta}_3 > 0$ in equation (3.7), as demonstrated below. Let $\boldsymbol{X}' = (X_1, X_2 = 1)$, then

$$\mathbb{E}[Y|\boldsymbol{X}'] > h(\boldsymbol{X}'),$$

$$\text{that is} \qquad \sum_{d \in D} \mathbb{P}(D = d|X_2 = 1)\mu(\boldsymbol{X}', d) > \sum_{d \in D} \mathbb{P}(D = d)\mu(\boldsymbol{X}', d),$$

$$\Rightarrow 0.94\mu(\boldsymbol{X}', D = 1) + 0.06\mu(\boldsymbol{X}', D = 0) > 0.40\mu(\boldsymbol{X}', D = 1) + 0.60\mu(\boldsymbol{X}', D = 0),$$

$$\Rightarrow \qquad \mu(\boldsymbol{X}', D = 1) > \mu(\boldsymbol{X}', D = 0),$$

$$\Rightarrow \qquad \beta_3 > 0.$$

Despite the causal structure, in this example there is a case for using the discrimination-free premium, as a social policy issue. The application of the discrimination-free premium would, on average, favor minority racial groups which have suffered historically through redlining; a social problem discussed in Section 3.6.2 (also see Chapter 1). As mentioned in Lindholm et al. (2022a), insurance can engineer socially beneficial outcomes by preventing further penalization of groups that have suffered historical injustices.

We remark that we have made opposing conclusions with respect to the application of the discrimination-free premium in Examples B and C, in spite of the fact that they are mathematically identical. The causal analysis allows us to determine the most appropriate premium in the absence of social fairness or legal anti-discrimination requirements. But the actual premium charged should be determined taking the social and legal context into consideration. In this example, we have shown that, although the discrimination-free premium is not indicated from a causal perspective, it may be the most suitable from a social and/or legal perspective because of well-documented, ongoing impacts of historical policies restricting the access to housing and insurance of group minorities. In other contexts, the inequity that might trump a purely causal approach might be based on other discriminatory factors, for example, gender, age, or socio-economic status.

3.5.4 Criteria for applying the discrimination-free premium

The three insurance pricing examples presented above suggest a general guide to create criteria that can aid practitioners and researchers decide if, under a causal framework, the discrimination-free premium should be used. We argue that:

(1) When the discrimination-free premium results in the fulfillment of a solidarity mechanism or it benefits a historically disadvantaged group, then it should be used. This approach is motivated by Rawls' Difference Principle (Rawls, 2001), which states unequal treatment is appropriate if and only if it benefits the least advantaged members in society. (Example C)

(2) Otherwise, when the discriminatory covariate is not deemed to be a risk factor of the loss random variable, an approach which ignores the discriminatory covariate, is preferable to one that takes it into consideration. (Example B)

(3) If there is reason to assign a causal connection between the discriminatory covariate and the loss random variable, then using the discrimination-free pricing formula is appropriate, provided it does not perpetuate or exacerbate social disadvantage. (Example A)

To apply these criteria, we need to identify potential rating factors, and separate the risk factors from the non-risk factors. A focus on risk factors, rather than on a comprehensive set of rating factors, within a DAG, should allow a more accurate allocation of the expected loss.[4] Finally, the social and legal context should be considered in order to determine if there is a case for overriding the allocation based on the DAG.

[4] Frees and Huang (2021) point out that policyholders are more likely to accept the use of rating factors that are understood to have a causal association with the loss.

3.6 Social Context

The criteria listed in the previous section point to the relevance of the social context in determining whether the discrimination-free premium should be used. In this section, we discuss in more detail how social policy and social context impacts pricing.

3.6.1 EU gender-neutral pricing

The construction of the discrimination-free premium in Lindholm et al. (2022a) is influenced by the EU legal requirement (European Commission, 2012), which, since 2012, has banned the use of gender as a rating factor in insurance. The regulation states:

> *The use of risk factors which might be correlated with gender therefore remains possible, as long as they are true risk factors in their own right. For example, price differentiation based on the size of a car engine in the field of motor insurance should remain possible, even if statistically men drive cars with more powerful engines. This would not be the case for differentiation based on the size or weight of a person in relation to motor insurance.*

The legislation uses the term "true risk factor" in an implicitly causal way, consistent with our Definition 3.3. From the assumptions implied in this quote, we construct a DAG, denoted by $\mathcal{G}_6$ in Figure 3.10. We denote the set of variables in this example by $\boldsymbol{V} = \{Y, D, X_1, X_2\}$, where D is gender, X_1 is the size of a car engine, X_2 is the size or weight of the policyholder and Y are motor insurance losses. In this case, the set of nondiscriminatory covariates is $\boldsymbol{X} = (X_1, X_2)^\top$ and the discriminatory covariate is gender.

In $\mathcal{G}_6$, although the size or weight of the insured (X_2) might be associated with gender (D), it is not a risk factor for a driving accident. Using the size or weight of the insured (X_2) as a rating factor, regardless of its predictive power, would be prohibited on the grounds of indirect discrimination, since it can result in a different treatment of men or women with no causal justification. However, the use of engine size (X_1) is justified, in spite of the predictive power that it could have over the discriminatory covariate.

The discrimination-free premium complies with this example, and also avoids indirect discrimination. However, as in Example B, in this case gender is not a risk factor, so the correct premium would be the unawareness premium, which may (spuriously) differ from the discrimination-free premium.

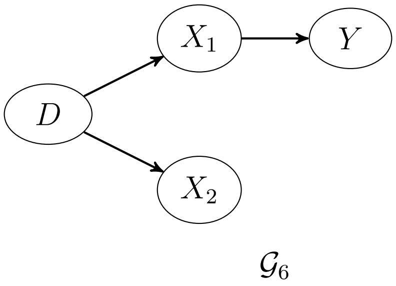


Figure 3.10: The DAG, denoted by $\mathcal{G}_6$, we believe represents the assumptions implied in the motor insurance pricing example in European Commission (2012). This example shows the causal structure over motor insurance losses (Y), gender (D), size of a car engine (X_1) and size or weight of the policyholder (X_2).

3.6.2 Redlining

In the previous section we proposed that historical disadvantage could outweigh contemporary causal arguments in determining the appropriate approach to the discrimination-free premium. In this section, we expand on the reasoning behind that proposal.

We first described redlining in Chapter 1. It originated as a business practice where residents of certain neighborhoods were discriminated against in availability and costs of loans and mortgage insurance, compared with residents of bordering areas. Redlining denied millions of African Americans the opportunity to purchase homes, to improve their neighbourhoods, or to move to more affluent areas. Homeowner's insurance rates were higher than similar houses outside the redlined areas, so that those who did manage to buy their homes were often unable to maintain them. Despite being banned more than 50 years ago, the impacts of redlining are still apparent today, in the demographics of the formerly redlined areas, and in national housing wealth disparities by race. Furthermore, as shown in Section 1.1 redlining in insurance is still an ongoing issue.

The study of redlining shows that racism in insurance, explicitly or implicitly, has had a pernicious influence on racial disparity in insurance pricing. Given the common insurance practice of using zip code as a rating factor, the discrimination-free premium in this social context would benefit a group that has been historically disadvantaged by insurers. Auto insurance is required by law; society (through its lawmakers) has created a demand for a product that is supplied by private companies. This means that the companies that participate in auto insurance have an implicit social contract – a responsibility to consider the needs of society as well as the wealth of their shareholders.

3.7 Challenges

The examples described in this chapter are highly simplified. In this section we discuss some of the challenges on implementing causal inference for insurance pricing in practice.

3.7.1 Variable selection

Among our objectives in this chapter is to emphasize the importance of sound rating factor selection. At present, regardless of the application, variable selection depends very significantly on human judgment (Spirtes et al., 2000). Further, variable aggregation and discretization (a common practice in insurance to avoid sparsity problems) can effect the reliability of causal inference (Spirtes et al., 2000). For example, two random variables A and B, both caused by C, that are statistically independent conditional on C, may be statistically dependent conditional on a (poor) proxy $C^{(p)}$ of C.

3.7.2 Estimating a DAG from observational data

Lindholm et al. (2022a) point out that one of the problems of assessing a graphical model, such as a DAG, is that actuarial pricing models are typically high dimensional. Table 3.5 lists the number of variables requested by insurers in an online application form for a personal automobile insurance policy in the state of Wisconsin, for five of the ten largest insurers by market share (Wisconsin OCI, 2020). Insurers are presented in decreasing order of market share, as measured by premiums written in 2020 for the state of Wisconsin. The numbers are approximate for several reasons. For example, (i) the insurer may use covariates beyond the direct information proffered, such as meta-data from the online application process; (ii) the insurer may collect information that is not used in the premium calculation; (iii) the insurer can request additional information to provide a more accurate quote after the online form is submitted; (iv) the insurer can engineer additional covariates by using interactions among the information requested. Nevertheless, taking only the online application variables into consideration, we find around 20 covariates involved, on average.

With 20 covariates and, typically, a large data sample, it is plausible to estimate a DAG through existing algorithms. A popular algorithm for estimating high dimensional and sparse DAGs (a DAG is sparse if each node is not connected through a directed edge to a large number of other nodes) is the PC algorithm, which may be implemented in `R` through the `pcalg` package (Kalisch et al., 2021). In a simulation study, Kalisch and Bühlmann

Insurer	Number of covariates
State Farm Group	25
The General Insurance	14
Progressive	24
GEICO	23
Erie Insurance	15
Average	20

Table 3.5: Number of variables requested by insurer in an online application form for a personal automobile insurance policy in the state of Wisconsin. Retrieved on 21-01-22.

(2007) show that the PC algorithm is asymptotically consistent and that the estimation process takes a reasonable amount of time (less than an hour even when the number of covariates is 1000).

The PC algorithm starts from a complete undirected graph (a DAG where each node is connected to all other nodes through undirected edges), tests an increasing order of possible conditional independence relations and updates the DAG, accordingly. The DAG can be constrained using background knowledge, for example, if prior belief requires one variable to be connected with another through a directed edge, then that causal link is imposed – and assumed – in the algorithm (Spirtes et al., 2000). Further, since insurance problems commonly involve binary, discrete and continuous data, a feasible modification of the PC algorithm can be implemented as shown in Cui et al. (2016).

Algorithms, such as the PC algorithm, that uncover causal structures from observational data implicitly impose a restriction on the probability distribution called stability (Pearl, 2009a), also called faithfulness in Spirtes et al. (2000). This assumption is the converse of the parental Markov condition presented in Section 3.3 and allows inference of causal graphs by the observed independencies in distribution.[5]

3.7.3 Availability of the Discriminatory Covariate

The application of the discrimination-free premium can face another obstacle in certain scenarios due to legal constraints. In the EU, the insurer is allowed to collect and use gender in the pricing process, provided it does not result in price or benefit differences based purely

[5] A Markovian model coupled with a stable distribution with respect to a DAG entail the minimality condition. See Peters et al. (2017) for a proof.

on gender (European Commission, 2012). However, where the discriminatory covariate is race, it is likely that the insurer would not have necessary information to implement the discrimination-free premium, which relies on knowing the discriminatory covariates, $\boldsymbol{D}$, in order to eliminate the impact of indirect discrimination.

The challenge of obtaining racial and ethnic information when the data cannot be collected directly, has been explored extensively in literature related to credit lending and credit scores. For example, Zhang (2018) explores this additional layer of complexity in fair lending of non-mortgage credit products using a proxy of race and ethnicity based on the borrower's surname and location. This results in a proxy of the discriminatory covariates, denoted by $\boldsymbol{D}^{(p)}$, which could then be used in the computation of the discrimination-free premium. The effectiveness of this method depends on the accuracy of the proxy; a poor proxy may create more problems than it solves.

Another alternative when only partial information on the discriminatory covariates is available, is explored in Lindholm et al. (2022b). Their recommendation is to compute $\mu(\boldsymbol{X}, \boldsymbol{d})$ through a multi-task feed-forward neural network that simultaneously predicts $\boldsymbol{D}$ for policyholders with the missing covariate, through the nondiscriminatory covariates $\boldsymbol{X}$. This scenario is explored in a simulation study where the discriminatory covariate is missing at random and not completely at random. The simulation study shows that the algorithm outperforms the case where the premium is fitted solely through the subset of policyholders with complete information. The authors suggest that partial information could be gathered by offering discounts to policyholders who are willing to share information on their protected characteristics.

3.8 Conclusion

In this chapter, we use causal inference to establish an insurance pricing model through directed acyclic graphs. Using a causal framework, we contrast the discrimination-free premium in Lindholm et al. (2022a) with an expected value that relies on the adjustment formula. We make use of the causal framework to more formally define risk factors, and we apply actuarial and causal analysis to three different examples.

From these examples we extrapolate more general criteria to determine whether the discrimination-free premium should be applied, based on causal and social impact arguments. The prevailing social context of gender-neutral pricing regulation in the EU is very different from that of redlining in the United States. We note that our criteria could be applied to other discriminatory characteristics, and to other social contexts. To conclude, we briefly mention some issues and suggestions for applying directed acyclic graphs and the discrimination-free premium formula in practice.

The calculation of premium rates cannot be considered as a purely scientific endeavor. It is influenced by and influences social policy and practice. The role of the actuary is to understand and integrate, as far as possible, the science of estimating future losses with the social principles sanctioned by the communities in which we live. The adoption of causal frameworks for modeling insurance losses allows actuaries to focus on the underlying assumptions of our models, re-assess their soundness in the scientific context (conscious that the assumption of causal relationships may be influenced by personal biases), and re-evaluate the results in the social context, to ensure that the premium rating process does not promote or exacerbate social inequality.

Chapter 4

Discrimination in Insurance Pricing: A Microsimulation

In this chapter we describe a microsimulation model which can generate a simulated population of the United States. It is designed to match in aggregate selected characteristics of the target population. We focus on a 2020 pseudo-population from Wisconsin, which we use to explore personal automobile insurance premium ratings. We contrast four pricing models, in terms of prediction accuracy, and in terms of their discriminatory impact over race, using four different definitions of discrimination proposed in the actuarial and machine learning literature. By adapting definitions for disparate impact and proxy discrimination to a statistical test we show that the traditional assumption of independence between frequency and severity cannot only result in reduced prediction performance, but can also be detrimental to racial minorities.

4.1 Introduction

In this chapter we make two contributions to the objective of this thesis. First, we circumvent the lack of public insurance data by generating a pseudo-population attributed with commonly used insurance rating factors and feasible insurance losses that are calibrated with data from the United States. An analysis of discrimination in insurance pricing requires modeling both severity and frequency. Although Chapter 2 provides a source for the largest national and publicly available database of MVTAs in the United States, its analysis is limited in giving insights on accident severity. Through the microsimulation, we incorporate a frequency component that can be traced for each individual of an even larger population.

Second, we expand the comprehension of discrimination in insurance pricing by exploring a possible source of discrimination. It is traditional in the actuarial science literature to assume independence of the frequency and severity components (Frees, 2014; Bolancé and Vernic, 2020; Oh et al., 2021). We show that this independence modeling assumption can have unintended consequences with respect to some discrimination metrics. Specifically, we determine if there is direct or indirect discrimination as defined in Lindholm et al. (2022a), if there is disparate impact as defined in Xin and Huang (2022) and if there is proxy discrimination as defined in Kilbertus et al. (2017) (these discrimination definitions are presented in Chapter 1). We also tackle the challenge of statistical identification that is carried by these discrimination definitions. A numerical illustration of these methods is presented using insurance data from our pseudo-population.

We simulate individual level data in order to have (1) full knowledge of our population, (2) total control over frequency and severity in personal automobile insurance and (3) the ability to evaluate the discriminatory impact of incorrect modeling assumptions. Microsimulation is a tool to answer 'what-if' questions (Li and O'Donoghue, 2013), which is exactly our objective with respect to discrimination.

Our microsimulation is based on a large sample of a population of individuals with a number of attributes, such as gender, age, marital status, education and driving record. The methodology we implement here is a static and non-longitudinal microsimulation-that is, no spatial or time transition probabilities are considered, nor are interactions between individuals. Our purpose is to analyze the immediate and short-term impact of changes in pricing policies. Our first step is to obtain data from a representative sample of the target population with selected properties. Then, a pseudo-population of individuals is simulated and calibrated under the key features of the target population (Gilbert and Troitzsch, 2005). Next we extrapolate the selected characteristics in order to obtain data on automobile claims.

The methodology under which microsimulations are calibrated requires a sequential assignment of attributes through a set of stochastic or deterministic rules defined by the modeller. Arnold et al. (2019) points out that, conceptualized this way, there are parallels between the microsimulation methodology and the data generating process represented by a DAG, as described in Chapter 3. In the structural causal model of a DAG (see Section 3.3.1), if one knows the values of the exogenous variables, then all endogenous variables can be known. Due to this conceptual link, we describe our microsimulation design through a DAG and use notation from causal inference to denote variables.

The chapter is organized as follows. In Section 4.2, we describe the base dataset that has the key features of the target population and the causal model that generated the pseudo-population. In Section 4.3, we introduce the frequency-severity relationship governing the distribution of the insurance losses. In Section 4.4, we show with an empirical example how a modeling assumption can lead to unintended discriminatory consequences. We summarize our conclusions in Section 4.5.

4.2 Dataset and Causal Model

4.2.1 Base dataset

Our target population for the microsimulation is the set of drivers in Wisconsin. Information used to calibrate the microsimulation comes from the 2020 American Community Survey (ACS), which was collected by the United States Census Bureau (USCB). The 2020 ACS contains a representative sample of 2.87 million housing units[1] out of the over 138 million occupied housing units in the United States (U.S. Census Bureau, 2021b). It is important to note that people who receive the survey are required by law to complete it. In order to protect the privacy of U.S. residents, data is aggregated by the USCB at the Zip Code Tabulation Areas (ZCTAs) (see definition in Appendix B) level for each of the 50 states, the District of Columbia, and Puerto Rico.

We highlight two of the important ways in which the ACS differs from the U.S. Census (U.S. Census Bureau, 2020). First, the census is conducted every ten years, whereas the ACS is collected on a yearly basis. Second, the census provides an official count of the population for the purpose of congressional apportionment, while the ACS has questions regarding housing characteristics like education, income, transportation, among others that are essential information for an insurance application.

Using the 2020 ACS data, we simulate a pseudo-population calibrated with some key features matching those in the state of Wisconsin. Compared to the national analysis in Chapter 2, in this chapter we focus only on one state since each state has different legal requirements established by that state's insurance commissioner and also, demographic characteristics between states differ substantially. Thus, we have chosen to analyze the simulated data of Wisconsin individually. The assumed causal structure, as defined in Section 3.3, of the microsimulation is described in the following Section 4.2.2 and the corresponding algorithm for the variable assignment to our pseudo-population is explained sequentially in Appendix D.

[1] The numbers of addresses in the 2020 sample is comparably lower than other recent years as a result of the COVID-19 pandemic (U.S. Census Bureau, 2021d).

4.2.2 Causal model

In this section we describe the attributes that were assigned to the pseudo-population and the reason for their selection. We also present the assumptions on the connections between these characteristics. The attributes and connections are used to define the causal structure of the microsimulation. The set of functions, $\boldsymbol{F}$, that generate the data, and the distribution of the exogenous variables, $\boldsymbol{U}$, are described in Appendix D.

The first step to define a causal structure is to select the nodes that compose the endogenous set, denoted by $\boldsymbol{V}$. The endogenous set was chosen to contain 16 characteristics that are endowed to the pseudo-population: race, age, gender, marital status, if the person had liability insurance in the last year (LI1Y), education level, risk aversion, driving ability, driving record, income level, a car make-model-year combination (MMY), zip code, mileage, type of driving, frequency and severity. These variables can be grouped into 5 different categories, based on the reasons we chose to include them in $\boldsymbol{V}$, as explained in Table 4.1.

#	Variable names	Reason for selection
9	Age, gender, marital status, LI1Y, education level, driving record, MMY, zip code and mileage	Most common variables requested in the online application forms of the 5 insurers in Table 3.5 for the state of Wisconsin.
1	Income level	This variable is not requested directly for auto insurance but two proxies of income are often used such as credit score and if the driver owns his/her home.
3	Risk aversion, driving ability and type of driving	Latent variables that are of interest to an insurer. Influenced by the empirical results on the severity distribution in Chapter 2.
1	Race	Discriminatory covariate
2	Frequency and severity	Response variables

Table 4.1: The 16 selected characteristics for the pseudo-population of Wisconsin of 2020 grouped by the reason for their selection.

The Faar Isaac Corporation (FICO) estimates that 95% of auto insurers use credit-based insurance in the states where it is legally allowed as a rating factor (NAIC, 2022). This common practice is due to the fact that credit history can be used effectively to predict insurance losses (American Academy of Actuaries, 2002), in particular, credit scores are statistically linked to the filing (or not) of a claim and the size of the claim (Golden et al., 2016). We use income level as a selected characteristic in Table 4.1, instead of the credit score, because we believe the DAG shown in Figure 4.1 explains the correlation between credit scores and insurance claims.

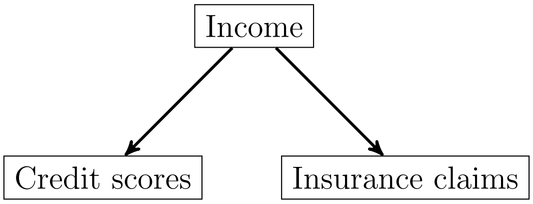


Figure 4.1: Hypothesized DAG that explains from a causal perspective the correlation between credit scores and insurance claims.

In the hypothesized causal structure in Figure 4.1, income is treated as a common causal effect of credit scores and insurance claims, but we assume that there are no causal effects between credit scores and insurance claims. This DAG presented in Kiviat (2019) originated from the comments of the former Arkansas Commissioner of Insurance that said people with low income cannot afford to absorb losses so they have to use their insurance, whereas people with higher income can afford to self-insure claims if they are in the range of \$400-\$700 (NAIC, 2009). Under the DAG in Figure 4.1 credit scores and insurance claims are independent conditional on income level, unconditionally they are potentially dependent. In consequence, the function in the SCM that assigns a value to insurance losses responds (either directly or indirectly) to income level and not to credit scores, and therefore, the latter is ignored in this analysis.

The three latent variables in Table 4.1 named ‘risk aversion’, ‘driving ability’ and ‘type of driving’ are in the endogenous set because we assign them values through specific functions in $\boldsymbol{F}$. Although their values are known to us within the simulation structure, in practice these variables are unobservable to the insurer. In spite of this, it is common practice to try and quantify them using proxies. For example, ‘risk aversion’ and ‘driving ability’ are conceptually similar to the variables called ‘aggressiveness’ and ‘inattention’ in Grari et al. (2022). The latent variable meant to capture the driving ability of the policyholder is called ‘driving experience’ in Ayuso et al. (2016), and the total time the driver has held a license is used as its proxy (a variable also requested by 4 out of the 5 insurers in Table 3.5).

Furthermore, these three latent variables are included in our generative model as a consequence of the observational results in Chapter 2. In Figure 2.10, we see that ‘speed related' and 'drugs/drinking' are among the most important covariates to predict MVTA severity. The problem with these two variables is that they are only known after the event. To include 'drugs/drinking' as an *ex-ante* variable, we use a similar rationale as for income in Figure 4.1, where we assume risk aversion is a common causal effect for insurance losses and driving under the influence or speeding. This approach has been previously presented in Golden et al. (2016) as the summary of risk-taking characteristics of sensation-seeking personalities. The 'speed-related' variable is included through 'type of driving' since we assume that policyholders who drive more frequently on roads with higher speed limits, tend to drive faster and therefore, have a greater probability of being involved in a speeding-related accident. As stated in the definition of the variable 'speed-related' in Appendix A, this variable conveys that the crash had any of the following: the speed was greater than reasonable (not necessarily over the speed limit), driving too fast for conditions, speed was above the speed limit or any of the drivers involved was racing. See Appendix D.14 for more details on 'type of driving'.

The set of possible acyclic graphs with 16 nodes has a size of 8.3×10^{46}, following Robinson (1977). Although not all DAGs are plausible, the subset of feasible DAGs that could have generated a pseudo-population calibrated with key features from the 2020 ACS is still large. In Figure 4.2 we show one example of a DAG we think is feasible out of many other choices that are possible and could have been analyzed. We label this DAG by $\mathcal{G}_m$, with subscript m for microsimulation, and the 16 nodes considered. The node for an insurance loss is deterministic since it is fully determined once we know the frequency and severity, which are explained in more detail in Section 4.3. The DAG $\mathcal{G}_m$ is Markovian by construction, since in addition to its acyclic structure, the variables $\boldsymbol{U}$ are jointly independent. See Appendix D for more details.

Considerations for $\mathcal{G}_m$

The following are considerations that influenced our selection of $\mathcal{G}_m$ out of the many other possible choices:

- **Insurance practice**. As pointed out in Table 4.1, in our selection of variables we consider the most common covariates that some P&C insurers ask in their online application form that could generate a feasible scenario of accident frequency and severity, described in detail in Section 4.3. We also consider some latent variables such as 'risk aversion' and 'driving ability' that are commonly used in the actuarial literature.

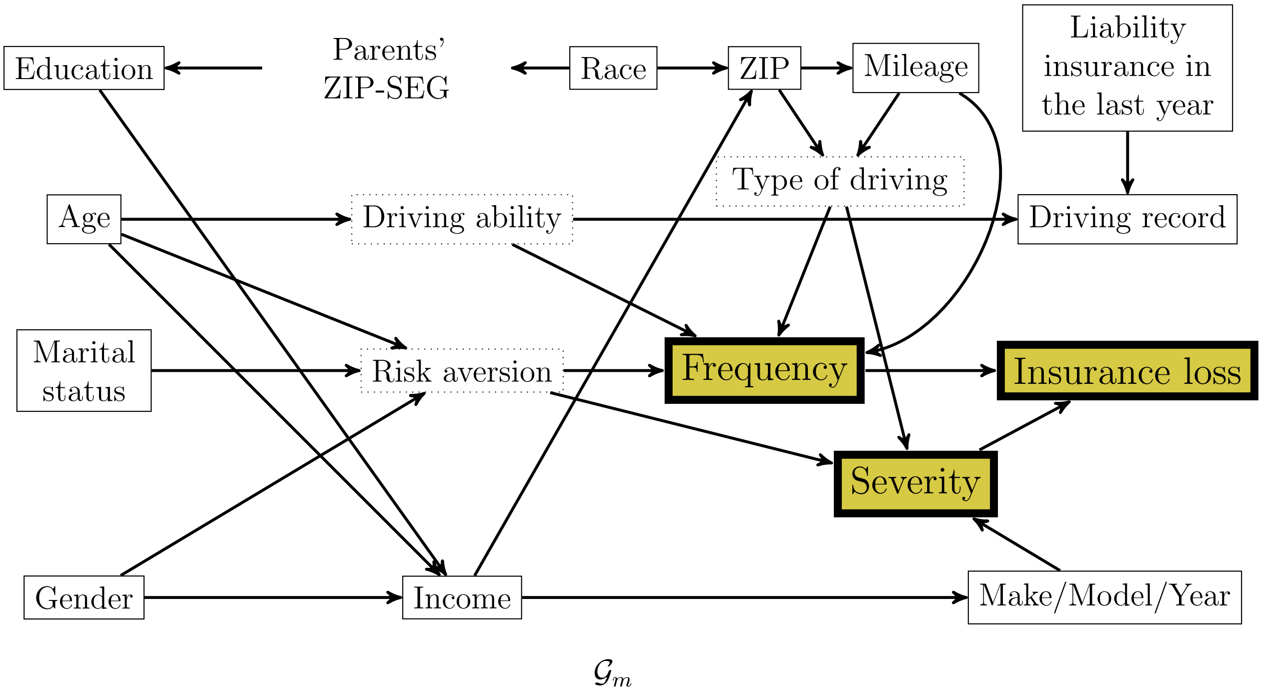


Figure 4.2: DAG, labeled $\mathcal{G}_m$, used to generate 16 selected individual characteristics for a pseudo-population of Wisconsin for 2020. Latent variables are drawn inside a dotted rectangle. Response variables have a higher line width and are color-filled.

- **Data limitations**. Once we selected the variables that are in $\boldsymbol{V}$ in accordance with insurance practice, we had to assign directed connections between them. Unfortunately, the set $\boldsymbol{V}$ cannot be assigned simultaneously to our pseudo-population since there exists no data source that describes their multivariate joint distribution. This means that variables need to be assigned one at a time in a sequential process, in line with microsimulation methodology. Additionally, given that we desire the pseudo-population to have key features that match those of the population of Wisconsin for 2020, we are constrained by the available marginal distribution of variables in the ACS, as described in Section 4.2.1.

- **Judgment**. Our judgment cannot be excluded from the modeling process. For example, we assume race has no direct causal effect over driving ability or on any other variable that directly impacts the frequency or severity of an automobile accident. Furthermore, we omitted variables or connections that could have had an important impact. For example, we did not include a variable for weather or driving conditions, which could have an effect on the frequency of crashes, in spite of not showing a significant impact on the severity distribution in Chapter 2. Another example of an omission is a connection between race and marital status, where if a race shows a higher probability of being married, then they would have had an indirect discount in their premium (through a higher probability of being risk averse).

4.3 Frequency and Severity

In this section we describe the assumptions behind our pricing model used to generate a distribution of insurance losses for the pseudo-population of Wisconsin in 2020. The definition of an insurance risk factor in Chapter 3 is used here, relative to the causal structure described by $\mathcal{G}_m$. We use a GLM to describe the underlying relationship between the risk factors and the response variables. Frequency and severity are calibrated to satisfy two requirements:

1. Parameters are adjusted such that the premium distribution of the pseudo-population is in a similar range to the automobile pricing data of Wisconsin in Chapter 1.

2. The balance between accident frequency and severity roughly approximates the observational automobile accident data of the United States.

To elaborate on the second requirement, in particular, we assume crashes in urban areas are frequent but not typically severe, while on average, crashes in rural areas are less frequent but more severe. Empirical evidence for this assumption is presented in Miller et al. (1991, p. 83) and more recently in Blincoe et al. (2010, p. 204). The latter states that

> *Urban roadway environments are characterized by high population densities. This typically produces higher traffic volumes and, on average, lower average travel speeds than are found in more rural areas. These conditions affect crash impacts in a variety of ways. Slower travel speeds can reduce the severity of crashes when they occur, but higher traffic volume creates more opportunities for exposure to distracted or alcohol impaired drivers, as well as more complex driving interactions in general [...]. By contrast, the higher speeds typically encountered on less congested rural roadways can lead to more serious injury outcomes in the event of a crash.*

Three key features are taken into consideration when calibrating our model with the empirical evidence in Miller et al. (1991) and Blincoe et al. (2010). These are: (1) the ratio of rural accidents to the total number of accidents, (2) the ratio of aggregate rural severity to aggregate severity, and (3) the rural to urban average severity ratio. A comparison of these key features at the national level with the observed features for our pseudo-population of Wisconsin can be seen in Table 4.2.

	Key feature	Miller et al. (1991)	Blincoe et al. (2010)	Wisconsin pseudo-population of 2020
(1)	$\frac{\text{Rural frequency}}{\text{Total frequency}}$	22%	27% in the U.S., 38% in WI	40%
(2)	$\frac{\text{Aggregate rural severity}}{\text{Aggregate severity}}$	43%	38%	52%
(3)	$\frac{\text{Average rural severity}}{\text{Average urban severity}}$	2.6	1.7	1.7

Table 4.2: Contrast of key features that describe the balance between frequency and severity for the United States in the literature with those generated by $\mathcal{G}_m$ for the pseudo-population of Wisconsin of 2020.

In Table 4.2, the pseudo-population is calibrated to ratio (1) obtained using only data from Wisconsin in Blincoe et al. (2010), which is 11% higher than the national average. To be consistent, since we are observing more crashes in rural areas we increase ratio (2) for the pseudo-population by an average of 11.5% compared to the two sources of key features, which correspond to ratios at the national level. Ratio (3) in the pseudo-population is the same as in Blincoe et al. (2010), which is the most recent source of public MVTA data in the United States.

Although the mathematical specifications for the insurance losses are introduced in subsequent sections, in Figure 4.3 we show the geographical distribution for a realization of insurance losses generated from our model by zip codes in Wisconsin that satisfy the two requirements for frequency and severity. On the left-hand side, the insurance losses per capita for the pseudo-population of Wisconsin reflect an approximation to the premium that should be charged to the policyholders by zip codes. By construction, these values are approximately in the same range as the real premiums in Figure 1.4. On the right-hand side, we observe the distribution of aggregate insurance losses. Zip codes that comprise (or neighbor) the highly populated cities of Milwaukee, Madison and Green Bay have low insurance losses per capita but high aggregate losses, as compared to zip codes in the rest of the state (we refer the reader to Figure 1.2 to recall the location of the aforementioned cities). This is a consequence of calibrating the balance of frequency and severity to the key features in Table 4.2 under the constraints of $\mathcal{G}_m$.

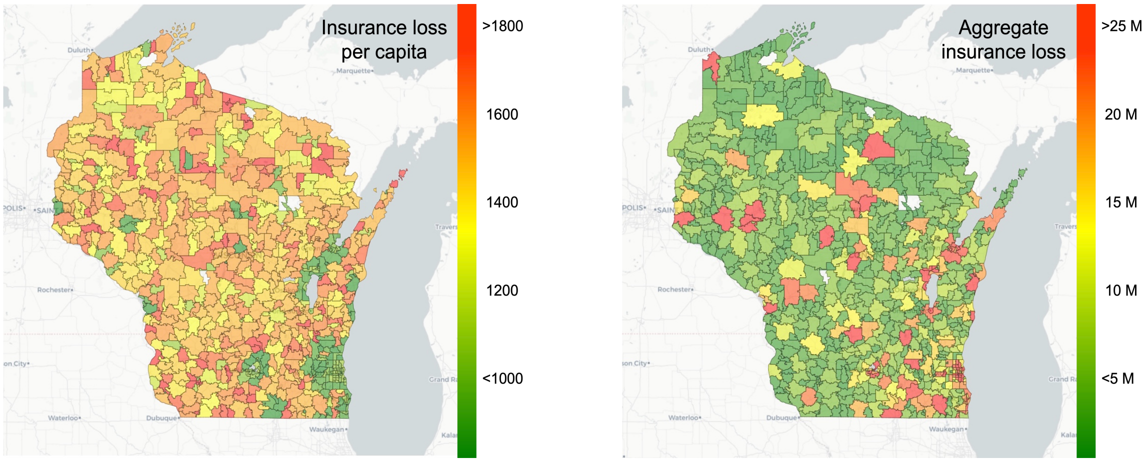


Figure 4.3: Distribution of insurance losses per capita (left-hand side) and in aggregate (right-hand side) by zip code for the state of Wisconsin.

4.3.1 Insurance loss

We use n to denote the number of policyholders in the insurance portfolio. The aggregate insurance loss for policyholder $i = 1, 2, \ldots, n$, denoted by S_i, is the sum of a random number of claims, N_i, of i.i.d. random variables $Y_{i,j}$, $j = 1, 2, \ldots$, where $Y_{i,j}$ denotes the individual loss amount associated with the j-th accident for policyholder i during the policy's duration. In the insurance industry, N_i is known as the frequency component and the common distribution of the $Y_{i,j}$'s as the severity. The insurance loss is

$$S_i = \begin{cases} Y_{i,1} + Y_{i,2} + \ldots + Y_{i,N_i} & N_i > 0, \\ 0 & N_i = 0. \end{cases} \tag{4.1}$$

The relationship between frequency and severity differs from traditional aggregate risk models in insurance due to our choice of underlying DAG for the model. For example, in the widely used collective risk model, frequency and severity are assumed to be independent (Klugman et al., 2019). However in our model, independence statements can only be made relative to the graph $\mathcal{G}_m$ in Figure 4.2. Here, frequency and severity are unconditionally potentially dependent, and they are independent conditional on at least six different sets of variables. From these, two noteworthy sets that (each on its own) guarantee their conditional independence are the frequency risk factors, $\mathbf{Pa}(N_i)$, and the severity risk factors, $\mathbf{Pa}(Y_{i,j})$. Moreover, as shown in subsequent sections, conditioning on these two sets has the additional advantage of unbiased frequency and severity estimates.

4.3.2 Frequency

We assume $N_i \sim \text{Poisson}(\lambda_i)$, where the conditional mean, λ_i, is a function of the risk factors of frequency; the function increases through the log-link the mean number of accidents by setting

$$\begin{aligned}\log(\lambda_i) = {} & \beta_0 + \begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix} \text{Mileage} + \begin{bmatrix}\beta_4\\ \beta_5\\ \beta_6\\ \beta_7\end{bmatrix} \text{Type of driving} + \begin{bmatrix}\beta_8\\ \beta_9\\ \beta_{10}\end{bmatrix} \text{Risk aversion} \\ & + \begin{bmatrix}\beta_{11}\\ \beta_{12}\\ \beta_{13}\end{bmatrix} \text{Driving ability,}\end{aligned} \tag{4.2}$$

where the risk factors are dummy variables. The parameter coefficients chosen and their relative impact are visualized in Figure 4.4. We assume higher mileage, risk-seeking behaviors and lower driving ability increase the expected frequency. Type of driving, described in Appendix D.14, plays two roles by increasing the expected mean in urban and suburban areas as compared to rural areas, but also by considering our assumption that roads with higher speed limits increase the expected frequency due to speeding-related accidents.

4.3.3 Severity

The assumptions for the severity model encoded in $\mathcal{G}_m$ are different from the assumptions in the process we used to generate its distribution. We now explain the differences, implications and our reasons for this approach. The differences are depicted in Figure 4.5, where on the left-hand side we have the DAG represented by $\mathcal{S}_1$, which is a subgraph of $\mathcal{G}_m$. This DAG shows our assumptions for the severity distribution which include its three risk factors,

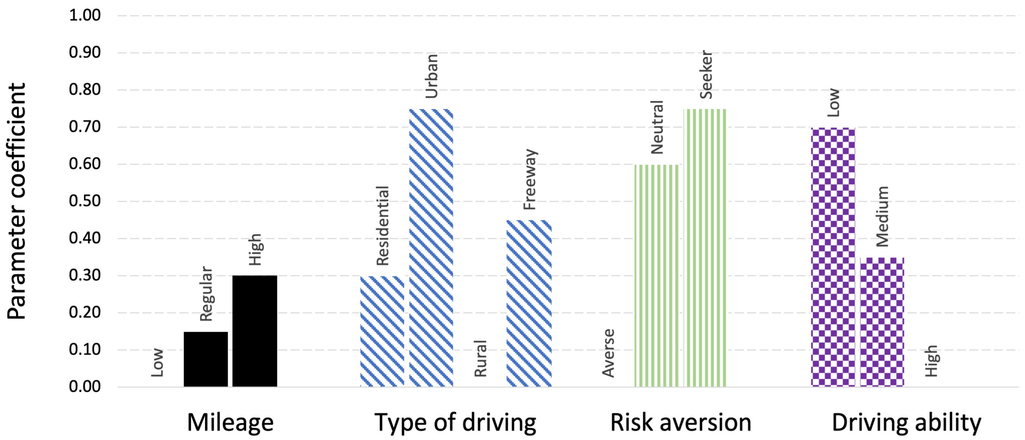


Figure 4.4: Relative impact of the parameter coefficients in the frequency mean equation (4.2). The base case is $\beta_0 = -3.2$ and parameters are shown in the same ordering as in the mean equation.

$\mathbf{Pa}(Y_{i,j})$, and the four independent noise terms in the set $\boldsymbol{U}$ affecting each of the random variables. In $\mathcal{S}_2$ (right-hand side) we have the DAG that represents how the data for severity is generated. There are three additional random variables (roles and notation explained below) as compared to $\mathcal{S}_1$ and each of them has its own (independent) noise term.

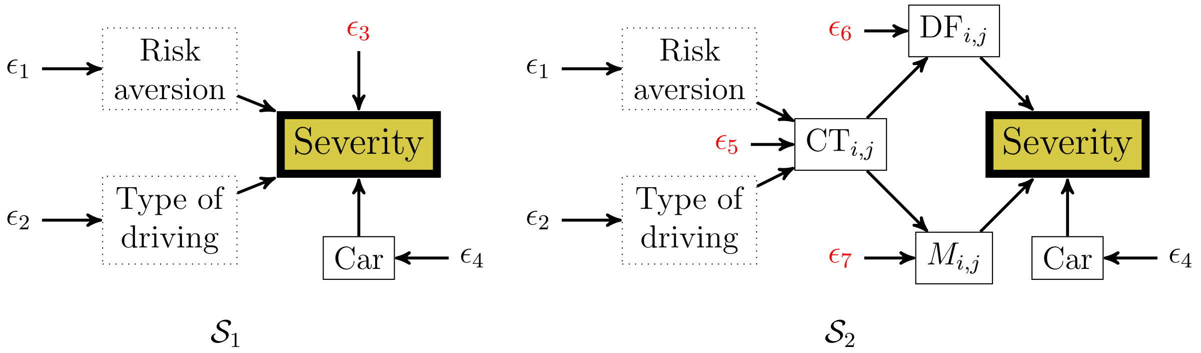


Figure 4.5: $\mathcal{S}_1$ is a subgraph from $\mathcal{G}_m$ that includes severity and its risk factors, along with their noise terms in $\boldsymbol{U}$. Subgraph $\mathcal{S}_2$ shows the process used in the severity simulation.

Endogenous variables in both $\mathcal{S}_1$ and $\mathcal{S}_2$ in Figure 4.5 are completely determined once the values for all noise terms are known. From the perspective of a modeler trying to estimate severity the differences between these subgraphs become irrelevant if, we know the function F in equation $\epsilon_3 = F_{\epsilon_3}(\epsilon_5, \epsilon_6, \epsilon_7)$. There are four things to note from the existence of this function:

1. The differences between the assumptions encoded in $\mathcal{G}_m$ and our generative process do not have an impact that makes it absolutely necessary for us to show $\mathcal{S}_2$ instead of $\mathcal{S}_1$.

2. We use the approach encoded in $\mathcal{S}_2$ to generate severity amounts by clusters in a way that is easy to understand (similar to the severity distribution by injury severity observed in Chapter 2).

3. We highlight the distinction between the generative process from insurance practice. Severity is a positive and continuous random variable which means it would not be sensible from a practical point of view to create clusters *a priori* based on the type of crashes. Otherwise, we would provide premiums to policyholders based on their estimated likelihood (through their rating factors) of being involved in a certain type of crash which is a difficult approach to defend.

4. The variability of severity in practice is large because it is a function of three independent noise terms.

We now explain the notation and roles encoded in $\mathcal{S}_2$. We assume severity is a function of medical expenses and property damage and is determined for every i and $j \in \{1, 2, \dots, N_i\}$ by

$$Y_{i,j} = \text{Car value}_i \, \text{DF}_{i,j} + M_{i,j} \tag{4.3}$$

where $\text{DF}_{i,j}$ stands for damage factor, a percentage representing the damage the car sustained. $M_{i,j}$ denotes the medical expenses incurred by policyholder i in the j-th accident. The first summand in Equation (4.3) is the amount corresponding to property damage and is measured as the value of the car – a deterministic function of the car's make-model-year combination, as displayed in $\mathcal{G}_m$ – times the percentage that is considered damaged.

For simplicity, we group the distribution of property damage and medical expenses by the type of crash, we denote this variable by $\text{CT}_{i,j}$. We assign 4 mutually exclusive categories to this variable $\boldsymbol{K} = \{\text{Bump, Mild, Severe, Very Severe}\}$, and it has a multinomial distribution where the parameters depend on the type of driving and risk aversion of the policyholder – two of the severity risk factors in $\mathcal{G}_m$. The parameters can be seen in Table D.7 of the appendix and are calibrated to obtain, on average, more severe accidents in rural areas as compared to urban or suburban areas. The implied assumptions are that more severe accidents occur as a result of driving in higher speed roads and risk-seeking behaviors; as observed for the severity distribution in Chapter 2.

The conditional distribution of the damage factor and the medical expenses given the type of crash are shown in Table 4.3. For the least severe accidents, we assume crashes labeled as ‘bump’ result in no medical expenses and, on average, 10% of vehicle damage, while mild crashes result in a constant $2000 medical expense with 40% of expected vehicle damage. On the other hand, severe crashes are expected to result, in most cases, in a total loss of the car and medical expenses that have a heavy-tailed distribution with a mean of $10,000. The most severe crashes always result in a total loss of the car and their medical expenses also has a heavy-tailed distribution with a mean of $50,000. The logarithm of the empirical average severity distribution, given by $\hat{S}_i/\hat{N}_i$, for those policyholders that had an accident can be seen in Figure 4.6. By construction, this distribution is similar to the one from the nationally representative MVTA data in Chapter 2 in Figure 2.9.

$\mathrm{CT}_{i,j}$	Conditional distribution		Expected value	Standard deviation	99.9% C.I.
Bump	$\mathrm{DF}_{i,j}$	Beta	10%	1%	(7 %, 14%)
	$\mathrm{M}_{i,j}$	Constant	0	-	-
Mild	$\mathrm{DF}_{i,j}$	Beta	40%	4%	(27%, 53%)
	$\mathrm{M}_{i,j}$	Constant	2,000	-	-
Severe	$\mathrm{DF}_{i,j}$	Beta	99%	2%	(82%, 100%)
	$\mathrm{M}_{i,j}$	Gamma	10,000	2,000	(4,692, 17,912)
Very severe	$\mathrm{DF}_{i,j}$	Constant	100%	-	-
	$\mathrm{M}_{i,j}$	Gamma	50,000	8,000	(27,779, 80,573)

Table 4.3: Distribution characteristics of the damage factor and medical expenses given the crash type for policyholder i in the j-th accident.

We determine the expected value of severity conditional on its risk factors relative to

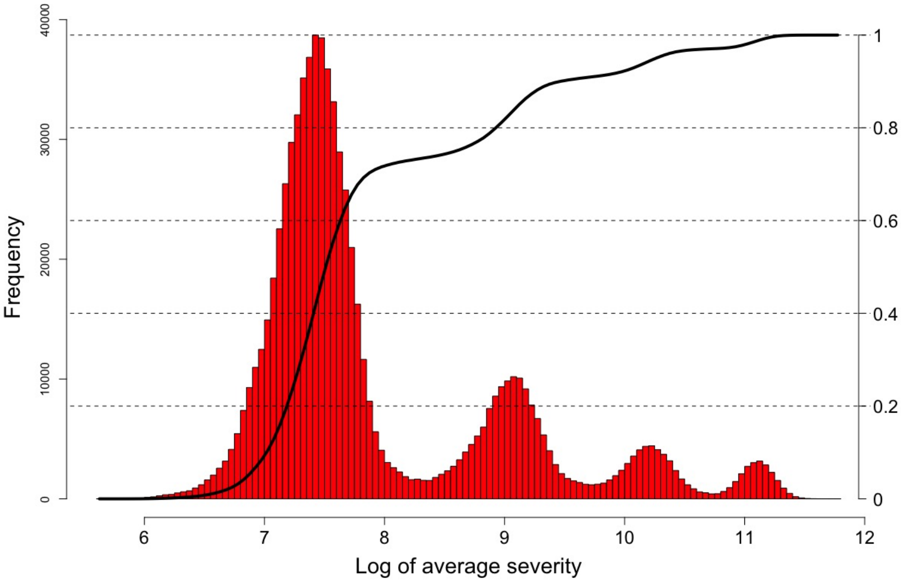


Figure 4.6: Histogram and cumulative density function of the log of average severity, $\log\left(\hat{S}_i/\hat{N}_i\right)$.

$\mathcal{G}_m$, and using the intermediary variables in $\mathcal{S}_2$ as

$$\begin{aligned}
\kappa_i &:= \mathbb{E}[Y_{i,j} \,|\, \mathbf{Pa}(Y_{i,j})] \\
&= \mathbb{E}[\text{Car value}_i \, \text{DF}_{i,j} + M_{i,j} \,|\, \mathbf{Pa}(Y_{i,j})] \\
&= \text{Car value}_i \, \mathbb{E}[\text{DF}_{i,j} \,|\, \mathbf{Pa}(Y_{i,j})] + \mathbb{E}[M_{i,j} \,|\, \mathbf{Pa}(Y_{i,j})] \\
&= \sum_{k \in \boldsymbol{K}} \Big(\text{Car value}_i \, \mathbb{E}[\text{DF}_{i,j} \,|\, \text{CT}_{i,j} = k] \\
&\quad + \mathbb{E}[M_{i,j} \,|\, \text{CT}_{i,j} = k] \Big) \, \mathbb{P}(\text{CT}_{i,j} = k \mid \mathbf{Pa}(Y_{i,j})).
\end{aligned} \tag{4.4}$$

The conditional expected values of property damage and the medical expenses is available in Table 4.3 and the weighting vector of probabilities $\mathbb{P}(\text{CT}_{i,j} = k \mid \mathbf{Pa}(Y_{i,j}))$, for all k, is in Table D.7, depending on the severity risk factors of policyholder i.

4.3.4 True premium

We can compute an expected insurance loss for each policyholder using our assumptions on the distribution of frequency and severity. As mentioned in Section 4.3.1, frequency and severity are independent conditional on different sets of variables. Here, to generate unbiased estimates of the two components in the true premium, additional to the conditional independence benefit, we condition on both the risk factors of frequency and severity, which we denote by $\boldsymbol{\mathcal{F}}_i = \mathbf{Pa}(N_i) \bigcup \mathbf{Pa}(Y_{i,j})$. Then, by the equivalence premium principle (discussed in Chapter 3; the same notation is used here for consistency), the conditional expected value of the insurance loss for policyholder i is

$$\begin{aligned}
\mu_i(\boldsymbol{\mathcal{F}}_i) &= \mathbb{E}[S_i \,|\, \boldsymbol{\mathcal{F}}_i] \\
&= \mathbb{E}\left[\mathbb{E}[S_i \,|\, \boldsymbol{\mathcal{F}}_i, N_i] \,|\, \boldsymbol{\mathcal{F}}_i\right] \\
&= \mathbb{E}\left[\mathbb{E}[Y_{i,1} + \ldots + Y_{i,N_i} \,|\, \boldsymbol{\mathcal{F}}_i, N_i] \,|\, \boldsymbol{\mathcal{F}}_i\right] \\
&= \mathbb{E}\left[N_i \, \mathbb{E}[Y_{i,j} \,|\, \boldsymbol{\mathcal{F}}_i, N_i] \,|\, \boldsymbol{\mathcal{F}}_i\right] \\
&= \mathbb{E}[N_i \,|\, \boldsymbol{\mathcal{F}}_i] \, \mathbb{E}[Y_{i,j} \,|\, \boldsymbol{\mathcal{F}}_i, N_i] \\
&= \mathbb{E}[N_i \,|\, \mathbf{Pa}(N_i)] \, \mathbb{E}[Y_{i,j} \,|\, \mathbf{Pa}(Y_{i,j})] \\
\mu_i(\boldsymbol{\mathcal{F}}_i) &= \lambda_i \, \kappa_i \qquad (4.5)
\end{aligned}$$

where λ_i and κ_i are obtained from equations (4.2) and (4.4), respectively.

Equation (4.5) corresponds to the true premium. It is the value that the insurer should charge the policyholder before loadings for expenses, profit or adverse experience. Although in practice $\mu_i(\boldsymbol{\mathcal{F}}_i)$ is unattainable, the microsimulation methodology allows us to compute it exactly. In the following section, different pricing estimates of $\mu_i(\boldsymbol{\mathcal{F}}_i)$ are used as a basis to determine actuarially fair risk premiums. An advantage of this two-part model is that we allow for possible dependence of the frequency and severity distributions through shared variables (Frees, 2014).

4.4 Numerical Results

In this section, we evaluate the prediction accuracy and potential discrimination of different modeling assumptions, with respect to race. To mimic a realistic scenario, we limit the set of covariates available in $\mathcal{G}_m$ in Figure 4.2, to those variables that are observable and commonly used as rating factors for automobile insurance. The set of available and nondiscriminatory variables for policyholder i, which we denote by $\boldsymbol{X}_i$, includes: age, marital status, gender, car value (through make-model-year), driving record, zip code, and mileage. In this case the discriminatory covariate is race, denoted by D_i, which can take values $d \in \boldsymbol{\mathcal{D}}$ where

$\boldsymbol{\mathcal{D}} = \{\text{White, Black, Asian, Hispanic}\}$. We assume the insurer has access to this variable, although in most jurisdictions this might not be the case, as mentioned in Section 3.7.3.

To simulate a feasible insurance portfolio in Wisconsin, we determine the number of policyholders, n, by sampling randomly 5% of the pseudo-population. Table 4.4 shows the some descriptive statistics of the portfolio by race. The one-year experience of this portfolio shows 83.2% policyholders with no claims, 15% with one claim and 1.8% with two or more claims. There were 42,642 claims that sum up to a total insurance loss of $273 million, with $143 million (52.6%) coming from rural zip codes, while the rural frequency represents 40% of the claims. Table 4.4 shows that White policyholders are a majority in the state of Wisconsin (83.4% of the portfolio). It also shows that racial minorities have, on average, a higher frequency and lower severity (explained by their lack of presence in rural zip codes), resulting in a lower cost per policyholder, which is an approximation to the pure premium. The average cost per policyholder of $1,206 for the insurance portfolio is 7% to 40% cheaper as compared to the average comprehensive gross premium in Wisconsin in 2022, depending on the insurer (Beardsley and Cohen, 2022). This is a reasonable approximation, considering that the microsimulation is calibrated to 2020 data, which means there are two years of inflation difference, and that our costs do not account for vandalism, natural disasters, the insurer's expenses and other risk loadings. In the following section, the distribution of the estimates of the pure premium for all policyholders is compared with $\mu_i(\boldsymbol{\mathcal{F}}_i)$ using a fixed frequency model, and we evaluate the discriminatory impact resulting from different severity model assumptions.

	White	Black	Asian	Hispanic	Total
Number of policyholders	189,010	14,404	6,518	16,512	226,444
Rural areas (%)	47.5	4.9	12.2	19.4	41.8
Suburban areas (%)	28.9	9.3	34.6	17.8	27.0
Frequency (%)	18.7	20.0	18.9	19.0	18.8
Severity ($)	6,536	5,282	6,207	5,803	6,388
Cost per policyholder ($)	1,227	1,058	1,158	1,115	1,206

Table 4.4: Descriptive statistics of the insurance portfolio by race.

4.4.1 Frequency model

For the frequency, we take the common approach taken by actuarial analysts and build the most predictive model possible by using all available data (Werner and Guven, 2007). We foremost assume that the discriminatory covariate has a causal effect over the distribution of zip codes but not directly over frequency. In this case, the insurer's frequency DAG, denoted by $\mathcal{G}_N$, can be seen in the left-hand side of Figure 4.7. We assume a Poisson regression model and a log-link for $\mathbb{E}[N_i \,|\, \boldsymbol{X}_i] = \exp\left([1 \;\; \boldsymbol{X}_i]^\intercal \boldsymbol{\beta}\right)$, where $\boldsymbol{\beta}$ is a 16-dimensional vector, where one dimension is used for the intercept, and the other 15 correspond to categories of the covariates mentioned in Section 4.4. Transformed parameter estimates with the link function for each variable category can be seen on the right-hand side of Figure 4.7.

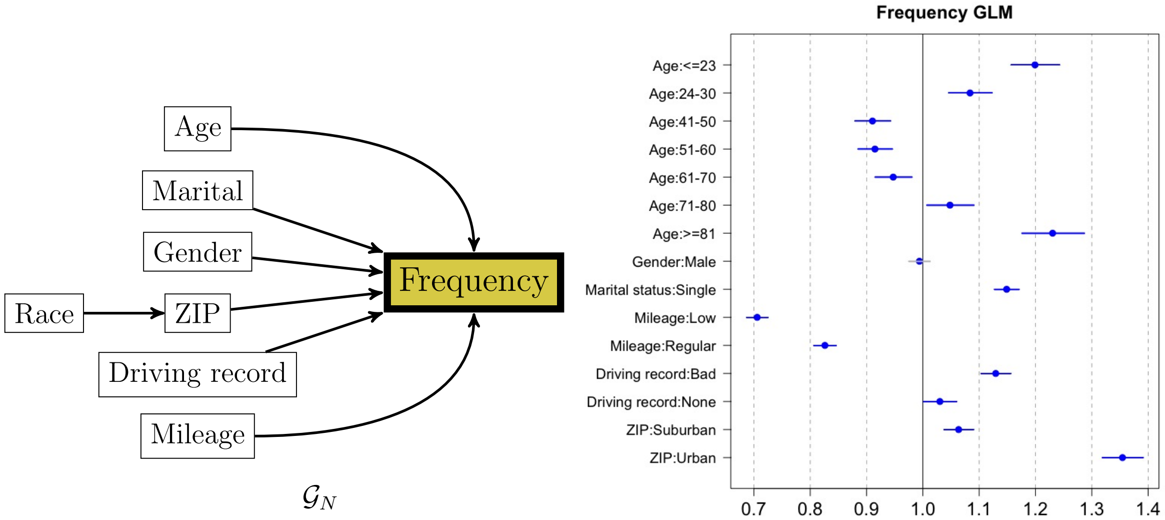


Figure 4.7: Assumed DAG for the frequency model (left-hand side). Parameter estimates and their confidence intervals after the transformation with the link function (right-hand side). Blue solid-lined (gray dotted-lined) denotes estimates that are (non) significant at the 95% level. The intercept is 0.1785.

The parameter estimates of the observable variables in Figure 4.7 are a reflection of our assumptions in the microsimulation. For example, in spite of being arbitrarily discretized, the age categories capture the convex shape of driving ability (see Figure D.4). This convexity is an indirect effect of age on frequency, now captured directly due to the unavailability of the latent variable (also causing driving record to be spuriously statistically significant). In a similar way, marital status and zip code capture the high accident frequency of policyholders that are risk seeking, and of those that tend to drive in urban areas, respectively (both unobservable variables). Further, since $\mathcal{G}_m$ in Figure 4.2 is Markovian, we can state conditional independence relationships from the DAG. One of them is that gender is conditionally independent of frequency given zip code and risk aversion. This can be observed in Figure 4.7 through the non-significance of the gender parameter estimate.

In this chapter the pure premium is a product of two terms, an estimate of expected frequency and an estimate of expected severity. This means that if there is a discriminatory impact caused by the pure premium, it has to come from one or both of these terms. In what follows we show the discriminatory impact over race by the estimated frequency that results from this GLM. Specifically, we determine if there is direct or indirect discrimination as defined in Lindholm et al. (2022a), if there is disparate impact defined in Xin and Huang (2022) and if there is proxy discrimination as defined in Kilbertus et al. (2017) (see Section 1.2.2 for a discussion of these definitions). The estimated frequency is

$$\hat{\lambda}_i = \hat{\mathbb{E}}[N_i \mid \boldsymbol{X}_i] = \exp\left([1 \;\; \boldsymbol{X}_i]^\intercal \, \hat{\boldsymbol{\beta}}\right), \tag{4.6}$$

and if it were the only factor in the computation of the pure premium, that price would avoid direct discrimination, as per Definition 3.1. That premium would also avoid indirect discrimination (Definition 3.2) which is less obvious, but it can be explained by the conditional independence relationships encoded in $\mathcal{G}_m$. In this DAG, the influence of race on frequency is through the paths:

- Race → ZIP → Type of driving → Frequency.
- Race → ZIP → Mileage → Frequency.
- Race → ZIP → Mileage → Type of driving → Frequency.
- Race → Education → Income → ZIP → Mileage → Frequency.
- Race → Education → Income → ZIP → Type of driving → Frequency.
- Race → Education → Income → ZIP → Mileage → Type of driving → Frequency.

Because $\mathcal{G}_m$ is Markovian, we have that $\mathbb{E}[N_i \mid \boldsymbol{X}_i, D_i] = \mathbb{E}[N_i \mid \boldsymbol{X}_i]$ as long as $\boldsymbol{X}_i$ contains zip code, which is the only observable variable shared in these six causal paths. As a consequence, when we compute $\hat{\mathbb{E}}[N_i \mid \boldsymbol{X}_i, D_i]$ the parameter estimates for the categories of race are not significant at the 95% confidence level as shown in Table 4.5. This holds in practice because we have:

Estimator	Indirect discrimination statistical test	Black	Asian	Hispanic
$\hat{\lambda}_i$	GLM estimate	-0.006	-0.024	-0.033
	p-value	0.741	0.409	0.077

Table 4.5: Parameter estimates for race when included in the frequency GLM.

1. No measurement errors.
2. No model misspecification.
3. No incorrect discretization that could create a spurious correlation between frequency and race. Age is incorrectly discretized but the only path between age and race is closed given that we conditioned on zip code.
4. No sampling bias.
5. No incorrect variable selection that could create a spurious correlation between frequency and race. Driving record is the only variable that should not be in the GLM but it does not open spurious paths between race and frequency.

This conditional independence implies that the discrimination-free premium and the unawareness premium coincide, and therefore, there is no indirect discrimination as defined in Lindholm et al. (2022a).

In practice, and as discussed in Section 1.2, to measure disparate impact we have to compare the expected value of the estimated premium for the minority groups relative to that of the non-minority group, in this case White policyholders. But since we know the true premium, it is more appropriate to compute instead the ratio of the expected relative bias. We report the relative bias – rather than absolute bias – to make comparisons across races possible since premium scales vary widely for different policyholders. This means we conclude there is disparate impact if our model is on average overcharging a minority race, where the comparison is relative to the true premium and the majority race.

We use the notation $\phi(X, Y)$ to denote a function $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that computes the operation $X/Y - 1$. For the frequency model, $\phi(\hat{\lambda}, \lambda)$ is the random variable of relative bias of the frequency estimator. Then, the estimated expected value of the relative bias for each race $d \in \boldsymbol{\mathcal{D}}$ is

$$\hat{\mathbb{E}}\left[\phi(\hat{\lambda}, \lambda) \,\middle|\, D = d\right] = \frac{1}{c_d} \sum_{i=1}^{n} \left(\frac{\hat{\lambda}_i}{\lambda_i} - 1 \right) \mathbb{I}_{(D_i = d)}, \tag{4.7}$$

where c_d is the sample size for race d (values in Table 4.4) and the values of λ_i and $\hat{\lambda}_i$ are determined by equations (4.2) and (4.6), respectively. Values of Equation (4.7) for the insurance portfolio can be seen in Table 4.6. The expected relative bias by race is positive because the distribution of the relative bias is right-skewed, in spite of the fact that 43% of the policyholders have a negative relative bias. That positive skew is caused by the model's inability to capture the type of driving (latent variable) of the policyholders. For example, drivers that live in urban zip codes and tend to drive in rural areas should have a low frequency parameter (low λ_i), but they are being treated as all other drivers living in urban zip codes which have on average a higher accident frequency (high $\hat{\lambda}_i$); this results in a large positive value of $\phi(\hat{\lambda}, \lambda)$.

	White	Black	Asian	Hispanic
$\hat{\mathbb{E}}\left[\phi(\hat{\lambda}, \lambda) \,\middle|\, D = d\right]$	24.08	24.53	24.48	25.03

Table 4.6: Estimated expected relative bias by race of the frequency estimator.

To statistically determine if the ratio of the values presented in Table 4.6 result in a disparate impact by race, we have to find the threshold τ in Equation (4.8), for all $d \in \boldsymbol{\mathcal{D}} \setminus \{\text{White}\}$. In this case, τ represents the proportion the expected relative bias of the White policyholders is relative to the expected relative bias for the racial minorities. As discussed in Section 1.2.2, τ is a regulatory threshold that in employment laws from the U.S. is determined by the four-fifths rule which sets $\tau = 0.8$ as the minimum acceptable value.[2]

$$\frac{\hat{\mathbb{E}}\left[\phi(\hat{\lambda}, \lambda) \,\middle|\, D = \text{White}\right]}{\hat{\mathbb{E}}\left[\phi(\hat{\lambda}, \lambda) \,\middle|\, D = d\right]} = \tau. \tag{4.8}$$

The problem with this formulation is that deriving the distribution of the ratio of expected relative bias is cumbersome. Instead, let $\tau = 1 + \Delta/\hat{\mathbb{E}}\left[\phi(\hat{\lambda}, \lambda) \,\middle|\, D = d\right]$ for non-zero expected relative bias, which results in

$$\hat{\mathbb{E}}\left[\phi(\hat{\lambda}, \lambda) \,\middle|\, D = \text{White}\right] - \hat{\mathbb{E}}\left[\phi(\hat{\lambda}, \lambda) \,\middle|\, D = d\right] = \Delta. \tag{4.9}$$

The formulation in Equation (4.9) is equivalent to the location-shift model (as presented in Holland (1986)) in the distribution-free rank sum test, first proposed in Wilcoxon (1945) but using the statistic from Mann and Whitney (1947). The assumptions for this nonparametric test (henceforth referred to as the WMW test), which are satisfied here, include two i.i.d.

[2] In employment law, the hiring probability for the majority group is in the denominator because it is assumed they have a higher probability of being hired and hence, they have the most favorable circumstances. Here, the expected value of the majority group is in the numerator because we assume they have, on average, the lowest premiums, which corresponds to the most favorable scenario.

samples that are mutually independent with continuous distributions. The null hypothesis for the one-sided lower-tail test states that $\Delta = 0$ ($\tau = 1$) versus the alternative $\Delta < 0$ ($\tau < 1$).[3] In this case, we conclude there is disparate impact if two conditions are met: we reject the null hypothesis, and the upper bound in the confidence interval of τ is below 0.8. The results for the estimator $\hat{\lambda}_i$ in Equation (4.6) are presented in Table 4.7 which allows us to conclude there is no evidence to reject the null hypothesis with a 95% confidence level, which means our frequency estimator does not result in disparate impact by race. It is noteworthy to mention that $\tau = 0.8$ is based on a regulatory parameter, but we use this concept more naturally through the WMW test which allows for sampling variability.

Estimator	Disparate impact statistical test	Black	Asian	Hispanic
$\hat{\lambda}_i$	Upper bound for τ	1.041	1.037	1.020
	p-value	0.781	0.434	0.368

Table 4.7: Results of the WMW test for disparate impact by race for the frequency estimator with a 95% confidence level.

It is clear that zip code is a proxy variable for race in $\mathcal{G}_N$. To test if there is proxy discrimination in expectation, as defined in Kilbertus et al. (2017) (described in Section 1.2), we can use the same nonparametric approach as with disparate impact and seek to statistically determine values of Δ (and by consequence τ) for which

$$\hat{\mathbb{E}}\left[\phi(\hat{\lambda}, \lambda) \,\middle|\, do(\text{ZIP} = \text{zip}_1)\right] - \hat{\mathbb{E}}\left[\phi(\hat{\lambda}, \lambda) \,\middle|\, do(\text{ZIP} = \text{zip}_2)\right] = \Delta. \tag{4.10}$$

In the microsimulation DAG, $\mathcal{G}_m$, which determines λ, we have that using the *do*-operator over zip code is the same as just conditioning over zip code since frequency and the variable zip code are conditionally independent given the risk factors of frequency. In the insurer's DAG, $\mathcal{G}_N$, used to determine $\hat{\lambda}_i$, the *do*-operator also has no consequence because it assumes the effect of zip code is not confounded. This means that we can compute the values of Equation (4.10) through traditional conditional expected values. In this test we say there

[3] Note that we do not require any assumption over the variance of the two samples. If the distributions for the two samples had equal variance, then we could also conclude that their underlying distributions are equal. Although this clearly is a stronger hypothesis, differences in the distribution of the bias is a second order problem that could be further explored. We do not make assertions about the distribution of the relative bias across races, only that on average they are treated the same. However, in general, the Flinger-Policello test is better suited for heteroscedastic data (a solution to the Behrens-Fisher problem). For more details see Brunner et al. (2018).

is proxy discrimination for each pairwise statistical test if the upper bound for τ is lower than 1 (i.e. if we reject the null hypothesis)[4], implying that the expected relative bias is significantly higher for the zip code with a predominantly minority population.

In the insurance portfolio there are 765 different zip codes which translate to 292,230 unique pair-wise comparisons through (4.10). Not all of these values are informative since among them, there would be comparisons between urban zip codes with thousands of policyholders and rural zip codes with less than 10 policyholders, for example. Instead, we choose to compare the 56 urban zip codes that have more than 500 policyholders. This selection of zip codes with 500 or more policyholder is arbitrary, but all results and conclusions that are drawn with respect to proxy discrimination (shown in what follows) are consistent if we change the number of policyholders to 100, 250 and 1000. From these 56 urban zip codes, 14 have a greater number of non-White policyholders than White policyholders (12 are in Milwaukee and 2 in Racine, a south-neighboring city to Milwaukee). This selection of zip codes can be seen in Figure 4.8.

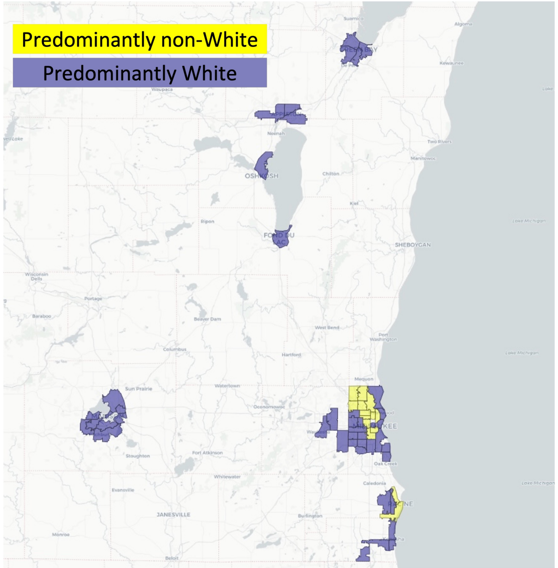


Figure 4.8: Selection of 56 urban zip codes that have more than 500 policyholders categorized by their racial composition.

For each of the 14 zip codes that are predominantly non-White, we determine values of τ with respect to the other 42 urban zip codes, resulting in a total of 588 comparisons. The

[4] We use 1 as the threshold because there is no parameter defined by law for this definition of discrimination, as opposed to disparate impact which is determined by the four-fifths rule.

results of the lower-tailed one-sided WMW tests for $\hat{\lambda}$ can be seen in Figure 4.9. The 42 zip codes are ordered – from left to right – by decreasing proximity to Milwaukee. With a 95% confidence level, this results in $57/588 = 10\%$ of comparisons with proxy discrimination by the frequency estimator. This result is driven by the inability of the frequency model to capture the low frequency parameter of drivers that live in urban zip codes and have the following characteristics: tend to drive in rural roads; are risk averse; have high driving ability. Specifically, there are four minority zip codes (53208, 53209, 53215 and 53216) that have a larger than average number of drivers with the aforementioned characteristics.

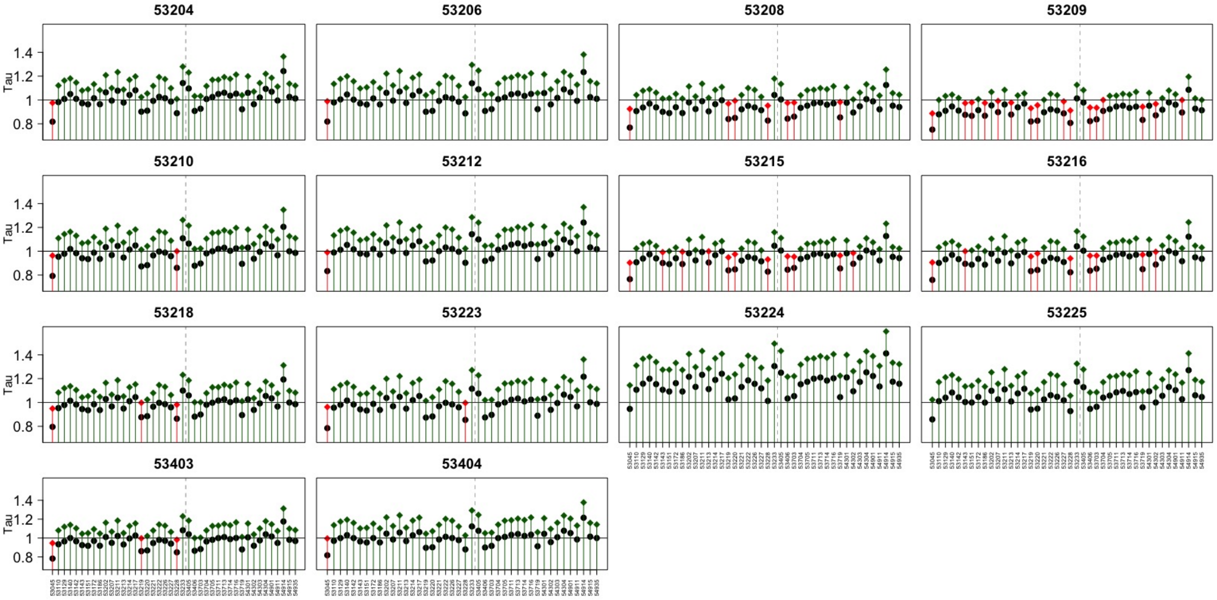


Figure 4.9: Upper bound (diamond) and estimator (circle) of τ for each of the 588 pairwise tests in Equation (4.10) for the frequency estimator with a 95% confidence level. Zip codes to the right of the vertical dotted line are outside of Milwaukee.

A summary of the discriminatory impact of the frequency estimator $\hat{\lambda}_i$ is available in Table 4.8. These results imply that if there is direct discrimination, indirect discrimination or disparate impact arising from the pure premiums estimators (presented in the next section), it has to be a result of the severity modeling assumption, as the fixed frequency estimator avoids these definitions of discrimination. If there is proxy discrimination in the pure premium estimators, it is a consequence of both frequency and severity models, or just the frequency estimator.

Estimator	Direct discrimination Lindholm et al. (2022a)	Indirect discrimination Lindholm et al. (2022a)	Disparate impact Xin and Huang (2022)	Proxy discrimination Kilbertus et al. (2017)
$\hat{\lambda}_i$	No	No	No	Yes

Table 4.8: Summary of the discriminatory effects on race by the frequency estimator $\hat{\lambda}_i$.

4.4.2 Four severity modeling assumptions

To compute estimates of the pure premium, we make four different modeling assumptions for severity. The structure of these four models is not data-driven, rather they are simply feasible actuarial pricing structures. In the first model, we assume the expected severity is equal for all policyholders, as depicted by $\mathcal{G}_1$ in Figure 4.10. In the second model, denoted by $\mathcal{G}_2$, we assume that expected severity is a deterministic function of the vehicle's value, since more expensive cars tend to have more expensive repairs. In both $\mathcal{G}_1$ and $\mathcal{G}_2$ we are implicitly assuming that frequency and severity are independent. The third and fourth are GLM models, denoted by $\mathcal{G}_3$ and $\mathcal{G}_4$, where we use another deterministic connection between the car value and severity, and use exactly the same covariates as for the frequency model. The only difference between them is that in $\mathcal{G}_3$ we replace zip code with a new proxy variable, denoted by P, explained below. These last two models assume that frequency and severity are unconditionally dependent, but independent conditional on their risk factors. The four modeling assumptions can be contrasted through their respective DAGs in Figure 4.10.

We denote estimates of severity relative to DAG $\mathcal{G}$ by $\hat{\kappa}_i^{(\mathcal{G})}$. Its functional form depends on the model assumptions:

- For the constant severity model we have that $\hat{\kappa}_i^{(\mathcal{G}_1)} = 6,388$, obtained as the average severity of the portfolio, previously presented in Table 4.4.
- Values of $\hat{\kappa}_i^{(\mathcal{G}_2)}$ can be seen in Table 4.9 where 10 mutually exclusive groups are created as a function of the car value. Monotonically higher average severity estimates are obtained for more expensive vehicles (as a consequence of the underlying severity assumptions described in Section 4.3.3). In Figure 4.11 we present the marginal empirical cumulative distribution function of the car groups by race, showing that on average Asian (Black and Hispanic) policyholders own more (less) expensive cars than other races. The value of the vehicle is assigned as a function of the income

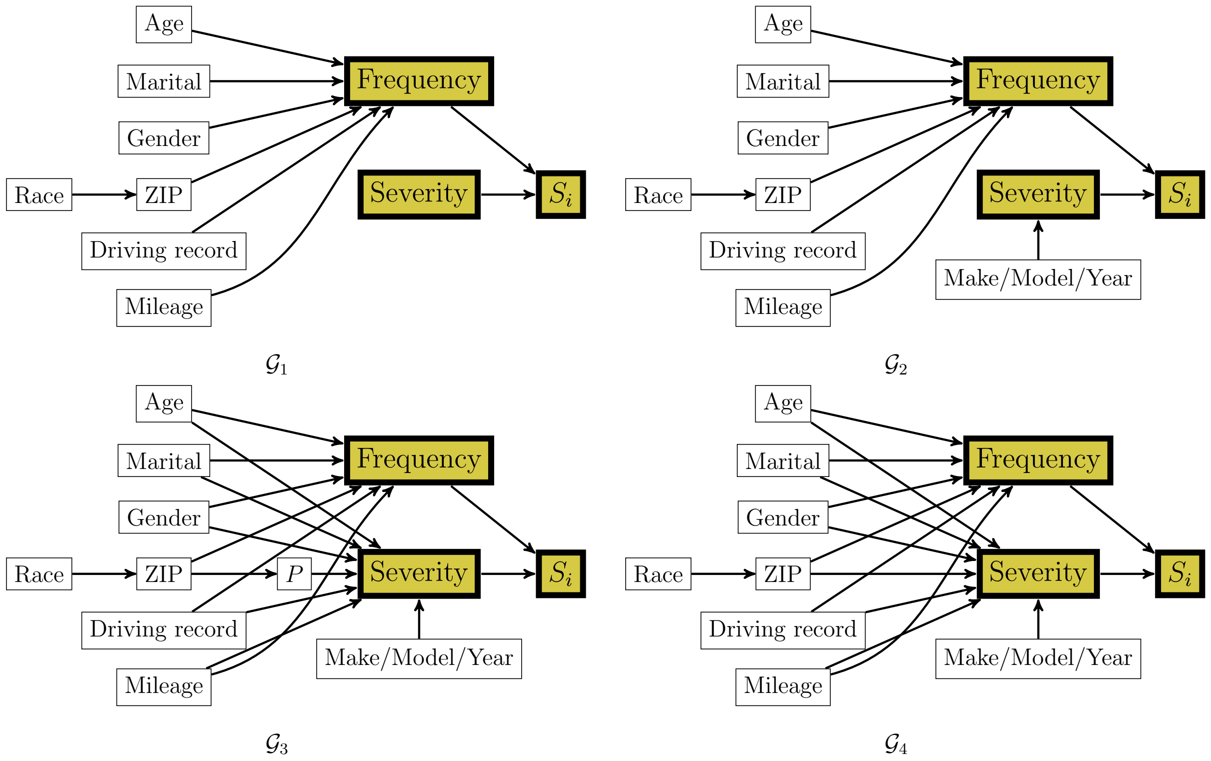


Figure 4.10: DAGs that represent four modeling assumptions for severity resulting in different estimates of the pure premium.

level, which is simulated to match ACS statistics (both variables explained in D.10 and D.11 of the Appendix).

- To determine $\hat{\kappa}_i^{(\mathcal{G}_3)}$ we assume a gamma GLM with a log-link using as a response variable the severity divided by the car value. The covariates are $\boldsymbol{X}_i$ (as in the frequency model) with the sole difference that we replace zip code by a proxy, P. As opposed to using the zip code categories which solely depend on the population density, P additionally considers the income distribution, as shown in Figure 4.12. Four groups are considered and obtained through the K-means algorithm, for more information on this clustering methodology see James et al. (2013) or Gan and Valdez (2020). By using P we change the clustering of zip codes from Figure D.9 (used in the underlying model) to a closer approximation of the historical insurance pricing practice in Wisconsin (as seen in Figure 1.4 of the Introduction), since rural areas are now split into two categories depending on their income instead of treating them as homogeneous zip codes, and by reducing the number of suburban zip codes neighboring the three main cities in Wisconsin.

- To determine $\hat{\kappa}_i^{(\mathcal{G}_4)}$ we assume a gamma GLM with a log-link using as a response variable the severity divided by the car value. The covariates for model are $\boldsymbol{X}_i$ (as in the frequency model).

Group number	Car value	Portfolio (%)	$\hat{\kappa}_i^{(\mathcal{G}_2)}$
1	10,000 or less	6.76	4,590
2	10,001 - 12,500	10.22	5,339
3	12,501 - 15,000	19.45	5,662
4	15,001 - 17,500	22.37	6,065
5	17,501 - 20,000	18.12	6,702
6	20,001 - 22,500	11.87	7,448
7	22,501 - 25,000	6.68	8,208
8	25,001 - 30,000	3.11	9,194
9	30,001 - 40,000	1.37	10,929
10	40,001 or more	0.04	13,061

Table 4.9: Severity as a function of car value.

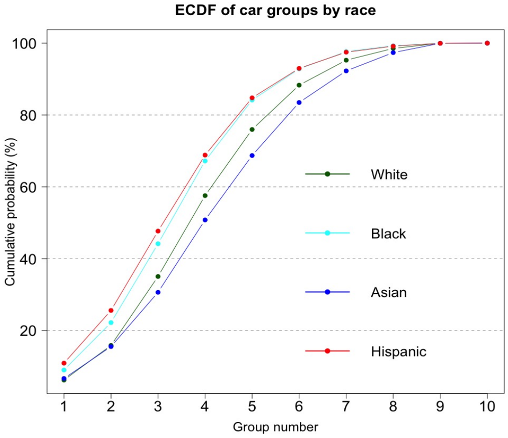


Figure 4.11: Marginal ECDF of groups by race.

The parameter estimates of severity under $\mathcal{G}_3$ and $\mathcal{G}_4$ can be compared in Figure 4.13. The age and gender estimates reflect the assumption that young males have more severe accidents through their risk seeking behavior. The transformed parameter estimate for males being lower than one can be explained by two paths in $\mathcal{G}_m$:

- Gender → Risk aversion → Severity.
- Gender→ Income → Car value → Severity.

Since the response variable in the two GLMs is the observed severity divided by the car value, the parameter estimate tells us that the ratio tends to be smaller for males. This can be explained by our data generating process which resulted in males owning, on average, more expensive vehicles. Thus, the estimates of severity are larger for males than females, on average. Due to $\mathcal{G}_m$, we know driving record is spuriously significant. Mileage and zip code (or the categories of P in the case of $\mathcal{G}_3$) are capturing the effects of type of driving which reflect the higher severity observed for drivers that tend to drive on freeways and rural roads as opposed to residential and urban streets. The main difference between the

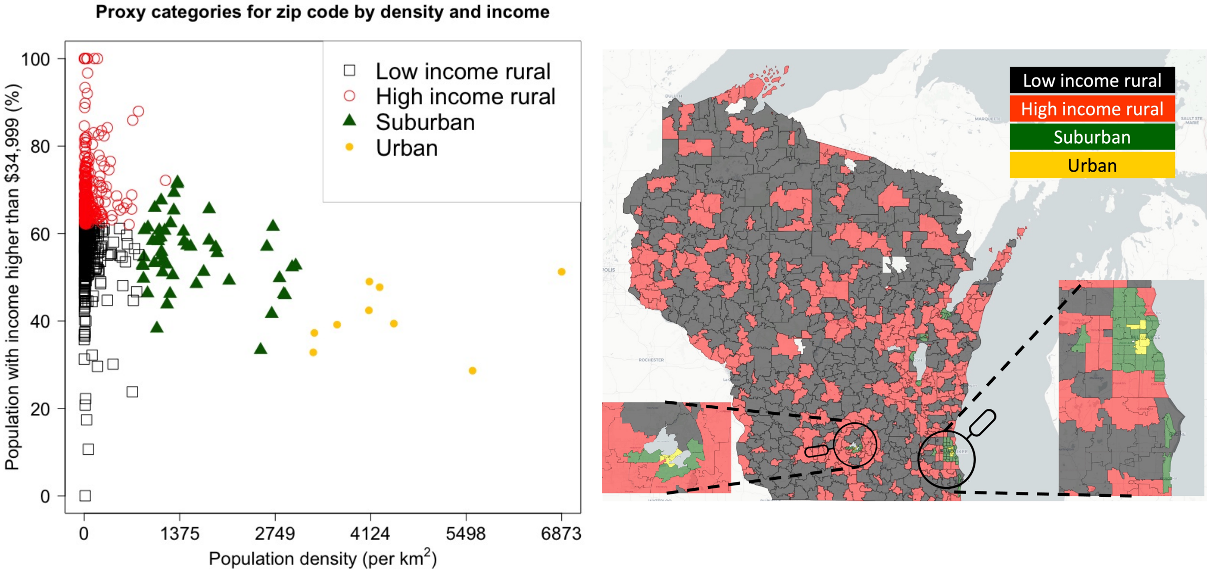


Figure 4.12: Distribution of zip codes by population density and the percentage of the population that have an income higher than $34,999 (left-hand side). Zip codes are grouped by four different categories in P. On the right-hand side a map of Wisconsin shows the geographical distribution of P.

two models is that $\mathcal{G}_4$ accurately reflects that the driver's address has a stronger influence on the type of driving than the annual mileage (this is described in the variable generation of type of driving in D.14), while $\mathcal{G}_3$, by using new zip code categories through P, cannot capture this accurately.

4.4.3 Discriminatory impact of the pure premium estimators

The four severity modeling assumptions allow us to obtain estimates of $\mu_i(\boldsymbol{\mathcal{F}}_i)$, given by $\hat{\mu}_i^{(\mathcal{G})} = \hat{\lambda}_i \, \hat{\kappa}_i^{(\mathcal{G})}$, relative to each DAG in Figure 4.10. We denote by μ and $\hat{\mu}^{(\mathcal{G})}$ the random variables for the true premium, $\mu_i(\boldsymbol{\mathcal{F}}_i)$, and premium estimator $\hat{\mu}_i^{(\mathcal{G})}$ for all i. We contrast the performance of the four premium estimators in Table 4.10 through the percentage increase in expected relative bias relative to a baseline as

$$\frac{\hat{\mathbb{E}}\left[\phi\left(\hat{\mu}^{(\mathcal{G})}, \mu\right) \,\middle|\, D = d\right]}{\hat{\mathbb{E}}\left[\phi\left(\hat{\mu}^{(\mathcal{G}_4)}, \mu\right) \mid D = d\right]},$$

where we select $\mathcal{G}_4$ as the baseline since it is the model with the smallest relative bias. The baseline premium expected relative bias for the portfolio is 51.44%, this positive value

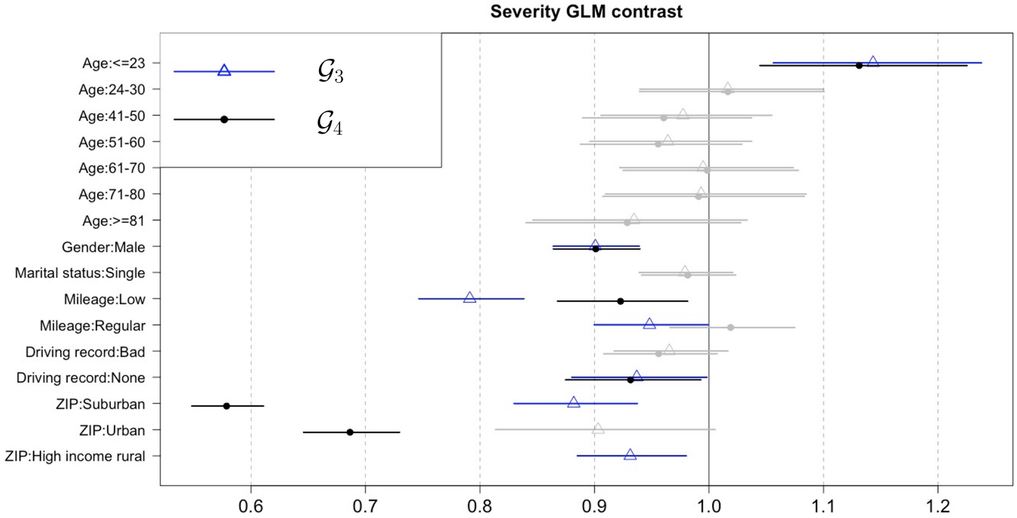


Figure 4.13: Parameter estimates and their confidence intervals after the transformation with the link function. Blue solid-lined (gray dotted-lined) denotes estimates that are (non) significant at the 95% level. The transformed intercepts are 0.487 and 0.554.

can be explained by the right-tailed distribution of the severity component, added to the positive skew observed for the frequency estimator in Table 4.6.

By this measure of performance $\mathcal{G}_1$ is the least accurate model. Moreover, it disproportionately affects racial minorities, explained by the fact that their average severity is below the portfolio average (as seen in Table 4.4). Black and Hispanic policyholders are affected more severely by this model as compared to Asian policyholders, because on average they own less expensive cars (see Figure 4.11). Model $\mathcal{G}_2$, which assigns severity through the value of the car, decreases the portfolio bias by 10 percentage points compared to the worst performer by creating more homogeneous groups with respect to average severity (mainly advantageous to Black and Hispanic policyholders). Lastly, model $\mathcal{G}_3$, by conditioning on P instead of zip code, increases the relative bias of the pure premium by 13.3% as compared to $\mathcal{G}_4$. This has a greater impact on Asian policyholders because of their large presence in suburban areas (see Table 4.4), coupled with the inability of P to capture the lower severity of the suburbs as seen in Figure 4.13.

The four pure premium estimators $\hat{\mu}_i^{(\mathcal{G}_1)}$, $\hat{\mu}_i^{(\mathcal{G}_2)}$, $\hat{\mu}_i^{(\mathcal{G}_3)}$ and $\hat{\mu}_i^{(\mathcal{G}_4)}$ avoid direct discrimination as defined in Lindholm et al. (2022a) since none of them uses the discriminatory covariate in the computation of the expected value of the insurance losses. To determine if each premium avoids indirect discrimination we have to statistically test if severity is conditionally independent of race. We can do this because in the frequency model we already conclude

Severity model	DAG	White	Black	Asian	Hispanic	Total
GLM (Baseline)	$\mathcal{G}_4$	51.42	51.63	56.00	49.62	51.44
Constant	$\mathcal{G}_1$	16.90	83.05	48.25	80.74	26.59
Car groups	$\mathcal{G}_2$	9.30	58.60	46.74	52.71	16.68
GLM with P	$\mathcal{G}_3$	12.13	13.43	34.41	17.60	13.30

Table 4.10: Percentage increase in the expected relative bias of the pure premium estimates compared to $\mathcal{G}_4$.

that there is no indirect discrimination (with a 95% confidence level). Note that the four DAGs assume that severity is conditionally independent of race (unconditionally for $\mathcal{G}_1$ and $\mathcal{G}_2$), but now we test for it.

Under $\mathcal{G}_1$ there is no distributional assumption over $Y_{i,j}$. For this reason we construct a nonparametric distribution of $\hat{\mathbb{E}}[Y_{i,j} \mid D_i]$ through 10,000 bootstrap samples (see Appendix E.1 for details). Not all of the confidence intervals overlap, meaning the expected severity is statistically dependent on race under $\mathcal{G}_1$. Using the same approach under $\mathcal{G}_2$, we obtain three car groups where the expected severity is statistically dependent on race even after conditioning on the car value (see Appendix E.1). Therefore, there is evidence of indirect discrimination as defined in Lindholm et al. (2022a) using premiums $\hat{\mu}_i^{(\mathcal{G}_1)}$ and $\hat{\mu}_i^{(\mathcal{G}_2)}$. For the two severity GLM models, we test for indirect discrimination similarly to the frequency model. The results in Table 4.11 leads us to conclude there is no statistical evidence of indirect discrimination arising from premiums $\hat{\mu}_i^{(\mathcal{G}_3)}$ and $\hat{\mu}_i^{(\mathcal{G}_4)}$. The explanation for this result is similar to that of the frequency model. In this case, the use of zip code and the car value closes all causal paths between race and severity.

Estimator	Indirect discrimination statistical test	Black	Asian	Hispanic
$\hat{\kappa}_i^{(\mathcal{G}_3)}$	GLM estimate	0.044	-0.015	0.046
	p-value	0.327	0.801	0.274
$\hat{\kappa}_i^{(\mathcal{G}_4)}$	GLM estimate	0.085	0.087	0.071
	p-value	0.059	0.171	0.088

Table 4.11: Parameter estimates for race when included in the respective GLM.

We test for disparate impact with the same methodology used for the frequency model by determining, for each DAG $\mathcal{G}$, and for each $d \in \boldsymbol{\mathcal{D}} \setminus \{\text{White}\}$, if the null hypothesis of the WMW test is rejected, and determining if the upper bound of τ is below 0.8 with 95% confidence in the equation

$$\frac{\hat{\mathbb{E}}\left[\phi\left(\hat{\mu}^{(\mathcal{G})}, \mu\right) \mid D = \text{White}\right]}{\hat{\mathbb{E}}\left[\phi\left(\hat{\mu}^{(\mathcal{G})}, \mu\right) \mid D = d\right]} = \tau.$$

The results of the statistical tests for each of the four premiums are presented in Table 4.12. As suspected by the percentage increases of expected relative bias in Table 4.10, we conclude there is disparate impact arising from premiums $\hat{\mu}_i^{(\mathcal{G}_1)}$ and $\hat{\mu}_i^{(\mathcal{G}_2)}$, but not from the GLM models. For premium $\hat{\mu}_i^{(\mathcal{G}_3)}$ and $\hat{\mu}_i^{(\mathcal{G}_4)}$ the null hypothesis ($\tau = 1$) is rejected for Asian policyholders (also for Black policyholders in $\mathcal{G}_3$) relative to White policyholders but the upper bound in the confidence interval of τ is greater than the regulatory parameter of 80% in these cases. This result, coupled with the lack of disparate impact arising from the frequency model, leads us to conclude that models $\mathcal{G}_1$ and $\mathcal{G}_2$ fail to capture accurately the accident severity of racial minorities and consequently overcharges their premiums significantly more than for White policyholders.

Estimator	Disparate impact statistical test	Black	Asian	Hispanic
$\hat{\mu}_i^{(\mathcal{G}_1)}$	Upper bound for τ	0.755	0.825	0.791
	p-value	0.000	0.000	0.000
$\hat{\mu}_i^{(\mathcal{G}_2)}$	Upper bound for τ	0.779	0.784	0.834
	p-value	0.000	0.000	0.000
$\hat{\mu}_i^{(\mathcal{G}_3)}$	Upper bound for τ	0.997	0.854	1.011
	p-value	0.032	0.000	0.317
$\hat{\mu}_i^{(\mathcal{G}_4)}$	Upper bound for τ	1.054	0.984	1.076
	p-value	0.998	0.004	0.999

Table 4.12: Results of the WMW test for disparate impact by race for the pure premium estimators with a 95% confidence level.

To test for proxy discrimination we use the same methodology as for the frequency model and show the details of the test results for premiums $\hat{\mu}_i^{(\mathcal{G}_1)}$, $\hat{\mu}_i^{(\mathcal{G}_2)}$, $\hat{\mu}_i^{(\mathcal{G}_3)}$ and $\hat{\mu}_i^{(\mathcal{G}_4)}$ in Appendix E.2. Out of the 588 pairwise tests of urban zip codes, we obtained 60% for $\mathcal{G}_1$, 12% for $\mathcal{G}_2$, 7% for $\mathcal{G}_3$ and 2% for $\mathcal{G}_4$ that resulted in proxy discrimination. Because we did multiple tests, we expect 5% of them to be spuriously wrong. This means that only $\mathcal{G}_4$ is free of proxy discrimination. According to this measure, the worst model is the constant severity model, which significantly overcharges urban zip codes with predominantly minority populations. Estimating severity as a function of the car value dramatically improves the performance relative to the constant severity model, by increasing the premium estimation accuracy in most minority zip codes. The 12% result for $\mathcal{G}_2$ is mostly driven by the four zip codes that also resulted in proxy discrimination for the frequency model (53208, 53209, 53215 and 53216), and its explanation is identical. The proxy discrimination under model $\mathcal{G}_3$ results specifically from the premium overestimation in zip code 53403; an urban zip code by its population density (a racial minority, predominantly) but classified as 'low income rural' through the variable P because of its income distribution. This classification difference in the severity GLM increases significantly the zip code premiums by not obtaining the severity discount shared with urban zip codes. Notably, model $\mathcal{G}_4$ does not result in proxy discrimination, because the shortcomings of the frequency model were fixed by the severity GLM in the bias estimation for the pure premium $\hat{\mu}_i^{(\mathcal{G}_4)}$.

The four discrimination measures are summarized for each of the four premium estimators in Table 4.13. Premiums $\hat{\mu}_i^{(\mathcal{G}_1)}$ and $\hat{\mu}_i^{(\mathcal{G}_2)}$, which assume that frequency and severity are unconditionally independent, are the worst performers both in terms of accuracy and discriminatory impact over race. It should be noted that the magnitude of the impact is considerably different, as using the car value to estimate severity has an overall improvement over the constant severity approach. The premium inaccuracy is to be expected due to the differences between $\mathcal{G}_1$ and $\mathcal{G}_2$ compared with the true model $\mathcal{G}_m$, but the significantly disproportionate impact on racial minorities over this incorrect modeling assumption is a contribution that has not been mentioned in the actuarial literature.

4.5 Conclusion

In this chapter we use concepts from causal inference to build a simulation tool, to generate a pseudo-population characterized by key attributes of a target population. We use this to study a feasible insurance portfolio in the state of Wisconsin and evaluate the discriminatory impact of different pricing policies. We provide a statistical methodology to determine disparate impact and proxy discrimination that are in line with actuarial practice. Ultimately, we show that incorrectly assuming independence between frequency and severity cannot only result in inaccurate premiums but can also be significantly detrimental to racial minorities. Moreover, this situation can occur regardless of the insurer's intention, since in

Estimator	Direct discrimination Lindholm et al. (2022a)	Indirect discrimination Lindholm et al. (2022a)	Disparate impact Xin and Huang (2022)	Proxy discrimination Kilbertus et al. (2017)
$\hat{\mu}_i^{(\mathcal{G}_1)}$	No	Yes	Yes	Yes
$\hat{\mu}_i^{(\mathcal{G}_2)}$	No	Yes	Yes	Yes
$\hat{\mu}_i^{(\mathcal{G}_3)}$	No	No	No	Yes
$\hat{\mu}_i^{(\mathcal{G}_4)}$	No	No	No	No

Table 4.13: Summary of the discriminatory effects on race by the four pure premium estimators.

this numerical illustration the *a priori* assumption is that race is not a risk factor of either frequency or severity.

This static microsimulation has a large number of assumptions on each of the characteristics that are attributed to the pseudo-population of Wisconsin for the year 2020. Nevertheless, many attributes are calibrated to statistics from the U.S. Census Bureau and, the balance between frequency and severity is calibrated to national automobile accident data reported in the literature. Bearing these considerations in mind, our numerical illustration offers a feasible explanation for the 70% premium overcharge in Milwaukee as compared to Madison and Green Bay, as shown in Figure 1.4 for an insurer in Wisconsin. If that were to be the case, the moral of the story for an insurer in Wisconsin is that an incorrect independence assumption between frequency and severity is causing indirect discrimination and disparate impact to racial minorities.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis we have addressed some of the main issues around discrimination in insurance pricing. We focused on statistical methods to identify and mitigate discrimination in the calculation of premiums. We also explored some data-driven ways in which discrimination can unintentionally arise. Such was the case where, inadvertently, standard actuarial models were particularly harmful to minority groups.

Directed acyclic graphs and other relevant tools from causal inference played a major role in our findings. These tools aided in framing graphically the pricing problem and in distinguishing risk factors from rating factors; a task which should be considered crucial for insurers. Models with maximum predictive accuracy cannot be the gold standard when some of the variables are not a cause of the risk, and more so if the use of these rating factors results in a premium that is especially damaging to a historically unfavored population.

Throughout this thesis we have assumed every person involved in the policy pricing process is well-intentioned. Nevertheless, as shown in some federal redlining cases in the United States, discrimination in insurance pricing can arise or be compounded by ill-intent or unconscious biases in the claims settlement or underwriting processes. Minority groups are adversely affected by more frequent and rigorous claims investigations, a less comprehensive policy coverage or by delayed settlements. All of which, to the detriment of society, creates a negative feedback loop to reinforce discriminatory outcomes.

We hope the tools and examples presented here can aid lawmakers and regulators to establish unambiguous constraints in the computation of discrimination-free insurance premiums. In consequence, this would light a path for actuaries to unequivocally eliminate the detrimental consequences of past and ongoing discrimination, which threaten to become more complex with the implementation of machine learning algorithms and the automation of underwriting processes.

5.2 Future Work

There are some avenues of research that remain to be explored.

5.2.1 Premium principles and risk loadings

Throughout this thesis we assumed the benchmark method to calculate premiums: the equivalence premium principle. This serves as a starting point by mitigating discrimination from the pure premium, which corresponds to the expected value of the future loss. However, other premium principles such as the portfolio percentile premium (Dickson et al., 2020), and the gross premium charged to the policyholder account for the variance of the future expected loss and other risk loadings. The variance and risk loadings are not considered in this thesis. Nevertheless, even if the expected value is discrimination-free according to any of the mathematical definitions stated in this thesis, the gross premium paid by the policyholder could be discriminatory if the variables used to calculate the risk loadings are left unchecked. In view of this, it is necessary to extend the results shown in this thesis to the calculation of a gross premium.

5.2.2 Dynamic microsimulation

In Chapter 4 we studied a one year insurance portfolio for a static pseudo-population from Wisconsin calibrated to statistics from the 2020 ACS. The objective was to study the discriminatory impact of incorrectly assuming independence of frequency and severity. To remedy the observed discriminatory impact, any of the mathematical solutions proposed in Kilbertus et al. (2017), Lindholm et al. (2022a) or Xin and Huang (2022) could be implemented. If those premiums were to be charged, it is of interest to study the dynamic behavior of both the composition and performance of the insurance portfolio over the years.

This dynamic behavior can be modeled by extending our proposed static microsimulation through the consideration of yearly transition probabilities. For example, for every year of the insurance portfolio, the pseudo-population has to be aged, and variables such as marital status, education, number of people living in each zip code and income have to be tuned to the statistics of the ACS of the corresponding year. Driving record would evolve based on the (non)-occurrence of accidents in the past year and vehicle prices have to be inflation-adjusted. A lapse rate for the portfolio has to be assumed and new policyholders could join the portfolio based on the premiums charged in the last year. This would create a dynamic composition of the portfolio, which can be analyzed to address the problem of adverse selection; a common concern when implementing premiums that are discrimination-free. Furthermore, portfolio performance over a longer time frame can by analyzed with more realistic market conditions through the assumption of multiple competing insurers in the state.

References

American Academy of Actuaries (2002). *The use of credit history for personal lines of insurance: Report to the National Association of Insurance Commissioners.* Washington, DC: American Academy of Actuaries.

Ansfield, B. (2021). The crisis of insurance and the insuring of the crisis: Riot reinsurance and redlining in the aftermath of the 1960s uprisings. *Journal of American History*, 107(4):899–921.

Araiza Iturria, C. A., Hardy, M., and Marriott, P. (2021a). A consolidated database of police-reported motor vehicle traffic accidents in the United States for actuarial applications [Database]. Zenodo. https://doi.org/10.5281/zenodo.7121281. Accessed: 02-12-2021.

Araiza Iturria, C. A., Hardy, M., and Marriott, P. (2021b). A consolidated database of police-reported motor vehicle traffic accidents in the United States for actuarial applications [Figure]. Zenodo. https://doi.org/10.5281/zenodo.7121442. Accessed: 02-12-2021.

Araiza Iturria, C. A., Hardy, M., and Marriott, P. (2021c). A consolidated database of police-reported motor vehicle traffic accidents in the United States for actuarial applications [Software]. Zenodo. https://doi.org/10.5281/zenodo.7120835. Accessed: 02-12-2021.

Arnold, K. F., Harrison, W. J., Heppenstall, A. J., and Gilthorpe, M. S. (2019). DAG-informed regression modelling,agent-based modelling and microsimulationmodelling: a critical comparison of methodsfor causal inference. *International Journal of Epidemiology*, 48(1):243–253.

Arrow, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review*, 53(5):941–973.

Avraham, R. (2017). *Discrimination and Insurance.* In: Kasper Lippert-Rasmussen. The Routledge Handbook of the Ethics of Discrimination, Ch. 28.

Ayuso, M., Guillén, M., and Marín, A. M. P. (2016). Using gps data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C*, 68:160–167.

Beardsley, A. and Cohen, J. (2022). Car Insurance in Wisconsin. https://insurify.com/car-insurance/wisconsin/. Insurify. Updated 01-11-2022, accessed: 03-11-2022.

Blincoe, L., Miller, T., Zaloshnja, E., and Lawrence, B. (2010). *The economic and societal impact of motor vehicle crashes (Revised in 2015) (Report No. DOT HS 812 013)*. National Highway Traffic Safety Administration.

Bolancé, C. and Vernic, R. (2020). Frequency and Severity Dependence in the Collective Risk Model: An Approach Based on Sarmanov Distribution. *Mathematics*, 8(92).

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.

Brunner, E., Bathke, A. C., and Konietschke, F. (2018). *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs: Using R and SAS*. Springer, first edition.

CarGurus (2022a). Used Car Price Trends. https://www.cargurus.com/Cars/price-trends/. Accessed: 10-08-2021.

CarGurus (2022b). Used Car Price Trends. https://www.cargurus.com/Cars/price-trends/. Accessed: 19-04-2022.

Chen, F. and Chen, S. (2011). Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. *Accident Analysis and Prevention*, 43:1677–1688.

Cheng, S. and Long, J. S. (2007). Testing for IIA in the Multinomial Logit Model. *Sociological Methods and Research*, 35(4):583–600.

Chibanda, K. F. (2022). Defining Discrimination in Insurance. Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing.

Civil Rights Act (1964). 42 U.S.C. 2000e.

Consumer Federation of America (2015). High Price of Mandatory Auto Insurance in Predominantly African American Communities. https://consumerfed.org/reports/high-price-of-mandatory-auto-insurance-in-predominantly-african-american-communities/. Accessed: 30-11-2022.

Council, F., Zaloshnja, E., Miller, T., and Persaud, B. (2005). Crash Cost Estimates by Maximum Police-Reported Injury Severity Within Selected Crash Geometries (FHWA-HRT-05-051). https://www.fhwa.dot.gov/publications/research/safety/05051/05051.pdf. Accessed: 01-09-2021.

Cui, R., Groot, P., and Heskesm, T. (2016). Copula PC Algorithm for Causal Discovery from Mixed Data. In *Machine Learning and Knowledge Discovery in Databases*, pages 377–392. Springer International Publishing.

Dickson, D., Hardy, M., and Waters, H. (2020). *Actuarial Mathematics for Life Contingent Risks*. Cambridge University Press, third edition.

Dolman, C., Appleton, N., Atkins, G., Baker, L., Beenders, V., Breitenbach, C., McLenaghan, J., Storozhev, M., and Yellowlees, C. (2020). The Australian Anti-Discrimination Acts: Information and Practical Suggestions for Actuaries. *Presented to the Actuaries Institute 20/20 All-Actuaries Virtual Summit*.

Dunn v. Midwestern Indem. (1979). 472 F. Supp. 1106 (S.D. Ohio 1979).

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.

EEOA (1972). Equal Employment Opportunity Act. Public Law 92-261.

Equal Employment Opportunity Commission (1979). Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures. https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines. Accessed: 19-12-2022.

European Commission (2012). Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats). https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2012:011:0001:0011:en:PDF. Accessed: 13-07-2020.

Fair Housing Act (1968). 42 U.S.C. 3601.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268.

FHWA (2016). Kabco injury classification scale and definitions. https://safety.fhwa.dot.gov/hsip/spm/conversion_tbl/pdfs/kabco_ctable_by_state.pdf. Accessed: 01-09-2021.

Frees, E. W. (2014). *Frequency and Severity Models*, volume 1 of *Predictive Modeling Applications in Actuarial Science*, page 138–164. Cambridge University Press.

Frees, E. W. and Huang, F. (2021). The Discriminating (Pricing) Actuary. *North American Actuarial Journal*, pages 1–23.

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

Gan, G. and Valdez, E. A. (2020). Data clustering with actuarial applications. *North American Actuarial Journal*, 24(2):168–186.

Gilbert, N. and Troitzsch, K. G. (2005). *Simulation for the Social Scientist*. Open University Press, second edition.

Glenn, B. J. (2000). The shifting rhetoric of insurance denial. *Law & Society Review*, 34(3):779–808.

Golden, L. L., Brockett, P. J., Ai, J., and Kellison, B. (2016). Empirical Evidence on the Use of Credit Scoring for Predicting Insurance Losses with Psycho-social and Biochemical Explanations. *North American Actuarial Journal*, 20(3):233–251.

Grari, V., Charpentier, A., Lamprier, S., and Detyniecki, M. (2022). A fair pricing model via adversarial learnig. *arXiv preprint arXiv:2202.12008*.

Griggs v. Duke Power Company (1971). 401 U.S. 424.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *30th Conference on Neural Information Processing Systems (NIPS)*.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The Elements of Statistical Learning*. Springer, second edition.

Hausman, J. and McFadden, D. (1984). Specification Tests for the Multinomial Logit Model. *Econometrica*, 52(5):1219–1240.

Hoffman, F. L. (1896). The Race Traits and Tendencies of the American Negro. *American Economic Association*, 11(1):1–329.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81:945–960.

Hu, F. and Zidek, J. V. (2002). The weighted likelihood. *The Canadian Journal of Statistics*, 30(3):347–371.

Huskey v. State Farm Fire & Casualty Co. (2022). C.A. No 22-CV-7014, N.D. Illinois 2022.

Insurance Information Institute (2019). Estimated Percentage Of Uninsured Motorists By State, 2019. https://www.iii.org/table-archive/20641. Accessed: 03-06-2022.

Insurify (2020). Insuring the American Driver. https://insurifycdn.com/files/state-of-auto-2020/Insuring_the_American_Driver.pdf. Accessed: 23-08-2022.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Application in R*. Springer, first edition.

Kalisch, M. and Bühlmann, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, 8:613–636.

Kalisch, M., Hauser, A., and Maechler, M. (2021). *Package 'pcalg'*. R package version 2.7-4.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding Discrimination through Causal Reasoning. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 656–666.

Kiviat, B. (2019). The Moral Limits of Predictive Pratices: The Case of Credit-Based Insurance Scores. *American Sociological Review*, pages 1–25.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2019). *Loss Model: From Data to Decisions*. Wiley, fifth edition.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford, first edition.

Li, J. and O'Donoghue, C. (2013). A survey of dynamic microsimulation models: uses, model structure and methodology. *International Journal of Microsimulation*, 6(2):3–55.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2022a). Discrimination-free insurance pricing. *ASTIN Bulletin*, 52(1):55–89.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2022b). A multi-task network approach for calculating discrimination-free insurance prices. *arXiv preprint https://arxiv.org/abs/2207.02799*.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

McFadden, D. (1977). Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments. Cowles Foundation Discussion Papers 474, Cowles Foundation for Research in Economics, Yale University.

Miller, M. J. (2009). Disparate Impact and Unfairly Discriminatory Insurance Rates. Casualty Actuarial Society E-Forum.

Miller, T., Viner, J., Rossman, S., Pindus, N., Gellert, W., Douglass, J., Dillingham, A., and Blomquist, G. (1991). *The Costs of Highway Crashes (Report No. FHWA-RD-91-055)*. National Highway Traffic Safety Administration.

Mosley, R. and Wenman, R. (2022). Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance. Casualty Actuarial Society Research Paper Series on Race and Insurance Pricing.

NAACP v. American Family Mut. Ins. Co. (1992). 978 F.2d 287 (7th Cir. 1992).

NAIC (2009). *The Use of Credit-Based Insurance Scores.* Hearing held June 15, 2009, Minneapolis, MN.

NAIC (2010). Property and Casualty Model Rating Law. https://content.naic.org/sites/default/files/inline-files/GDL-1775.pdf. Accessed: 03-03-2021.

NAIC (2019). House Financial Services Subcommittee on Oversight and Investigations Hearing on Examining Discrimination in the Automobile Loan and Insurance Industries. https://www.naic.org/documents/government_relations_190501_statement_for_record_discrimination_auto.pdf. Accessed: 25-02-2021.

NAIC (2020a). Milestones in Racial Discrimination within the Insurance Sector. https://content.naic.org/sites/default/files/inline-files/HistoricalContextOutline_Final_0.pdf. Accessed: 22-04-2021.

NAIC (2020b). NAIC Announces Special Committee on Race and Insurance. https://content.naic.org/article/news_release_naic_announces_special_committee_race_and_insurance.htm. Accessed: 25-02-2021.

NAIC (2022). Credit-based insurance scores. https://content.naic.org/cipr-topics/credit-based-insurance-scores. Accessed: 07-10-2022.

National Center for Statistics and Analysis (2013). *National Automotive Sampling System (NASS) General Estimates System (GES) analytical user's manual 1988-2010 (Report No. DOT HS 811 532).* National Highway Traffic Safety Administration.

National Center for Statistics and Analysis (2019). *National Automotive Sampling System (NASS) General Estimates System (GES) analytical user's manual 1988-2015 (Report No. DOT HS 812 320).* National Highway Traffic Safety Administration.

National Center for Statistics and Analysis (2020). *Traffic safety facts 2018 annual report: A* *compilation of motor vehicle crash data (Report No. DOT HS 812 981).* National Highway Traffic Safety Administration.

National Center for Statistics and Analysis (2021). *Fatality Analysis Reporting System (FARS) analytical user's manual, 1975-2019 (Report No. DOT HS 813 023).* National Highway Traffic Safety Administration.

National Center for Statistics and Analysis (2022a). *2020 FARS/CRSS coding and validation manual (Report No. DOT HS 813 251).* National Highway Traffic Safety Administration.

National Center for Statistics and Analysis (2022b). *Crash Report Sampling System CRSS analytical user's manual 2016-2020 (Report No. DOT HS 813 236).* National Highway Traffic Safety Administration.

National Center for Statistics and Analysis (2022c). *Fatality Analysis Reporting System (FARS) analytical user's manual, 1975-2020 (Report No. DOT HS 813 254).* National Highway Traffic Safety Administration.

National Safety Council (2017). *Manual on Classification of Motor Vehicle Traffic Accidents (ANSI D 16.1-2017).* Eigth Edition.

Neal, B. (2020). *Introduction to Causal Inference from a Machine Learning Perspective.* Course Lecture Notes.

NHTSA (2021). 2020 fatality data show increased traffic fatalities during pandemic. https://www.nhtsa.gov/press-releases/2020-fatality-data-show-increased-traffic-fatalities-during-pandemic. Accessed: 29-09-2022.

O' Neil, C. (2016). *Weapons of Math Destruction.* Crown Publishers, New York, NY, USA, first edition.

Oh, R., Jeong, H., Ahn, J. Y., and Valdez, E. A. (2021). A multi-year microlevel collective risk model. *Insurance: Mathematics and Economics*, 100:309–328.

Ohlsson, E. (2010). *Non-Life Insurance Pricing with Generalized Linear Models.* Springer-Verlag Berlin Heidelberg, first edition.

Osborne, J. W. (2015). *Best Practices in Logistic Regression.* SAGE Publications, first edition.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

Pearl, J. (2009b). *Causality: Models, Reasoning and Inference.* Cambridge University Press, second edition.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer.* John Wiley & Sons, first edition.

Pebesma, E., Bivand, R., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Sumner, M., MacQueen, D., Lemon, J., Lindgren, F., O'Brien, J., and O'Rourke, J. (2020). *Classes and Methods for Spatial Data.* R package version 1.4-5.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms.* The MIT Press, first edition.

Pope, D. G. and Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, 3(3):206–231.

ProPublica (2017). Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas With the Same Risk. https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk. Co-published with Consumer Reports. Accessed: 30-11-2022.

Qureshi, B., Kamiran, F., Karim, A., Ruggieri, S., and Pedreschi, D. (2020). Causal inference for social discrimination reasoning. *Journal of Intelligent Information Systems*, 54:425–437.

Rawls, J. (2001). *Justice as Fairness: A Restatement*. Harvard University Press, second edition.

Ripley, B. and Venables, W. (2021). *Feed-Forward Neural Networks and Multinomial Log-Linear Models*. R package version 7.3-16.

Robinson, R. W. (1977). Counting unlabeled acyclic digraphs. In *Combinatorial Mathematics V*, pages 28–43, Berlin, Heidelberg. Springer Berlin Heidelberg.

Romei, A. and Ruggieri, S. (2013). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638.

Rothstein, R. (2017). *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright Publishing Corporation, first edition.

Schutt, R. K. (2011). *Investigating the Social World: The Process and Practice of Research*. Sage, seventh edition.

Scism, L. and Maremont, M. (2010). Insurers Test Data Profiles to Identify Risky Clients. *The Wall Street Journal*. https://www.wsj.com/articles/SB10001424052748704648604575620750998072986. Accessed: 25-02-2021.

Scot J. Paltrow (2000). Insurers Stopped Offering Dual Rates In the '60s, but Didn't Tell Customers. https://www.wsj.com/articles/SB956793999358472506. Accessed: 22-04-2021.

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

SOA (2022). Actuarial Toolkit: Risk Factor. https://actuarialtoolkit.soa.org/tool/glossary/risk-factor. Accessed: 01-02-2022.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search*. The MIT Press, second edition.

Stanton, J. F. (2002). The Fair Housing Act and Insurance: An Update and the Question of Disability Discrimination. *Hofstra Law Review*, 31(1):141–206.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(307).

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25).

Tutz, G., Pößnecker, W., and Uhlmann, L. (2015). Variable selection in general multinomial logit models. *Computational Statistics and Data Analysis*, 82:207–222.

United States v. American Family Mut. Ins. Co. (1995). C.A. No 95-C-0327, E.D. Wisc. 1995.

United States v. Erie Ins. Co. (2008). C.A. No 08-CV-0945-S, W.D.N.Y 2008.

United States v. Nationwide Mut. Ins. Co. (1997). C.A. No C2-97-291, S.D. Ohio 1997.

U.S. Census Bureau (2015). ZCTA Brochure. https://www2.census.gov/geo/pdfs/education/brochures/ZCTAs.pdf. Accessed: 19-02-2021.

U.S. Census Bureau (2018). DP05 ACS Demographic and housing estimates. 2018 American Community Survey 5-Year Estimates. ftp.census.gov. Accessed: 19-12-2019.

U.S. Census Bureau (2020). Differences between the ACS and the decennial census. https://www.census.gov/content/dam/Census/library/publications/2020/acs/acs_general_handbook_2020_ch09.pdf. Accessed: 08-04-2021.

U.S. Census Bureau (2021a). 2010 census urban and rural classification and urban area criteria. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html. Accessed: 23-08-2022.

U.S. Census Bureau (2021b). B25001: Universe of housing units in the 2020 ACS 5-year estimates. https://data.census.gov/cedsci/table?q=Housing%20Units&tid=ACSDT5Y2020.B25001. Accessed: 04-10-2022.

U.S. Census Bureau (2021c). Census bureau releases experimental 2020 american community survey 1-year data. https://www.census.gov/newsroom/press-releases/2021/experimental-2020-acs-1-year-data.html. Press Release Number CB21-TPS.139. Accessed: 22-09-2022.

U.S. Census Bureau (2021d). Sample Size. https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/sample-size/#note2020. Accessed: 04-10-2022.

U.S. Federal Home Loan Bank (1938a). Map of Milwaukee County, Wisconsin. https://collections.lib.uwm.edu/digital/collection/agdm/id/3028/rec/1. Accessed: 14-04-2021.

U.S. Federal Home Loan Bank (1938b). Text to accompany the Map of Milwaukee County, Wisconsin. https://uwm.edu/libraries/wp-content/uploads/sites/59/2015/03/Pages-1-59-Text-to-accompany-Map-of-Milwaukee-County-Wisconsin-residential-security-map-19381.pdf. Accessed: 14-04-2021.

USBLS (2022). Consumer price index. https://www.bls.gov/cpi/. Accessed: 20-07-2022.

Verbelen, R., Antonio, K., and Claeskens, G. (2017). Unraveling the predictive power of telematics data in car insurance pricing. *Available at SSRN: https://ssrn.com/abstract=2872112 or http://dx.doi.org/10.2139/ssrn.2872112*.

Vijverberg, W. (2011). Testing for IIA with the Hausman-Mcfadden Test. IZA Discussion Paper No. 5826, Available at SSRN https://ssrn.com/abstract=1882845.

Werner, G. and Guven, S. (2007). GLM Basic Modeling: Avoiding Common Pitfalls. *Casualty Actuarial Society Forum, Winter*, pages 257–272.

Werner, G. and Modlin, C. (2016). *Basic Ratemaking*. Casualty Actuarial Society.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

Wisconsin OCI (2020). 2020 Financial and Statistical Data. https://oci.wi.gov/Documents/AboutOCI/2020WisconsinInsuranceReport_Market_Share.pdf. Accessed: 01-02-2022.

WisDOT (2021). Traffic engineering, operations and safety manual. https://wisconsindot.gov/dtsdManuals/traffic-ops/manuals-and-standards/teops/13-05.pdf. Accessed: 21-09-2022.

Wolff, M. J. (2006). The Myth of the Actuary: Life Insurance and Frederick L. Hoffman's Race Traits and Tendencies of the American Negro. *Public Health Reports*, 121:84–91.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Xin, X. and Huang, F. (2022). Anti-discrimination insurance pricing: Regulations, fairness criteria, and models. *Available at SSRN: https://ssrn.com/abstract=3850420*.

Yasmin, S., Eluru, N., and Pinjari, A. R. (2015). Pooling data from fatality analysis reporting system (FARS) and generalized estimates system (GES) to explore the continuum of injury severity spectrum. *Accident Analysis and Prevention*, 84:112–127.

Zaloshnja, E., Miller, T., Cocuncil, F., and Persaud, B. (2006). Crash costs in the United States by crash geometry. *Accident Analysis and Prevention*, 39:644–651.

Zhang, F., Noh, E. Y., Subramanian, R., and Chen., C.-L. (2019a). *Crash Report Sampling System: Sample design and weighting (Report No. DOT HS 812 706)*. National Highway Traffic Safety Administration.

Zhang, F., Subramanian, R., Chen, C.-L., and Noh, E. Y. (2019b). *Crash Report Sampling System: Design Overview, Analytic Guidance, and FAQs (Report No. DOT HS 812 688)*. National Highway Traffic Safety Administration.

Zhang, Y. (2018). Assessing Fair Lending Risks Using Race/Ethnicity Proxies. *Management Science*, 64(1):1–20.

Appendix A

Variable Standardization

Over the period of study, some variable names and coding practices in both GES/CRSS and FARS have been changed by the NHTSA. In our database for the period 2001-2020, the variable names and coding practices were standardized to those in the 2020 FARS user's manual (National Center for Statistics and Analysis, 2022a,c). For the GES and CRSS databases, names and coding practices are standardized using National Center for Statistics and Analysis (2013, 2019, 2022b).

Variables from the GES/CRSS and FARS databases that we deemed relevant for insurance purposes or that are traditionally considered for insurance underwriting can be found in Table A.1. This table shows which variables are included in each of the GES/CRSS or FARS databases. In Table A.2 we show for each of the variables included in both GES/CRSS and FARS a brief definition, the standardized code categories along with some additional notes. For variables not included natively in the GES/CRSS database, we imputed them according to the description presented in Section 2.2.3. We remark that the variables in both Table A.1 and A.2 are available in the final database but some variables are not in the final statistical analysis presented in Section 2.4.

We grouped some variable categories to avoid data sparsity problems. All variables with unknown or not reported values are replaced with NaN. The records key identifiers in the databases, denoted by `ST_CASE`, `VEH_NO` and `PER_NO`, are not modified in any way. The number of records described in Table A.2 matches the number of records after the filtering procedure described in Section 2.2.1.

No.	Variable	FARS	CRSS	GES	No.	Variable	FARS	CRSS	GES
1	ST_CASE	x	x	x	18	BODY_TYP	x	x	x
2	VEH_NO	x	x	x	19	DEFORMED	x	x	x
3	PER_NO	x	x	x	20	SPEC_USE	x	x	x
4	AGE	x	x	x	21	TRAV_SP	x	x	x
5	GENDER	x	x	x	22	DR_ZIP	x	x	x
6	YEAR	x	x	x	23	CDL_STAT	x		
7	PER_TYP	x	x	x	24	PREV_ACC	x		
8	INJ_SEV	x	x	x	25	PREV_SUS	x		
9	DRINKING	x	x	x	26	PREV_DWI	x		
10	DRUGS	x	x	x	27	PREV_SPD	x		
11	HISPANIC	x			28	SPEEDREL	x	x	x
12	RACE	x			29	DR_SF1	x	x	x
13	NUMOCCS	x	x	x	30	STATE	x		
14	MAKE	x	x	x	31	COUNTY	x		
15	MODEL	x	x	x	32	HARM_EV	x	x	x
16	MOD_YEAR	x	x	x	33	HOUR	x	x	x
17	HIT_RUN	x	x	x	34	WEATHER	x	x	x

Table A.1: Variables deemed relevant for personal automobile insurance from the FARS, GES and CRSS databases. An ‘x’ denotes if the variable is natively included in the respective database.

Variable	Code categories	Standardization notes
AGE	Ranges in $[0, 97]$.	Discrete age categories. Due to historical coding practices, people aged 97 or older are coded as 97.
GENDER	1 = Male, 2 = Female.	Gender of the individual.
PER_TYP	1 = Driver, 2 = Passenger, 3 = SNO, 5 = Pedestrians, 6 = Pedalcyclists.	This variable describes the role of the individual. Stationary non-occupants (SNO) are people in a working vehicle, transport device used for assistance or recreation (such as wheelchairs or skateboards) or standing in buildings.
INJ_TYP	0 = No Injury, 1 = Possible Injury, 2 = Minor Injury, 3 = Serious Injury, 4 = Fatal Injury.	The 9,325 and 2,648 records in GES/CRSS and FARS, respectively, that were reported as injured but their injury severity is unknown (historically coded with 5) are not useful to quantify insurance losses. Therefore, these records were randomly reassigned with equal probabilities to the categories: possible injury, minor injury and serious injury.
DRINKING	0 = No, 1 = Yes.	This variable records whether the individual was recorded as having been drinking. In 2001-2008 the coding for no and yes was different in the GES dataset.
DRUGS	0 = No, 1 = Yes.	This variable records whether the individual was under the influence of drugs. In 2001-2008 the coding for no and yes was different in the GES dataset.

`RACE`	White, Black, Asian, Hispanic.	We define four major categories using both `HISPANIC` and `RACE` for the FARS dataset. Using `RACE`, we group Asian as those records with categories: Chinese, Japanese, Filipino, Asian Indian, Korean, Samoan, Vietnamese, Guamanian, Other Asian or Pacific Islander, Combined Other Asian. The variable `HISPANIC` is divided as Cuban, Mexican, Puerto Rican, Spanish, among others. To simplify, records with code values from 1 to 6 are categorized as 'Hispanic' and the variable `RACE` is ignored for these records. In practice, the Hispanic category is not mutually exclusive with race, but this is done to differentiate between Hispanics and non-Hispanics. Other race categories that are small in numbers are not included into the previous four major categories. Records with other race categories are: Native Hawaiian, Multiple Races, Other Indian or Other Race. For GES/CRSS and those records in FARS without a fatal injury, race is randomly assigned using 2018 ACS Data given the person's zip code, gender and age group.
`NUMOCCS`	Ranges in $[1, 80]$.	Discrete number of occupants in the vehicle.
`MAKE`	Ranges in $[1, 98]$	Discrete vehicle's make categories. Coding has been standard since 1988 and 1991 for GES/CRSS and FARS, respectively. In the FARS user's manual, code 77 corresponds to the make Victory which is omitted in both user's manual for GES/CRSS. Regardless, this code appears in 52 records for GES/CRSS, which we assume corresponds to Victory and therefore, omitted in the NHTSA notes.

`MODEL`	Ranges in $[1, 63]$.	Discrete vehicle's model categories. Models for 'non-standard' cars are recoded as NaN. FARS and CRSS have the same coding practice. GES uses the same as FARS for the period 2011-2015 but there is a different coding standard during 2001-2010. To standardize, the Make-Model tables were checked for the records that make up 80% of the data. Differences were standardized with some models of: Volkswagen, KIA and Oldsmobile.
`MOD_YEAR`	Ranges in [1900, 2021].	Discrete number for the vehicle's model year.
`HIT_RUN`	0 = No, 1 = Yes.	Different hit and run categories in FARS were grouped to the 'Yes' classification.

BODY_TYP	1 = convertible, 2 = 2-door sedan, 3 = (2,3)-door hatchback, 4 = 4-door sedan, 5 = (4,5)-door hatckback, 6 = station wagon, 7 = hatchback (unknown door number), 8 = sedan (unknown door number), 9 = other automobile, 10 = auto-based pickup, 11 = auto-based panel, 12 = large limousine, 13 = 3-wheel automobile, 14 = 3-door coupe, 15 = utility vehicles, 16 = van-based trucks, 17 = light trucks, 18 = buses, 19 = medium/heavy trucks, 20 = motorcycles, 21 = other vehicles.	Classification of the vehicle based on its configuration, shape, size and doors. For this variable the coding practices are equal in FARS, GES and CRSS. The codes 1 to 13 automobile (and automobile derivatives) categories left as in the historical coding practices.
DEFORMED	0 = no damage, 2 = minor damage, 4 = moderate damage, 6 = severe damage.	This variable records the amount of damage sustained by the vehicle. Coding values were different for GES during the period 2001-2008.
SPEC_USE	0 = no special use, 1 = special use.	Example of a vehicle with a special use are taxi, military vehicle, police vehicle, ambulance, fire truck, among others.
TRAV_SP	Ranges in $(0, 97)$.	Discrete number for travel speeds in miles per hour. Values greater than 96 coded as 97. The maximum speed codes have changed over the years for the three datasets.

DR_ZIP	XXXXX numeric.	Driver's address U.S. zip codes. Non-residents of the U.S. coded as NaN.
SPEEDREL	0 = No, 1 = Yes.	This variable records whether the driver's speed was related to the crash. If positive, it means any of the following: the speed was greater than reasonable or prudent (not necessarily over the limit), driving too fast for conditions, speed above the speed limit or racing. Different speed related categories in all datasets grouped to the 'Yes' classification. FARS data prior to 2009 did not include this variable and instead, the variable DR_SF1 had speeding categories with codes 43, 44 and 46. Thus, from 2001 to 2008, the aforementioned codes are standardized to have a value of SPEEDREL equal to 1.
DR_SF1	0 = None, 1 = Careless driving, 2 = Police related, 3 = Miscellaneous.	Factors related to the driver expressed in the case materials. Careless driving includes: improper driving, road rage or driving in an emotional state (fatigued, depressed, among others). Police related factors include: police pursuit, alcohol and or drug test refused and non-traffic violation charged (manslaughter, homicide, among others). Since 2020, this variable is now collected in a separate excel file.
HARM_EV	1 = Collision with MVT, 2 = Non-collision, 3 = Collision with object not fixed, 4 = Collision with fixed object.	This data element describes the first injury or damage producing event of the crash. MVT stands for motor vehicle in transport. Non-collision includes rollover, fire or explosion, gas inhalation, surface irregularities, among others. GES coding was different for the period 2001-2010, then it was standardized to the FARS and CRSS coding practice.
HOUR	Ranges in [0, 23].	Discrete number denoting the hour of the accident. Accidents that occurred at 12:00 am standardized to 0 hours.

`WEATHER`	0 = Clear, 1 = Atmospheric condition.	An ‘atmospheric condition’ includes rain, snow, cloudy, fog/smoke, sand, among others.

Table A.2: Variable definitions, code categories and standardization notes for variables from the GES/CRSS and FARS databases deemed relevant for actuarial applications.

Appendix B

ZIP Code Tabulation Areas

ZIP Code Tabulation Areas (ZCTAs) are generalized areal representations of U.S. Postal Service (USPS) ZIP codes. The USPS ZIP codes are not geographical areas, they are a collection of mail delivery routes. The algorithm for the creation of ZCTAs by the Census Bureau is as follows:

1. In a census block, the most repeated ZIP code was assigned to the entire census block as the preliminary ZCTA code.

2. Blocks are aggregated by code to create larger areas.

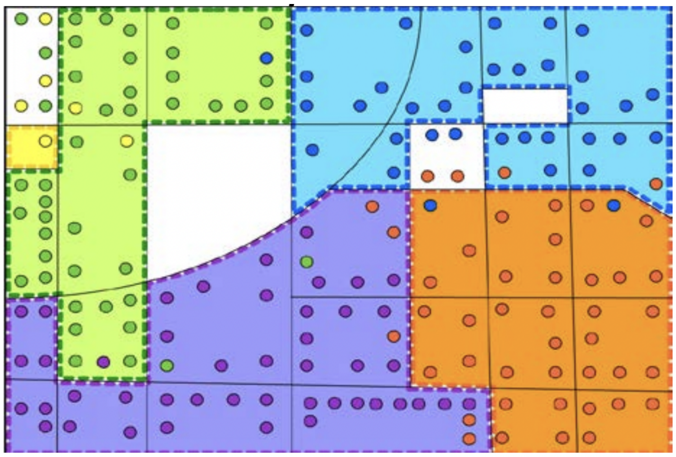


Figure B.1: Step 1 and 2 for the creation of ZCTA codes. Each dot represents an address and each color represents a different ZIP code. Retrieved from U.S. Census Bureau (2015)

3. If a block did not have a single most frequent ZIP code then it is assigned to the ZCTA with which it shared the longest boundary.

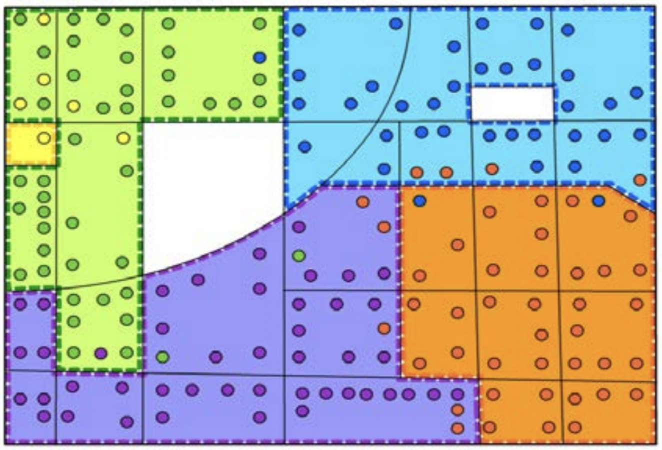


Figure B.2: Step 3 for the creation of ZCTA codes. The upper left block is assigned to the green ZCTA. Retrieved from U.S. Census Bureau (2015)

4. Unassigned enclaves less than 2 square miles are assigned to the surrounding ZCTA based on the length of shared boundary.

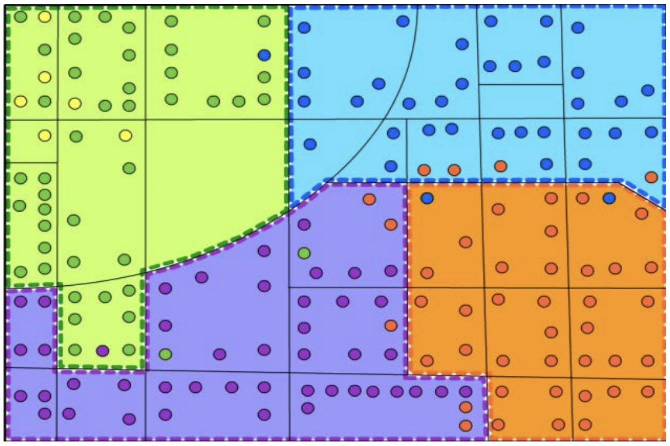


Figure B.3: Step 4 for the creation of ZCTA codes. The two unassigned enclaves are assigned to the ZCTA based on the length of shared boundary. Retrieved from U.S. Census Bureau (2015)

Appendix C

Discrimination-free Premium Example

To illustrate the discrimination-free pricing formula, and the bias correction methods, consider the following simple example. Suppose there is one discriminatory covariate with two groups, $D \in \{d_1, d_2\}$, and one nondiscriminatory covariate with two categories, $X \in \{x_1, x_2\}$. The losses and probabilities associated with the covariates are expressed in the following tables. The left-hand table gives the joint probability distribution for (D, X) and the right-hand table gives the conditional expected loss, $\mu(Y|X, D)$, for all combinations of X and D.

Probabilities

	d_1	d_2	$\mathbb{P}(X)$
x_1	0.20	0.40	0.60
x_2	0.35	0.05	0.40
$\mathbb{P}(D)$	0.55	0.45	

Expected Loss,
$\mu(Y|X, D)$

	d_1	d_2
x_1	100	200
x_2	50	70

We see a direct impact of the discriminatory covariate, in that costs for the d_2 group are higher than for the d_1 group for each value of X. We also see an indirect impact, in that the d_1 group is more likely to be in the x_2 group than the x_1 group ($\mathbb{P}(D = d_1|X = x_2) = 0.875$), and vice versa for the d_2 group ($\mathbb{P}(D = d_2|X = x_1) = 0.667$).

Using the $\mu(Y|X, D)$ values for the premiums would be directly discriminatory, as premiums are different for the d_1 and d_2 groups, even when they have the same X covariate.

Note that the average premium paid by the d_1 group is $\mathbb{E}[Y|X, d_1] = 68.1$, compared with $\mathbb{E}[Y|X, d_2] = 185.5$ for the d_2 group.

We can avoid direct discrimination by using an unawareness premium, $\mathbb{E}[Y|X = x]$. This gives

$$\mathbb{E}[Y|X = x_1] = 100 \times \frac{0.20}{0.60} + 200 \times \frac{0.40}{0.60} = 166.67$$
$$\mathbb{E}[Y|X = x_2] = 50 \times \frac{0.35}{0.40} + 70 \times \frac{0.05}{0.40} = 52.50$$

The expected income, using this premium, is

$$0.6\,\mathbb{E}[Y|X = x_1] + 0.4\,\mathbb{E}[Y|X = x_2] = 121.0 = \mathbb{E}[Y]$$

This approach avoids direct discrimination, as the premium does not vary with D, but there is an indirect discriminatory impact since the premium $\mathbb{E}[Y|x_2]$ is almost equal to $\mathbb{E}[Y|x_2, d_1]$ due to the fact that the d_1 group is more likely to be in x_2. A similar effect can be observed for the d_2 group given x_1. This means the unawareness premium, in spite of not using D, still allows inference of the discriminatory covariate due to the conditional distribution of $D|X$. The average premium paid by the d_1 group is 94.0, compared with 153.9 for the d_2 group.

Compare this with the discrimination-free premium from equation (3.2), which avoids indirect discrimination. We can arbitrarily choose a probability measure for D; suppose we use the real world measure, $\mathbb{P}(D)$. Then we have

$$h(x_1) = 100 \times 0.55 + 200 \times 0.45 = 145.0 < \mathbb{E}[Y|X = x_1]$$
$$h(x_2) = 50 \times 0.55 + 70 \times 0.45 = 59.0 > \mathbb{E}[Y|X = x_2]$$

The expected income, using this premium, is

$$0.6 \times h(x_1) + 0.4 \times h(x_2) = 110.6 < \mathbb{E}[Y]$$

This approach avoids both direct and indirect discrimination by not charging different premiums based on the discriminatory covariate and by not allowing inference of the discriminatory covariate given the nondiscriminatory covariate. As expected from the commentary in Section 3.1, there is a portfolio bias since the expected premium income is not equal to the expected loss. In order to ensure that the premium income is sufficient, on average, to meet the claims, we must make an adjustment. The three adjustment methods proposed in Section 3.1 are illustrated using the example above.

1. Using a fixed additive adjustment, c, to all the premiums, such that the expected premium is equal to the expected claims gives

$$(h(x_1)+c)\times 0.6+(h(x_2)+c)\times 0.4=121.0 \implies c=10.4$$
$$\implies h(x_1)+c=155.4; \qquad h(x_2)+c=69.4$$

   On average, members of the d_1 group pay 100.6, compared with 145.8 for the d_2 group.

2. Using a fixed proportional adjustment, $\gamma=(1+c)$, to all the premiums, such that the expected premium is equal to the expected claims, gives

$$h(x_1)(1+c)\times 0.6+h(x_2)(1+c)\times 0.4=121.0 \implies c=0.094$$
$$\implies h(x_1)(1+c)=158.6; \qquad h(x_2)(1+c)=64.5$$

   On average, members of the d_1 group pay 98.7, compared with 148.1 for the d_2 group.

3. Adjust the probability measure for D such that the expected income matches expected losses: let p_1^* denote the adjusted marginal probability for d_1, so $1-p_1^*$ is the adjusted marginal probability for d_2. Then we have

$$\begin{aligned}&\big(100\times p_1^*+200\times(1-p_1^*)\big)\times 0.6+\big(50\times p_1^*+70\times(1-p_1^*)\big)\times 0.4=121.0\\ &\implies p_1^*=0.397\\ &\implies h^*(x_1)=100\times 0.397+200\times 0.603=160.3\\ &\text{and }\ h^*(x_2)=50\times 0.397+70\times 0.603=\ 62.0\end{aligned}$$

   On average, members of the d_1 group pay 97.7, compared with 149.3 for the d_2 group.

Appendix D

Microsimulation Algorithm

In this appendix we explain in detail the data generation process and assumptions involved in the microsimulation algorithm. Each variable of $\mathcal{G}_m$ in Figure 4.2 and its corresponding assumptions are presented in the same order as in the code that sequentially generated a pseudo-population with selected characteristics. Examples of the distribution of some variables attributed to the pseudo-population are included for the state of Wisconsin with 2020 data. To specify the provenance, we show the table's ID code assigned by the USCB for those variables generated with parametric assumptions and empirical probabilities obtained from a table in the ACS.

D.1 Race

This categorical variable is generated from a multinomial distribution with parameters equal to the percentage that each race represents in a given state and year according to the ACS data, specifically in table DP05. The racial information from the ACS is based on self-identification. We considered four groups: White, Black, Asian and Hispanic. The concept of Hispanic origin is separate of the concept of race, which means people can self-identify as any race and to be of Hispanic origin. In this thesis, we group race and Hispanic origin; those who have a Hispanic origin are categorized as Hispanic regardless of the race they self-identify with.

Other race categories that are relatively small are not included into the previous four major categories. Examples of these other race categories are: Native Hawaiian, American Indian, or Other Race. The percentages of the four major categories were proportionally increased to match the population size of the state. For the state of Wisconsin, the four major race groups represented 97.2% of the population.

D.2 Age

Integer in the range $[18, 100]$. The variable matches the ACS data from table S0101, which is presented as the percentage of people that belong to a certain five-year age bracket, except for the last bracket called '85 years and over'. We assume a uniform distribution within each five-year age bracket and an exponential decay at a rate of 10% for people within the last age bracket.

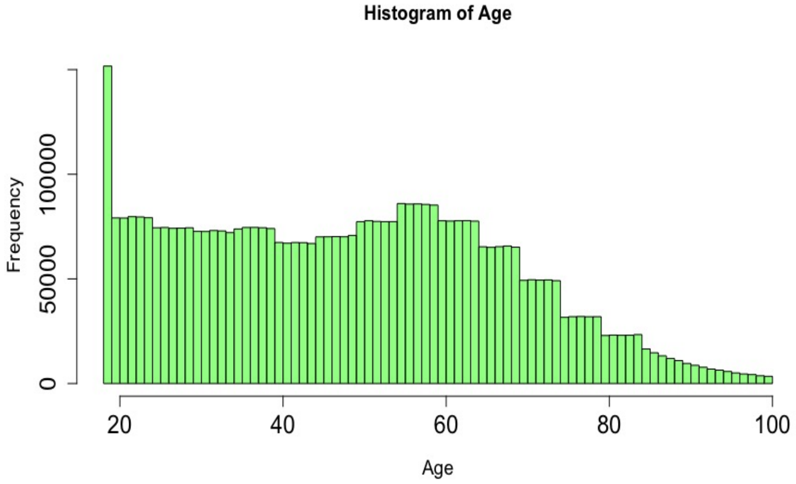


Figure D.1: Histogram of age for the pseudo-population based on the 2020 ACS data for the state of Wisconsin.

D.3 Gender

Categorical variable with possible labels {Male, Female}, generated from a binomial distribution with parameter equal to the percentage of men in table S0101 from the ACS.

D.4 Marital status

Categorical variable with possible labels {Single, Married}, generated from a binomial distribution with parameter equal to the percentage of currently married people according to table S1201 from the ACS.

D.5 Liability insurance in the last year

Categorical variable representing if the simulated individual had liability insurance while driving in the last year. The possible labels are {Yes, No} and it is generated from a binomial distribution with parameter equal to the percentage of uninsured motorists in that state according to Insurance Information Institute (2019).

D.6 Education level

Categorical variable with the five major groups considered in table S1501 from the ACS, which are: {Less than high school (LHS), High school (HS), College or associate degree (COL), Bachelor's degree (BACH), Graduate or professional degree (GRAD)}. In the generation DAG 4.2, education is based on the parents' zip code and socioeconomic group (SEG) which depends on race. For the generation of the pseudo-population the parent's zip code and SEG are not generated, but are included in $\mathcal{G}_m$ because we believe it is important to remark the influence that these variables have on people's education level. Thus, for data generation purposes education depends directly on race.

The five categories in table S1501 are available by race and presented as 3 groups: LHS, HS or COL, BACH or GRAD. In order to split the latter two groups into their specific category, we used the overall population numbers of HS, COL, BACH and GRAD. This implies we assume the percentage of people with education level equal to HS and BACH are the same across races. Another limitation is that the number of people with graduate studies or professional exams is only available for people age 25 and older. We assume the same percentages for people aged 18 to 24. The education level variable for the pseudo-population of Wisconsin calibrated with 2020 data is shown in absolute and relative terms in Figure D.2.

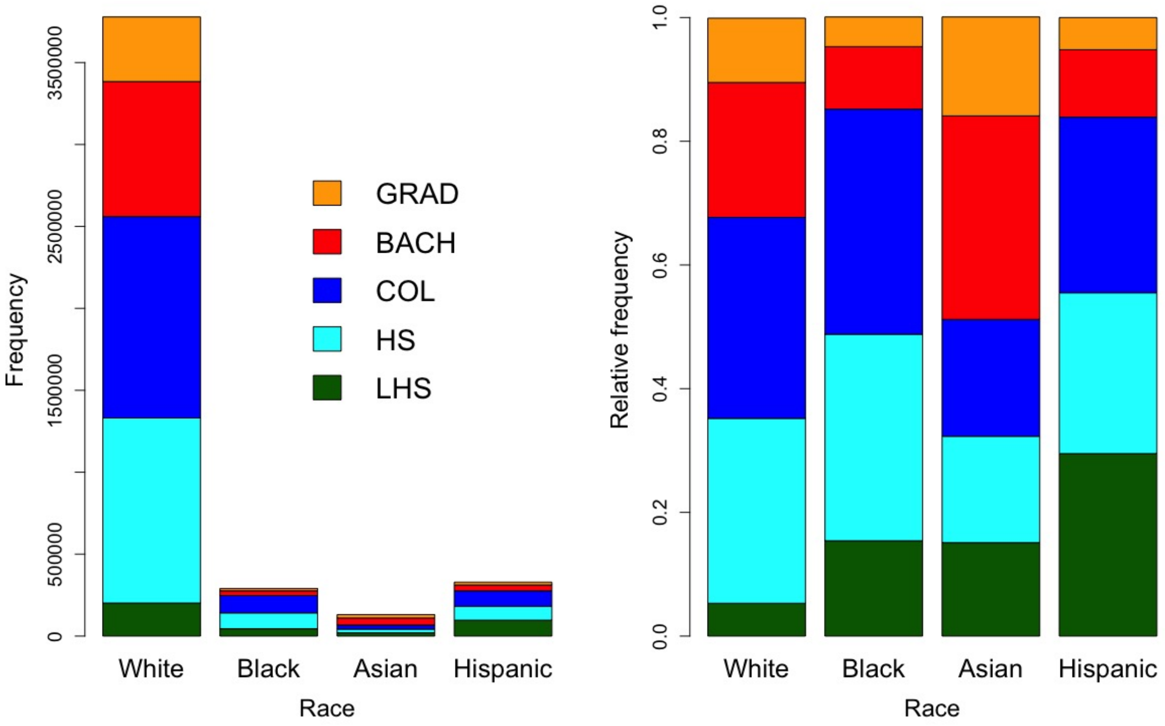


Figure D.2: Absolute frequency (left-hand side) of education level by race and its relative frequency (right-hand side) for the pseudo-population of Wisconsin in 2020.

D.7 Risk aversion

Risk aversion is a latent variable that is commonly assumed in insurance applications and it is modeled in different ways. This variable is meant to capture the policyholder's attitude towards the risk. For example, Werner and Modlin (2016) mention that in personal automobile insurance a policyholder that chooses high policy limits tend to be more risk averse. In Grari et al. (2022), also for a car insurance policy, the 'aggressiveness' of the policyholder is taken into account, which is a concept parallel to risk aversion.

We assume risk aversion can take three different categories, which are {Averse, Neutral, Seeker} and its assignment if a function of age, marital status and gender. We generate risk aversion using a multinomial distribution with parameters that vary across three groups: single men aged 18 to 25, any married policyholder, and all others. These parameters are shown in Table D.1 and the resulting distribution of risk aversion can be seen in Figure D.3.

Risk aversion	18-25 single men	Married	All others
Averse	25%	50%	25%
Neutral	25%	25%	50%
Seeker	50%	25%	25%

Table D.1: Risk aversion probabilities based on age, marital status and gender.

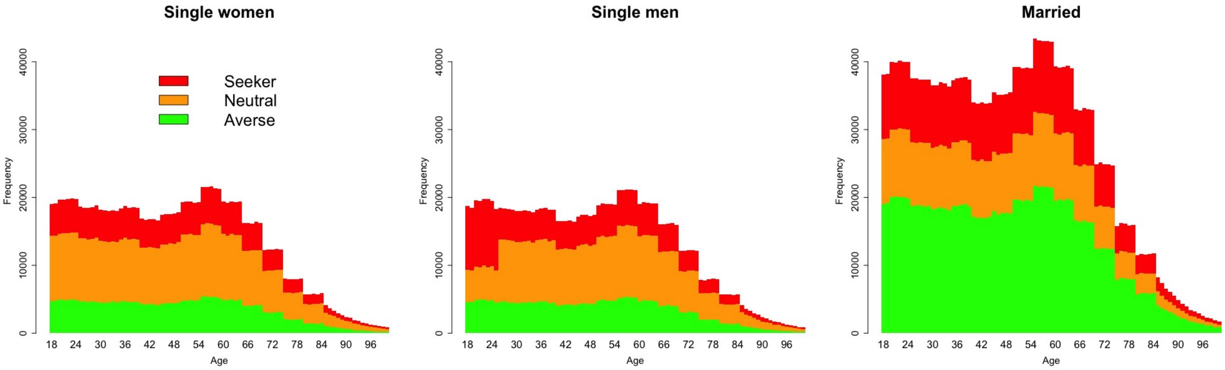


Figure D.3: Distribution of risk aversion by gender, marital status and age for the pseudo-population of Wisconsin in 2020.

D.8 Driving ability

Driving ability is another latent variable meant to classify the insurance portfolio by their skill at the wheel. This variable is commonly assumed a function of either the number of past accidents the driver has had, or the years of experience the driver has at the wheel. We assume three driving ability categories: {Low, Medium, High}, and its assignment depends directly on age. As can be seen in Figure D.4, we assume young and old drivers are more likely to have a low driving ability due to inexperience and decreasing motor ability, respectively, which tends to occur as part of the aging process.

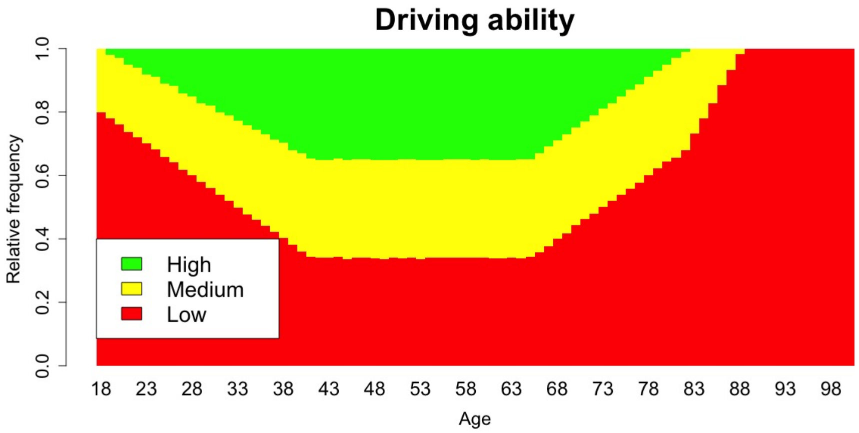


Figure D.4: Relative frequency of driving ability by age for the pseudo-population of Wisconsin in 2020.

D.9 Driving record

Driving record is a commonly asked question in personal automobile insurance policies. The insurer asks if the prospective policyholder has had in the past year one of the following: a car accident, an administrative offense, driven under the influence of alcohol or illegal drugs, or a speeding ticket.

We assume three possible categories for driving record: {None, Good, Bad}, and its assignment depends deterministically in variable D.5 and by a probabilistic relationship with driving ability. If the driver did not have liability insurance in the past year we directly assign no driving record. If the driver had liability insurance in the past year, then driving record gets sampled from a binomial distribution with parameters shown in Table D.2. The resulting distribution for Wisconsin is shown in Figure D.5.

Driving ability	Probability of good driving record
High	90.0%
Medium	83.3%
Low	66.6%

Table D.2: Probability of having a good driving record based on the driving ability of the policyholder.

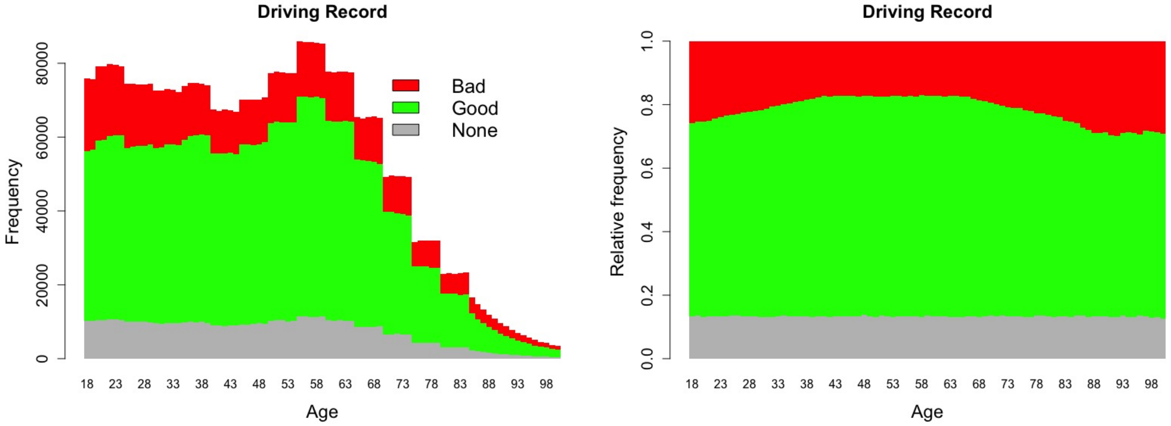


Figure D.5: Absolute (LHS) and relative (RHS) frequency of driving record by age (which has an indirect causal effect) for the pseudo-population of Wisconsin in 2020.

D.10 Income level

Categorical variable with the ten income groups considered in the ACS:

(1) '<10,000'

(2) '10,000 to 14,999'

(3) '15,000 to 24,999'

(4) '25,000 to 34,999'

(5) '35,000 to 49,999'

(6) '50,000 to 74,999'

(7) '75,000 to 99,999'

(8) '100,000 to 149,999'

(9) '150,000 to 199,999'

(10) '>200,000'

In $\mathcal{G}_m$ we assume the income level is a function of gender, education and age. In table S1501 from the ACS data, there is a variable named 'median earning in the past 12 months for 25 years old or more (inflation-adjusted)' which is available by gender and education for each zip code (we assume the same distribution for those aged 18 to 24). We make the simplifying assumption that earnings are equal to the income level. In general, earnings are higher or equal to income because the former considers other sources of revenue such as investments. This means that income level is overestimated for the pseudo-population in comparison to the target population.

The variable for median earnings in the past 12 months collected by the ACS is not paired with a dispersion measure, which we seek to account for by weighting through the population size for each possible combination of gender and education level. An example follows in Table D.3 for males with education level equal to LHS. The obtained empirical probabilities are used to sample an income level using a multinomial distribution by gender and education level.

MALE LHS	Population	Median earnings ($)
ZIP 1	100	8,340
ZIP 2	20	13,410
ZIP 3	2	24,820
	122	

Income level	Probability
<10,000	82%
10,000 - 14,999	16%
15,000 - 24,999	2%

Table D.3: On the left-hand side we have a fictional example of median earnings in the past 12 months on table S1501 as presented in the ACS. On the right-hand side, we have a weighted empirical probability of being in a subset of three income levels given gender and education level.

After assigning an income level based on gender and education level, we incorporate the following age assumptions:

- If education level is LHS or HS, and
  - age is less than 24: 60% probability of being bumped down 2 income levels.
  - age is 25 to 29: 60% probability of being bumped down 1 income levels.
- If education level is COL, BACH or GRAD, and
  - age is less than 24: 30% probability of being bumped down 2 income levels.
  - age is 25 to 29: 50% probability of being bumped down 1 income levels.
  - age is 30 to 39: 30% probability of being bumped down 1 income levels.
  - age is 45 to 49: 20% probability of being bumped down 1 income levels.

A simulation of the resulting distribution of income levels for the pseudo-population of Wisconsin in 2020 can be seen in Figure D.6. This assignment of income levels to policyholders is limited since we are using the average per zip code of the median earnings in the past 12 months. An example of these limitations can be seen in the top-right plot where there are zero female policyholders with an income greater than 200,000 dollars. This is not accurate as compared to what we would observe in the target population. Also, in reality, the connection between income level and education level is not as clear as this figure makes it seem, but it shows the strength and importance of the causal assumption made in DAG 4.2. The abnormal high relative frequency of males with the highest income level and less than high school is due to the fact that the COVID-19 pandemic disrupted ACS data collection in 2020 (U.S. Census Bureau, 2021c). Specifically in this case, the median earnings variable for zip code 53718 had a significant increase from a 9-year average of $23,384 to over $250,000 for males with less than high school.

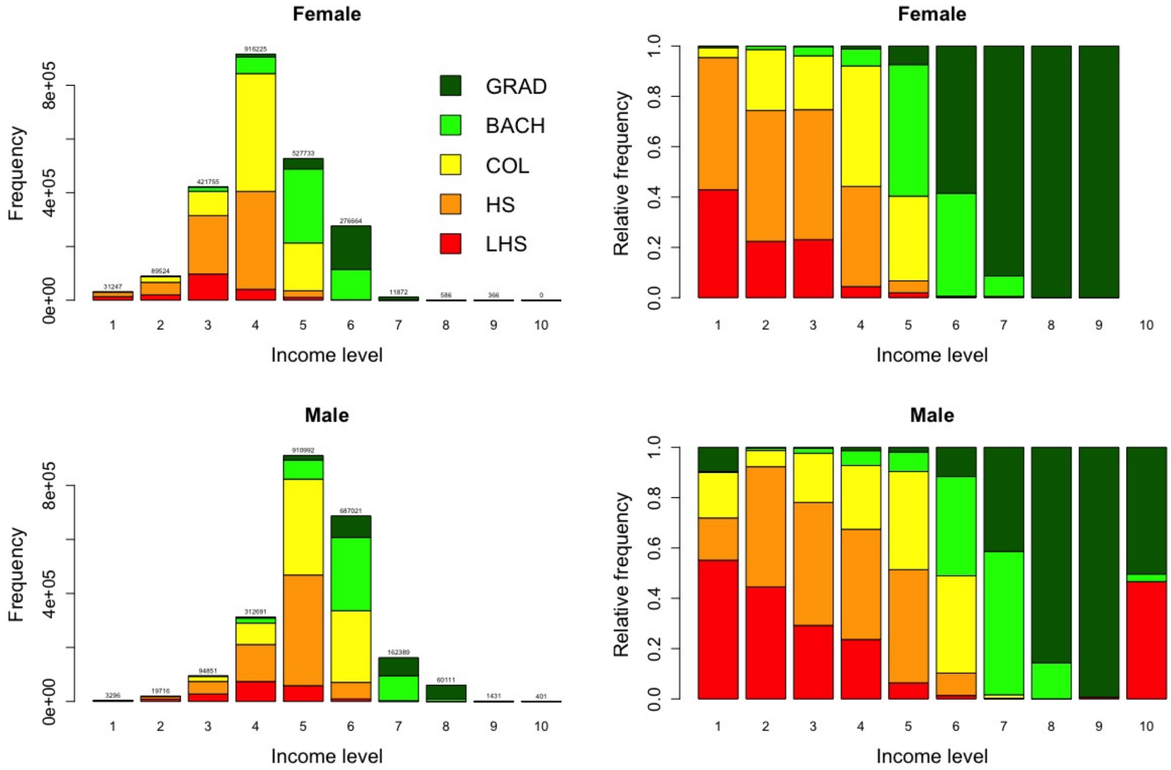


Figure D.6: Distribution of income levels for females (top row) and males (bottom row) by education level (absolute frequency in the LHS and relative frequency in the RHS).

D.11 Car: make-model-year

There are three separate categorical variables to denote the car's make, model and year. There are 6,089 possible combinations of these three variables from cars obtained from CarGurus (2022b). In $\mathcal{G}_m$, the car assigned to the pseudo-population is assumed a function of income. To do this, we use an ordered list of cars by their prices to create deciles, which we then match with an income level from those described in variable D.10. Once each person is matched through their income level with a car group, a random car (which is a combination of make-model-year) is assigned from that group. The top 5 makes and make-year combinations for the pseudo-population of Wisconsin of 2020 can be seen in Figure D.7.

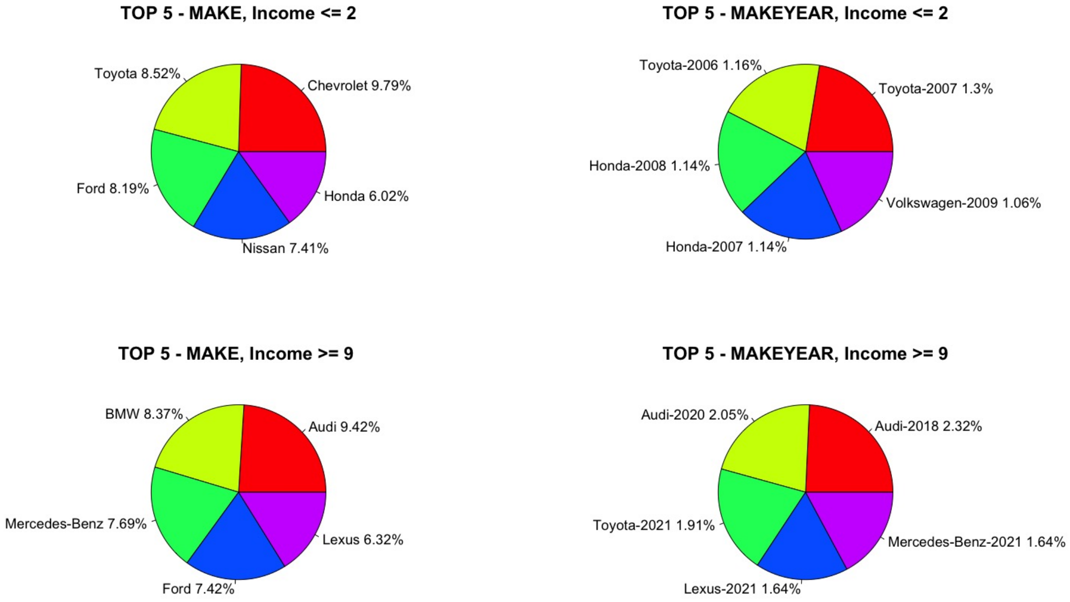


Figure D.7: Top 5 make and make-year combinations of the pseudo-population of Wisconsin of 2020. The top (bottom) row shows people with the lowest (highest) two income categories.

D.12 Zip code

Numeric integer with format XXXXX. In $\mathcal{G}_m$, zip code is assumed a function of race and income. This variable is generated using table DP05 and S1901 from the ACS. The former is used to obtain the total number of habitants by race and zip code, and then, this number is multiplied by the percentage of households in each income level – a value obtained from the latter table. This methodology assumes implicitly two things: (1) the percentage of households is equivalent to the percentage of people in each zip code, and (2) the percentage of households in each income level is uniform across races. A fictitious example to illustrate the methodology follows:

	DP05: Population				S1901: Households (%)	
	White	Black	Asian	Hispanic	<10,000	10,000 to 14,999
ZIP 1	100	10	20	50	50%	50%
ZIP 2	50	40	30	20	10%	90%

Table D.4: Total number of habitants by race and zip code along with the percentage of households in each income level as presented in tables DP05 and S1901 in the ACS.

ZIP - Income level	White	Black	Asian	Hispanic
ZIP 1 - <10,000	50	5	10	25
ZIP 2 - <10,000	5	4	3	2
ZIP 1 - 10,000 to 14,999	50	5	10	25
ZIP 2 - 10,000 to 14,999	45	36	27	18

Table D.5: Number of habitants by race for each zip code and income level after multiplying values in Table D.4.

Using numbers from Table D.5, we obtain empirical estimates of the probability of living in each zip code given race and income level. We use these values to sample a zip code for the pseudo-population using a multinomial distribution with these empirical probabilities from the ACS data. The distribution of the difference in race proportions for each zip code between the pseudo-population of Wisconsin of 2020 and the ACS data can be seen in Figure D.8. Considering our assumptions, the White population is overrepresented by 0.8% on average. While the minority races such as Black and Hispanic are underrepresented by

0.2% and 0.5% on average, respectively. The Asian pseudo-population of Wisconsin has no difference in proportion with the ACS data on average.

There are three sources of variability to explain the differences in Figure D.8. First, the income variable in table S1901 used to generate the empirical proportions of the multinomial sampling denotes an income category, while the variable, called income level, conditioned on to sample a zip code is based on median earnings in the last 12 months as explained in D.10. Second, the income level for the pseudo-population is adjusted with an assumption to add age dependency, also explained in D.10. The third reason is sampling variability from the multinomial distribution.

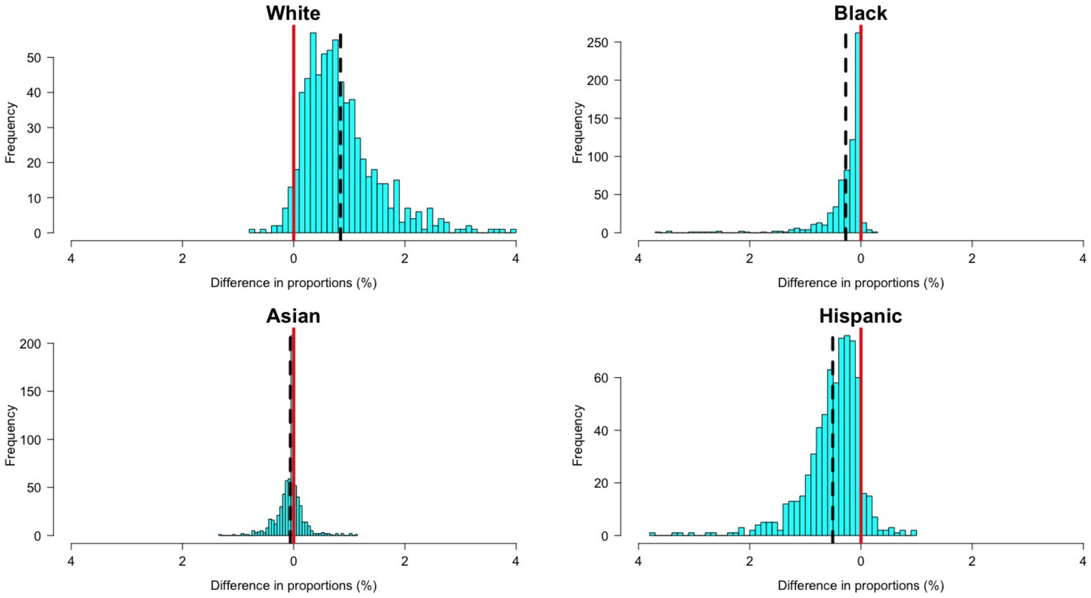


Figure D.8: Distribution of the difference in race proportions for each zip code between the pseudo-population of Wisconsin of 2020 and the ACS data. Dotted vertical line represents the mean of the difference vector and the solid vertical line denotes zero.

D.13 Mileage

This variable is commonly obtained in an insurance application form by requiring an estimate of the the annual mileage driven by the policyholder which is then categorized relative to what is considered a high mileage in that jurisdiction. We represent this variable through three categories: {Low, Regular, High} and is assigned as a function of the policyholder's zip code. We assume people in high density zip codes (such as cities and other urban areas) are more likely to have a lower mileage and vice versa. To do this, we classify zip codes as rural, suburban and urban based on their population density. For the state of Wisconsin, the geographical distribution of population density categories can be seen in Figure D.9. We classify as urban areas the 10% most densely populated zip codes of Wisconsin which includes those that comprise the cities of Milwaukee, Madison, Green Bay, Kenosha, Racine, among others. We categorized as suburban areas those zip codes that have a population density above the 75th percentile and below the 90th percentile. This criterion resulted in most suburban areas being neighbors to urban areas inside and outside of Wisconsin. The mileage variable is sampled for the pseudo-population using a multinomial distribution with the parameter vectors on Table D.6.

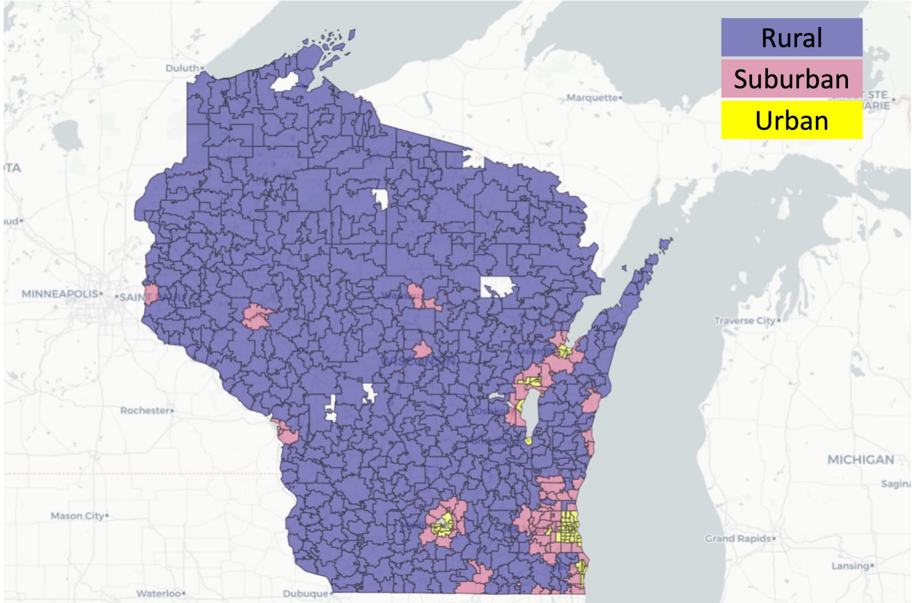


Figure D.9: Classification of zip codes into rural, suburban and urban categories based on population density of 2020 for the state of Wisconsin. Full interactive map available here.

D.14 Type of driving

This variable is assumed to be latent and categorical to denote the most frequent type of road driven by the policyholder. The categories are {Residential, Urban, Rural, Freeway}

	Low	Regular	High
Rural	10	30	60
Suburban	33	33	34
Urban	70	20	10

Table D.6: Multinomial distribution parameters (in percentages) to assign mileage categories based on the zip code density category.

and each has its own speed limit depending on the jurisdiction. For Wisconsin, the speed limits are 15, 25, 45 and 65 miles per hour, respectively (WisDOT, 2021). We assume that policyholders that drive more frequently in freeways than other types of roads tend to drive faster due to the higher speed limits. This categorical variable is considered given the importance that speeding has on predictive accuracy as shown for the observational data in Chapter 2.

In $\mathcal{G}_m$, type of driving is a function of zip code and mileage. For zip code we use the same population density classifications shown in Figure D.9. We assume the driver's address has a stronger influence than the annual mileage. This variable is simulated from a multinomial distribution with parameter distribution as shown in Figure D.10.

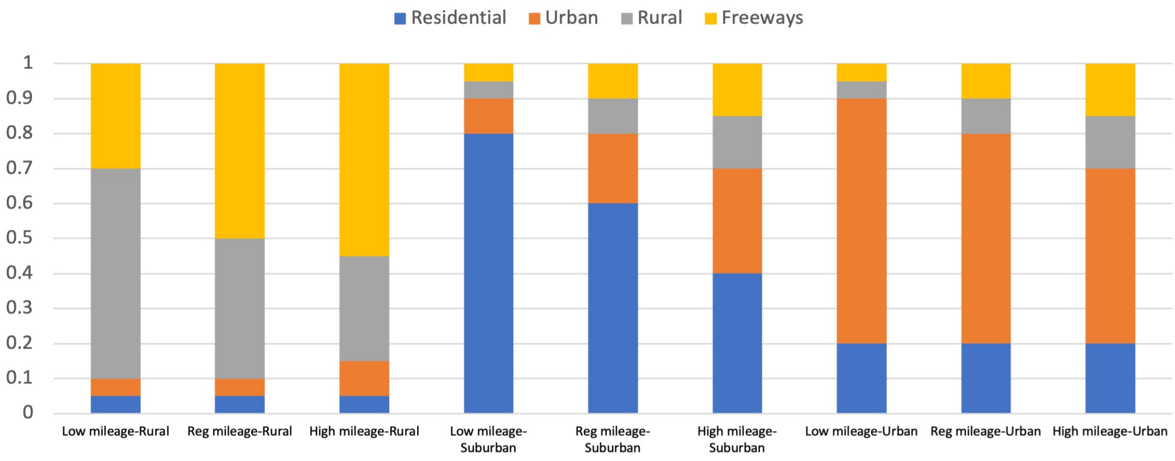


Figure D.10: Multinomial parameter distribution for type of driving based on zip code and mileage.

D.15 Severity

This response variable is described in detail in Section 4.3.3. This appendix is included to show the conditional probabilities $\mathbb{P}(\mathrm{CT}_{i,j} = k \mid \mathbf{Pa}(Y_{i,j}))$ in Table D.7 which is needed for the calculation of the expected value of severity.

		Crash type			
Type of driving	Risk aversion	Bump	Mild	Severe	Very severe
Residential	Averse	94.0%	4.0%	1.5%	0.5%
	Neutral	92.0%	5.0%	2.0%	1.0%
	Seeker	90.0%	6.0%	2.5%	1.5%
Urban	Averse	85.0%	9.0%	4.0%	2.0%
	Neutral	83.0%	10.0%	4.5%	2.5%
	Seeker	81.0%	11.0%	5.0%	3.0%
Rural	Averse	75.0%	15.0%	6.5%	3.5%
	Neutral	73.0%	16.0%	7.0%	4.0%
	Seeker	71.0%	17.0%	7.5%	4.5%
Freeway	Averse	55.0%	31.0%	9.0%	5.0%
	Neutral	53.0%	32.0%	9.5%	5.5%
	Seeker	50.0%	34.0%	10.0%	6.0%

Table D.7: Multinomial distribution parameters for the type of crash conditional on type of driving and risk aversion.

Appendix E

Microsimulation Statistical Test Results

E.1 Indirect Discrimination

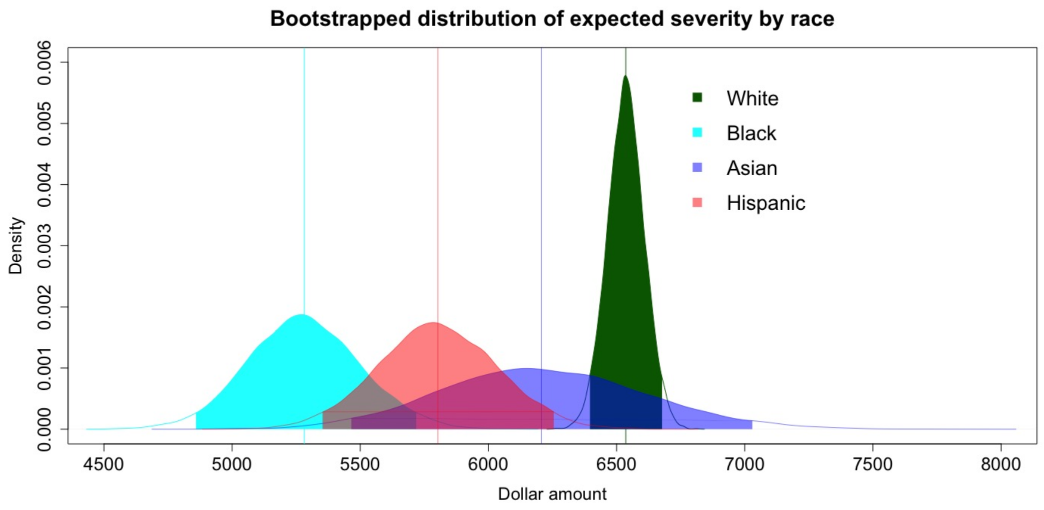


Figure E.1: Distribution of expected severity by race under $\mathcal{G}_1$, and obtained through 10,000 bootstrap samples with 95% confidence intervals.

Group number	Car value	White	Black	Asian	Hispanic
		95% Confidence interval			
1	10,000 or less	(4056, 4933)	(4479, 8060)	(2258, 6876)	(3103, 5324)
2	10,001 - 12,500	(5089, 5919)	(3349, 5431)	(3317, 8110)	(3846, 6114)
3	12,501 - 15,000	(5469, 6083)	(4315, 6333)	(3251, 6820)	(4214, 5997)
4	15,001 - 17,500	(5963, 6529)	(4290, 5966)	(3301, 6158)	(4656, 6529)
5	17,501 - 20,000	(6521, 7193)	(4451, 6513)	(4371, 7967)	(5130, 7472)
6	20,001 - 22,500	(7059, 7921)	(4332, 7143)	(5492, 9862)	(6731, 10,801)
7	22,501 - 25,000	(7795, 9016)	(4223, 7308)	(5222, 10,807)	(5670, 10,333)
8	25,001 - 30,000	(8192, 10,086)	(2911, 4547)	(8255, 19,132)	(6917, 16,210)
9	30,001 - 40,000	Unreliable estimates due to small sample size			
10	40,001 or more				

Table E.1: 95% confidence intervals of expected severity by race and car group under $\mathcal{G}_2$, obtained with 10,000 bootstrap samples.

E.2 Proxy Discrimination

The following 4 figures are the results of the lower-tailed one-sided WMW tests to conclude if there is proxy discrimination arising for the four pure premium estimators. The 588 pairwise tests for each estimator are presented by the 14 urban zip codes that are predominantly inhabited by a minority race and contrasted with respect to the 42 zip codes predominantly inhabited by White policyholders. The 42 zip codes are ordered – from left to right – by decreasing proximity to Milwaukee.

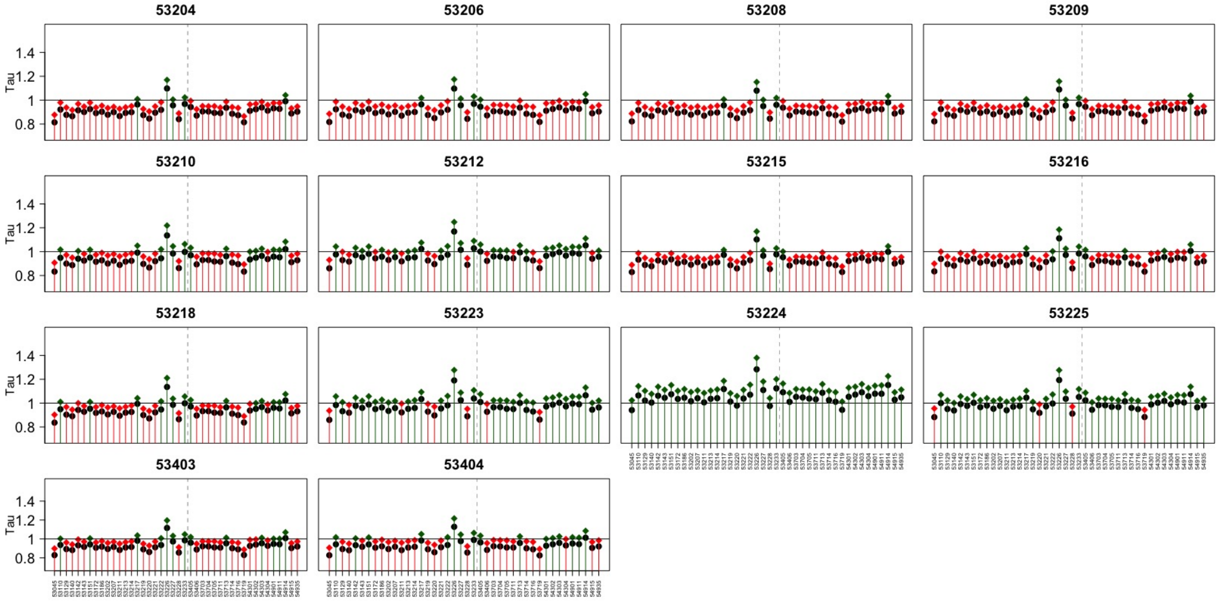


Figure E.2: Upper bound (diamond) and estimator (circle) of τ for each of the 588 pairwise proxy discrimination tests for $\hat{\mu}_i^{(\mathcal{G}_1)}$ with a 95% confidence level. Zip codes to the right of the vertical dotted line are outside of Milwaukee.

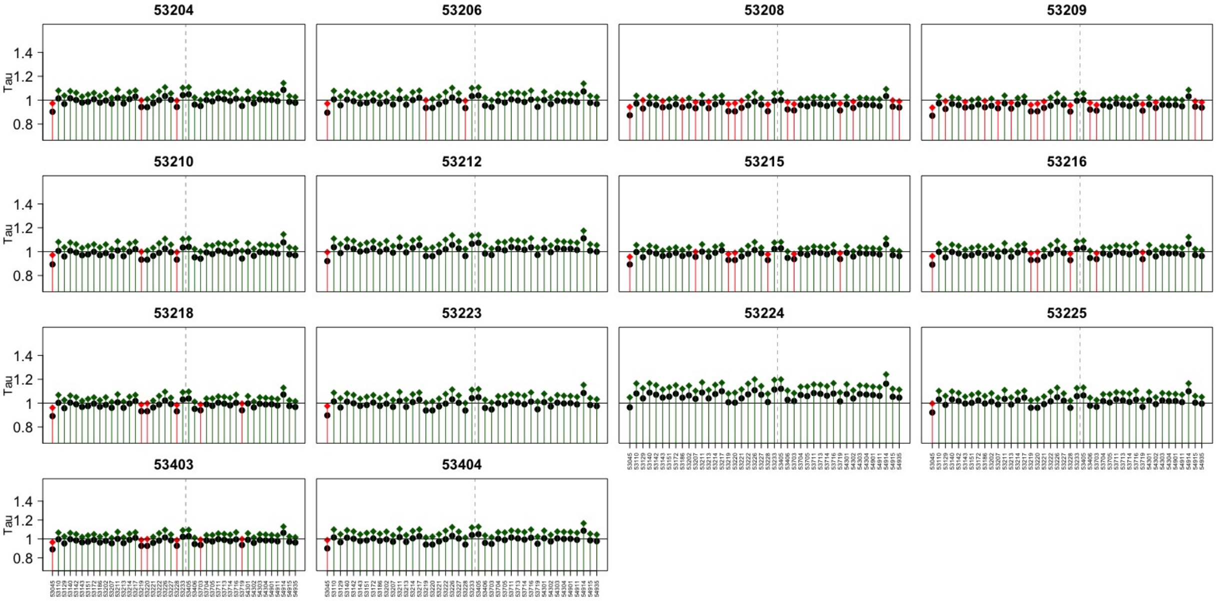


Figure E.3: Upper bound (diamond) and estimator (circle) of τ for each of the 588 588 pairwise proxy discrimination tests for $\hat{\mu}_i^{(\mathcal{G}_2)}$ with a 95% confidence level. Zip codes to the right of the vertical dotted line are outside of Milwaukee.

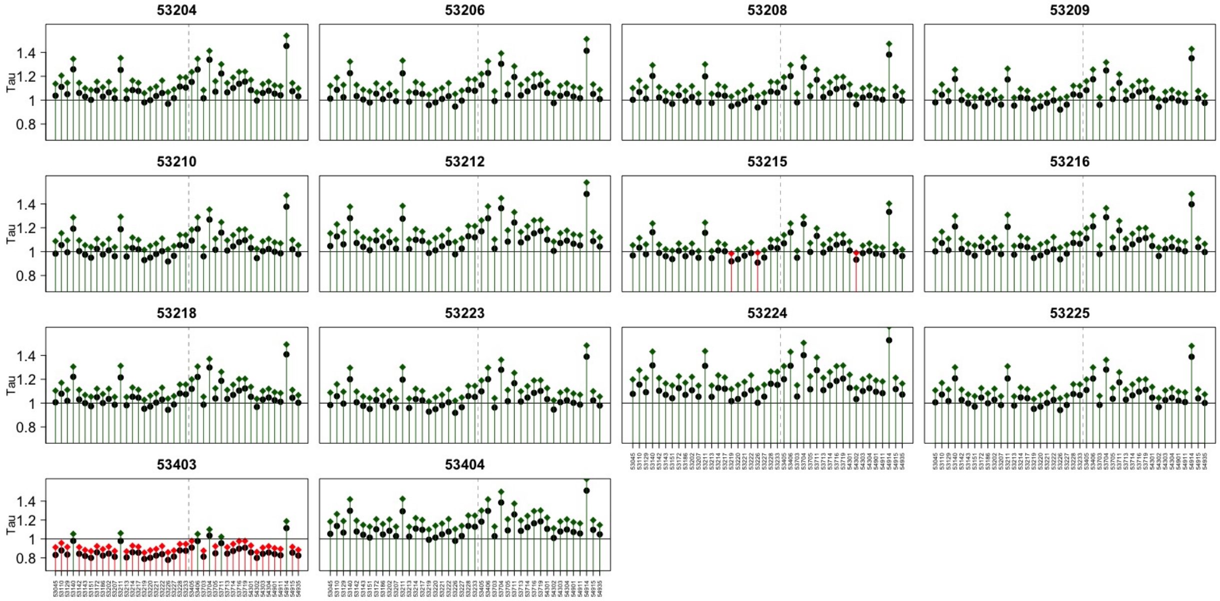


Figure E.4: Upper bound (diamond) and estimator (circle) of τ for each of the 588 pairwise proxy discrimination tests for $\hat{\mu}_i^{(\mathcal{G}_3)}$ with a 95% confidence level. Zip codes to the right of the vertical dotted line are outside of Milwaukee.

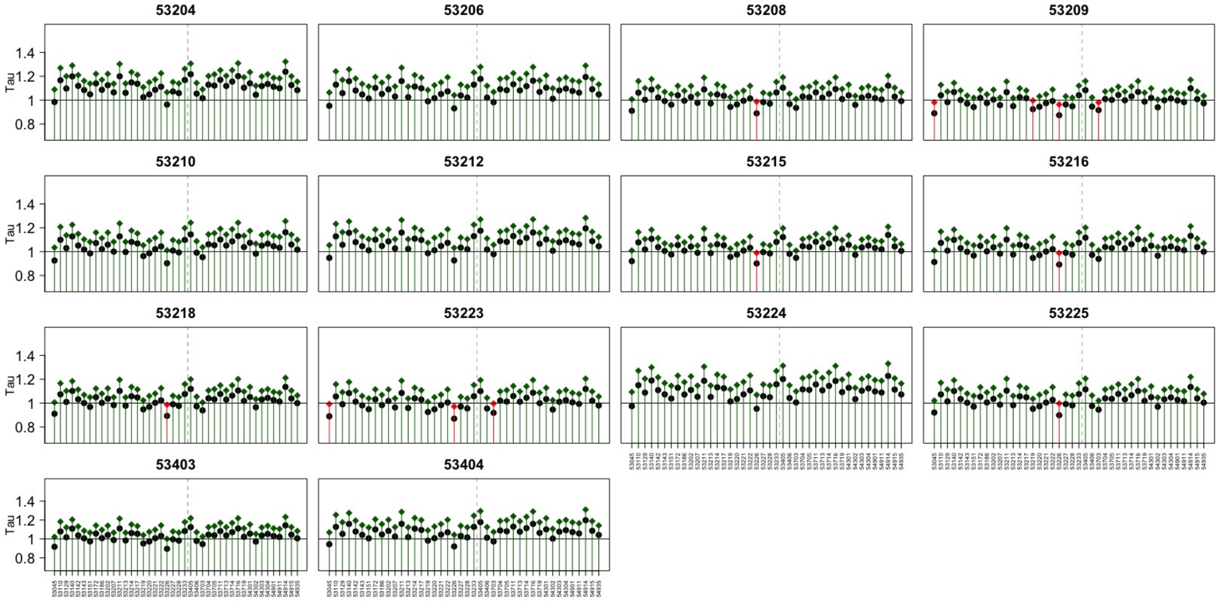


Figure E.5: Upper bound (diamond) and estimator (circle) of τ for each of the 588 pairwise proxy discrimination tests for $\hat{\mu}_i^{(\mathcal{G}_4)}$ with a 95% confidence level. Zip codes to the right of the vertical dotted line are outside of Milwaukee.