Optimal Driver Risk Modeling

Huiying Mao

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Feng Guo, Chair
Xinwei Deng, Co-chair
Inyoung Kim
Shyam Ranganathan

July 23, 2019
Blacksburg, Virginia

Keywords: Decision-adjusted Modeling, Driver Risk, Kinematic,
Naturalistic Driving Study, Rare Events, Traffic Safety

Optimal Driver Risk Modeling

Huiying Mao

(ABSTRACT)

The importance of traffic safety has prompted considerable research on predicting driver risk and evaluating the impact of risk factors. Driver risk modeling is challenging due to the rarity of motor vehicle crashes and heterogeneity in individual driver risk. Statistical modeling and analysis of such driver data are often associated with Big Data, considerable noise, and lacking informative predictors. This dissertation aims to develop several systematic techniques for traffic safety modeling, including finite sample bias correction, decision-adjusted modeling, and effective risk factor construction.

Poisson and negative binomial regression models are primary statistical analysis tools for traffic safety evaluation. The regression parameter estimation could suffer from the finite sample bias when the event frequency (e.g., the total number of crashes) is low, which is commonly observed in safety research. Through comprehensive simulation and two case studies, it is found that bias adjustment can provide more accurate estimation when evaluating the impacts of crash risk factors.

I also propose a decision-adjusted approach to construct an optimal kinematic-based driver risk prediction model. Decision-adjusted modeling fills the gap between conventional modeling methods and the decision-making perspective, i.e., on how the estimated model will be used. The key of the proposed method is to enable a decision-oriented objective function to properly adjust model estimation by selecting the optimal threshold for kinematic signatures and other model parameters. The decision-adjusted driver-risk prediction framework can outperform a general model selection rule such as the area under the curve (AUC), especially when predicting a small percentage of high-risk drivers.

For the third part, I develop a Multi-stratum Iterative Central Composite Design (miCCD) approach to effectively search for the optimal solution of any "black box" function in high dimensional space. Here the "black box" means that the specific formulation of the objective function is unknown or is complicated. The miCCD approach has two major parts: a multi-start scheme and local optimization. The multi-start scheme finds multiple adequate points to start with using space-filling designs (e.g. Latin hypercube sampling). For each adequate starting point, iterative CCD converges to the local optimum. The miCCD is able to determine the optimal threshold of the kinematic signature as a function of the driving speed.

Optimal Driver Risk Modeling

Huiying Mao

(GENERAL AUDIENCE ABSTRACT)

When riding in a vehicle, it is common to have personal judgement about whether the driver is safe or risky. The drivers' behavior may affect your opinion, for example, you may think a driver who frequently hard brakes during one trip is a risky driver, or perhaps a driver who almost took a turn too tightly may be deemed unsafe, but you do not know how much riskier these drivers are compared to an experienced driver. The goal of this dissertation is to show that it is possible to quantify driver risk using data and statistical methods.

Risk quantification is not an easy task as crashes are rare and random events. The wildest driver may have no crashes involved in his/her driving history. The rareness and randomness of crash occurrence pose great challenges for driver risk modeling. The second chapter of this dissertation deals with the rare-event issue and provides more accurate estimation.

Hard braking, rapid starts, and sharp turns are signs of risky driving behavior. How often these signals occur in a driver's day-to-day driving reflects their driving habits, which is helpful in modeling driver risk. What magnitude of deceleration would be counted as a hard brake? How hard of a corner would be useful in predicting high-risk drivers? The third and fourth chapter of this dissertation attempt to find the optimal threshold and quantify how much these signals contribute to the assessment of the driver risk. In Chapter 3, I propose to choose the threshold based on the specific application scenario. In Chapter 4, I consider the threshold under different speed limit conditions.

The modeling and results of this dissertation will be beneficial for driver fleet safety management, insurance services, and driver education programs.

Acknowledgments

I would like to express my sincere gratitude to my advisors Dr. Feng Guo and Dr. Xinwei Deng for their mentoring and guidance during my research. I could never do this without their consistent support.

Thank you to Dr. Inyoung Kim and Dr. Shyam Ranganathan for their time and advice as my committee members. Thanks for all the help, comments, positive input, and caring me all the time.

Thank you to the faculty and staff from the Department of Statistics at Virginia Tech for all the great courses and creating such a friendly and mutual-respectful community.

Thank you to the Virginia Tech Transportation Institute (VTTI) for providing the data, research opportunities, study space, and computing resources.

Thank you to my fellow students and friends for making my doctoral life be such a happy and memorable journey.

Thank you to my beloved fiancé for his company and encouragement all the time.

Finally, thank you to my family for their love and support to make it all possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Traffic safety is critical to people's everyday lives. In the year of 2016, over two thirds of the US population are licensed drivers, traveled a total of about 3.2 trillion miles. At the same time, there are up to $242 billion economic loss due to traffic accidents, including over 34,000 fatal crashes [62]. It is reported that motor vehicle traffic crash was the leading cause of death for the people between age 8 to 24 in the United States [63].

Driver is one key component in the transportation system. This dissertation focuses on modeling the risk of drivers from two aspects. The first aspect is accurately evaluating the impact of crash risk factors and safety countermeasures. Extensive research has been conducted on this aspect [22, 38, 64, 78]. For example, according to the World Health Organization, the risk factors for road traffic can be summarized as four categories: factors influencing exposure to risk, risk factors influencing crash involvement, risk factors influencing crash severity, and risk factors influencing severity of post-crash injuries [29]. For safety countermeasures, the US federal highway administration started to provide a list of infrastructure-oriented Proven Safety Countermeasures since 2008 and update it every four to five years [1].

The second aspect is driver-level risk prediction, e.g., predicting the probability of one driver being involved in a crash or predicting the crash occurrence rate for one driver. Some existing

research can be found at Guo and Fang [31] and Arbabzadeh and Jafari [6]. Explanatory variables play an important role in a driver-level risk prediction model. Explanatory variable is also named as factor, covariate, predictor, or feature in different contexts. Research interest and domain knowledge in transportation safety are often required to select the covariate candidates in the first place. Statistical methods and analysis tools then provide guidance in how the covariates should be constructed and which are the ones to keep in the model.

1.2 Naturalist driving data

Data is an important element of traffic safety modeling. As the data collection and storage technology greatly advance, it is feasible to establish a precise and optimal driver risk modeling procedure using the multi-dimensional and large-scale dataset. Currently, one major data resource is the crash databases: the Crashworthiness Data System, the General Estimation System (GES), the Fatality Analysis Reporting System (FARS), and various state-level crash databases. However, it is biased to look at only one side of the information – driving associated with crashes. Crash-free drivers and crash-free driving segments also matter to studying traffic safety.

Naturalistic driving study (NDS) provides researchers with a wealth of resources to have both the crash-related and crash-free information [30]. In an NDS, participants drive as they normally would. A data acquisition system (DAS) was installed in each participant's vehicle to continuously record data. Thanks to the development of in-vehicle data collection technology and lower cost, more and more NDSs are conducted worldwide, e.g., the 100-Car NDS, the second Strategic Highway Research Program (SHRP 2) NDS, and Europe's UDRIVE NDS [20, 21, 23]. These NDSs provide a great opportunity to address questions about driver performance and behaviour, as well as traffic safety [22, 36, 38, 61, 83].

The SHRP 2 NDS is the largest NDS so far with more than 3,500 drivers from six data collection sites in Florida, Indiana, North Carolina, New York, Pennsylvania, and Washington. The study collected a wide range of information for drivers under natural driving conditions. DAS for SHRP 2 includs radar, multiple cameras, accelerometers, and other equipments. The driving data were collected continuously from ignition-on to ignition-off, and the data were collected asynchronously; for example, videos at 10 Hz, GPS at 1 Hz, and acceleration at 10 Hz. The crashes in SHRP 2 were identified thorough a comprehensive process. An automated algorithm is first employed to screen through the kinematic time series data. The driving segments that were potentially crashes were manually examined through the videos to confirm whether a crash had actually occurred [34]. SHRP 2 categorizes the identified crashes into four severity levels: level 1 – the most severe crashes involving airbag deployment or potential injury, level 2 – moderate severe crashes, level 3– minor crashes in which the vehicle makes physical contact with another object or departs the road, and level 4 – tire strike crash. SHRP 2 also collected abundant drivers information, including demographic information, sleep habits, personality measure (from surveys they took at the beginning of the study), and self-reported driving history.

1.3 Statistical methods for driver risk modeling

Statistical modeling and analysis of driver-related data plays an important role for enhancing traffic safety. Two types of regression models are generally employed to evaluate driver risk. One is logistic regression, which considers binary response representing whether a driver has

been involved in a crash. That is,

$$Y = \begin{cases} 1, & \text{driver involved in at least one crash,} \\ 0, & \text{driver not involved in any crashes,} \end{cases}$$

Logistic regression assumes whether a driver has been involved in a crash follows a Bernoulli distribution, i.e.,

$$Y \sim \text{Bernoulli}(p),$$

where p is the probability of the driver being involved in a crash. Logistic regression models the relationship between covariates and the logit of probability

$$\log(\frac{p}{1-p}) = \text{logit}(p) = f(\boldsymbol{X}), \tag{1.1}$$

where $\boldsymbol{X}$ is the vector of explanatory variables.

The other type is Poisson and NB regression whose response are count-based, representing the number of crashes. Poisson and NB regression assume the count response follows a Poisson distribution, i.e.,

$$Y \sim \text{Poisson}(\lambda \cdot E),$$

where λ is a driver's crash occurrence rate and E is the corresponding exposure. The exposure can be either miles traveled or the driving time. The unit of crash occurrence rate follows the unit of exposure, which could be miles or hours depending on the application scenario. Poisson regression assumes the rate λ to be constant while NB regression assumes that λ follows a Gamma distribution. Similar to the logistic regression, Poisson and NB regression

estimates the relationship between covariates and the logarithm of crash rate.

$$\log(\lambda) = f(\boldsymbol{X}). \tag{1.2}$$

The logarithm of exposure works as an offset in model estimation. $f(\cdot)$ can be of different forms for model (1.1) and model (1.2). Models are linear models when $f(\cdot)$ is linear in terms of $\boldsymbol{X}$. Generalize Additive Model (GAM) can be implemented to estimate the non-linear relationship between covariates and the logit of p or the logarithm of λ. Maximum likelihood method (MLE) is often used for model estimation. Regularized maximum likelihood can be used for variable selection [27, 80, 90].

In this dissertation, "driver risk" is referring to either the crash occurrence rate (λ) or the probability of crash involvement. The selection of appropriate model type depends on the study of interest. If the crash occurrence rate is of more interest, Poisson and NB regression should be used. Logistic regression is used for estimating the probability of involving a crash. It does not account for the exposure as an offset as in the Poisson and NB models but the exposure can be used as a covariate in the logistic regression. It is intuitive that drivers with larger exposure are more likely to be involved in crashes.

In both two types of models, motor vehicle collisions, i.e. crashes, are the primarily safety measurement. However, *crashes are rare events*. For most of the driving time, there are no crashes happened. Only a small percentage of the drivers has crash experience. The rareness of crashes in a safety data depends on the study unit of interest. For example, the percentage of drivers have been involved in a crash in recent three months is smaller than that in a year. In a broader driving risk modeling scenario, if trips or short time segments are of interest, the percentage of trips/segments have a crash involved would be much smaller than that of drivers. The rareness of crash also depends on the crash severity. A more severe crash is

less likely to occur. It is reported an average of 1.13 fatalities per 100 million vehicle miles traveled, and the corresponding rate for crash with injury is about 70 times of the fatality rate [62].

Since crashes are rare events, accurate prediction and inference is of great challenge in the driver risk evaluation. The conventional regression methods often do not work properly in modeling rare events. For example, logistic regression underestimates the occurrence probability of an event when handling extremely unbalanced data, i.e., dominating number of zeros for the binary response [35]. Another key challenge is on appropriate performance measure for rare-event model evaluation. Accuracy is apparently unacceptable as the model can still achieve high accuracy even when predicting every instance as a nonevent [11, 18, 67]. A proper performance measurement should relate to the model performance under the specific study objective, which is closely connected to decision-making.

Accurate prediction of driver risk is challenging also due to the heterogeneity in individual driver risk. The rapid development of connected vehicles and automatic driving system technology offers great potential to improve traditional driver risk prediction through telematics data. Telematics data are usually high-frequency data, creating a data-rich environment. It brought challenges as well as opportunities. How to extract meaningful patterns from the telematics data and how to utilize the signatures in modeling with their most efficiency become urgent problems to solve. This dissertation uses kinematic signature as an example for the usage of telematics data.

1.4 Dissertation structure

The rest of this dissertation is organized as follows. Chapter 2 considers a bias adjustment strategy for the Poisson and NB regression to achieve more accurate estimation when eval-

uating the impacts of crash risk factors. Chapter 3 proposes a decision-adjusted driver risk prediction model, which incorporates how the model will be used in model estimation. The proposed model fills the gap between decision making and model estimation. In particular, chapter 3 provides guidance in determining the optimal threshold of kinematic signatures. Chapter 4 proposes an effective Multi-stratum Iterative Central Composite Design (miCCD) optimization algorithm to identify kinematic signature's optimal threshold under different driving speed conditions. Chapter 5 provides the concluding remarks.

Chapter 2

Adjusting Finite Sample Bias in Traffic Safety Modeling[1]

2.1 Introduction

Accurately evaluating the crash risk associated with transportation infrastructure and driver characteristics is essential to improving safety. Traffic safety is usually measured by crash frequency, which can be the number of crashes that occurred in a roadway segment, or the number of crashes a driver experienced, over a specified period (e.g., one year) [2]. As crashes are rare events, it is not uncommon to see crash-count data with low sample mean and excessive zero responses [49, 50]. Poisson and negative binomial (NB) regression have been the fundamental statistical analysis tools for count data; however, a small number of events brings challenges to parameter estimation and inference. This chapter focuses on the finite sample bias for parameter estimation caused by a small number of events when fitting a Poisson or NB regression model.

Poisson and NB regression models are important methods in transportation safety studies. For instance, safety performance function, an essential tool to evaluate crash risk provided by *The Highway Safety Manual*, is based on the Poisson/NB regression [2]. NB regression assumes that the rate in a Poisson model follows a Gamma distribution and can accommodate

[1] A modified version of this chapter has been published at *Accident Analysis and Prevention*

over-dispersion, a common issue for crash-count data. The coefficient estimates of the two models can be used to evaluate risk factors associated with crashes.

There are other models developed for transportation safety studies by relaxing certain assumptions in the Poisson and NB regressions. For example, the generalized estimating equation (GEE) and random/mixed effect models can be used when data violate the independence assumption. The GEE model takes into account the within-subject correlation, such as spatial-temporal correlation or observations with repeated measures, by an empiric covariance matrix [52, 53]. The random/mixed effect model assumes a distribution for the unobserved effect over subjects, and it can handle multiple sources of correlation [15, 31, 32]. Semi-parametric models and generalized additive models relax the linear assumption of Poisson and NB models to accommodate more-complicated relationships between crash rate and risk factors [45, 87].

As crashes are rare events, the number of crashes for a specific road segment or a driver is usually limited, leading to a low average occurrence rate and/or a large number of zeros in the crash-count data [41, 71]. The low event frequency can cause biased estimation and inaccurate inference for the count models [25, 47, 51]. Lord [46] showed that a low sample mean and small sample size can seriously affect the estimation of the dispersion parameter of the NB model. The commonly used goodness-of-fit test statistics (scaled deviance and Pearson's χ^2) are also inappropriate for the low mean value problem [84]. When excessive zero responses exist in the dataset, both the Poisson model and the NB model will produce inaccurate prediction [48]. For a thorough review of the low-occurrence problem in transportation safety, please refer to Lord and Mannering [50] and Lord and Geedipally [49].

Zero-inflated models, proposed by Lambert [42], are commonly used in analyzing crash counts with an excessive number of zeros [4, 5, 14, 41, 44, 60, 66, 70, 71, 77]. Nevertheless, they have also been subject to several criticisms [54, 55, 57]. In fact, Lord et al. [54] argued

that the assumed safe period with an event rate being zero in zero-inflated models does not reflect the actual crash-generating process. Alternatively, many methods have been proposed and applied. For instance, Malyshkina and Mannering [56] proposed a zero-state Markov switching model and Park and Lord [65] applied finite mixture models to such datasets. Recently, researchers have proposed more-flexible models, including Sichel (SI), Negative Binomial-Lindley (NB-L), Negative Binomial-Generalized Exponential (NB-GE), and Negative Binomial-Crack (NB-CR). These models incorporate more parameters into the underlying count distribution so that the extra degree of freedom can account for the excessive number of zeros [48, 49, 82, 91, 92]. These models may better fit the data, but they are hard to estimate and difficult for practitioners to understand.

Poisson and NB regression models are generally estimated using the maximum likelihood method. The resultant estimators for the unknown parameters are called maximum likelihood estimators (MLEs). The MLEs have the consistency property that ensures the MLEs converging to the true values when the sample size is sufficiently large. However, McCullagh and Nelder [59] showed that the MLEs could be biased and the bias is not negligible for a modest sample size. When the number of events is limited, a bias adjustment to the MLEs can improve the parameter estimation. Generally, there are two types of approaches to bias reduction for MLEs. One approach is based on applying the Jefferys invariant prior to the likelihood function to directly generate an improved estimator [24, 39, 40]. The other approach reduces the bias by subtracting the approximated bias from the regular MLE [16, 43, 59]. McCullagh and Nelder [59] determined a specific correction formula for the coefficient estimation of generalized linear models (GLMs).

The finite sample bias of Poisson and NB regression models has been sporadically investigated in the literature [28, 69]. Saha and Paul [69] studied the bias-corrected dispersion parameter estimation of the NB regression, which showed less bias and superior efficiency compared to

the MLE, the method of moments estimator, and the maximum extended quasi-likelihood estimators in most instances. Giles and Feng [28] derived the bias-correction formula for the parameter estimation of Poisson regression from Cox and Snell's general definition of residuals. Although considerable research has been devoted to reducing the bias of MLE under Poisson and NB models, limited research has been conducted in transportation safety to identify the situations where the bias correction is necessary and factors affecting the magnitude of bias.

This chapter aims to study the finite sample bias for the parameter estimation of Poisson and NB regression models in the context of traffic safety modeling. I demonstrate a bias-correction procedure based on the approximated bias provided by McCullagh and Nelder [59], followed by deriving the explicit bias correction formula for one special scenario, Poisson regression with a single binary explanatory variable. Using a Monte Carlo simulation study, I quantitatively evaluate the magnitude of bias and identify factors affecting the bias. I apply the bias-correction method to an infrastructure safety evaluation, which involves a three-year crash dataset collected from road segments with different pavement types. I also examine the relationship between the bias correction magnitude and crash counts by hypothetically reducing the number of crashes in the pavement data.

The remainder of this chapter is organized as follows. Section 2 derives the explicit bias-correction formula for four illustrative scenarios using the analytic bias results of McCullagh and Nelder [59]. Section 3 examines the benefit of bias correction through a Monte Carlo simulation study, which also elucidates when the bias correction starts to make a difference and to what extent the bias-adjustment procedure is beneficial. Section 4 demonstrates the bias correction through two real-case safety applications and two hypothetical situations by reducing the number of crashes. Section 5 summarizes this chapter with some discussion.

2.2 Method

Poisson and NB regression models assume that the frequency of events Y_i, e.g., the crash count, follows a Poisson distribution,

$$Y_i \sim \text{Poisson}(\lambda_i \cdot E_i), \quad i = 1, 2, \cdots, n, \tag{2.1}$$

where λ_i is the crash occurrence rate for the i^{th} road segment or the i^{th} driver. The λ_i is a constant in the Poisson regression and a random variable in the NB regression, respectively. In the NB regression, random variable λ_i follows a Gamma distribution, i.e.,

$$\lambda_i \sim \text{Gamma}(k, \mu_i),$$

where μ_i is the mean of the crash occurrence rate λ_i and $1/k$ is the dispersion parameter of the NB regression. The NB regression is more dispersed for smaller k. The Poisson regression is a special case of the NB regression when $k = \infty$. In Equation (2.1), the E_i is the corresponding exposure, which could be the length of the observation period or the total vehicle miles traveled.

A logarithm link function is used to link the event rate λ_i in the Poisson regression or the expected event rate μ_i in the NB regression with a linear transformation of p explanatory variables, $X_{i1}, X_{i2}, \cdots, X_{ip}$. That is,

$$\begin{gathered} \log(\lambda_i) = \eta_i \quad \text{or} \quad \log(\mu_i) = \eta_i, \\ \eta_i = \boldsymbol{X}_i' \boldsymbol{\beta}, \end{gathered} \tag{2.2}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$; $\boldsymbol{X}_i$ is the covariates vector for entity i, $\boldsymbol{X}_i = (1, X_{i1}, \cdots, X_{ip})'$. The coefficient β_j indicates the impact of the j^{th}

variable on crash risk, $j = 1, \cdots, p$. The estimation of $(\beta_1, \cdots, \beta_p)$ is the focus of the safety evaluation.

2.2.1 Adjusting finite sample bias for regression coefficient estimation

The MLE for the regression coefficient $\boldsymbol{\beta}$ is obtained by maximizing the log-likelihood function

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[- e^{\boldsymbol{X}_i'\boldsymbol{\beta}} \cdot E_i + Y_i\big(\boldsymbol{X}_i'\boldsymbol{\beta} + \log(E_i)\big) - \log(Y_i!)\Big]. \tag{2.3}$$

Denote the MLE as $\widehat{\boldsymbol{\beta}}$, which, in general, has no closed-form expression. The estimation is typically based on a numerical method, such as the Newton-Raphson method.

The MLE $\widehat{\boldsymbol{\beta}}$ is asymptotically unbiased and normally distributed. However, $\widehat{\boldsymbol{\beta}}$ in general is a biased estimator and the difference between $\widehat{\boldsymbol{\beta}}$ and the true value $\boldsymbol{\beta}$ might not negligible for a small sample size [24, 59]. The approximation of MLEs' bias can be traced at least as far back as Bartlett [9], being a side-product of when Bartlett studied the confidence interval for one unknown parameter from a random sample. Cox and Snell [17] extended Barlett's result to multidimensional MLEs. McCullagh [58] derived the bias using a similar procedure but with tensor notation and showed that the bias is of order $\mathcal{O}(n^{-1})$. McCullagh and Nelder [59] approximated the bias of GLMs with a canonical link using the leading $\mathcal{O}(n^{-1})$ term. Since Poisson regression and NB regression are GLMs and the log link is a canonical link, I apply the bias approximation for $\widehat{\boldsymbol{\beta}}$ provided by McCullagh and Nelder [59] as

$$\text{bias}(\widehat{\boldsymbol{\beta}}) \approx (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\boldsymbol{\xi}, \tag{2.4}$$

where $\mathbf{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n)'$, $\mathbf{W} = \text{cov}(\boldsymbol{Y})$, and $\boldsymbol{\xi}$ is an $n-$dimensional vector with

the i^{th} element being $\xi_i = -\frac{1}{2}Q_{ii}\frac{\kappa_{3i}}{\kappa_{2i}}$. Q_{ii} is the i^{th} diagonal element of the matrix $\mathbf{Q} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'$, $\kappa_{3i} = \partial^2\lambda_i/\partial\eta_i^2$, and $\kappa_{2i} = \partial\lambda_i/\partial\eta_i$. With log-link, $\kappa_{3i} = e^{\eta_i}$, and $\kappa_{2i} = e^{\eta_i}$. Therefore, the ξ_i is reduced to $\xi_i = -\frac{1}{2}Q_{ii}$. An estimate of the approximated bias in Equation (2.4) can be

$$\widehat{\text{bias}(\widehat{\boldsymbol{\beta}})} = (\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\,\widehat{\mathbf{W}}\,\widehat{\boldsymbol{\xi}}, \tag{2.5}$$

where $\widehat{\mathbf{W}}$ and $\widehat{\boldsymbol{\xi}}$ are obtained by plugging in the regular MLE $\widehat{\boldsymbol{\beta}}$. The bias-corrected coefficient estimate $\widetilde{\boldsymbol{\beta}}$ can be calculated as

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \widehat{\text{bias}(\widehat{\boldsymbol{\beta}})} = \widehat{\boldsymbol{\beta}} - (\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\,\widehat{\mathbf{W}}\,\widehat{\boldsymbol{\xi}}. \tag{2.6}$$

The bias-correction procedure in (2.6) is applicable to both the Poisson and NB models. The only difference is that the $\widehat{\mathbf{W}}$ of an NB model involves the estimated dispersion parameter while that of a Poisson model does not. The procedure can be applied for both continuous and discrete explanatory variables. In both cases, the covariates matrix $\mathbf{X}$ represents the corresponding explanatory variable data.

To provide a more illustrative understanding for the finite sample bias of $\boldsymbol{\beta}$, I obtain the explicit bias formula for the following special case of the Poisson/NB regression.

Special case 1: Poisson regression with one explanatory variable

Consider the case with only one exploratory variable, i.e.,

$$\log(\lambda_i) = \beta_0 + \beta_1 X_i.$$

The coefficient β_1 represents the impact of $\{X_i\}_{i=1}^n$ on the response variable $\{Y_i\}_{i=1}^n$ and is the focus of the evaluation.

From Equation (2.4), the approximated bias for $\widehat{\beta}_1$ is

$$\text{bias}(\widehat{\beta}_1) = \frac{1}{n} \cdot \frac{3abc - 2b^3 - a^2 d}{2(ac - b^2)^2}, \tag{2.7}$$

where

$$a = \frac{1}{n}\sum_{i=1}^{n} \lambda_i E_i,\ b = \frac{1}{n}\sum_{i=1}^{n} X_i \lambda_i E_i,\ c = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \lambda_i E_i,\ d = \frac{1}{n}\sum_{i=1}^{n} X_i^3 \lambda_i E_i.$$

A detailed derivation can be found in the Appendix A.1.

From Equation (2.5), the estimated bias approximation for $\widehat{\beta}_1$ is

$$\widehat{\text{bias}(\widehat{\beta}_1)} = \frac{1}{n} \cdot \frac{3\overline{Y} \cdot \overline{XY} \cdot \overline{X^2 Y} - 2(\overline{XY})^3 - (\overline{Y})^2 \cdot \overline{X^3 Y}}{2\left[\overline{Y} \cdot \overline{X^2 Y} - (\overline{XY})^2\right]^2}, \tag{2.8}$$

where

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,\ \overline{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i,\ \overline{X^2 Y} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 Y_i,\ \overline{X^3 Y} = \frac{1}{n}\sum_{i=1}^{n} X_i^3 Y_i.$$

The derivation is very similar to the procedure in the Appendix A.1.

Equation (2.8) shows that the bias is related to the number of events in the dataset. If all the Y_i's were lowered to their 1/10 with X_i's being fixed, the bias would increase tenfold. (The denominator is to the power of three for Y, and the numerator is to the power of four for Y.) In other words, the biases are amplified when event occurrences are rare.

Additionally, the approximated bias in Equation (2.7) is consistent with the bias correction formula provided by Giles and Feng [28], which was derived from Cox and Snell's general definition of residuals.

Special case 2: Poisson regression with one binary explanatory variable

A Poisson regression with a single dichotomous predictor could shed more light on what factors affect the magnitude of bias. We examine the finite sample bias when the single explanatory variable is dichotomous: X_i is either 0 or 1. In epidemiology, the observation group with $X_i = 0$ is typically referred to as the reference group and the group with $X_i = 1$ is the treatment group. The MLE of β_1, in this case, has a closed-form solution:

$$\widehat{\beta}_1 = \log\Big(\frac{C_1/T_1}{C_0/T_0}\Big), \tag{2.9}$$

where C_0 and C_1 are the number of events in the reference group and treatment group respectively, i.e., $C_0 = \sum_{\{X_i=0\}} Y_i, C_1 = \sum_{\{X_i=1\}} Y_i$; T_0 and T_1 are the respective total exposure in the two groups: $T_0 = \sum_{\{X_i=0\}} E_i, T_1 = \sum_{\{X_i=1\}} E_i$.

From Equation (2.7), the bias approximation is

$$\text{bias}(\widehat{\beta}_1) = \frac{1}{2\lambda_0 T_0} - \frac{1}{2\lambda_1 T_1}, \tag{2.10}$$

as $a = \frac{\lambda_0 T_0 + \lambda_1 T_1}{n}$ and $b = c = d = \frac{\lambda_1 T_1}{n}$.

From Equation (2.8), the estimated bias approximation is

$$\widehat{\text{bias}(\widehat{\beta}_1)} = \frac{1}{2C_0} - \frac{1}{2C_1}, \tag{2.11}$$

as $\overline{Y} = \frac{C_0 + C_1}{n}$ and $\overline{XY} = \overline{X^2 Y} = \overline{X^3 Y} = \frac{C_1}{n}$.

The estimated bias is related to the total number of events in each group. In addition, the absolute value of bias would be large when the data are unbalanced; i.e., one group has a substantially larger number of observations than the other group.

Special case 3: Poisson regression with two binary explanatory variables

We also derive the explicit formula for the bias of two dichotomous explanatory variables in a Poisson regression model. For a Poisson regression with multiple explanatory variables, MLEs and their bias would be different depending on the inclusion of an interaction term or not.

When there are two binary explanatory variables, $X_{i1} \in \{0,1\}, X_{i2} \in \{0,1\},\ i = 1, \cdots, n$, the model without the interaction term is

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}. \tag{2.12}$$

The corresponding estimated bias approximation for MLEs is

$$\begin{aligned}
\widehat{\text{bias}(\widehat{\beta}_1)} &= \frac{C_{00}C_{10}C_{\cdot 1}^2}{2(C_{00}C_{10}C_{\cdot 1} + C_{01}C_{11}C_{\cdot 0})^2}(C_{10} - C_{00}) + \frac{C_{01}C_{11}C_{\cdot 0}^2}{2(C_{00}C_{10}C_{\cdot 1} + C_{01}C_{11}C_{\cdot 0})^2}(C_{11} - C_{01}) \\
\widehat{\text{bias}(\widehat{\beta}_2)} &= \frac{C_{00}C_{01}C_{1\cdot}^2}{2(C_{00}C_{01}C_{1\cdot} + C_{10}C_{11}C_{0\cdot})^2}(C_{01} - C_{00}) + \frac{C_{10}C_{11}C_{0\cdot}^2}{2(C_{00}C_{01}C_{1\cdot} + C_{10}C_{11}C_{0\cdot})^2}(C_{11} - C_{10}),
\end{aligned} \tag{2.13}$$

where $C_{00}, C_{01}, C_{10}, C_{11}$ represent the number of events in the four strata divided by the value of X_{i1} and X_{i2}. For example, C_{01} is the number of events in the group of X_{i1} being 0 and X_{i2} being 1 ($\{i : X_{i1} = 0 \text{ and } X_{i2} = 1\}$). The biases depend on the number of events in all the strata.

Similarly, the model with an interaction term and corresponding bias is

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{12} X_{i1} X_{i2} \tag{2.14}$$

and the corresponding estimated bias correction is

$$\begin{aligned}
\widehat{\text{bias}(\widehat{\beta}_1)} &= \frac{1}{2C_{00}} - \frac{1}{2C_{10}}, \\
\widehat{\text{bias}(\widehat{\beta}_2)} &= \frac{1}{2C_{00}} - \frac{1}{2C_{01}}, \\
\widehat{\text{bias}(\widehat{\beta}_{12})} &= \frac{1}{2C_{11}} - \frac{1}{2C_{10}} - \frac{1}{2C_{01}} + \frac{1}{2C_{00}}.
\end{aligned} \tag{2.15}$$

The biases for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are similar to the bias format for a single explanatory variable. They are related to the number of events in the different strata, rather than the number of data points.

Special case 4: Negative Binomial regression with one binary explanatory variable

Consider a NB model with a single dichotomous exploratory variable, i.e.,

$$\log(\mu_i) = \beta_0 + \beta_1 X_i,$$

where X_i is either 0 or 1. In epidemiology, the observation group with $X_i = 0$ is typically referred to as the reference group and the group with $X_i = 1$ is the treatment group. The coefficient β_1 represents the treatment effect, the impact of $\{X_i\}_{i=1}^n$ on the response variable $\{Y_i\}_{i=1}^n$, and it is the focus of the evaluation.

From Equation (2.4), the approximated bias for $\widehat{\beta}_1$ is

$$\text{bias}(\widehat{\beta}_1) = \frac{1}{2V_0} - \frac{1}{2V_1}, \tag{2.16}$$

where

$$V_0 = \mathrm{Var}\big(\sum_i Y_i | X_i = 0\big) = \sum_{X_i=0} \Big[\mu_0 E_i + \frac{(\mu_0 E_i)^2}{k}\Big],$$
$$V_1 = \mathrm{Var}\big(\sum_i Y_i | X_i = 1\big) = \sum_{X_i=1} \Big[\mu_1 E_i + \frac{(\mu_1 E_i)^2}{k}\Big].$$

A detailed derivation can be found in the Appendix A.2. It is difficult to estimate V_0 and V_1, the variance of crash counts in each stratum. The estimation of V_0 and V_1 depends on the estimation of μ_0, μ_1, and k, which often requires an iterative estimation procedure. Practically, under the finite sample situation, $\mu_0 E_i$ or $\mu_1 E_i$ for one observation is typically small. The magnitude of higher order terms $\frac{(\mu_0 E_i)^2}{k}$ and $\frac{(\mu_1 E_i)^2}{k}$ are much smaller than $\mu_0 E_i$ and $\mu_1 E_i$, respectively. Thus I consider to approxiamte V_0 and V_1 by C_0 and C_1, the expected number of crashes in the reference group and treatment group.

$$C_0 = \mathbf{E}\big(\sum_i Y_i | X_i = 0\big) = \sum_{X_i=0} \mu_0 E_i \approx V_0,$$
$$C_1 = \mathbf{E}\big(\sum_i Y_i | X_i = 1\big) = \sum_{X_i=1} \mu_1 E_i \approx V_1.$$

Therefore, the bias of $\widehat{\beta}_1$ can be approximated by

$$\mathrm{bias}(\widehat{\beta}_1) \approx \frac{1}{2C_0} - \frac{1}{2C_1}, \tag{2.17}$$

where the expected crash counts in each stratum, C_0 and C_1, can be estimated by the observed number of crashes in the reference group and treatment group. The balance of crash counts in different stratum also matters to the magnitude of bias. The magnitude would be larger when the number of crashes in different stratum were more unbalanced.

2.2.2 Adjusting finite sample bias for event frequency prediction

Predicting the event frequency is commonly required in safety modeling. The Crash Modification Factor is based on the predicted crash frequency [2]. Given a new set of predictors $\boldsymbol{X}_0$, a reasonable crash rate prediction should be the corresponding expected crash rate. Assume the underlying true parameter is $\boldsymbol{\beta}$, the proper predicted crash rate should be $\exp(\boldsymbol{X}_0\boldsymbol{\beta})$, denoted as λ_0. Since $\boldsymbol{\beta}$ is unknown, the prediction can be made by plugging in either the regular MLE $\widehat{\boldsymbol{\beta}}$ or the bias-corrected coefficient estimate $\widetilde{\boldsymbol{\beta}}$. One natural question is whether $\exp(\boldsymbol{X}_0\widehat{\boldsymbol{\beta}})$ or $\exp(\boldsymbol{X}_0\widetilde{\boldsymbol{\beta}})$ would provide a more accurate prediction of crash frequency and be closer to $\exp(\boldsymbol{X}_0\boldsymbol{\beta})$.

Denote the predicted occurrence rate based on $\widetilde{\boldsymbol{\beta}}$ as $\widetilde{\lambda}_0 = \exp(\boldsymbol{X}_0\widetilde{\boldsymbol{\beta}})$. This predictor overestimates the actual expected event rate, as

$$\mathbb{E}(\widetilde{\lambda}_0) = \mathbb{E}\big(\exp(\boldsymbol{X}_0\widetilde{\boldsymbol{\beta}})\big) \geq \exp\big(\mathbb{E}(\boldsymbol{X}_0\widetilde{\boldsymbol{\beta}})\big) \approx \exp\big(\boldsymbol{X}_0\boldsymbol{\beta}\big) = \lambda_0. \tag{2.18}$$

The inequality is obtained by applying Jensen's inequality since the exponential function is a convex function.

An alternative way to predict crash rate is by plugging in the regular MLE, $\widehat{\boldsymbol{\beta}}$. In this case, denote the corresponding predicted crash rate as $\widehat{\lambda}_0 = \exp(\boldsymbol{X}_0\widehat{\boldsymbol{\beta}})$. It can be shown that this estimator is unbiased when the new predictor $\boldsymbol{X}_0$ is from the data points used in parameter estimation $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$. When $\boldsymbol{X}_0 \notin \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$, the property of predicted crash frequency by plugging in the MLE is difficult to decide.

Suppose $\boldsymbol{X}_0$ has the same value as the explanatory variable of the i'-th entity, i.e., $\boldsymbol{X}_0 = \boldsymbol{X}_{i'}$. Then, $\lambda_0 = \lambda_{i'}$, $\widehat{\lambda}_0 = \widehat{\lambda}_{i'}$, and $\widehat{\lambda}_{i'} = Y_{i'}/E_{i'}$. The expectation of the predicted crash rate is

$$\mathbb{E}(\widehat{\lambda}_0) = \mathbb{E}(\widehat{\lambda}_{i'}) = \mathbb{E}\big(Y_{i'}/E_{i'}\big) = \lambda_{i'} = \lambda_0. \tag{2.19}$$

The third equal sign is because $\mathbb{E}\big(Y_{i'}\big) = \lambda_{i'}E_{i'}$, as $Y_{i'} \sim \text{Poisson}(\lambda_{i'}E_{i'})$. Therefore, the predicted crash rate $\widehat{\lambda}_0$ by plugging in $\widehat{\boldsymbol{\beta}}$ is unbiased when $\boldsymbol{X}_0 \in \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$.

To sum up, predicting crash frequency based on the bias-corrected estimator $\widetilde{\lambda}_0 = \exp(\boldsymbol{X}_0\widetilde{\boldsymbol{\beta}})$ would lead to over-prediction. The prediction based on MLE, $\widehat{\lambda}_0 = \exp(\boldsymbol{X}_0\widehat{\boldsymbol{\beta}})$ is a better choice under certain conditions; for example, when the given explanatory variable $\boldsymbol{X}_0$ has appeared in the dataset for parameter estimation ($\boldsymbol{X}_0 \in \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$).

2.3 Simulation

I conducted a Monte Carlo simulation study to evaluate the performance of the regular MLE $\widehat{\boldsymbol{\beta}}$ and the bias-corrected MLE $\widetilde{\boldsymbol{\beta}}$. The objectives were to examine if the bias-correction procedure could lead to a more accurate estimation, to identify the non-negligible finite sample bias situations, the magnitude of bias, and the factors affecting bias.

Without loss of generality, I consider observations generated from an NB regression, whose expected event rate is associated with a binary predictor variable. That is,

$$\begin{aligned} Y &\sim \text{Poisson}\big(\lambda \cdot E\big) \\ \lambda &\sim \text{Gamma}(k,\ \mu) \end{aligned} \tag{2.20}$$

where

$$\mu = e^{\beta_0 + \beta_1 X}, \quad X = 0 \text{ or } 1.$$

I generated $n = 5,000$ observations from the above NB model. The distribution of the exposure, E_i, is similar to the application described in Section 4. The specific simulation setup for β_0 and β_1 can be found in Table 2.1, where μ_0 is the expected event rate for the reference group and μ_1 is the expected event rate for the treatment group. The simulation

parameters are setup such that the expected number of crashes within each group, C_0 and C_1, range from 4 to 1000. Within which, 2,500 observations come from the reference group ($X = 0$) and 2,500 observations are from the treatment group ($X = 1$). I enumerate the parameter k in model (2.20) within $\{0.5, 50, \infty\}$. For each scenario of model setting, the simulation is repeated 1,000 times. Note that the NB model will get close to a Poisson model when the value of k becomes large. When $k = \infty$, the simulation is generated from the Poisson regression.

Table 2.1: Simulation setup for β_0 and β_1, and $\mu_0 = \exp(\beta_0), \mu_1 = \exp(\beta_0 + \beta_1)$

β_0	β_1	μ_0	μ_1
1.5	-1.0	4.48	1.65
1.5	-0.5	4.48	2.72
1.5	-0.1	4.48	4.06
1.5	0.1	4.48	4.95
1.5	0.5	4.48	7.39
1.5	1.0	4.48	12.18

Figure 2.1 shows the probability of both the reference group and treatment group having at least one events when C_0 and C_1 range from 1 to 1000. When the expected number of total crashes in either group is greater than four, with probability approximately equal to one, there would be at least one event within each group. The shaded area indicates the scenarios included in the simulation. x-axis is the logarithm of the expected total number of crashes for the reference group C_0 to the base 10 ; y-axis is the logarithm of the expected total number of crashes for the treatment group C_1 to the base 10.

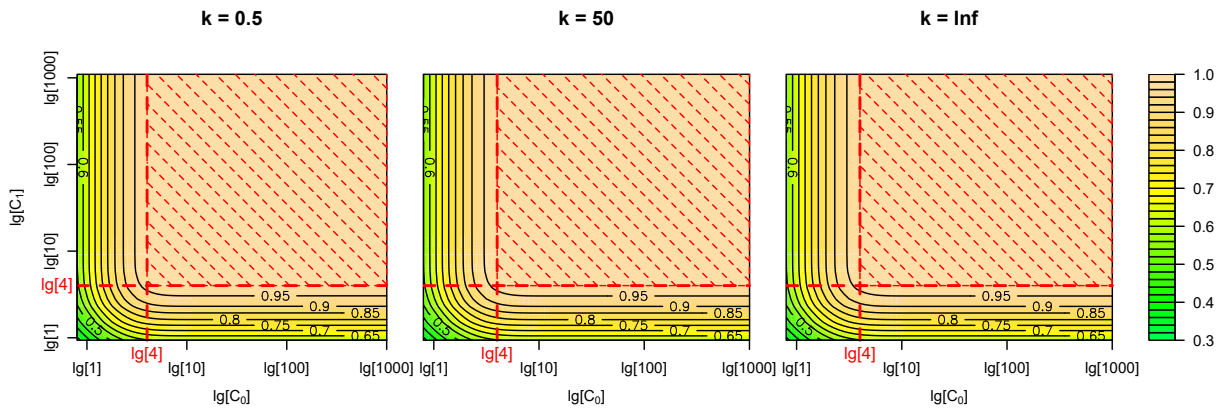

Figure 2.1: The heatmap for the probability of at least one event in both the reference group and the treatment group (C_0 is the expected total number of crashes in the reference group and C_1 is that of the treatment group.)

2.3.1 Simulation results

I present the simulation results in three parts. Recall that the regular MLE of β is denoted as $\widehat{\boldsymbol{\beta}}$ and the bias-corrected MLE of β is denoted as $\widetilde{\boldsymbol{\beta}}$. For each model setting, the simulation is repeated 1,000 times. Firstly, I use the percentage bias of $\widehat{\beta}_1$ and $\widetilde{\beta}_1$ to show the effect of bias correction using some example cases. The percentage bias of $\widehat{\beta}_1$ and $\widetilde{\beta}_1$ are calculated as

$$\text{pct bias}(\widehat{\beta}_1) = \frac{1}{1000}\sum \frac{\widehat{\beta}_1 - \beta_1}{\beta_1} \times 100\%, \quad \text{pct bias}(\widetilde{\beta}_1) = \frac{1}{1000}\sum \frac{\widetilde{\beta}_1 - \beta_1}{\beta_1} \times 100\%,$$

where β_1 is the underlying true parameter as set in Table 2.1. Secondly, I use the side-by-side plot to compare the bias of $\widehat{\boldsymbol{\beta}}_1$ and the bias of $\widetilde{\boldsymbol{\beta}}_1$ for comprehensive simulation scenarios, where the biases are calculated as

$$\text{bias}(\widehat{\beta}_1) = \frac{1}{1000}\sum \widehat{\beta}_1 - \beta_1, \quad \text{bias}(\widetilde{\beta}_1) = \frac{1}{1000}\sum \widetilde{\beta}_1 - \beta_1.$$

Lastly, I use the difference between the variance of $\widehat{\boldsymbol{\beta}}_1$ and the variance of $\widetilde{\boldsymbol{\beta}}_1$ to show that the bias-corrected MLE has smaller variance than the regular MLE. The variance difference of the two estimators are calculated as

$$\text{var}(\widehat{\beta}_1) - \text{var}(\widetilde{\beta}_1) = \frac{1}{1000}\sum\left(\widehat{\beta}_1 - \text{ave}(\widehat{\beta}_1)\right)^2 - \frac{1}{1000}\sum\left(\widetilde{\beta}_1 - \text{ave}(\widetilde{\beta}_1)\right)^2,$$

where $\text{ave}(\widehat{\beta}_1) = \frac{1}{1000}\sum\widehat{\beta}_1$ and $\text{ave}(\widetilde{\beta}_1) = \frac{1}{1000}\sum\widetilde{\beta}_1$.

Figure 2.2 plots the percent bias of $\widehat{\beta}_1$ and $\widetilde{\beta}_1$ when $\beta_1 = 0.1$ (event rate in treatment group is 10% higher than the reference group), $k = \infty$, and the expected number of crashes in the treatment group (C_1) is 9, 28, and 89. It shows that the bias-corrected estimator is more close to the true parameter value than the regular MLE for the majority cases. The percent bias of the bias-corrected estimator $\widetilde{\beta}_1$ are around 0% except for some unlikely scenarios that the expected number of crashes in the reference group (C_0) is smaller than five, while the percent bias of the uncorrected MLE $\widehat{\beta}_1$ are further away from zero percent, ranging from -60% to 100%. The left plot also shows that the bias of $\widehat{\beta}_1$ decreases when the expected number of crashes in the reference group increases, which testifies our approximation of $\text{bias}(\widehat{\beta}_1)$ in Equation (2.16).

Figure 2.3 comprehensively shows the bias of $\widehat{\boldsymbol{\beta}}$ and the bias of $\widetilde{\boldsymbol{\beta}}$ side-by-side for all the simulation scenarios. From the plot, I can see the regular MLE $\widehat{\beta}_1$ underestimates β_1 when the expected number of events in the reference group is larger than the expected number of events in the treatment group ($C_0 > C_1$), and it overestimates β_1 when the expected number of events in the reference group is smaller than the expected number of events in the treatment group ($C_0 < C_1$). The bias is more severe when the difference between the expected numbers of events in the two groups is larger. In other words, the bias is smaller when the expected event counts in the two groups are more balanced.

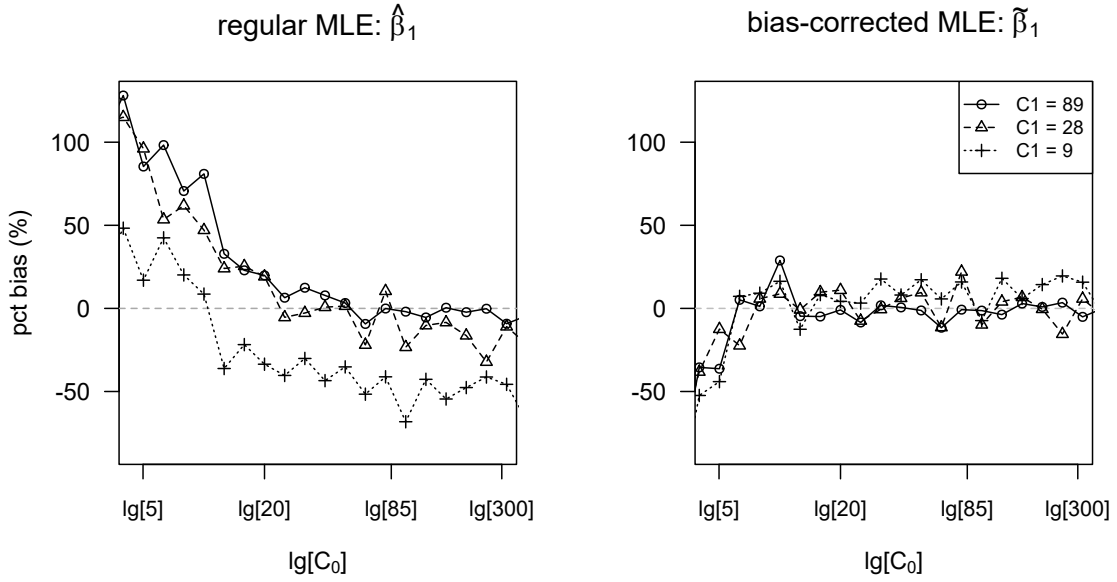

Figure 2.2: The comparison between the percent bias of $\widehat{\beta}_1$ (left) and the percent bias of $\widetilde{\beta}_1$ (right) when $\beta_1 = 0.1$ and $k = \infty$ (C_0 is the expected total number of crashes in the reference group and C_1 is that in the reference group.)

For the bias-corrected MLE $\widetilde{\beta}_1$, the bias is relatively small for a "larger area" in the plot, which again shows the effectiveness of bias-correction procedure for the majority scenarios. The procedure over-corrects for small expected number of crashes in either the reference group or the treatment group (the bottom and left "edges" in the right plots). Higher-order correction can help the estimation when the number of events for one group is small. In addition, the two "edges" get narrower as the k increases, which means the benefit of bias correction is more prominent when the k becomes larger.

Figure 2.4 shows the difference between the sample variance of $\widehat{\beta}_1$ and the sample variance of $\widetilde{\beta}_1$. The difference are non-negative for all the scenarios plotted. Non-negative means that the bias-corrected coefficient estimate $\widetilde{\beta}_1$ has smaller variance compared to the regular MLE $\widehat{\beta}_1$, especially when the expected number of events in either group is limited. (The bottom-left corner has darker yellow.) A smaller sample variance means a more stable estimate, so

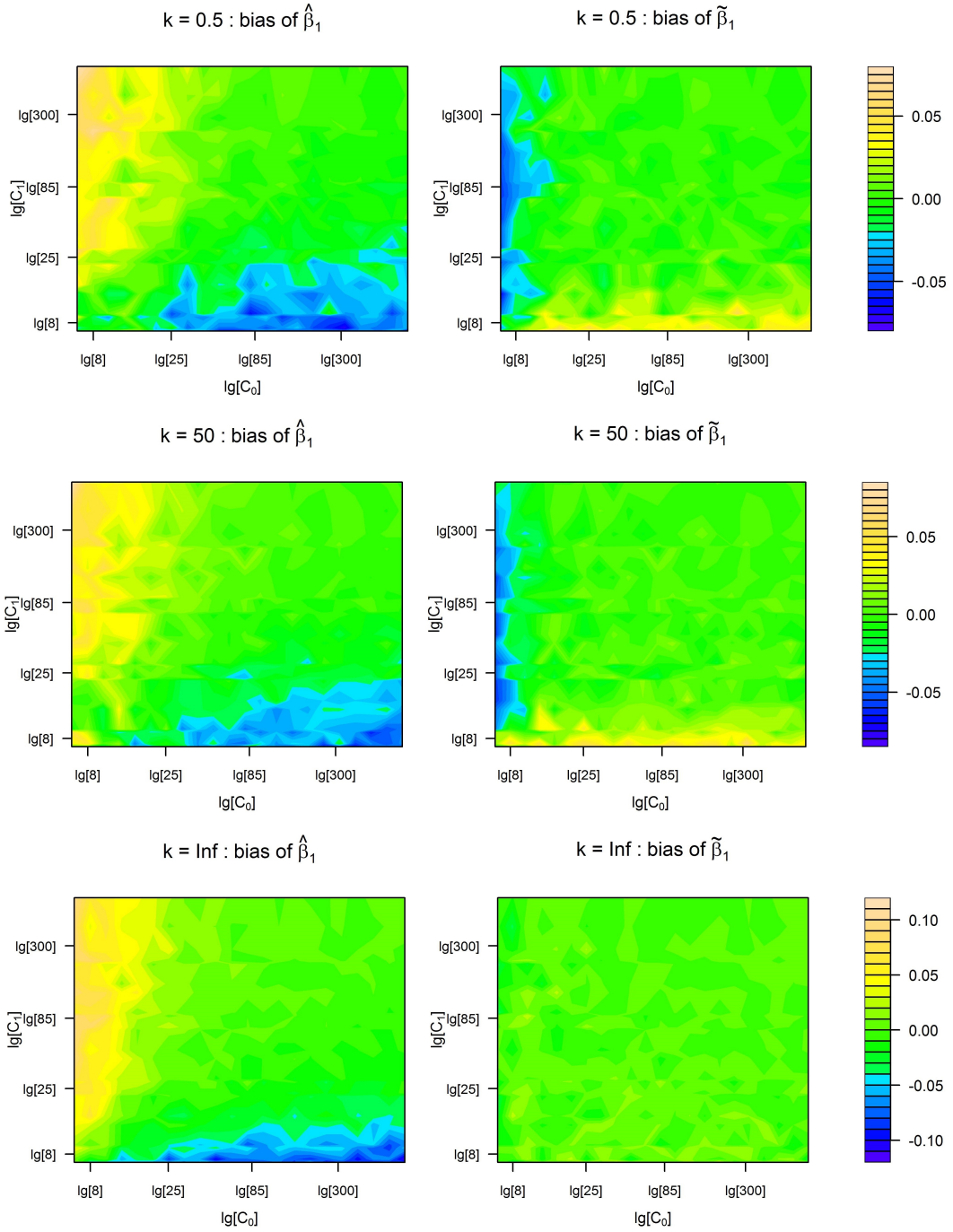

Figure 2.3: The contour plots of the bias of $\widehat{\beta}_1$ (left) and the bias of $\widetilde{\beta}_1$ (right) (C_0 and C_1 is the expected total number of crashes in the reference and treatment group, respectively.). Here "yellow" represents positive bias (overestimation), "blue" represents negative bias (underestimation), and "green" means relatively small bias.

the bias-corrected estimator $\widetilde{\beta}_1$ is better.

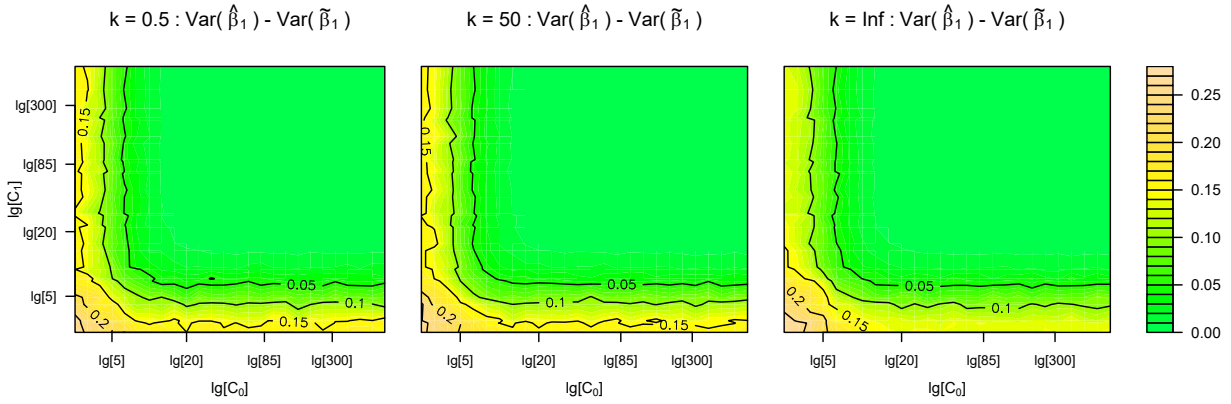

Figure 2.4: The heatmap for the difference between the sample variance of $\widehat{\beta}_1$ and the sample variance of $\widetilde{\beta}_1$

For a short summary, the simulation study shows that (1) bias-corrected MLE $\widetilde{\boldsymbol{\beta}}$ is less biased and has smaller variance compared to the regular MLE $\widehat{\boldsymbol{\beta}}$; (2) the benefit is more substantial as the k increases; (3) the effect of bias correction is more pronounced when the number of crashes in one stratum of a categorical explanatory variable is less than 50; (4) the bias-corrected MLE $\widetilde{\boldsymbol{\beta}}$, as well as the regular MLE, is unstable with too few events in one stratum of a categorical explanatory variable (i.e. when the number of crashes is less than five for Poisson and less than seven for Negative Binomial); (5) the balance of the crash counts within strata of a categorical explanatory variable also matters to the magnitude of bias.

2.4 Case study

To illustrate the benefit of bias correction and examine the magnitude of bias, I applied the bias-correction procedure to the SHRP2 driver dataset and an infrastructure safety evaluation dataset.

2.4.1 Case study 1: SHRP 2 driver risk assessment

The driver data is sourced from the SHRP 2 NDS. There are 3,437 drivers from six data-collection sites in our analysis. Information such as age, gender, and other demographic characteristics were recorded at the beginning of the study. We considered the number of crashes with property damage or higher severity as the safety response for each driver. There are 1,149 crashes for the 3,437 drivers with a total 1,161,712 driving hours. The average crash occurrence rate is 9.89×10^{-4} crashes/driving hour.

Table 2.2 shows the seven predictors we used for the risk analysis, including one continuous variable, years of driving. For continuous variables, one standard deviation is used as the unit for parameter estimation. Poisson regression is used because there is no over-dispersion. The descriptive statistics, magnitude of bias correction , specific regression coefficient estimates, and rate ratio (RR) estimates are shown in Table 2.2, Table 2.3, Figure 2.5, Figure 2.6, Figure 2.7, and Figure 2.8.

The magnitude of the bias correction is relatively large for the levels with small event counts. For instance, the level of "unmarried partners" and the "other" level of the "average annual mileage" variable. (Based on the simulation, the correction for the "other" group of the "marital status" variable may be a little over-corrected, as the number of events for this group is under five.) The magnitude of bias correction is also affected by the difference between the event counts of the estimation group and the reference group. For example, the absolute bias correction for "15 K - 20 K miles" is smaller than that of "10 K - 15 K miles" even though the latter has more events in its group. The "15 K - 20 K miles" group needs smaller bias correction because its number of events is about the same as the number of events for the reference group. The percent correction in terms of regression coefficients ranged from -25.75% to 19.29%, and the percent correction for RR can be as high as 10.31%.

Table 2.2: Descriptive statistics and corresponding bias magnitude for the explanatory variables of SHRP 2 driver data

Categorical vars	**Freq (pct)**	**No. of events (pct)**	$\widetilde{\beta}-\widehat{\beta}$	$\frac{\widetilde{\beta}-\widehat{\beta}}{\widehat{\beta}}$ ×100(%)	$\widetilde{RR}-\widehat{RR}$	$\frac{\widetilde{RR}-\widehat{RR}}{\widehat{RR}}$ ×100(%)
age group						
30-64	1,015 (29.5%)	228 (19.8%)				
16-19	534 (15.5%)	284 (24.7%)	-0.002	-0.2	-0.004	-0.2
20-29	1,021 (29.7%)	403 (35.1%)	-0.002	-0.5	-0.003	-0.2
65+	867 (25.2%)	234 (20.4%)	-0.001	-0.4	-0.001	-0.1
sex						
F	1,794 (52.2%)	598 (52%)				
M	1,643 (47.8%)	551 (48%)	0.000	1.2	0.000	0.0
income						
low	972 (28.3%)	373 (32.5%)				
middle	1,502 (43.7%)	467 (40.6%)	0.000	0.3	0.000	0.0
upper	746 (21.7%)	237 (20.6%)	0.000	1.7	0.000	0.0
other	217 (6.3%)	72 (6.3%)	0.005	-25.7	0.005	0.5
site						
FL	767 (22.3%)	321 (27.9%)				
IN	274 (8%)	95 (8.3%)	0.004	-16.3	0.004	0.4
NC	558 (16.2%)	171 (14.9%)	0.001	-0.6	0.001	0.1
NY	772 (22.5%)	270 (23.5%)	0.000	-0.1	0.000	0.0
PA	262 (7.6%)	54 (4.7%)	0.007	-2.2	0.005	0.7
WA	804 (23.4%)	238 (20.7%)	0.000	-0.2	0.000	0.0
marital status						
single	1,514 (44.1%)	664 (57.8%)				
married	1,381 (40.2%)	305 (26.5%)	0.000	-0.2	0.000	0.0
divorced	199 (5.8%)	71 (6.2%)	0.006	2.5	0.007	0.6
unmarried partners	111 (3.2%)	43 (3.7%)	0.011	19.3	0.012	1.1
widow(er)	205 (6%)	61 (5.3%)	0.006	-7.0	0.006	0.6
other	27 (0.8%)	5 (0.4%)	0.098	-23.0	0.067	10.3
average annual mileage						
less than 5K miles	411 (12%)	174 (15.1%)				
5K - 10K miles	900 (26.2%)	259 (22.5%)	0.000	0.1	0.000	0.0
10K - 15K miles	1,078 (31.4%)	350 (30.5%)	-0.001	0.3	-0.001	-0.1
15K - 20K miles	457 (13.3%)	144 (12.5%)	0.001	-0.1	0.000	0.1
20K - 25K miles	219 (6.4%)	74 (6.4%)	0.004	-0.8	0.002	0.4
25K - 30K miles	121 (3.5%)	57 (5%)	0.006	-2.7	0.005	0.6
more than 30K miles	194 (5.6%)	70 (6.1%)	0.005	-1.2	0.003	0.5
other	57 (1.7%)	21 (1.8%)	0.020	-8.6	0.016	2.1
Continuous vars	**Mean (sd)**	$\overline{xy}$ ($\overline{x^2y}$)	$\widetilde{\beta}-\widehat{\beta}$	$\frac{\widetilde{\beta}-\widehat{\beta}}{\widehat{\beta}}$ ×100%	$\widetilde{RR}-\widehat{RR}$	$\frac{\widetilde{RR}-\widehat{RR}}{\widehat{RR}}$ ×100%
years of driving	25.58 (22.47)	6.78 (305.57)	0.0002	0.93	0.0002	0.02

Table 2.3: Specific regression coefficient estimates and rate ratio estimates for SHRP 2 driver data

	$\widetilde{\beta}$(CI)	$\widehat{\beta}$(CI)	$\widetilde{RR}$(CI)	$\widehat{RR}$(CI)
age 16-19	0.816 (0.499, 1.134)	0.818 (0.501, 1.136)	2.262 (1.647, 3.108)	2.267 (1.65, 3.114)
age 20-29	0.402 (0.13, 0.675)	0.404 (0.132, 0.676)	1.495 (1.139, 1.963)	1.498 (1.141, 1.967)
age 65+	0.277 (-0.019, 0.573)	0.278 (-0.018, 0.574)	1.319 (0.981, 1.774)	1.321 (0.982, 1.776)
sex M	0.01 (-0.11, 0.129)	0.01 (-0.11, 0.129)	1.01 (0.896, 1.138)	1.01 (0.896, 1.138)
income middle	-0.079 (-0.221, 0.063)	-0.078 (-0.22, 0.064)	0.924 (0.802, 1.065)	0.925 (0.802, 1.066)
income upper	0.023 (-0.153, 0.198)	0.022 (-0.154, 0.198)	1.023 (0.858, 1.219)	1.022 (0.858, 1.219)
income other	-0.015 (-0.277, 0.247)	-0.02 (-0.282, 0.242)	0.985 (0.758, 1.28)	0.98 (0.754, 1.273)
site IN	-0.019 (-0.249, 0.212)	-0.022 (-0.253, 0.208)	0.981 (0.779, 1.236)	0.978 (0.777, 1.231)
site NC	-0.196 (-0.385, -0.007)	-0.197 (-0.386, -0.008)	0.822 (0.68, 0.993)	0.821 (0.68, 0.992)
site NY	-0.074 (-0.237, 0.089)	-0.074 (-0.237, 0.089)	0.929 (0.789, 1.093)	0.929 (0.789, 1.093)
site PA	-0.33 (-0.621, -0.04)	-0.337 (-0.628, -0.047)	0.719 (0.538, 0.961)	0.714 (0.534, 0.954)
site WA	-0.223 (-0.392, -0.053)	-0.223 (-0.392, -0.054)	0.8 (0.676, 0.948)	0.8 (0.676, 0.948)
married	-0.208 (-0.424, 0.008)	-0.208 (-0.424, 0.008)	0.812 (0.655, 1.008)	0.812 (0.654, 1.008)
divorced	0.228 (-0.082, 0.539)	0.223 (-0.088, 0.533)	1.256 (0.921, 1.714)	1.249 (0.916, 1.704)
unmarried partners	0.068 (-0.246, 0.382)	0.057 (-0.257, 0.371)	1.071 (0.782, 1.466)	1.059 (0.774, 1.45)
widow(er)	-0.085 (-0.429, 0.26)	-0.091 (-0.435, 0.253)	0.919 (0.651, 1.296)	0.913 (0.647, 1.288)
marriage other	-0.329 (-1.236, 0.579)	-0.427 (-1.334, 0.481)	0.72 (0.291, 1.784)	0.653 (0.263, 1.618)
annual 5-10 K miles	-0.322 (-0.519, -0.124)	-0.321 (-0.519, -0.124)	0.725 (0.595, 0.883)	0.725 (0.595, 0.884)
annual 10-15 K miles	-0.361 (-0.553, -0.168)	-0.36 (-0.552, -0.167)	0.697 (0.575, 0.846)	0.698 (0.576, 0.846)
annual 15-20 K miles	-0.408 (-0.641, -0.175)	-0.408 (-0.642, -0.175)	0.665 (0.527, 0.84)	0.665 (0.526, 0.839)
annual 20-25 K miles	-0.501 (-0.781, -0.222)	-0.505 (-0.785, -0.226)	0.606 (0.458, 0.801)	0.603 (0.456, 0.798)
annual 25-30 K miles	-0.218 (-0.527, 0.09)	-0.224 (-0.533, 0.084)	0.804 (0.591, 1.095)	0.799 (0.587, 1.088)
annual more than 30 K miles	-0.382 (-0.669, -0.095)	-0.387 (-0.673, -0.1)	0.682 (0.512, 0.909)	0.679 (0.51, 0.905)
annual other	-0.216 (-0.676, 0.245)	-0.236 (-0.697, 0.225)	0.806 (0.508, 1.278)	0.79 (0.498, 1.252)
year_driving	0.026 (-0.167, 0.218)	0.025 (-0.167, 0.218)	1.026 (0.846, 1.243)	1.026 (0.846, 1.243)

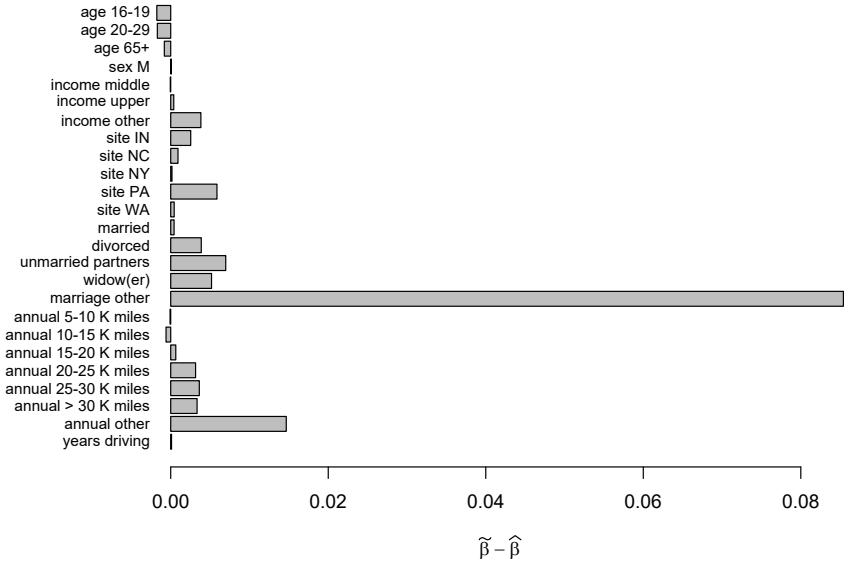

Figure 2.5: Correction magnitude of regression coefficients, $\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$, for SHRP 2 driver data

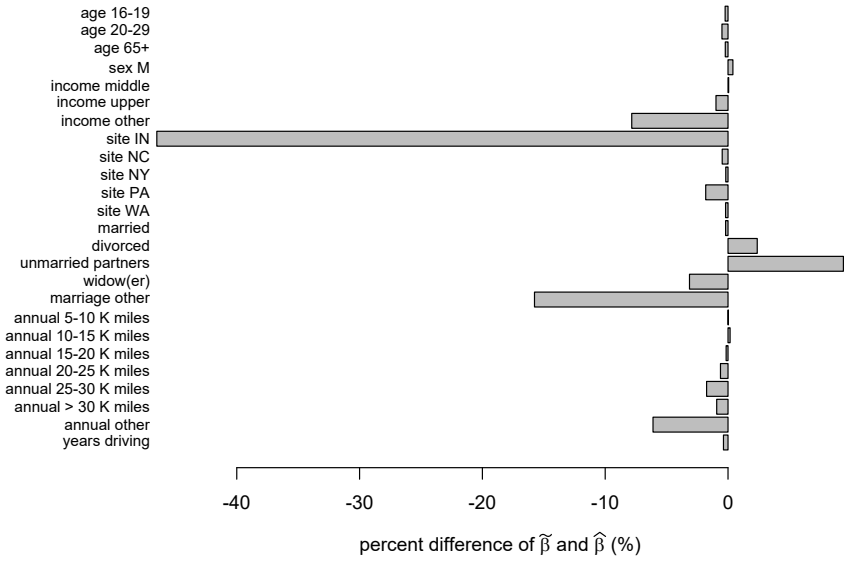

Figure 2.6: Percent change of regression coefficients, $(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})/\widehat{\boldsymbol{\beta}} \times 100\%$, for SHRP 2 driver data

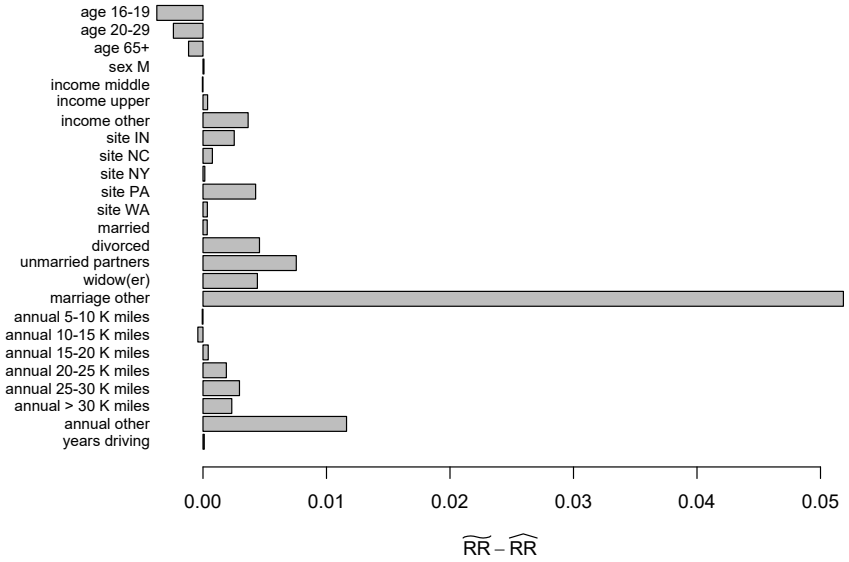

Figure 2.7: Correction magnitude of rate ratios, $\widetilde{RR} - \widehat{RR}$, for SHRP 2 driver data

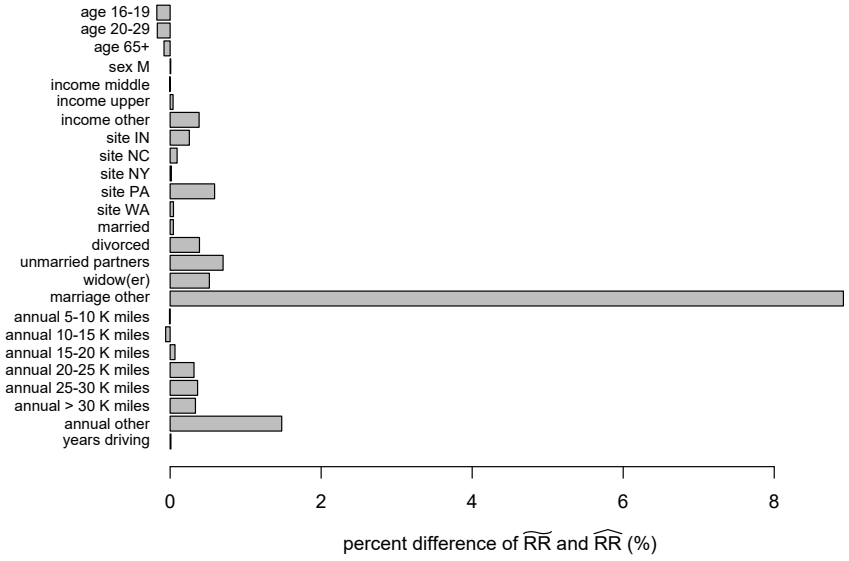

Figure 2.8: Percent change of rate ratios, $(\widetilde{RR} - \widehat{RR})/\widehat{RR} \times 100\%$, for SHRP 2 driver data

2.4.2 Case study 2: road infrastructure safety evaluation

The road infrastructure dataset includes information from 5,238 short road segments, which are collected from 2012 to 2014 in the State of Washington. The length for each segment is 0.1 mile. The number of crashes with property damage is the safety response. A total of 32,298 crashes was observed during the study period. There is 59.9 percent of zero responses in the dataset. Million vehicle miles traveled is used as the exposure. For the 5,238 short road segments, the overall average crash rate is 2.7 per million vehicle miles traveled.

Table 2.4 lists the 12 covariates used in the analysis, including route type, whether the road segment is an entrance/exit, whether it is an intersection, whether it is a ramp, whether it is a wye connection, whether it is a divided highway, rural/urban, number of lanes, pavement type, friction, gradient, and horizontal curvature. For categorical variables, I list their number of observations and percentage in each stratum. For continuous variables, I present their mean and standard deviation.

The NB regression was implemented because of the existence of overdispersion. The estimated dispersion parameter is 2.06. The difference between bias-corrected coefficient estimates $\widetilde{\boldsymbol{\beta}}$ and regular MLEs $\widehat{\boldsymbol{\beta}}$, as well as the percentage change $\frac{\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}}{\widehat{\boldsymbol{\beta}}} \times 100\%$, are given in Table 2.5. The first stratum of each categorical variable is treated as the reference level, so there is no coefficient estimation and hence no bias correction for it.

Comparing the magnitude of bias correction along with the number of crashes (two columns under the "original dataset" tab), the bias correction is generally larger for a stratum having a smaller number of events. For example, the bias correction is the largest for the coefficient of 'BST' pavement type $(14.8 \times 10^{-3}, -2.5\%)$, which only has 20 crashes in its stratum. In addition, it is the number of events rather than the sample size that matters to the bias magnitude. For instance, there is only one observation for five lanes, but it has 103 crashes.

The bias correction magnitude for this stratum is trivial.

To test if the number of crashes affects the magnitude of bias, I also conducted bias correction for two hypothetical pavement datasets where the crash count and exposure of each road segment is reduced to only 1/6 and 1/12, respectively, of the original pavement dataset. The covariates used in the two hypothetical pavement datasets are the same as the original dataset. The original data were collected for three years. Reducing the crash count and exposure to 1/6 is thus like the data being collected for only half a year, resulting in the total number of crashes being 5,192. The number of crashes in each stratum of the categorical covariates and the bias correction magnitude can be found under the " 'half-year' dataset" tab in Table 2.5. Similarly, reducing the crash count and exposure to 1/12 is like the data collection only lasting for a quarter of a year, resulting in a total crash count of 2,502. Its results are under " 'quarter-year' dataset." After reducing the number of crashes, the percentage of zero responses are 76.7 percent and 82.5 percent for the "half-year" dataset and the "quarter-year" dataset, respectively. There is no longer a crash for the 'BST' pavement type, so "NA" (not available) appears in the corresponding bias magnitude places. By comparing the results from the original dataset and the two hypothetical datasets, I find that the magnitude of the correction gets larger when the number of crashes decreases. This testifies that the number of crashes is the factor that influences the magnitude of bias rather than the number of observations.

It is seen that the balance of event counts in one stratum compared to the reference stratum also matters to the magnitude of bias correction. For example, the magnitude for the coefficient of five lanes is smaller than that for three lanes, even though the number of crashes for five-lane road segments is rarer. The reason is that the number of crashes happening on five-lane roads is more comparable to the number of events occurring within the reference level. It is worth pointing out that how the bias correction will change the significance of

certain covariate is case-dependent. The significance can be directly related to the confidence interval, which is automatically adjusted based on the bias-corrected point estimator.

To sum up, the number of events is a key statistic affecting the magnitude of bias rather than the number of observations collected. The balance of event counts within different strata also plays a significant role to the bias magnitude.

2.5 Summary and discussion

The proposed decision-adjusted modeling framework uses a tailored model evaluation criterion to optimize the model according to the specific objective. Unlike conventional approaches that treat modeling and decision as two distinct modules, the proposed framework integrates decision objective into the modeling procedure. Under the decision-adjusted modeling framework, the optimal model selected is more suitable to the application in need. Case study shows the proposed model outperforms general model selection rules such as AUC. This is especially beneficial in imbalanced data scenarios, in which the number of one class dominates the other, or two classes have different cost unities.

The study confirms that using high g-force event rates can improve individual driver risk prediction, and that DEC and LAT events have a larger contribution than ACC events. To reach the maximum prediction power, the thresholds in determining high g-force events must be selected carefully under each decision rule. What covariates are included in the model also influence the optimal threshold selected. In addition to identifying risky drivers, the decision-adjusted framework could provide guidance to industry for algorithm development in applications such as determining when to activate an in-vehicle warning system. The triggers should be selected based on specific application objective. An offline driver risk prediction model optimization with a driver sample is recommend before the implementation

of in-vehicle warning system. If offline selection is not feasible, we suggest to use $0.30g$ as the trigger for an elevated acceleration event (ACC), $0.46g$ for an elevated acceleration event (DEC), and $0.50g$ for an elevated lateral acceleration event (LAT).

There is a trade-off between computational cost and prediction performance. Previous sections show that utilizing the decision-adjusted approach improves prediction precision. However, decision-adjusted modeling is more computationally expensive compared to the traditional model evaluation criteria. Whenever there is a new study objective, the tuning parameters have to be calculated again, as opposed to the off-the-shelf thresholds given by the conventional approach.

The relationship between speed and high g-force events can be incorporated into the proposed decision-adjusted modeling framework. Intuitively, it is less likely to have an event with large g-force at high speed compared to low speed. It is not sufficient to derive the high g-force events from constant thresholds. Therefore, selecting the threshold as a function of velocity is one direction for next-step improvement. Discretization of the velocity region into different speed bins or use of functional models could provide possible solutions.

The decision-adjusted modeling framework provides a tailored solution to optimize modeling outputs for a specific research objective. The study also confirms that driving kinematic signatures provide useful information for crash risk prediction and that the thresholds for kinematic events are critical for driver-level crash risk prediction.

Table 2.4: Descriptive statistics of pavement data

Categorical variable	**Levels**	**Frequency**	**Percentage**
Route type	Interstate	2,236	42.7%
	State route	1,160	22.1%
	United States route	1,842	35.2%
Entrance/Exit	No	5,163	98.6%
	Yes	75	1.4%
Intersection	No	4,906	93.7%
	Yes	332	6.3%
Ramp	No	4,759	90.9%
	Yes	479	9.1%
Wye Connection	No	5,211	99.5%
	Yes	27	0.5%
Divided highway	No	1,883	35.9%
	Yes	3,355	64.1%
Rural/Urban	Rural	3,467	66.2%
	Urban	1,771	33.8%
No. of lanes	1	1,752	33.4%
	2	2,481	47.4%
	3	624	11.9%
	4	38	7.3%
	5	1	0.0%
Pavement type	Asphalt Concrete (ACP)	3,846	73.4%
	Portland Cement Concrete (PCCP)	1,229	23.5%
	Bituminous Surface (BST)	119	2.3%
	ACP/PCCP[1]	44	0.8%
Continuous variable	**Range**	**Mean**	**Std. deviation**
Friction	5.0 - 85.2	51.8	13.4
Gradient	0.0 - 6.4	1.4	1.4
Horizontal Curvature	0.0 - 10.5	0.9	1.1

[1] Half of the road segment is ACP and the other half is PCCP.

Table 2.5: Bias magnitude for the explanatory variables of pavement data

Variables	**original dataset**		**"half-year" dataset**		**"quarter-year" dataset**	
	No. of crashes	$\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$ ($\times 10^{-3}$) **(pct change)**	**No. of crashes**	$\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$ ($\times 10^{-3}$) **(pct change)**	**No. of crashes**	$\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}$ ($\times 10^{-3}$) **(pct change)**
Route type						
I	29543		4851		2377	
SR	1053	0.2 (0.0%)	134	8.0 (-0.8%)	51	39.1 (-2.1%)
US RTE	978	0.2 (-0.1%)	110	9.0 (-1.0%)	35	43.4 (-2.5%)
Entrance/Exit						
0	31047		5006		2423	
1	527	0.0 (0.0%)	89	0.6 (0.1%)	40	4.0 (1.3%)
Intersection						
0	30836		4986		2410	
1	738	0.0 (0.0%)	109	0.8 (0.1%)	53	-3.0 (-0.3%)
Ramp						
0	22641		3610		1738	
1	8933	0.0 (0.0%)	1485	0.0 (0.0%)	725	0.1 (0.1%)
Wye Connection						
0	31381		5063		2447	
1	193	0.0 (0.0%)	32	-1.2 (-0.1%)	16	-3.0 (-0.2%)
Divided highway						
0	1384		174		69	
1	30190	0.0 (0.0%)	4921	3.6 (-0.2%)	2394	19.7 (-0.9%)
Rural/Urban						
Rural	2422		291		101	
Urban	29152	0.0 (0.0%)	4804	-1.4 (-0.1%)	2362	-4.7 (-0.3%)
No. of lanes						
1	910		100		33	
2	6270	0.0 (0.0%)	943	-1.6 (-0.5%)	433	-6.7 (-1.2%)
3	12904	0.0 (0.0%)	2132	-1.8 (-0.3%)	1038	-7.4 (-0.8%)
4	11387	0.0 (0.0%)	1903	-1.8 (-0.5%)	950	-7.4 (-1.4%)
5	103	0.0 (0.0%)	17	0.0 (0.0%)	9	1.3 (0.1%)
Pavement type						
ACP	9911		1519		692	
ACP/PCCP	611	0.0 (0.0%)	101	0.7 (0.4%)	51	3.0 (1.2%)
BST	20	14.8 (-2.5%)	0	NA	0	NA
PCCP	21032	0.0 (0.0%)	3475	0.0 (-0.1%)	1720	-0.1 (-0.1%)
Friction		0.0 (0.0%)		0.0 (0.0%)		0.0 (0.0%)
Gradient		0.0 (0.0%)		0.0 (0.0%)		0.1 (0.1%)
Horizontal Curvature		0.0 (0.0%)		0.1 (0.2%)		0.2 (0.8%)

Chapter 3

Decision-adjusted Driver Risk Predictive Models using Kinematics Information

3.1 Introduction

Predicting crash risk and identifying high-risk drivers are critical for developing appropriate safety countermeasures, driver education programs, and use-based insurance systems. Predicting driver-level risk is challenging due to the numerous factors contributing to individual crash risk coupled with the rarity of crash events. With the rapid advancements in in-vehicle data instrumentation and connected vehicle technology, high-frequency driving data collection becomes more accessible and cost-effective. This trend toward increasingly comprehensive and available data is expected to continue as vehicle technology advances. Indeed it is likely that forthcoming automated vehicles need to capture and analyze data in order to ensure continuous safe operation is maintained by providing the capability to predict potential threats to safe operation before they manifest (we presume vehicle-level risk for automated systems is comparable to estimating driver-level risk such that methods defined herein may have implications for current and future transportation). These requirements provides a strong opportunity to improve our current driver risk models by better utilizing kinematic driving information. However, there are unique challenges to the use of such data in prediction; for example, determining the threshold for an abnormal deceleration. Consequently, there is an need to develop a comprehensive driver risk prediction framework that

effectively uses kinematic driving data.

The crash risk of individual drivers varies substantially [19, 20, 22, 31, 33, 72, 81]. Predicting and identifying high-risk drivers can provide important information for improving transportation safety. Studies have shown that a small percentage of drivers represent a disproportion volume of the total crashes. For example, Guo and Fang found that 6% of drivers account for 65% of the crashes and near-crashes [31]. Based on a simulation study, Habtemichael and de Picado-Santos showed that limiting the risk behavior of 4% to 12% of high-risk drivers would reduce crashes by 9% to 27% in different traffic conditions [33].

The quality and quantity of predictors are essential for risk models. Traditional driver risk prediction models mainly use demographic features, personality factors, crash/citation/violation history, and observable risky driving behavior as predictors, but their prediction power is typically limited [13, 31, 33, 76, 86]. Age, gender, and other demographic features often lead to a large cohort of drivers, which are not suitable for applications focusing on the small percentage of high-risk drivers. Personality measures come from different surveys. Not every driver would take a survey and the survey data themselves are usually not very accurate. Information on past crash history has been shown to be a good crash indicator but suffers from the regression-to-the-mean effect; i.e., a unit with a high number of crashes in the past, when observed in a future period, tends to have a lower number of crashes [2]. Observable risky driving behavior, such as distraction, also indicates driver risk, but the identification procedure can be challenging [68, 79, 88].

The increasingly prevalent of connected vehicle technology (referred to herein specifically as telematics) offers a promising source of information for risk prediction. Telematics involves sharing vehicle and driver information from vehicles to vehicles, vehicles to the roadside through leveraging advancements from interdisciplinary fields (e.g., telecommunications, vehicular technology, and computer science). Telematics data can include a variety of content;

however, for this research we are particularly interested in the time history kinematic data such as longitudinal acceleration, deceleration, and lateral acceleration. Aggressive driving behavior, such as hard braking, harsh acceleration, and sharp turning, could lead to abnormally high kinematic values. It is broadly suggested in the literature that kinematic signatures are useful for crash risk prediction [3, 7, 8, 73, 74, 75, 89]. Although kinematic data offer potential to identify risky drivers, limited research has been conducted to systematically address issues concerning how to properly use such information to maximize prediction performance. For example, there is no standard on what constitutes a risky kinematic signature, which is often defined by a simple subjectively selected acceleration threshold value.

The threshold values for the kinematic signatures used in the literature are not consistent. They were usually selected based on researchers' expertise. For example, Simons-Morton et al. defined "elevated gravitational event" as when acceleration exceeds $0.35g$, deceleration exceeds $0.45g$, or lateral acceleration exceeds $0.05g$ [74]. Klauer et al. explore the prevalence of events when acceleration or deceleration exceeds 0.30g among driver groups with different crash and near-crash rates [37].

With the recent availability of large-scale naturalistic driving studies (NDS) such as the Second Strategic Highway Research Program (SHRP 2) NDS, it is feasible and necessary to fully investigate which thresholds can provide the most prediction power for identifying risky drivers. One commonly used approach to measure the performance of driver risk models is based on the area under the curve (AUC) of the receiver operating characteristics (ROC) curve [31, 74]. However, AUC and other general model selection methods do not necessarily provide the optimal solution for a specific objective, e.g., identifying the top 1% riskiest drivers.The AUC criterion is with respect to the entire spectrum of possible threshold settings (resulting true positive rate and false positive rate range from zero to one); therefore, the model with the largest AUC is not necessarily optimal for a specific

objective for decision making. This could have a substantial impact when the objective is to identify a small percentage of the riskiest drivers. A specific objective is often best served by a dedicated prediction model that is optimized with respect to the decision rule corresponding to the objective rather than by generic criteria, e.g, a decision rule can be to maximize the predictive precision for the top 1% highest risk drivers.

In this paper, we propose a decision-adjusted modeling approach to develop an optimal driver risk prediction model and to estimate kinematic thresholds by optimizing a prespecified decision rule rather than relying on subjectivity as with previous approaches. Under this framework, model estimation will be conducted to optimize a decision-based model evaluation criterion. This framework is applied to develop the optimal driver risk prediction models under different decision rules. Within our approach, we also focus on incorporating a broader set of fused telematics data in the risk prediction models, in which the parameters to be adjusted are the thresholds of kinematic signatures such as elevated longitudinal and lateral acceleration. We demonstrate the proposed framework using data from the SHRP 2 NDS. The proposed model is compared with a traditional driver-characteristics-based model as well as a model optimized using a generic AUC criterion.

The remainder of the paper is organized as follows. Section 2 details the proposed decision-adjusted predictive modeling framework. Section 3 introduces SHRP 2 NDS data, followed by a formal definition and a full exploration of kinematic signatures using different threshold values. Section 4 quantifies the model improvement by comparing our proposed model to a model using traditional features and a model selected by AUC. Section 5 provides a summary and discussion.

3.2 Decision-adjusted modeling framework

Traditional statistical model selection methods are based on certain statistical criteria, such as the likelihood ratio test or the AUC value. The resultant model may not be adequate for the specific decision goal. For example, in predicting high-risk drivers based on logistic regression, a model selected by AUC might perform poorly when the goal is used to identify a small percentage of the riskiest drivers because the AUC criterion selects a model with respect to the entire spectrum of possible decision points. Thus, our work focuses on overcoming this limitation by targeting a small percentage of riskiest drivers associated with just one particular decision point on the ROC curve. The decision-adjusted modeling framework directly formulates the study's specific goal through a decision-based objective function in the model selection/optimization process. The model selection/optimization process involves model form determination, variable selection, and parameter tuning. The model form depends on the response variable type. For binary response data, potential models include logistic regression, decision tree, and neural network, etc. Variable selection determines which covariates should be included. The parameter tuning refers to the hyperparameter tuning for the selected model form and to the critical value adjustment in building certain predictor variables. For example, in predicting driver risk using kinematic signatures, the critical value adjustment is with respect to the threshold values for defining a kinematic signature that will maximize the prediction power.

3.2.1 Decision-adjusted driver risk prediction

The specific objective of a given driver safety management program could vary substantially based on the specific use case. For example, in current fleets it is common that only a certain number of drivers can be trained or provided with advanced safety countermeasures

due to limited resources. A specific study goal in this context may be to identify a targeted number of the riskiest drivers for receiving such aforementioned safety enhancing approaches/technologies. This allows fleet for wide safety improvements while minimizing the total cost of implementing countermeasures by directly addressing the risky drivers. The decision-adjusted framework will provide better support for the specific objective than models based on generic criteria.

The kinematic information is incorporated in the driver risk prediction model in the form of high g-force event rates. The high g-force events are derived from kinematic signatures when acceleration (ACC), deceleration (DEC), or lateral acceleration (LAT) exceed certain thresholds. Therefore, determining the optimal threshold values for kinematic variables is a crucial component of the model selection/optimization process.

The driver risk prediction model selection/optimization process is specified as follows. Suppose that a prediction model $\mathcal{M}$ is developed from data $\{\boldsymbol{X}_i, \boldsymbol{R}_i(\boldsymbol{\delta}), Y_i\}_{i=1}^n$, where Y_i is the indicator variable for driver $i, i = 1, 2, \ldots, n$,

$$Y_i = \begin{cases} 1, & \text{driver } i \text{ involved in at least one crash,} \\ 0, & \text{driver } i \text{ not involved in any crashes;} \end{cases}$$

$\boldsymbol{X}_i$ is a vector of traditional explanatory variables such as age, gender, and other characteristics; and $\boldsymbol{R}_i(\boldsymbol{\delta})$ is a vector of high g-force event rates, which depends on threshold value $\boldsymbol{\delta} = (\delta_{ACC}, \delta_{DEC}, \delta_{LAT})$. That is,

$$\boldsymbol{R}_i(\boldsymbol{\delta}) = \big(RA_i(\delta_{ACC}), RD_i(\delta_{DEC}), RL_i(\delta_{LAT})\big).$$

The model $\mathcal{M}$ predicts the probability of driver i being involved in a crash, $\Pr(Y_i = 1)$, with

$\mathbf{X}$ and kinematic metrics $\mathbf{R}(\boldsymbol{\delta})$. The predicted label for this driver is as follows:

$$Z_i = Z_i\big(\mathbf{X}, \mathbf{R}(\boldsymbol{\delta}), \mathcal{M}, \tau\big) = \begin{cases} 1, & \text{if } \widehat{\Pr}(Y_i = 1) > \tau, \\ 0, & \text{if } \widehat{\Pr}(Y_i = 1) \leq \tau, \end{cases}$$

where τ is a cutoff value usually determined by the specific study objective. For example, in the application of providing safety improvement education for a targeted number of the highest risk drivers, τ will be selected such that the number of drivers with label $Z = 1$ satisfies the requirement of the decision of interest.

With the above setup, the decision-adjusted modeling approach can be described as an optimization problem.

$$\max_{\boldsymbol{\delta}, \mathcal{M}} (or \min) \quad \eta\Big(\boldsymbol{Z}\big(\mathbf{X}, \mathbf{R}(\boldsymbol{\delta}), \mathcal{M}, \tau\big), \boldsymbol{Y}\Big), \tag{3.1}$$

where $\eta(\cdot)$ is the decision-adjusted objective function, which is a function of $\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_n)$ and $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)$. If the specific goal is to identify the top 10% of riskiest drivers, the objective function would be to maximize the prediction precision, the percentage of correct identification among those 10% drivers with the predicted label $Z = 1$. That is,

$$\max \quad \eta(\boldsymbol{Z}, \boldsymbol{Y}) = \frac{\boldsymbol{Z}'\boldsymbol{Y}}{\mathbf{1}'\boldsymbol{Z}} = \frac{\sum_{i=1}^{n} Y_i \cdot Z_i}{\sum_{i=1}^{n} Z_i},$$

such that

$$\mathbf{1}'\boldsymbol{Z} = \sum_{i=1}^{n} Z_i = 10\% \cdot n.$$

Similarly, for the example of identifying at least 50% of the drivers who would have crashes, the decision-adjusted objective function is

$$\min \quad \eta(\boldsymbol{Z}, \boldsymbol{Y}) = \mathbf{1}'\boldsymbol{Z} = \sum_{i=1}^{n} Z_i,$$

such that

$$\frac{\boldsymbol{Z}'\boldsymbol{Y}}{\mathbf{1}'\boldsymbol{Y}} = \frac{\sum_{i=1}^{n} Y_i \cdot Z_i}{\sum_{i=1}^{n} Y_i} \geq 50\%.$$

As to the last example of minimizing the average cost of running a driver education program, the objective becomes to minimize the expected average cost, which can be expressed as

$$\begin{aligned} \min \quad \eta(\boldsymbol{Z}, \boldsymbol{Y}) = & \frac{1}{n}\big[\boldsymbol{Y}'(\mathbf{1} - \boldsymbol{Z}) \cdot C(-|+) \\ & + \boldsymbol{Z}'(\mathbf{1} - \boldsymbol{Y}) \cdot C(+|-)\big] \end{aligned}$$

where $C(-|+)$ is the cost of applying safety countermeasures to a driver who would have no crashes, and $C(+|-)$ is the cost of crashes for a driver who is labeled as low-risk. There would be no cost for correctly identified drivers (i.e., $Z_i = Y_i$).

The workflow of the proposed decision-adjusted modeling can be found in Figure 3.1. The choice of prediction model $\mathcal{M}$ can come from a variety of models, such as logistic regression, gradient boosting tree, support vector machine, or neural network.

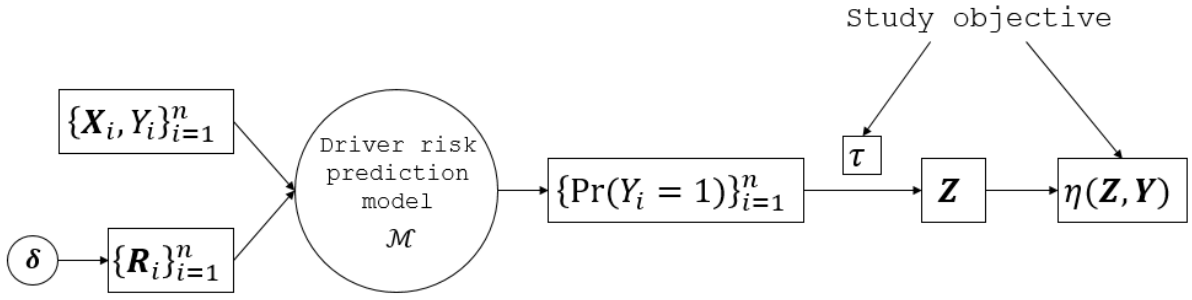

Figure 3.1: The workflow of decision-adjusted driver risk prediction model. (η is the objective function to be optimized; the optimization is achieved by adjusting threshold values $\boldsymbol{\delta}$ and the model $\mathcal{M}$.)

3.3 Application and case study

I apply the proposed method to the SHRP 2 NDS. The primary objective is to identify a targeted percentage of drivers as "high-risk." The corresponding decision-adjusted objective function is to maximize the prediction precision, the percentage of correctly identified high-risk drivers. Model performance was evaluated under the study objective of identifying certain percentage of the riskiest drivers, and I elaborate the prediction precision and the optimal threshold values of high g-force event rates when the target percentage of high-risk drivers is 1% to 20%.

3.3.1 The SHRP 2 NDS data

The SHRP 2 NDS was a large-scale observational study with more than 3,500 drivers from six data collection sites in Florida, Indiana, North Carolina, New York, Pennsylvania, and Washington. The study collected a wide range of information for drivers under natural driving conditions. A data acquisition system, including radar, multiple cameras, three dimensional accelerometers, and other equipment, was installed in each participant's vehicle [12]. The driving data were collected continuously from ignition-on to ignition-off. The data were collected asynchronously; for example, videos at 10 Hz, GPS at 1 Hz, and acceleration at 10 Hz. The SHRP 2 NDS provides a great opportunity to address questions about driver performance and behaviour, as well as traffic safety.

The crashes were identified thorough a comprehensive process. The kinematic time series data were screened through an automated algorithm. The driving segments that were potentially crashes were manually examined through the videos to confirm whether a crash had actually occurred [34]. This study includes 1,149 crashes from severity Level 1, the most severe crashes involving airbag deployment or potential injury, to Level 3, minor crashes in

which the vehicle makes physical contact with another object or departs the road.

After cleaning data, there are 3,440 drivers used for driver risk prediction modeling. Among the participants evaluated, 810 drivers (23.5%) experienced at least one crash (Level 3 or more severe) during the study period. The study also collected demographic and personality information such as age, gender, and sleep habits at the beginning of the study. Table 3.1 shows the descriptive statistics for the traditional predictors used in our study. Personality factors come from survey data. For example, the driving knowledge survey is a questionnaire compiled from a number of Department of Motor Vehicle (DMV) driving knowledge tests, and the score represents the number of questions answered correctly (out of 19). The clock drawing assessment is used as a screening tool to help identify possible signs of dementia or other neurological disorders. Mileage history, number of violations/crashes, insurance status, and specific driving behavior are self-reported. The number of violations/crashes refers to the past three years, and the participants reported their specific driving behavior history for the past 12 months.

Table 3.1: Description of covariates used in driver risk prediction models

Covariates	**Descriptive Statistics**
Demographic features	
Gender	F: 1796(52.21%); M: 1644(47.79%)
Age group	16-19: 534(15.52%); 20-24: 741(21.54%); 25-29: 280(8.14%); 30-34: 164(4.77%); 35-39: 129(3.75%); 40-44: 117(3.4%); 45-49: 149(4.33%); 50-54: 165(4.8%); 55-59: 144(4.19%); 60-64: 149(4.33%); 65-69: 209(6.08%); 70-74: 173(5.03%); 75-79: 262(7.62%); 80-84: 155(4.51%); 85-89: 59(1.72%); 90-94: 8(0.23%); 95-99: 2(0.06%)
Marital status	Single: 1514(44.01%); Married: 1384(40.23%); Divorced: 199(5.78%); Widow(er): 205(5.96%); Unmarried partners: 111(3.23%); NA: 27(0.78%)
Education level	Some high school: 272(7.91%); High school diploma or G.E.D.: 321(9.33%); Some education beyond high school but no degree: 980(28.49%); College degree: 913(26.54%); Some graduate or professional school, but no advanced degree (e.g., J.D.S., M.S. or Ph.D.): 352(10.23%); Advanced degree (e.g., J.D.S., M.S. or Ph.D.): 584(16.98%); NA: 18(0.52%)
Income level	Under $29K: 576(16.74%); $30K to $40K: 396(11.51%); $40K to $50K: 321(9.33%); $50K to $70K: 572(16.63%); $70K to $100K: 611(17.76%);$100K to $150K: 502(14.59%); $150K +: 245(7.12%); NA: 217(6.31%)
Work status	Not working outside the home: 1237(35.96%); Part-time: 945(27.47%); Full-time: 1214(35.29%); NA: 44(1.28%)
Having children at home or not	No: 2423(70.44%); Yes: 681(19.8%); NA: 336(9.77%)
Sleep duration	Sufficient: 1730(50.29%); Slightly insufficient: 1130(32.85%); Markedly insufficient: 212(6.16%); Very insufficient or did not sleep at all: 17(0.49%); NA: 351(10.2%)
Data collection site	FL: 768(22.33%); IN: 275(7.99%); NC: 558(16.22%); NY: 772(22.44%); PA: 263(7.65%); WA: 804(23.37%)
Personality factors	
Driving knowledge survey score	mean: 15.1; std. dev: 2.0; NA: 351(10.2%)
Clock drawing score	mean: 2.1; std. dev: 1.0; NA: 259(7.53%)
Barkley's ADHD Score	mean: 3.2; std. dev: 2.2; NA: 234(6.8%)
Sensation seeking survey score	mean: 14.6; std. dev: 6.9; NA: 242(7.03%)
Driving behavior	
Driving hours in the study	mean: 337.7; std. dev: 243.9
Mileage last year (mile)	mean: 12117.1; std. dev: 9530.8; NA: 251(7.3%)
Annual mileage	10K - 15K miles: 1078(31.34%); 15K - 20K miles: 458(13.31%); 20K - 25K miles: 219(6.37%); 25K - 30K miles: 121(3.52%); 5K - 10K miles: 901(26.19%); less than 5K miles: 412(11.98%); more than 30K miles: 194(5.64%); NA: 57(1.66%)
Driving experience (years)	mean: 25.6; std. dev: 22.5; NA: 23(0.67%)
Number of violations	0: 2452(71.28%); 1: 664(19.3%); 2 or More: 305(8.87%); NA: 19(0.55%)
Number of crashes	0: 2548(74.07%); 1: 667(19.39%); 2 or More: 197(5.73%); NA: 28(0.81%)
Number of crashes at fault	0: 458(13.31%); 1: 343(9.97%); 1 or More: 50(1.45%); NA: 2589(75.26%)
Insurance status	No: 47(1.37%); Yes: 3342(97.15%); NA: 51(1.48%)
Run red lights past 12 mo	Never: 2062(59.94%); Rarely: 1030(29.94%); Sometimes: 84(2.44%); Often: 7(0.2%); NA: 257(7.47%)
Drive sleepy past 12 mo	Never: 1479(42.99%); Rarely: 1409(40.96%); Sometimes: 279(8.11%); Often: 15(0.44%); NA: 258(7.5%)
Impatiently pass on the right	Never: 824(23.95%); Hardly Ever: 1022(29.71%); Occasionally: 1045(30.38%); Quite Often: 194(5.64%); Frequently: 74(2.15%); Nearly All the Time: 21(0.61%); NA: 260(7.56%)
Brake aggressively	Never: 1926(55.99%); Hardly Ever: 1114(32.38%); Occasionally: 109(3.17%); Quite Often: 5(0.15%); NA: 286(8.31%)
Involved in racing	Never: 2963(86.21%); Often: 6(0.17%); Rarely: 167(4.86%); Sometimes: 28(0.81%); NA: 273(7.94%)
Racing frequency	Frequently: 3(0.09%); Hardly Ever: 214(6.22%); Nearly All the Time: 1(0.03%); Never: 2904(84.42%); Occasionally: 41(1.19%); Quite Often: 0(0%); NA: 277(8.05%)
Nod off while driving	No: 2570(74.71%); Yes: 503(14.62%); NA: 367(10.67%)
Nod off while driving frequency	1-2 times per month: 41(1.19%); 1-2 times per week: 9(0.26%); 3-4 times per week: 2(0.06%); Nearly every day: 1(0.03%); Never: 10(0.29%); Rarely: 426(12.38%); NA: 2951(85.78%)

The high g-force event rates are calculated from a total of 1,161,112 driving hours. The longitudinal and lateral acceleration are measured in terms of the acceleration of gravity (g; $1g = 9.8m/s^2$)). Let a_x and a_y denote the longitudinal (x-axis) acceleration and lateral (y-axis) acceleration while driving. Positive a_x indicates acceleration and negative indicates deceleration. Positive and negative a_y mean lateral acceleration to the left and right. A high ACC g-force event is defined when $a_x > \delta_{ACC}$; a high DEC g-force event if $a_x < -\delta_{DEC}$; and a high LAT g-force event if $|a_y| > \delta_{LAT}$. The corresponding event rates per driving hour for driver i, RA_i, RD_i, RL_i, are calculated as

$$
\begin{aligned}
RA_i(\delta_{ACC}) &= \frac{1}{T_i} \sum_{driver\, i} \mathbb{1}(a_x > \delta_{ACC}); \\
RD_i(\delta_{DEC}) &= \frac{1}{T_i} \sum_{driver\, i} \mathbb{1}(a_x > -\delta_{DEC}); \\
RL_i(\delta_{LAT}) &= \frac{1}{T_i} \sum_{driver\, i} \mathbb{1}(|a_y| > \delta_{LAT}),
\end{aligned}
$$

where $\sum_{driver\, i} \mathbb{1}(\cdot)$ represents the number of high g-force events and T_i is the total driving time for driver i. For an actual hard brake or other maneuver, there is typically a sequence of multiple data points beyond the threshold value. To identify the number of events rather than the number of data points, a moving average filter was applied to the raw data and several criteria were used to cluster data points from a potential high g-force event into one event, including that the data points should be close to each other temporally and the smoothed data should be a local maximum pattern.

Setting the thresholds too low or too high could result in non-informative explanatory variables in the driver risk prediction model. The number of high g-force events varies substantially with respect to different thresholds. Figure 3.2 shows the total number of ACC, DEC, and LAT for all participants versus the thresholds chosen. The counts decrease exponentially as thresholds increase. If the thresholds for ACC, DEC, and LAT are all set at $0.30g$, there

will be millions of high g-force events. The g-force events based on low thresholds are likely to be dominated by normal driving behaviors, such as stopping before traffic signals. As such, low thresholds will mask true high-risk, aggressive driving behavior, which will be a detriment to its ability to predict individual driver risk. On the other hand, if the thresholds are set too high, very few events will be identified. High g-force events themselves become rare events, and even risky drivers may rarely encounter any in their driving history. There should be an optimal threshold value to best distinguish risky drivers from safe drivers. The decision-adjusted driver risk prediction framework proposed in this chapter can identify the optimal thresholds.

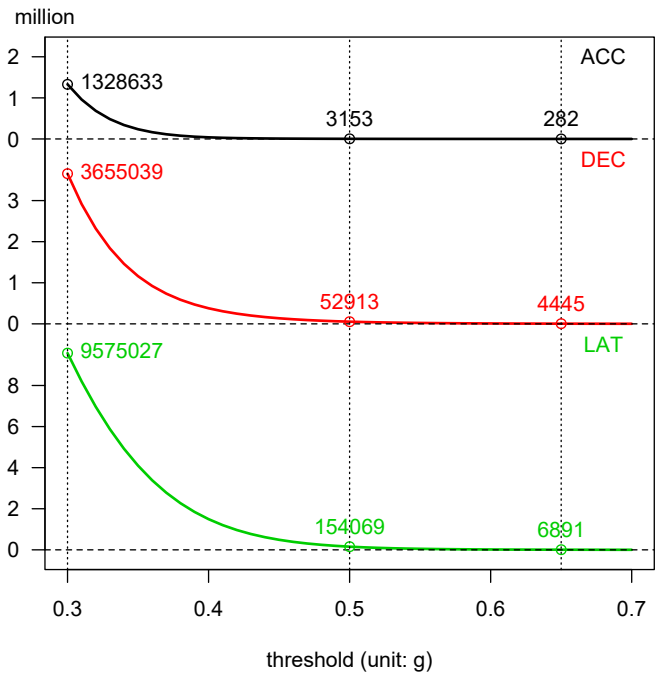

Figure 3.2: Total number of high g-force events (ACC, DEC, LAT) versus the thresholds chosen

3.3.2 Decision-adjusted driver risk prediction through regularized logistic regression

A variety of model formats, such as generalized linear, gradient boosting tree, and neural network, can be applied to driver risk prediction. Here I adopt a regularized logistic regression (elastic net) for model development and performance evaluation [27, 90]. Specifically, the risk for a driver with predictor information $\boldsymbol{X}_i$ and kinematic metrics $\boldsymbol{R}_i(\boldsymbol{\delta})$ is formulated as

$$\Pr(Y = 1|\boldsymbol{X}_i, \boldsymbol{R}_i(\boldsymbol{\delta})) = \frac{\exp\left((\boldsymbol{X}_i, \boldsymbol{R}_i(\boldsymbol{\delta}))^T\boldsymbol{\beta}\right)}{1 + \exp\left((\boldsymbol{X}_i, \boldsymbol{R}_i(\boldsymbol{\delta}))^T\boldsymbol{\beta}\right)}, \tag{3.2}$$

where $\boldsymbol{\beta}$ is the regression coefficient. The model fitting is accomplished by maximizing a penalized log-likelihood function

$$\ell(\boldsymbol{\beta}) - \lambda[(1-\alpha)\frac{1}{2}||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1], \tag{3.3}$$

where $\ell(\boldsymbol{\beta})$ is the log-likelihood function and $\lambda[(1-\alpha)\frac{1}{2}||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1]$ is the regularization part with $||\cdot||_2$ being the l_2 norm and $||\cdot||_1$ being the l_1 norm. $0 \leq \alpha \leq 1$ and $\lambda \geq 0$ are two hyperparameters specific to the regularized logistic regression model. Lasso regression and ridge regression are two special cases of elastic net when $\alpha = 1$ and $\alpha = 0$.

With elastic net, I consider the optimization problem of the proposed decision-adjusted method as

$$\max_{\boldsymbol{\delta},\lambda,\alpha} / \min \quad \eta\Big(\boldsymbol{Z}\big(\boldsymbol{X}, \boldsymbol{R}(\boldsymbol{\delta}), \hat{\boldsymbol{\beta}}(\lambda,\alpha), \tau\big), \boldsymbol{Y}\Big). \tag{3.4}$$

The optimal model can be obtained by adjusting the threshold values $\boldsymbol{\delta}$ and hyperparameters λ and α.

The traditional driver risk prediction model without telematics variables can be considered as a special case of the decision-adjusted model. If the threshold values $\delta_{ACC}, \delta_{DEC}, \delta_{LAT}$ are set to be sufficiently large, every driver would have the same high g-force event rates

(i.e., $\boldsymbol{R}_i(\boldsymbol{\delta}) = \mathbf{0}, i = 1, 2, \ldots, n$). The event rates become non-informative predictors, and the decision-adjusted driver risk prediction model degenerates to the traditional models.

I compared three driver risk prediction models: two benchmark models $\mathcal{M}_0$, $\mathcal{M}_1$, and the proposed model $\mathcal{M}_2$. Specifically, $\mathcal{M}_0$: traditional driver risk prediction without high g-force events; $\mathcal{M}_1$: driver risk prediction with high g-force event rates, optimized by AUC; $\mathcal{M}_2$: decision-adjusted driver risk prediction with high g-force event rates.

The predictors and modeling strategy for the three models are listed in Table 3.2. The difference between $\mathcal{M}_0$ and $\mathcal{M}_1$ is that $\mathcal{M}_1$ also includes the occurrence rates of ACC, DEC, and LAT as predictors, along with other traditional covariates listed in Table 3.1. $\mathcal{M}_1$ selects the threshold values for ACC, DEC, and LAT by maximizing the AUC of the prediction model. Its α is set to be 1 (lasso), and its λ is chosen to minimize the 10-fold cross-validation error ($\lambda = 0.0074$). $\mathcal{M}_2$ uses the same set of predictors as $\mathcal{M}_1$. $\mathcal{M}_1$ and $\mathcal{M}_2$ differ by the modeling strategies. $\mathcal{M}_2$ tunes its threshold values and model hyperparameters to optimize the specific study objective compared to the AUC by $\mathcal{M}_1$.

Table 3.2: Description of three comparison methods

	Predictors	Modeling Strategy
$\mathcal{M}_0$	traditional covariates[1]	elastic net
$\mathcal{M}_1$	traditional covariates[1] ACC, DEC, and LAT rates	elastic net adjusted by AUC
$\mathcal{M}_2$	traditional covariates[1] ACC, DEC, and LAT rates	decision-adjusted modeling

[1] The traditional covariates used in this study can be found in Table 3.1.

I conducted an exhaustive search of the parameter space to optimize $\mathcal{M}_1$ and $\mathcal{M}_2$. The candidate threshold values for $\delta_{ACC}, \delta_{DEC}, \delta_{LAT}$ range from $0.30\,g$ to $0.70\,g$, by $0.02\,g$ in equal steps. For the hyperparameters, $\log(\lambda)$ ranges from -8.5 to -2.5, by 0.4 in equal space,

and α ranges from 0 to 1, by 0.2 in equal steps.

R package 'glmnet' is used for the implementation of the elastic net [26]. For example, model $\mathcal{M}_1$ selects DEC rate when $\delta_{DEC} = 0.46g$, driving hours in the study, LAT rate when $\delta_{LAT} = 0.50g$, and 39 other variables. The coefficients are listed in Table 3.3, ordered by their magnitude. Coefficients for the other variables are shrunk to zero.

3.3.3 Prediction performance comparison

I use prediction precision to evaluate the performance of the three models. Prediction precision measures the percentage of correct identification among drivers who were labeled as high-risk, i.e. $\frac{\boldsymbol{Z}'\boldsymbol{Y}}{\boldsymbol{1}'\boldsymbol{Z}} = \frac{\sum_{i=1}^{n} Y_i \cdot Z_i}{\sum_{i=1}^{n} Z_i}$. Higher prediction precision indicates better prediction performance. Figure 3.3 shows the relative improvement of prediction precision for $\mathcal{M}_1$ and $\mathcal{M}_2$ compared to $\mathcal{M}_0$ when the targeted percentage of high-risk drivers is between 1% and 20%. Table 3.4 lists the corresponding number of true positives and prediction precision for the three models. The relative improvement of prediction precision for $\mathcal{M}$ compared to $\mathcal{M}_0$ is defined as $\frac{\text{precision}(\mathcal{M}) - \text{precision}(\mathcal{M}_0)}{\text{precision}(\mathcal{M}_0)}$.

Generally, our proposed method $\mathcal{M}_2$ performs the best among the three alternative models. The decision-adjusted model improves the prediction precision by 6.3% to 26.1% compared to the baseline model $\mathcal{M}_0$. It is also superior to $\mathcal{M}_1$ by 5.3% to 31.8%. The improvement is more prominent when identifying a small percentage of the riskiest drivers (e.g., $< 5\%$). $\mathcal{M}_1$ is also better than $\mathcal{M}_0$ when the target percentage of high-risk drivers is greater than 4%. The benefit can be credited to the inclusion of high g-force event rates. The results confirm that using kinematic information can improve individual driver risk prediction, and the improvement is more significant when a decision-adjusted modeling approach is applied. Furthermore, the substantial improvement of $\mathcal{M}_2$ when targeting a small percentage of high-risk drivers indicates that the decision-adjusted modeling is more desirable for highly

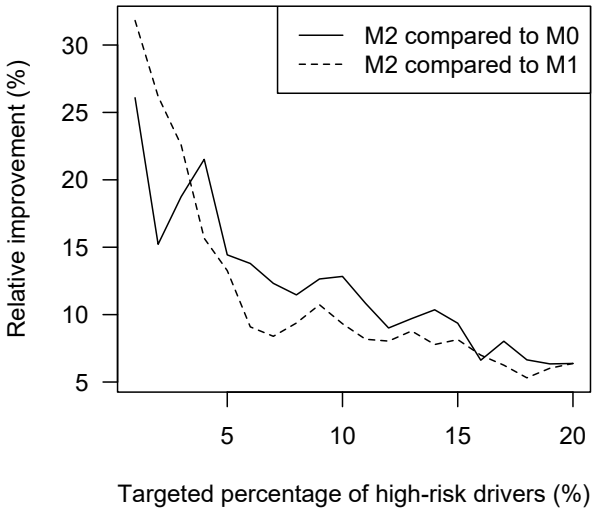

Figure 3.3: Relative improvement of prediction precision of the three models

imbalanced scenarios; for example, to predict a small percentage of high-risk drivers.

3.3.4 Optimal thresholds

The threshold values for ACC, DEC, and LAT have a substantial influence on the association between high g-force event rates and crash occurrence. Figure 3.4 shows the Spearman's rank-order correlation for the occurrence of a crash and three types of high g-force event rates under different threshold values. The correlation with DEC rate is the highest among the three types, followed by LAT rate, and the correlation with ACC rate is the weakest. The DEC rate reaches its maximum correlation with the occurrence of a crash when the threshold is selected at $0.50g$, resulting in a Spearman correlation of 0.226. The LAT rate and the ACC rate achieve their largest correlation with crash occurrence at $0.56g$ and $0.40g$, respectively.

The following two subsections illustrate in detail the optimal thresholds for $\mathcal{M}_1$ and $\mathcal{M}_2$,

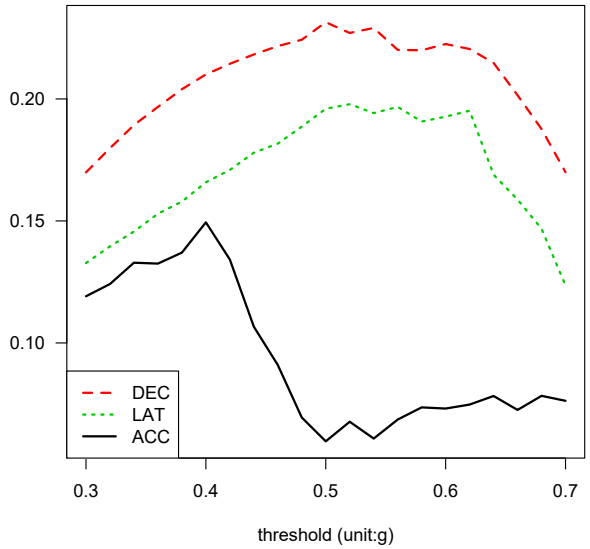

Figure 3.4: Spearman's rank-order correlation between high g-force event rates and crash occurrence at driver level

respectively. There is no threshold selection procedure for $\mathcal{M}_0$ as kinematic variables were not included.

Optimal thresholds for $\mathcal{M}_1$

Model $\mathcal{M}_1$ incorporates high g-force event rates and the traditional risk predictors. It chooses the threshold values of high g-force event rates to maximize AUC. For our driver risk prediction study, the optimal threshold values are

$$\boldsymbol{\delta} = (\delta_{ACC}, \delta_{DEC}, \delta_{LAT}) = (0.30g, 0.46g, 0.50g).$$

The two heat maps in Figure 3.5 represent the AUC values profiling $\delta_{ACC} = 0.30g$ and $\delta_{DEC} = 0.46g$, respectively. The green area means the corresponding threshold values would generate a relative small AUC value and the red to white area is for a relatively large AUC. The two black dots on the heat map represent the optimal threshold setting

$(0.30g, 0.46g, 0.50g)$. It generates a driver risk prediction model with an AUC value of 0.752. The color heterogeneity along the x-axis on the left heat map indicates that the resulting AUC value is sensitive to the selection of δ_{DEC}. On the right heat map, the major colors are yellow and red, which implies that when δ_{DEC} is $0.46g$, the resulting driver risk prediction model performs relatively acceptably in terms of AUC no matter what thresholds are chosen for ACC and LAT. For comparison, the AUC of the traditional driver risk prediction model, $\mathcal{M}_0$, is 0.742.

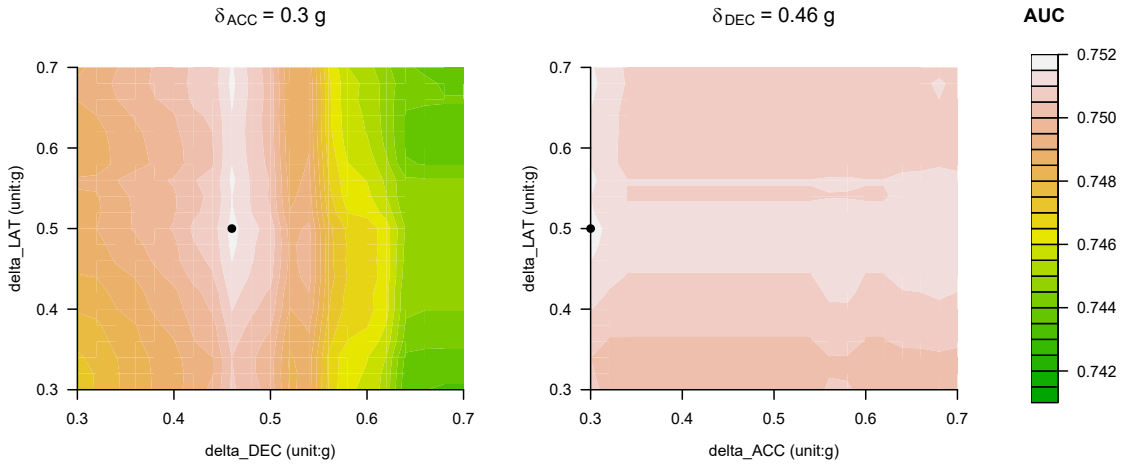

Figure 3.5: Heat map of $\mathcal{M}_1$'s AUC values evaluated on $\delta_{ACC} = 0.3\,g$ (left) and $\delta_{DEC} = 0.46\,g$ (right). The dots represent the optimal threshold setting of $\mathcal{M}_1$, $(\delta_{ACC}, \delta_{DEC}, \delta_{LAT}) = (0.30g, 0.46g, 0.50g)$, which maximizes the AUC value among all threshold settings evaluated.

Optimal thresholds for $\mathcal{M}_2$

For the proposed model, $\mathcal{M}_2$, the selected optimal thresholds could be different depending on the different decision rules under consideration. For example, when 20% is the target percentage of high-risk drivers, the optimal threshold setting is

$$\boldsymbol{\delta} = (0.62g, 0.48g, 0.60g), \quad \log(\lambda) = -6.9, \quad \alpha = 0.6.$$

Under such a setting, $\mathcal{M}_2$ can correctly identify 367 high-risk drivers out of the 688. When the preset target percentage is 10%, $\mathcal{M}_2$ performs the best at a different parameter combination,

$$\boldsymbol{\delta} = (0.64g, 0.70g, 0.30g), \quad \log(\lambda) = -8.1, \quad \alpha = 0.0.$$

Figure 3.6 shows the point-and-whisker-plot for $\delta_{ACC}, \delta_{DEC}$, and δ_{LAT} in the "top three" optimal settings when the target percentage of high-risk driver ranges from 1% to 20%. The connected points represent the mean of the "top three" optimal settings, and the length of the half whiskers represents their one standard deviation. Take the target percentage of 20% as an example: the optimal model can identify 367 high-risk drivers correctly (Table 3.4), and I plot the mean and $(+/-)$ one standard deviation for the settings that can correctly identify 367, 366, and 365 high-risk drivers. The "top three" optimal settings are considered to ensure the robustness of the settings I present. The optimal thresholds for $\delta_{ACC}, \delta_{DEC}$, and δ_{LAT} are calculated and presented separately.

The pattern differs for the three types of high g-force events. The means for the "top three" optimal δ_{ACC}s are about the same for different decision rules, and their whiskers are relatively long compared to the "top three" optimal δ_{DEC}s and δ_{LAT}s. This phenomenon indicates that the selection of δ_{ACC} is trivial for the driver risk prediction model. The prediction performance would be similar no matter what δ_{ACC} is chosen, while I should finely tune δ_{DEC} and δ_{LAT} to get achieve more precision.

3.4 Summary and discussion

Conventional approaches treat model development and decision making based on model outputs as two separate modules. The decision-adjusted modeling framework proposed in this study integrates decision objective and modeling procedure thus the models developed

is optimal for the specific research application. The approach is especially beneficial in imbalanced data scenarios, e.g. the number of observations from one category dominates the others for the model response. In the application of identifying a small percentage of highest risk drivers, the decision-adjust modeling framework demonstrates superior performance over general model selection criteria such as AUC of the ROC curve.

The rise of connected vehicle and automatic driving system makes high-frequency, high resolution telematics data widely available. Such telematics data provide rich information on vehicle kinematics, traffic environment, and drivers behavior. Driver's aggressive driving behavior identified via telematics, such hard brake, can provide crucial information on driver risk for many applications such as insurance, fleet safety management, and teenage driver risk management. Using the lastest NDS up-to-date, the SHRP 2 NDS, the study confirms that high g-force events, including longitudinal and lateral acceleration, provide crucial information on individual driver risk prediction.

I conduct a systematic and comprehensive evaluation of the optimal thresholds of high g-force events for driver risk evaluation using the decision-adjusted framework. The results demonstrate that the thresholds in determining high g-force events do vary under each decision rule. In addition, covariates included in the model also influence the optimal thresholds selected. In addition to identifying risky drivers, the decision-adjusted framework provides guidance to industry for algorithm development in applications such as algorithms for activate in-vehicle warning systems. A systematic optimization process as demonstrated in this paper is recommended for each new objective and implementation.

The decision-adjusted modeling framework provides a tailored solution to optimize models for a specific research objective. In the context of predicting a small percentage of highest risk drivers, the proposed model framework provide superior results compared to models based on conventional models. The results confirm that driving kinematic signatures provide

useful information for crash risk prediction and that the thresholds for kinematic events are critical for driver-level crash risk prediction. The methodology and results of this paper could provide crucial information for driver risk prediction, safety education, use-based insurance, driver behavior intervention, as well as connected-vehicle safety technology development.

Table 3.3: The non-zero coefficients for model $\mathcal{M}_1$

Variable	**Coefficient**
DEC rate when $\delta_{DEC} = 0.46g$	2.18
Driving hours in the study	2.17
LAT rate when $\delta_{LAT} = 0.50g$	0.74
Nod off while driving frequency. 1-2 times per week	0.48
Driving knowledge survey score	-0.37
Racing frequency. Nearly all the time	0.36
Education level. NA	-0.35
Nod off while driving frequency. Nearly every day	0.33
DEC rate when $\delta_{ACC} = 0.30g$	0.32
Age group. 80-84	0.27
Number of violations. 2 or more	0.26
Education level. Some high school	0.24
Sensation seeking scale survey score	0.22
Number of crashes at fault. 1 or more	0.19
Run red lights or not. NA	-0.18
Drive sleepy or not. Often	0.12
Age group. 45-49	-0.08
Education level. Some graduate or professional school, but no advanced degree (e.g., J.D.S., M.S., or Ph.D.)	-0.08
Marital status. Single	0.08
Insurance status.Yes	-0.08
Barkley's ADHD Score	0.07
Brake aggressively. Occasionally	0.06
Work status. NA	-0.06
Annual mileage. less than 5K miles	0.06
Nod off while driving frequency. 1-2 times per month	-0.05
Annual mileage. 15K - 20K miles	-0.05
Having children at home or not. Yes	-0.05
Age group.40-44	-0.05
Work status. Part-time	0.05
Mileage last year	-0.04
Education level. Some education beyond high school but no degree	0.04
Impatiently pass on the right. Never	0.04
Number of crashes at fault. NA	-0.03
Data collection site. WA	-0.03
Data collection site. PA	-0.03
Racing frequency. Rarely	0.03
Nod off while driving frequency. Rarely	0.03
Income level.Under $29,000	0.02
Education level. College degree	-0.02
Drive sleepy or not. Rarely	0.02
Run red lights or not. Rarely	0.01
Data collection site. NC	-0.01

Table 3.4: Prediction performance comparison of the three models

Risky pct	Number of True Positives			Prediction Precision[1]		
	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$
1%	23	22	29	67.6%	64.7%	85.3%
2%	46	42	53	66.7%	60.9%	76.8%
3%	64	62	76	62.1%	60.2%	73.8%
4%	79	83	96	57.7%	60.6%	70.1%
5%	97	98	111	56.4%	57%	64.5%
6%	116	121	132	56.3%	58.7%	64.1%
7%	138	143	155	57.3%	59.3%	64.3%
8%	157	160	175	57.1%	58.2%	63.6%
9%	174	177	196	56.3%	57.3%	63.4%
10%	187	193	211	54.4%	56.1%	61.3%
11%	203	208	225	53.7%	55%	59.5%
12%	222	224	242	53.9%	54.4%	58.7%
13%	237	239	260	53%	53.5%	58.2%
14%	251	257	277	52.2%	53.4%	57.6%
15%	267	270	292	51.7%	52.3%	56.6%
16%	287	286	306	52.2%	52%	55.6%
17%	299	304	323	51.2%	52.1%	55.3%
18%	316	320	337	51.1%	51.7%	54.4%
19%	331	332	352	50.7%	50.8%	53.9%
20%	345	345	367	50.2%	50.2%	53.4%

[1] Prediction precision $= \frac{\text{number of true positives}}{\text{number of predicted positives}}$

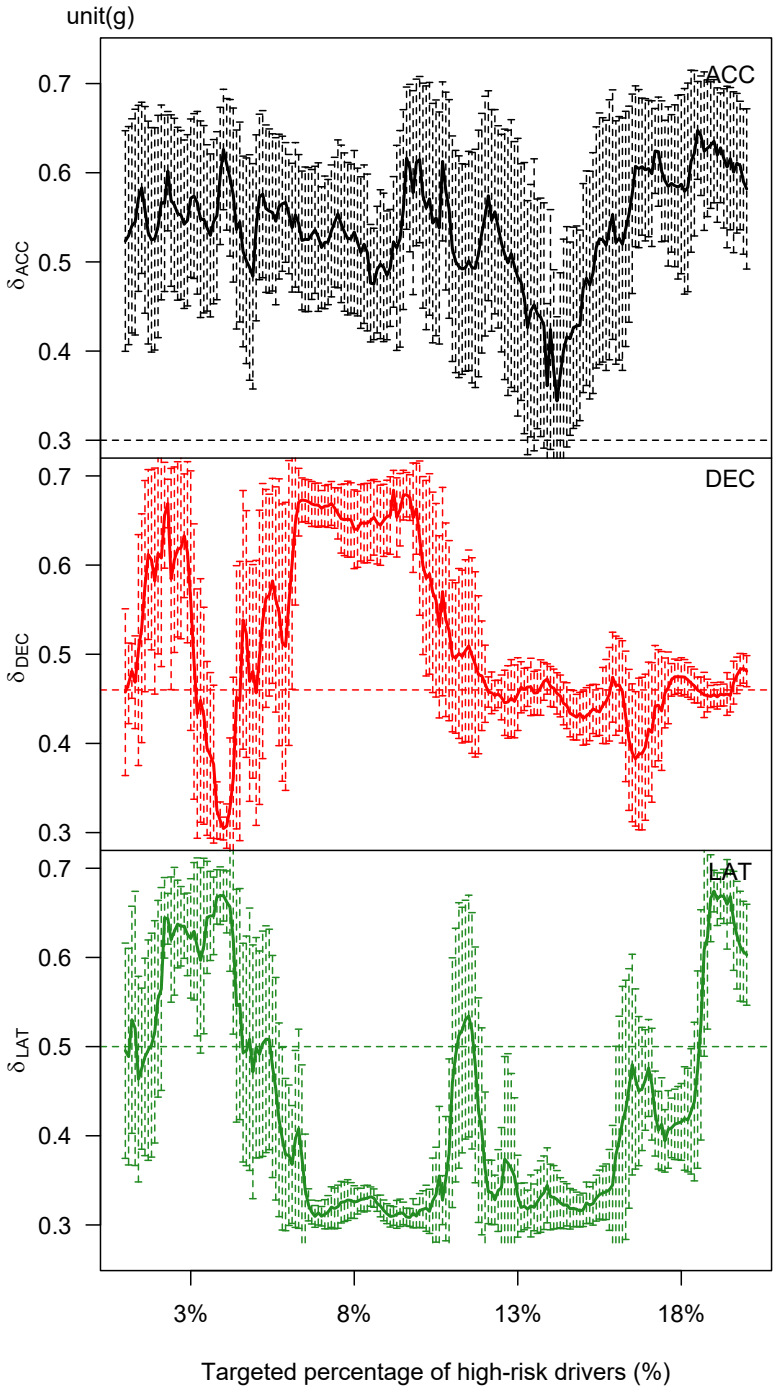

Figure 3.6: Point-and-whisker-plot of the "top three" optimal threshold settings under different decision rules for $\mathcal{M}_2$. The dashed lines show $\mathcal{M}_1$'s optimal threshold of for reference

Chapter 4

An Experimental Design Approach for Global Optimization with Applications in Kinematic Signature Identification

4.1 Introduction

Kinematic signatures have been broadly suggested in the literature in predicting traffic crash involvement and also as a measurement reflecting driver riskiness [3, 7, 8, 73, 74, 75, 89]. With the development of telecommunications and informatics, it is common and cost-effective to share vehicle's real-time information, including the kinematic signatures, to the surrounding infrastructures (V2I), to other vehicles (V2V), to cloud (V2C) in the traffic control center, to pedestrians (V2P) on the road, and to everything (V2X). The identification of informative kinematic signature has great potential in advancing fleet management, insurance pricing, and driver education, for a safer driving environment. In the literature, kinematic signature's identification heavily depends a predefined threshold value, and the threshold value usually does not reply on the driving speed. This chapter proposed a efficient search algorithm to determine the optimal threshold, especially the optimal threshold can be a function of the driving speed.

The threshold values for identifying the kinematic signatures are not consistent. Some researchers determine the threshold subjectively based on their expertise, and the kinematic

signature itself has different names within different literature. For example, Bagdadi and Várhelyi [8] used 'jerk' to denote the behavior of hard breaking, and the threshold for a jerk is $-9.9m/s^2$. Simons-Morton et al. [74, 75] adopted the name of 'elevated gravitational-force event' as the driving acceleration is usually measured in the unit of gravity (g), where $1g = 9.8m/s^2$. In their study, kinematic signatures in all directions are counted as elevated gravitational-force events, including rapid start when longitudinal acceleration exceeds $0.35g$, hard break when longitudinal deceleration exceeds $0.45g$, and sharp turns when lateral acceleration exceeds $0.05g$. In general, there is a trend from subjective to objective in selecting the threshold values for the kinematic signatures. For example, Klauer et al. [37] started to explore different thresholds such that a minimal amount of valid kinematic signatures lost and a reasonable amount of invalid signatures identified, where the validity of a signature is determined by human data reductionists. In Chapter 3, I use a decision-driven approach to determine the optimal threshold according to the study objective. I also started to adopt the name of 'high g-force (HGF) event.', and I specified the HGF event of acceleration as ACC, the HGF event of deceleration as DEC, and the HGF event of lateral acceleration as LAT. I herein use kinematic signature and HGF event indiscriminately.

The threshold values of HGF events in the literature are unrelated to speed. Intuitively, it is less likely to have a kinematic signature with large g-force at high speed compared to low speed, so the threshold should vary according to the driving speed. It is insufficient to extract HGF events from just constant thresholds regardless of the driving speed. Considering the threshold as a function of velocity can be more powerful in distinguishing the real risky drivers from the safe ones. There are multiple ways in constructing the functional form of the threshold in terms of velocity. In this chapter, I discretize the feasible speed region into different speed bins and write the threshold as a step function. Determining the optimal threshold value in each speed bin becomes a large-dimensional optimization problem.

In practice, the threshold value cannot be too low or too high. The g-force events based on low thresholds are likely to be dominated by normal driving behavior, such as stopping before traffic signals. As such, HGF events determined by low thresholds will mask true high-risk, aggressive driving behavior. Low threshold will be a detriment to HGF events' ability to accurately access driver risk. On the other hand, if the thresholds are set too high, very few events will be identified. HGF events become rare events, and even risky drivers may rarely encounter any in their driving history. HGF events would lost the distinguish power with high thresholds. The problem setting naturally determines the feasible region for the large-dimensional optimization problem.

There are two major directions in solving a large-dimensional optimization problem. The first direction is using gradient-based approaches, and the other is to utilize the response surface methodology [85]. Gradient-based approaches cannot work for unknown or non-differentiable target functions as gradient is the key component in this direction. Response surface methodology designs the points to evaluate in order to approximate the target function efficiently and accurately. A first-order model is used when the experimental region is far from the optimum region, followed by a second-order model when the experimental region gets closer. The traditional response surface method has an underlying assumption that there is only one single optimum, so it is easily to stuck in a local optimum.

Based on the traditional response surface method, I proposed a Multi-stratum Iterative Central Composite Design(miCCD) approach. The approach is constituted of two major parts: a multi-start scheme and local optimization. miCCD first finds multiple adequate points to start with using certain space-filling design. It works like a partition of the feasible region of the optimization problem. From all the space-filling points, miCCD selects several adequate points to further optimize within their vicinity. The iterative CCD part use a second-order design method to gradually converge to the local optimum. The multi-start

scheme prevents the optimization process stopping at a local optimum, and the iterative CCD ensures the convergence.

The rest of this chapter is organized as follows: Section 2 gives a literature review in gradient-based optimization approaches and traditional response surface method. Section 3 shows the specific steps of the proposed miCCD, followed by its application in determining the optimal threshold of kinematic signatures with driving speed taken into consideration in Section 4. Section 5 provides a summary and discussion.

4.2 Literature review

Gradient-based approach is not a uncommon way to solve an optimization problem. For example, Newton-Raphson method and Fisher Scoring method are commonly used to find the maximum likelihood estimate. They all need to calculate the first-order derivative and second-order derivative of the objective function. Gradient-based approaches solve the optimization problem efficiently. However, they cannot work for unknown or non-differentiable functions.

Response surface methodology provides another optimization strategy, which efficiently and accurately approximate the target function, denoted as $f(\cdot)$, and then finds the optimum. It involves iteratively selecting (design) points to evaluate, fitting a response surface based on the points and their corresponding response, and then marching to the optimum. Response surface methodology generally works in a sequential fashion, consisting mainly of two phases [85]. The main goal of the first phase is to get closer to the optimum of the response surface. Methods for the first phase include but not limit to steepest ascent search, rectangular grid search, and Latin hypercube sampling (LHS). When the experimental region is near the optimum, response surface methodology enters the second phase – using a second-order

model to approximate the response surface in the small region around the optimum and so to identify the optimum. A second-order model is able to capture the curvature effect of the response surface, while a first-order model (e.g., steepest ascent search) cannot. Commonly used second-order models are CCD, Box-Behnken design, and uniform shell design.

LHS is one of the space-filling designs, which is able to be close to the optimum with enough design points. LHS guarantees the spread of design points while the points are still being uniform and random. It divides each marginal coordinate into equal-sized segments and samples the design points in each segment so that each segment has just one point. Comparing to steepest ascent algorithm, LHS is more efficient in finding the region near the optimum. In a steepest ascent algorithm, the optimum point along the steepest ascent path will be taken as the center point for the next steepest ascent search. It iteratively get closer to the optimum region – the next center point depends on the result in the previous step. On the contrary, there is no order difference in evaluating the points in an LHS design. They can even be evaluated simultaneously in parallel computing.

The CCD is a commonly used second-order design in response surface studies. It is more efficient than a complete three-level factorial experiment. See Figure 4.1 for central composite design in two and three dimensions. Followed the notations in Wu and Hamada [85], a $K-$dimensional CCD contains

(i) n_f cube points, which come from a factorial design with K factors having two levels,

(ii) n_c center points,

(iii) $2K$ axial points with distance α from the center points along the coordinate axes.

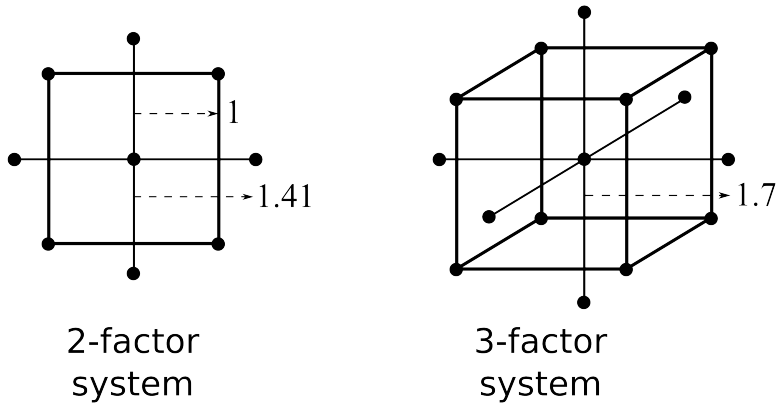

Figure 4.1: Central composite design (CCD) illustration

The factorial design can be a full factorial design (2^K design), a fraction factorial design (2^{K-p} design), or a Plackett-Burman design. The fraction factorial design and the Plackett-Burman design are more efficient than the full factorial design, i.e, having less design points to evaluate. A $K-$level CCD with a full factorial design has $2^K+2\times K+1$ distinct design points to evaluate. The number of evaluation points grows exponentially as K increases. When the function to optimize is large-dimensional, i.e., K is large, the more efficient fraction factorial design and Plackett-Burman design should be implemented.

CCD also needs to determine the distance between the axial points to the center point, α. It is usually chosen within 1 and $\sqrt{K}$, and Box et al. [10] recommended $\alpha = \sqrt[4]{n_f}$ as it guarantees rotatability. Other practical constraints also limit the selection of α. For example, if the objective function only take values at discrete design points (e.g., to certain accuracy), α can only have limited options. Similarly, when the center point of a CCD is near the boundary, α can only take values such that the axial points are still inside the feasible region.

To sum up, LHS partitions the whole feasible region uniformly and randomly, and the design points can be evaluated simultaneously. CCD approximates the local pattern of the response surface. It is used when the design point is near the optimum. The proposed miCCD

algorithm, details in Section 3, combines the advantages of LHS and CCD.

4.3 Multi-stratum Iterative Central Composite Design (miCCD)

A minimization problem is assumed herein without loss of generality. The problem to optimize is to find $\boldsymbol{x}^*$, such that

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in[a,b]^K}{\arg\min}\, f(\boldsymbol{x}), \tag{4.1}$$

or equivalently,

$$f(\boldsymbol{x}^*) = \min_{\boldsymbol{x}\in[a,b]^K} f(\boldsymbol{x}), \tag{4.2}$$

where $[a,b]^K$ is the feasible region, and $a, b \in \mathbf{R}$. K is the dimension of variable $\boldsymbol{x}$ and the dimension of the optimization problem.

The first characteristic of this optimization problem is that the functional format of $f(\cdot)$ is unknown, so the derivatives are not calculable. In other words, $f(\cdot)$ works like a "black box" – it produces output for any inputs within the feasible region while the exact underlying mechanism does not have an explicit expression. Under this circumstance, the gradient-based optimization method is not applicable.

The second characteristic for (4.1) is that multiple local optima may exist for $f(\cdot)$. Traditional response surface method searching for the global optimum could easily stuck at a local optimum. To avoid converging to a local minimum, I propose to use multi-stratum optimization with the aid of the space-filling design.

A space-filling design can be viewed as a way that partitions the feasible region into multiple sub-regions. Each sub-region corresponds to one design point in the space filling design. See Figure 4.2 for illustration. A sub-region could also be a line, a plane, or a hyper-plane. Denote $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ as the design points selected in the space filling design, and $\{R_1, R_2, ..., R_n\}$ as

the corresponding partition of the feasible region satisfying

$$R_i \cap R_j = \emptyset, \quad 1 \leq i \leq j \leq n,$$

and

$$R_1 \cup \cdots \cup R_n = [a, b]^K.$$

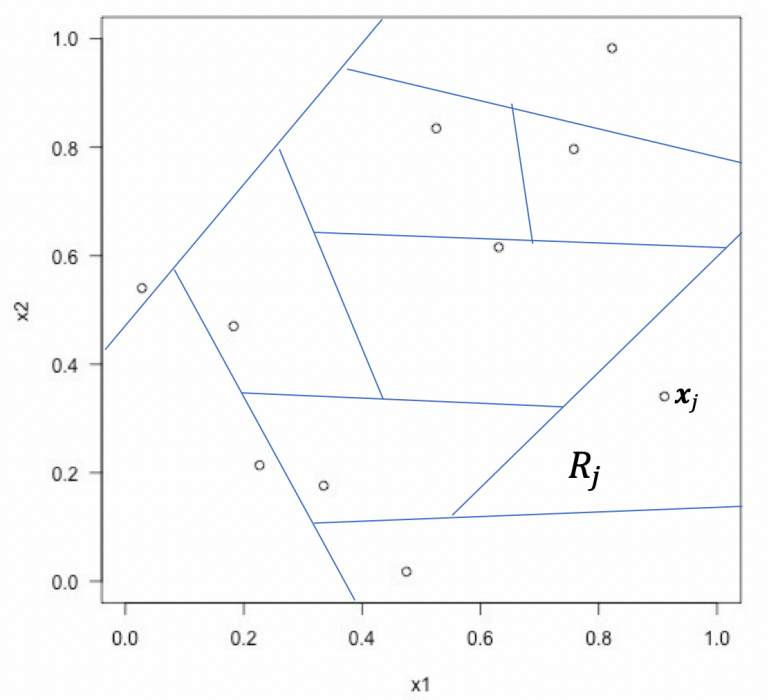

Figure 4.2: Space-filling design illustration–a partition of the feasible region

Denote $\boldsymbol{x}_i^*$ as the optimum point in the ith sub-region, i.e.,

$$\boldsymbol{x}_i^* = \arg\min_{\boldsymbol{x} \in R_i} f(\boldsymbol{x}), \quad i = 1, \cdots, n, \tag{4.3}$$

or equivalently,

$$f(\boldsymbol{x}_i^*) = \min_{\boldsymbol{x} \in R_i} f(\boldsymbol{x}), \quad i = 1, \cdots, n. \tag{4.4}$$

With the partition of the feasible region $[a, b]^K$, I can rewrite the optimization problem 4.2

as

$$f(\boldsymbol{x}^*) = \min_{\boldsymbol{x}\in[a,b]^K} f(\boldsymbol{x}) = \min_{\boldsymbol{x}\in R_1\cup\cdots\cup R_n} f(\boldsymbol{x}) \tag{4.5}$$

$$= \min\{\min_{\boldsymbol{x}\in R_1} f(\boldsymbol{x}), \cdots, \min_{\boldsymbol{x}\in R_n} f(\boldsymbol{x})\} \tag{4.6}$$

$$= \min\{f(\boldsymbol{x}_1^*), \cdots, f(\boldsymbol{x}_n^*)\}. \tag{4.7}$$

Therefore,

$$\boldsymbol{x}^* \in \{\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \cdots, \boldsymbol{x}_n^*\}. \tag{4.8}$$

It is not necessary to optimize in EVERY sub-region, as it is unlikely for a point generating large $f(\cdot)$ value to be near the global minimum. The space-filling design provides good guidance for further conducting optimization in the sub-region with large potential. It is enough to further explore the sub-regions that generate the m smallest $f(\cdot)$ values. Denote $\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(m)}$ as the points generating the m smallest $f(\cdot)$ values within $\{f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n)\}$, and $\boldsymbol{x}_{(1)}^*, \cdots, \boldsymbol{x}_{(m)}^*$ as the corresponding optimum point in each sub-region, i.e.,

$$\boldsymbol{x}_{(i)}^* = \arg\min_{\boldsymbol{x}\in R_{(i)}} f(\boldsymbol{x}), \quad 1 \le i \le m.$$

Then,

$$f(\boldsymbol{x}^*) = \min\{f(\boldsymbol{x}_1^*), \cdots, f(\boldsymbol{x}_n^*)\} \tag{4.9}$$

$$= \min\{f(\boldsymbol{x}_{(1)}^*), \cdots, f(\boldsymbol{x}_{(n)}^*)\} \tag{4.10}$$

$$= \min\{f(\boldsymbol{x}_{(1)}^*), \cdots, f(\boldsymbol{x}_{(m)}^*)\}. \tag{4.11}$$

Therefore,

$$\boldsymbol{x}^* \in \{\boldsymbol{x}_{(1)}^*, \boldsymbol{x}_{(2)}^*, \cdots, \boldsymbol{x}_{(m)}^*\}. \tag{4.12}$$

The selection of m depends on the roughness of the response surface. If $f(\cdot)$ is a bumpy surface, larger m is needed. If the surface is convex with one global optimum and no local

optima, $m = 1$ is enough. Usually, $m = 30\% \cdot n$ would be enough for any objective surface.

The optimization problem now becomes to find the optimum in each sub-region. In order to find the optimum in each sub-region, we then utilize the CCD to iteratively "decline" to a smaller point. With enough design points in the space-filling design step, it is reasonable to believe $\boldsymbol{x}_{(j)}$ is in the vicinity of the optimum $\boldsymbol{x}^*_{(j)}$, then a second-order response surface can well approximate $f(\boldsymbol{x})$ in the sub-region $R_{(j)}$. That is,

$$\widehat{f}(\boldsymbol{x}) \approx f(\boldsymbol{x}), \quad \text{when } \boldsymbol{x} \in R_{(j)}, \quad j = 1, \cdots, m.$$

Since $\widehat{f}(\boldsymbol{x})$ is quadratic in terms of $\boldsymbol{x}$, its minimum can be easily obtained through a canonical analysis. The minimum will then be the the center point for the next round of CCD until convergence.

Suppose $\boldsymbol{x}^t_{(j)}$ is the center of the t-th round of CCD for sub-region $R_{(j)}$, and the fitted second-order function is $\widehat{f}^t(\boldsymbol{x})$. $\boldsymbol{x}^{t+1}_{(j)}$ is the point that maximize $\widehat{f}^t(\boldsymbol{x})$ and the center of the $(t+1)$-th round of CCD. That is,

$$\boldsymbol{x}^{t+1}_{(j)} = \underset{\boldsymbol{x} \in R_{(j)}}{\arg\min}\, \widehat{f}^t(\boldsymbol{x}), \quad j = 1, \cdots, m. \tag{4.13}$$

Then,

$$f(\boldsymbol{x}^t_{(j)}) \approx \widehat{f}^t(\boldsymbol{x}^t_{(j)}) > \widehat{f}^t(\boldsymbol{x}^{t+1}_{(j)}) \approx f(\boldsymbol{x}^{t+1}_{(j)}), \quad j = 1, \cdots, m \tag{4.14}$$

Therefore, iterative CCD ensures "decline" for convergence, i.e.,

$$\lim_{t \to \infty} f(\boldsymbol{x}^t_{(j)}) = f(\boldsymbol{x}^*_{(j)}), \quad j = 1, \cdots, m. \tag{4.15}$$

After $\boldsymbol{x}^*_{(1)}, \cdots, \boldsymbol{x}^*_{(m)}$ have been identified, the global minimum is achieved using Formula 4.11.

Algorithm 1 summarizes the proposed miCCD approach, which leverages the experimental

design thinking for optimization. To sum up, the miCCD approach has two major parts: a multi-start scheme and local minimization. Multi-start scheme selects multiple adequate candidates to start with using certain space-filling design. For each adequate candidate, a second-order approximation of the vicinity helps iteratively decline to the minimum point. Multi-start strategy prevents the optimization process stuck at the local optima. Local minimization ensures decline to achieve convergence.

Algorithm 1 multi-stratum iterative CCD (miCCD)

1. Start from a Latin-hypercube sampling (LHS) with n design points

$$\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n,$$

whose corresponding $f(\cdot)$ values are $f_1, \cdots, f_n$.
2. Identify the m smallest $f(\cdot)$ values $f_{(1)}, \cdots, f_{(m)}$ and corresponding

$$\boldsymbol{x}_{(1)}, \cdots, \boldsymbol{x}_{(m)}$$

3. Start from $j = 1$:
   (a) create a central composite design (CCD) centered at $\boldsymbol{x}_{(j)}$
   (b) fit a second-order response surface using the points evaluated on step 3(a)
   (c) find $\boldsymbol{x}^*_{(j)} \in [a, b]^k$ that minimize the fitted surface in step 3(b) and evaluate $f(\boldsymbol{x}^*_{(j)})$
   (d) if $f(\boldsymbol{x}^*_{(j)}) < f(\boldsymbol{x}_{(j)})$, let $\boldsymbol{x}_{(j)} = \boldsymbol{x}^*_{(j)}$ and repeat 3(a)-3(d);
   else $j = j + 1$ and go to step 3 until $j = m$.

4.4 Case study

The proposed miCCD is applied to determine the optimal threshold for HGF events under different speed conditions. HGF events often caused by aggressive driving behavior. For example, rapid start, hard break, and sharp turns could lead to abnormally high kinematic value. The frequency of such event is informative in distinguishing high-risk drivers from the low-risk ones. In this chapter, I am to identify the threshold as a step function of speed and the selection criteria is that the driver risk model using the resulting HGF rate can have the most distinguish power.

4.4.1 Data

The kinematics data are from the second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study (NDS). In an NDS, participants drive as they normally would. In SHRP 2, there are 3,500+ drivers with over one million hours of driving being recorded. Accelerometers and Global Positioning System (GPS) sensors were installed on the participants' vehicle to continuously record kinematics data, including acceleration, deceleration, and lateral acceleration while participants drive. To identify the number of HGF events rather than the number of data points, a moving average filter was applied to the raw data and several criteria were used to cluster data points from a potential HGF event into one event, including that the data points should be close to each other temporally and the smoothed data should be a local maximum pattern.

In Chapter 3, I have shown that using HGF event rate improves individual driver risk prediction, and that HGF events of deceleration (DEC events) have the largest contribution among the three types of HGF events. This chapter only uses DEC events to illustrate the idea.

A relatively low constant threshold ($0.3g$) is first implemented to extract all possible DEC events from time series deceleration. These events herein are referred as "candidate" DEC events. Their smoothed local maximum deceleration all exceed $0.3g$. Figure 4.3 shows the occurrence rate of candidate DEC events within different speed bins. On overage, there are about three candidate DEC events per driving hour. The occurrence rate is bell-shaped with respect to the driving speed, except for very high speed (above 80 mph). At low speed, there is no need for a brake with deceleration greater than $0.3g$, so the occurrence rate within this speed bin is relatively low. As speed accumulates, harsh braking is more likely to occur so the occurrence rate increases. When in high speed, for example, driving on a highway,

vehicle with a sudden braking is possible to roll over. The chance of a candidate DEC event happening at high speed again drops. Speed relates to the occurrence of candidate DEC events, so the threshold in identifying a DEC event should take speed into consideration.

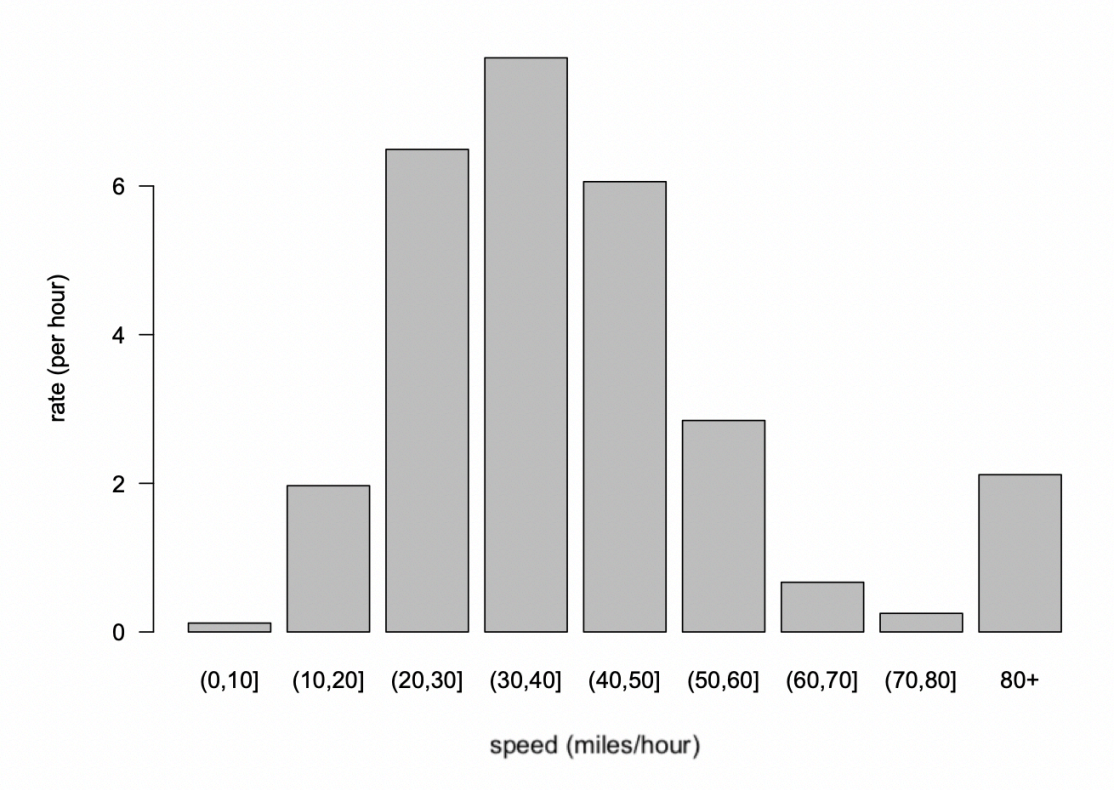

Figure 4.3: Average occurrence rate (per hour) of candidate DEC events (HGF events with smoothed maximum deceleration greater than $0.3g$) for different speed

Denote (a, v) as the smoothed local maximum deceleration and corresponding speed for a candidate DEC event. Suppose the threshold in identifying a DEC event is a function of speed, i.e.,

$$\delta = \delta(v).$$

Given $\delta(v)$, a candidate DEC event is a "qualified" DEC event when $a > \delta(v)$. "Qualified" means it will be counted as a DEC event to calculate the DEC occurrence rate for each driver. The DEC occurrence rate is used as a covariate in a driver risk prediction model.

Figure 4.4 shows the maximum deceleration and corresponding speed of all candidate DEC events for a single driver. Darker color means more events. The red curve shows an example of a functional threshold $\delta(v)$. Given the example threshold, the red dots are qualified DEC events.

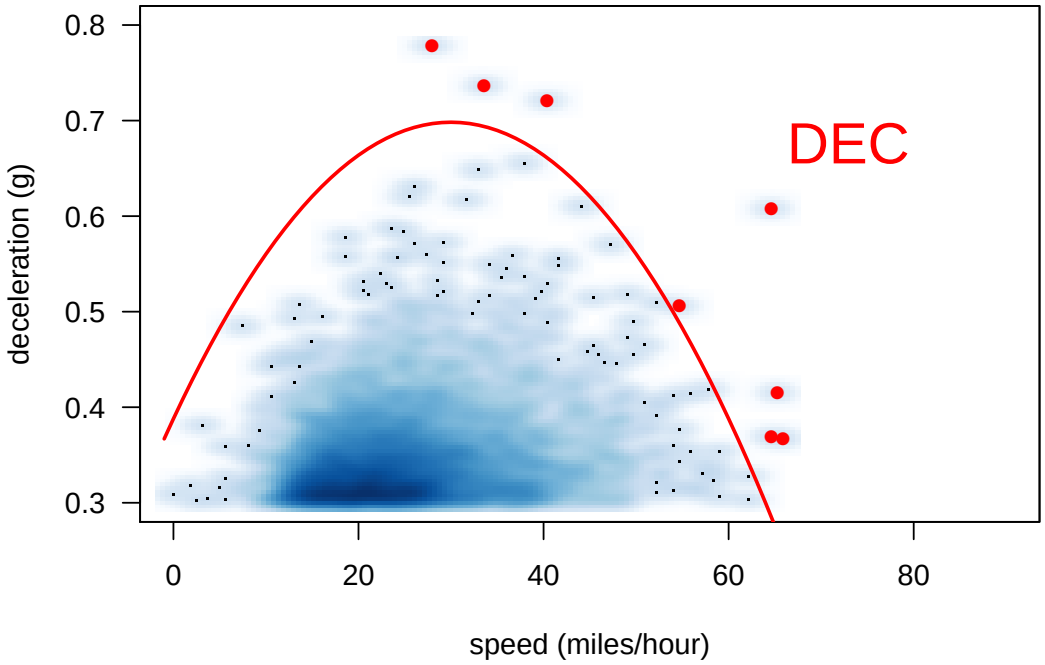

Figure 4.4: Peak deceleration versus speed for a single driver's candidate DEC events. Red dots represent "qualified" DEC events

4.4.2 Optimization problem specification

The model of using DEC rate in predicting high-risk drivers can be formulated as:

$$\mathcal{M}: \quad \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 \cdot G_i\big(\delta(v)\big) \tag{4.16}$$

where i means the i^{th} driver, $p_i = \Pr(Y_i = 1), Y_i$ is a crash indicator: Y_i, w/ crashes; Y_i, w.o. crash, and G_i is the DEC rate, determined by threshold $\delta(v)$.

Thresholds are chosen to optimize the Area Under the Curve (AUC) of model 4.16. That is,

$$\delta^*(v) = \underset{\delta(v)}{\arg\max}\ \mathrm{AUC}\Big(\mathcal{M}\big(\delta(v)\big)\Big) \tag{4.17}$$

A step function is used to approximate the optimal threshold function in terms of v. Suppose $\mathbb{V}$ is a partition of the feasible speed range: $[0, +\infty)$ miles per hour (mph), $\mathbb{V} = \{\mathbf{V}_1, \cdots, \mathbf{V}_K\}$, then

$$\delta(v) = \sum_{k=1}^{K} \delta_k \cdot \mathbb{1}(v \in \mathbf{V}_k). \tag{4.18}$$

For example, a partition of the speed range with three cuts at 10 mph, 30 mph, and 60 mph would result in four speed bins. Formula (4.18) can be written as

$$\delta(v) = \begin{cases} \delta_1, & v \in [0, 10] \text{ mph} \\ \delta_2, & v \in (10, 30] \text{ mph} \\ \delta_3, & v \in (30, 60] \text{ mph} \\ \delta_4, & v > 60 \text{ mph} \end{cases}$$

The optimal threshold cannot be too high or too low (it cannot be lower than 0.3g). Practitioners suggests 0.8g to the ceiling of the optimal threshold. Given all the setup, the problem has evolved to determine K threshold values to optimize the AUC of model (4.16), i.e.,

$$\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta} \in [0.3g, 0.8g]^K}{\arg\max}\ \mathrm{AUC}\Big(\mathcal{M}\big(\boldsymbol{\delta}\big)\Big), \tag{4.19}$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_K)$.

Qualified DEC events are then aggregated to the driver level in predicting drivers' crash risk. The aggregation process would take a relatively long time if extracting the qualified

events from all candidate events, so the number of events larger than a predetermined set of thresholds $\{0.30g, 0.31g, \cdots, 0.80g\}$ under different speed bins for each driver are prepared before taking them into the model. The optimization problem 4.19 then becomes

$$\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta} \in \{0.30g, 0.31g, \cdots, 0.80g\}^K}{\arg\max} \text{AUC}\Big(\mathcal{M}(\boldsymbol{\delta})\Big). \tag{4.20}$$

The objective function $\text{AUC}\Big(\mathcal{M}(\boldsymbol{\delta})\Big)$ has a complex "black box" structure in terms of $\boldsymbol{\delta}$, so the proposed miCCD is applied to solve the problem.

4.4.3 Results

Comparison with Grid Search

To measure miCCD's performance, I repeat the miCCD optimization algorithm 1000 times for each of the following n and m settings in Table 4.1. Each repetition starts with a random selected LHS, so the resulting $\boldsymbol{\delta}^*$ could be different for each repetition. Denote $\boldsymbol{\delta}^{(r)}$ as the optimal solution found in the r-th repetition of miCCD evaluation under each setting. Accuracy measures the percentage of times when the miCCD algorithm actually achieve the global optimum, i.e., $\frac{1}{1000} \cdot \sum_{r=1}^{1000} \mathbb{1}(\boldsymbol{\delta}^{(r)} = \boldsymbol{\delta})$. "Time used" is average time used for each repetition using a single Intel Xeon CPU E5-1650 v2 processor.

The global optimum for $K = 2, 3$, and 4 of problem (4.20) are known as brute-force grid search is plausible. There are 51 distinct values within $\{0.30g, 0.31g, \cdots, 0.80g\}$. With the prepared dataset, it usually takes less than 0.05 seconds to evaluate one design point for model 4.16. Grid search for $K = 2$ and $K = 3$ have 51^2 and 51^3 points to evaluate, and a personal computer takes approximately 1.5 minutes and 75 minutes if using a single Intel Xeon CPU E5-1650 v2 processor. When $K = 4$, 51^3 points to evaluate, grid search is doable on a computer cluster, taking around 20 hours using 24 Intel Xeon CPU E5-2600 v4

processors.

Table 4.1: Simulation results for miCCD algorithm

Dimension	n	m	**Accuracy**	**Time used**
K = 2	20	3	90.8%	12 sec
	20	6	98.8%	23 sec
	50	3	88.6%	10 sec
	50	6	98.9%	21 sec
	100	3	91.2%	12 sec
	100	6	99.2%	22 sec
K = 3	50	10	59.2%	73 sec
	50	20	79.9%	151 sec
	100	10	63.3%	65 sec
	100	20	85.8%	146 sec
	100	50	98.0%	391 sec
	200	10	69.8%	60 sec
	200	20	88.1%	129 sec
	200	50	98.7%	368 sec
K = 4	2000	10	31.5%	112 sec
	2000	20	42.4%	158 sec
	2000	50	64.4%	319 sec
	2000	100	80.5%	618 sec
	2000	200	95.0%	1251 sec
	2000	500	99.7%	3466 sec
	5000	200	96.0%	1122 sec
	5000	500	99.9%	2833 sec
	5000	1000	100.0%	6164 sec

Based on the simulation results shown in Table 4.1, miCCD can reach the true optimum 99% of the time with an appropriately chosen number of adequate points to continue (m). miCCD's accuracy slightly depend on the number of design points to evaluate in the LHS step (n) but largely on the number of design points to further explore (m). The accuracy increases as the number of design points to further explore, m, increases. The computational cost increases exponentially with respect to K. It is more efficient with larger n. This is caused by that more points to evaluate in the first step is more likely to land in "better"

points to start off, and the "better" points are associated with less iterations of CCD. When $m \approx 10^{K-1}/2$, miCCD can achieve about 99% accuracy. Therefore, I recommend m to be around $10^{K-1}/2$ and n to be the larger the better based on computational capacity.

Optimal threshold of DEC event

Figure 4.5 shows the optimal threshold result when the feasible speed range is partitioned as follows:

$$\mathbb{V} = \Big\{[0, 10] \text{ mph}, (10, 30] \text{ mph}, (30, 60] \text{ mph}, 60 + \text{mph}\Big\}$$

Specifically, $n = 2,000$ and $m = 500$ are used in miCCD to ensure converging to the global optimum. Bootstrap method with 100 bootstrap samples is used to obtain the confidence band. The grey area represents the 95% confidence band, which is obtained by the bootstrap method. Note that brute-force grid search is not feasible for obtaining the confidence band using bootstrap method (100 bootstrap samples would take 20×100 hours to run on a 24-core computer cluster).

As shown in Figure 4.5, the optimal threshold is of concave shape in terms of the driving speed. When the driving speed is small, there is not much space for a large deceleration, so the corresponding threshold is small. On the other hand, when driving in the highway ($v > 60$ mph), it is risky to have a kinematic signature event with large g-force. The corresponding threshold should be reduced accordingly.

4.5 Discussion

This chapter proposes a Multi-stratum Iterative Central Composite Design (miCCD) global optimization algorithm for a large-dimensional "black box" function. "Black box" means the specific formulation of the objective function is unknown or is complicated. The 'miCCD'

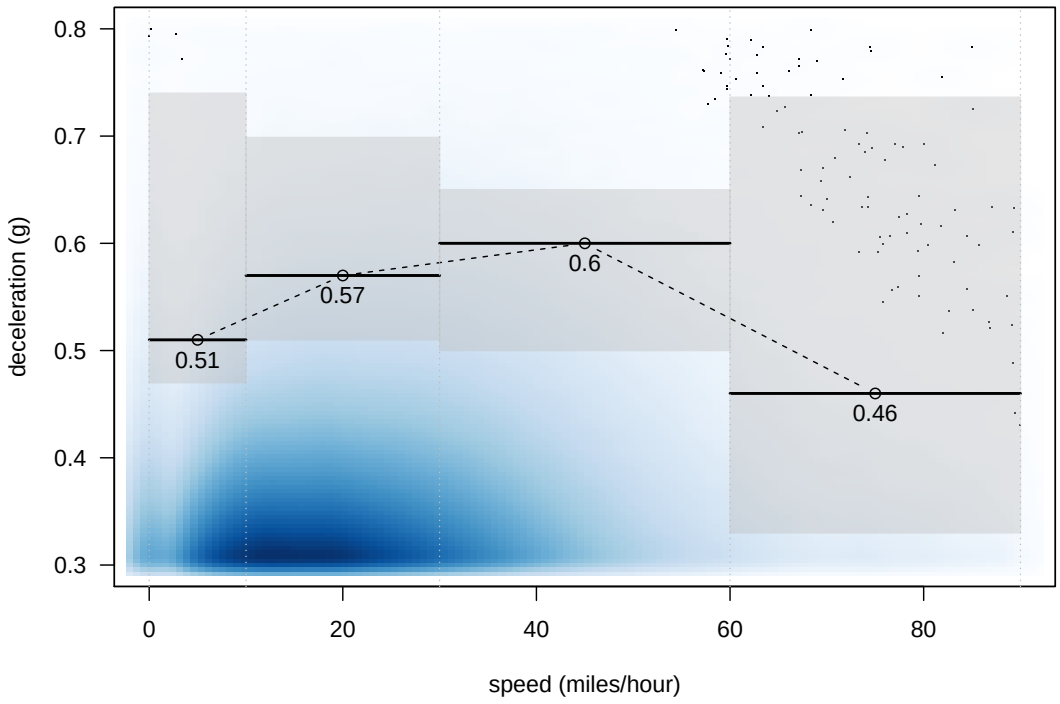

Figure 4.5: Optimal threshold for DEC event with 95% bootstrap confidence band when speed range is partitioned into $[0, 10]$ mph, $(10, 30]$ mph, $(30, 60]$ mph, 60 + mph. Background shows the candidate DEC events of all SHRP 2 drivers where darker area means more events.

approach has two major parts: a multi-start scheme and local optimization. The multi-start scheme finds multiple adequate points to start with using space-filling design (e.g. LHS). For each adequate starting point, iterative CCD keeps decreasing until converging to a local minimum. Enough multiple starting points ensure the algorithm achieve the global optimum.

The magnitude of K increases the computational cost exponentially. For a larger dimension, more efficient fraction factorial design and Plackett-Burman design can substitute the full factorial design in the iterative CCD part. In future research, accuracy and efficiency can be compared with other optimization methods, for example, Gaussian surrogate models.

In identifying the optimal threshold, instead of using step function to approximate $\delta(v)$, spline is another solution. If a quadratic spline with two knots ξ_1 and ξ_2 is used, $\delta(v)$ can be approximated by

$$\delta(v) \approx \beta_1 \cdot v + \beta_2 \cdot v^2 + \beta_3 \cdot (v - \xi_1)_+^2 + \beta_4 \cdot (v - \xi_2)_+^2.$$

The optimization problem then becomes

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta} \in \mathbf{R}^4}{\arg\max} \ \mathrm{AUC}\Big(\mathcal{M}\big(\boldsymbol{\beta}\big)\Big), \tag{4.21}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$. The resulting optimal threshold can be more smooth in terms of the driving speed, which the practitioners can understand more easily. On the other hand, there are two drawbacks of the spline approach. The first one is that there is no bounded feasible region for $\boldsymbol{\beta}$ as $\boldsymbol{\beta} \in \mathbb{R}^4$. The second disadvantage is that the evaluation time for one design point would be elongated, since for each given $\boldsymbol{\beta}$, the qualified DEC event rate has to be aggregated from all the candidate DEC events. The aggregation process would cost a long time.

The proposed miCCD algorithm can also be applied to tuning the hyper-parameters of machine learning models. The success of many machine learning models rely heavily on the selection of the right hyper-parameters. Hyper-parameter tuning requires expert experience, rules of thumb, or sometimes brute-force search. For example, adjusting the learning rate, maximum tree depth, and number of trees in a gradient boosting method such that it can minimize the misclassification error on the validation dataset. The misclassification error is not differentiable in terms of the hyper-parameters, so gradient-based optimization approach is not a solution. miCCD provides an option to search the appropriate hyper-parameters efficiently.

Chapter 5

Concluding Remarks

In this dissertation, I have focused on improving traffic safety modeling, especially driver risk prediction, from three aspects. Firstly, I addressed the finite sample bias issue when using Poisson and NB regression in traffic safety modeling as crashes are rare events. Secondly, I proposed a decision-adjusted modeling framework to incorporate how the model will be used into the model estimation procedure. The decision-adjusted modeling framework is applied to identify the optimal thresholds of kinematic signatures based on the particular application objective. Thirdly, I developed an effective miCCD algorithm for searching the global optimum of any objective function. With miCCD, I can provide the threshold recommendation of high g-force events for different driving speed. The miCCD has broad potential applications, e.g., tuning the hyper-parameters of machine learning models.

The major effort of this dissertation is that it established a decision-driven and kinematics-based driver risk evaluation framework. It is well known that high g-force events (representing hard breaks, rapid starts, and sharp turns) are informative when predicting driver risk. However, there was no systematic study about how much they contribute, and practitioners use constant and subjective threshold in real-world in triggering high g-force events recording. In this work, to make the best use of the high g-force events, I developed an efficient optimization algorithm to identify the optimal threshold under different driving speed. I systematically explored and quantified the benefit of high g-force events in a driver risk evaluation model. I also proposed that the optimization should consider the decision rule, i.e., how the estimated model to be used will influence the optimization objective. This

telematics-based driver risk evaluation framework can be helpful to fleet management, driver safety coach program, and insurance pricing.

This work can be further extended to applying the decision-adjusted modeling approach to identify the optimal functional threshold in terms of driving speed. In Chapter 4, the optimal threshold is selected based on maximizing the AUC of a driver risk prediction model with DEC rate being the only covariate. In future research, how would the optimal threshold vary when additional covariates are included in the driver risk prediction model is worth investigation. Additionally, how would the optimal threshold change with respect to different decision rules, i.e., how the estimated model will be used, is of interest.

Another direction is kinematics-based trip-level crash risk prediction and instantaneous risk prediction. Trip-level risk analysis has a board application prospect, for instance, use-based insurance for car-rental services and ride-sharing activity. Instantaneous risk prediction is even finer-grained risk analysis where the model output could be the probability of crash occurrence within the next five minutes. The fine-grained risk prediction brings opportunities as well as challenges. It contains more informative covariates for modeling: time of day, weather, geolocation, average speed to name a few. However, the very small percentage of trips/time segments that have been involved in a crash (less than one-thousandth or smaller) brings new challenges for statistical modeling.

Appendices

Appendix A

Adjusting finite sample bias in traffic safety modeling

A.1 Finite sample bias for Poisson regression with one explanatory variable

If there is only one exploratory variable in the Poisson regression model, then

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad W = \begin{bmatrix} \lambda_1 E_1 & & \\ & \ddots & \\ & & \lambda_n E_n \end{bmatrix}.$$

Let's denote

$$a = \frac{1}{n}\sum_{i=1}^{n} \lambda_i E_i,\ b = \frac{1}{n}\sum_{i=1}^{n} x_i \lambda_i E_i,\ c = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \lambda_i E_i,\ d = \frac{1}{n}\sum_{i=1}^{n} x_i^3 \lambda_i E_i,$$

then

$$X'WX = n \cdot \begin{bmatrix} a & b \\ b & d \end{bmatrix}, \quad (X'WX)^{-1} = \frac{1}{n(ac-b^2)} \begin{bmatrix} c & -b \\ -b & a \end{bmatrix},$$

and

$$Q = X(X'WX)^{-1}X' = \frac{1}{n(ac-b^2)} \begin{bmatrix} c - 2bx_1 + ax_1^2 & \cdots & c - bx_1 - bx_n + ax_1x_n \\ \vdots & \ddots & \vdots \\ c - bx_1 - bx_n + ax_1x_n & \cdots & c - 2bx_n + ax_n^2 \end{bmatrix},$$

$$\Rightarrow \quad \boldsymbol{\xi} = -\frac{1}{2}\begin{bmatrix} Q_{11} \\ \vdots \\ Q_{nn} \end{bmatrix} = \frac{-1}{2n(ac-b^2)} \cdot \begin{bmatrix} c - 2bx_1 + ax_1^2 \\ \vdots \\ c - 2bx_n + ax_n^2 \end{bmatrix}.$$

Therefore, the bias of the MLE $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ is

$$\text{bias}(\hat{\boldsymbol{\beta}}) = (X'WX)^{-1}X'W\boldsymbol{\xi} = \frac{1}{2n(ac-b^2)^2}\begin{bmatrix} -2ac^2 + bc^2 + abd \\ 3abc - 2b^3 - a^2d \end{bmatrix}$$

That is,

$$\text{bias}(\hat{\beta}_1) = \frac{1}{n} \cdot \frac{3abc - 2b^3 - a^2d}{2(ac-b^2)^2}.$$

A.2 Finite sample bias for NB regression with one binary explanatory variable

For the NB regression with only one binary exploratory variable , suppose there are n data points available for model estimation, $\{X_i, Y_i, E_i\}_{i=1}^n$, with X_i being either 0 or 1, and Y_i and E_i are the corresponding event count and exposure, respectively. Assume there are n_1 data points with $X_i = 1$ and n_0 data points with $X_i = 0$, and $n_0 + n_1 = n$.

Without the loss of generality, we arrange the data $\{X_i, Y_i, E_i\}_{i=1}^n$ with $X_i = 1$ instances before $X_i = 0$ instances. That is,

$$X_i = \begin{cases} 1 & \text{if } i = 1, \cdots, n_1, \\ 0 & \text{if } i = n_1 + 1, \cdots, n. \end{cases}$$

For the $X_i = 1$ group, denote μ_1 as the mean of event rate and V_1 as the variance for the

sum of crash counts, $V_1 = \text{Var}(\sum_i Y_i | X_i = 1)$.

$$V_1 = \text{Var}(\sum_i Y_i | X_i = 1) = \sum_{i=1}^{n_1} \Big[\mu_1 E_i + \frac{(\mu_1 E_i)^2}{k}\Big].$$

Similarly, denote the mean of event rate as μ_0 and V_0 as the variance for the sum of crash counts, $V_0 = \text{Var}(\sum_i Y_i | X_i = 0)$, for the $X_i = 0$ group.

$$V_0 = \text{Var}(\sum_i Y_i | X_i = 0) = \sum_{i=n_1+1}^{n} \Big[\mu_0 E_i + \frac{(\mu_0 E_i)^2}{k}\Big].$$

The data matrix $\mathbf{X}$ and covariance matrix $\mathbf{W}$ in Equation (2.4) are

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_1 \text{ rows} \\ \\ \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_0 \text{ rows} \end{matrix},$$

$$\mathbf{W} = \left[\begin{array}{ccc:ccc} \mu_1 E_1 + \frac{(\mu_1 E_1)^2}{k} & & & & & \\ & \ddots & & & & \\ & & \mu_1 E_{n_1} + \frac{(\mu_1 E_{n_1})^2}{k} & & & \\ \hdashline & & & \mu_0 E_{n_1+1} + \frac{(\mu_0 E_{n_1+1})^2}{k} & & \\ & & & & \ddots & \\ & & & & & \mu_0 E_n + \frac{(\mu_0 E_n)^2}{k} \end{array}\right] \begin{matrix} \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_1 \text{ rows} \\ \\ \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_0 \text{ rows} \end{matrix}.$$

Denote

$$V_1 = \mathrm{Var}(\sum_i Y_i | X_i = 1) = \sum_{i=1}^{n_1} \Big[\mu_1 E_i + \frac{(\mu_1 E_i)^2}{k}\Big],$$
$$V_0 = \mathrm{Var}(\sum_i Y_i | X_i = 0) = \sum_{i=n_1+1}^{n} \Big[\mu_0 E_i + \frac{(\mu_0 E_i)^2}{k}\Big].$$

That is, V_1 is the sum of the upper half of the diagonal elements of the matrix $\mathbf{W}$, and V_0 is the sum of the lower half of the diagonal elements.

Then

$$\mathbf{X'WX} = \begin{bmatrix} V_0 + V_1 & V_1 \\ V_1 & V_1 \end{bmatrix}, \quad (\mathbf{X'WX})^{-1} = \frac{1}{V_0 V_1} \begin{bmatrix} V_1 & -V_1 \\ -V_1 & V_1 + V_0 \end{bmatrix},$$

and

$$\mathbf{Q} = \mathbf{X}(\mathbf{X'WX})^{-1}\mathbf{X'} = \frac{1}{V_0 V_1} \left[\begin{array}{ccc:ccc} V_0 & \cdots & V_0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ V_0 & \cdots & V_0 & 0 & \cdots & 0 \\ \hdashline 0 & \cdots & 0 & V_1 & \cdots & V_1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & V_1 & \cdots & V_1 \end{array}\right].$$

$$\Rightarrow \quad \xi_i = -\frac{1}{2} Q_{ii} = \begin{cases} -\frac{1}{2V_0} & \text{if } i = 1, \cdots, n_1; \\ -\frac{1}{2V_1} & \text{if } i = n_1 + 1, \cdots, n. \end{cases}$$

Therefore, the bias of the MLE $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)$ is

$$\mathrm{bias}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X'WX})^{-1}\mathbf{X'W}\boldsymbol{\xi} = \begin{bmatrix} -\frac{1}{2V_0} \\ \frac{1}{2V_0} - \frac{1}{2V_1} \end{bmatrix},$$

that is,

$$\mathrm{bias}(\hat{\beta}_1) = \frac{1}{2V_0} - \frac{1}{2V_1}.$$

Bibliography

[1] Proven safety countermeasures, 2018. https://safety.fhwa.dot.gov/provencountermeasures/, Accessed: 2018-11-21.

[2] AASHTO. *Highway safety manual*, volume 1. American Association of State Highway and Transportation Officials, Washington DC, 2010. ISBN 1560514779.

[3] Anders E af Wåhlberg. Speed choice versus celeration behavior as traffic accident predictor. *Journal of safety research*, 37(1):43–51, 2006.

[4] J. Aguero-Valverde. Full bayes poisson gamma, poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. *Accident Analysis and Prevention*, 50:289–297, 2013. ISSN 0001-4575. doi: 10.1016/j.aap.2012.04.019.

[5] P. C. Anastasopoulos. Random parameters multivariate tobit and zero-inflated count data models: Addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. *Analytic Methods in Accident Research*, 11:17–32, 2016. ISSN 2213-6657. doi: 10.10161/j.amar.2016.06.001.

[6] Nasim Arbabzadeh and Mohsen Jafari. A data-driven approach for driving safety risk prediction using driver behavior and roadway information data. *IEEE Transactions on Intelligent Transportation Systems*, 19(2):446–460, 2018.

[7] Omar Bagdadi. Assessing safety critical braking events in naturalistic driving studies. *Transportation research part F: traffic psychology and behaviour*, 16:117–126, 2013.

[8] Omar Bagdadi and András Várhelyi. Jerky driving—an indicator of accident proneness? *Accident Analysis & Prevention*, 43(4):1359–1363, 2011.

[9] MS Bartlett. Approximate confidence intervals. *Biometrika*, 40(1/2):12–19, 1953.

[10] George EP Box, J Stuart Hunter, et al. Multi-factor experimental designs for exploring response surfaces. *The Annals of Mathematical Statistics*, 28(1):195–241, 1957.

[11] Paula Branco, Luis Torgo, and Rita Ribeiro. A survey of predictive modelling under imbalanced distributions. *arXiv preprint arXiv:1505.01658*, 2015.

[12] Kenneth L Campbell. The shrp 2 naturalistic driving study: Addressing driver performance and behavior in traffic safety. *TR News*, (282), 2012.

[13] David E Cantor, Thomas M Corsi, Curtis M Grimm, and Koray Özpolat. A driver focused truck crash prediction model. *Transportation Research Part E: Logistics and Transportation Review*, 46(5):683–692, 2010.

[14] Jodi Carson and Fred Mannering. The effect of ice warning signs on ice-accident frequencies and severities. *Accident Analysis & Prevention*, 33(1):99–109, 2001.

[15] Erdong Chen and Andrew P Tarko. Modeling safety of highway work zones with random parameters and random effects models. *Analytic methods in accident research*, 1:86–95, 2014.

[16] Gauss M Cordeiro and Peter McCullagh. Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 629–643, 1991.

[17] D. R. Cox and E. J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 30(2):248–275, 1968. ISSN 1369-7412.

[18] Sophia Daskalaki, Ioannis Kopanas, and Nikolaos Avouris. Evaluation of classifiers for an uneven class distribution problem. *Applied artificial intelligence*, 20(5):381–417, 2006.

[19] Hamish A Deery and Brian N Fildes. Young novice driver subtypes: Relationship to high-risk behavior, traffic accident record, and simulator driving performance. *Human factors*, 41(4):628–643, 1999.

[20] Thomas A Dingus, Sheila G Klauer, Vicki Lewis Neale, Andy Petersen, Suzanne E Lee, Jeremy Sudweeks, Miguel A Perez, Jonathan Hankey, David Ramsey, Santosh Gupta, et al. The 100-car naturalistic driving study. phase 2: results of the 100-car field experiment. Technical report, United States. Department of Transportation. National Highway Traffic Safety …, 2006.

[21] Thomas A Dingus, Jonathan M Hankey, Jonathan F Antin, Suzanne E Lee, Lisa Eichelberger, Kelly E Stulce, Doug McGraw, Miguel Perez, and Loren Stowe. *Naturalistic driving study: Technical coordination and quality control.* Number SHRP 2 Report S2-S06-RW-1. 2015.

[22] Thomas A Dingus, Feng Guo, Suzie Lee, Jonathan F Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113(10):2636–2641, 2016.

[23] Rob Eenink, Yvonne Barnard, Martin Baumann, Xavier Augros, and Fabian Utesch. Udrive: the european naturalistic driving study. In *Proceedings of Transport Research Arena*. IFSTTAR, 2014.

[24] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[25] Lasse Fridstrøm, Jan Ifver, Siv Ingebrigtsen, Risto Kulmala, and Lars Krogsgård Thomsen. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis & Prevention*, 27(1):1–20, 1995.

[26] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4), 2009.

[27] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[28] David E Giles and Hui Feng. Reducing the bias of the maximum likelihood estimator for the poisson regression model. *Economics Bulletin*, 31(4):2933–2943, 2011. ISSN 1545-2921.

[29] D. Gorman. World report on road traffic injury prevention. *Public Health*, 120(3): 280–280, 2006. ISSN 0033-3506. doi: 10.1016/j.puhe.2005.09.003.

[30] Feng Guo. Statistical methods for naturalistic driving studies. *Annual review of statistics and its application*, 6:309–328, 2019.

[31] Feng Guo and Youjia Fang. Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention*, 61:3–9, 2013.

[32] Feng Guo, Xuesong Wang, and Mohamed A Abdel-Aty. Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis & Prevention*, 42(1): 84–92, 2010.

[33] Filmon G Habtemichael and Luis de Picado-Santos. The impact of high-risk drivers and benefits of limiting their driving degree of freedom. *Accident Analysis & Prevention*, 60:305–315, 2013.

[34] Jonathan M Hankey, Miguel A Perez, and Julie A McClafferty. Description of the shrp 2 naturalistic database and the crash, near-crash, and baseline data sets. Technical report, Virginia Tech Transportation Institute, 2016.

[35] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[36] Sheila G Klauer, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, David J Ramsey, et al. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. 2006.

[37] Sheila G Klauer, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, and David J Ramsey. Comparing real-world behaviors of drivers with high versus low rates of crashes and near crashes. Technical report, 2009.

[38] Sheila G Klauer, Feng Guo, Bruce G Simons-Morton, Marie Claude Ouimet, Suzanne E Lee, and Thomas A Dingus. Distracted driving and risk of road crashes among novice and experienced drivers. *New England journal of medicine*, 370(1):54–59, 2014.

[39] Ioannis Kosmidis and David Firth. Bias reduction in exponential family nonlinear models. *Biometrika*, 96(4):793–804, 2009.

[40] Ioannis Kosmidis, David Firth, et al. A generic algorithm for reducing bias in parametric estimation. *Electronic Journal of Statistics*, 4:1097–1112, 2010.

[41] SSP Kumara and Hoong Chor Chin. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. *Traffic Injury Prevention*, 4(1): 53–57, 2003.

[42] Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

[43] Bert Lambrecht, William Perraudin, and Stephen Satchell. Approximating the finite sample bias for maximum likelihood estimators using the score–solution. *Econometric Theory*, 13(2):310–312, 1997.

[44] Jinsun Lee and Fred Mannering. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. *Accident Analysis & Prevention*, 34(2):149–161, 2002.

[45] Xiugang Li, Dominique Lord, and Yunlong Zhang. Development of accident modification factors for rural frontage road segments in texas using generalized additive models. *Journal of Transportation Engineering*, 137(1):74–83, 2010.

[46] Dominique Lord. Modeling motor vehicle crashes using poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention*, 38(4):751–766, 2006.

[47] Dominique Lord and James Bonneson. Calibration of predictive models for estimating safety of ramp design configurations. *Transportation Research Record: Journal of the Transportation Research Board*, (1908):88–95, 2005.

[48] Dominique Lord and Srinivas Reddy Geedipally. The negative binomial–lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. *Accident Analysis & Prevention*, 43(5):1738–1742, 2011.

[49] Dominique Lord and Srinivas Reddy Geedipally. Safety prediction with datasets characterised with excess zero responses and long tails. In *Safe Mobility: Challenges, Methodology and Solutions*, pages 297–323. Emerald Publishing Limited, 2018.

[50] Dominique Lord and Fred Mannering. The statistical analysis of crash-frequency data:

a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5):291–305, 2010.

[51] Dominique Lord and Luis F Miranda-Moreno. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of poisson-gamma models for modeling motor vehicle crashes: a bayesian perspective. *Safety Science*, 46 (5):751–770, 2008.

[52] Dominique Lord and Bhagwant Persaud. Accident prediction models with and without trend: application of the generalized estimating equations procedure. *Transportation Research Record: Journal of the Transportation Research Board*, (1717):102–108, 2000.

[53] Dominique Lord, Abdelaziz Manar, and Anna Vizioli. Modeling crash-flow-density and crash-flow-v/c ratio relationships for rural and urban freeway segments. *Accident Analysis & Prevention*, 37(1):185–199, 2005.

[54] Dominique Lord, Simon P Washington, and John N Ivan. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, 37(1):35–46, 2005.

[55] Dominique Lord, Simon Washington, and John N Ivan. Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention*, 39(1):53–57, 2007.

[56] Nataliya V Malyshkina and Fred L Mannering. Zero-state markov switching count-data models: An empirical assessment. *Accident Analysis & Prevention*, 42(1):122–130, 2010.

[57] Nataliya V Malyshkina, Fred L Mannering, and Andrew P Tarko. Markov switching negative binomial models: an application to vehicle accident frequencies. *Accident Analysis & Prevention*, 41(2):217–226, 2009.

[58] Peter McCullagh. *Tensor methods in statistics*, volume 161. Chapman and Hall London, 1987.

[59] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[60] Shaw-Pin Miaou. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, 26(4):471–482, 1994.

[61] Vicki L Neale, Thomas A Dingus, Sheila G Klauer, Jeremy Sudweeks, and Michael Goodman. An overview of the 100-car naturalistic study and findings. *National Highway Traffic Safety Administration, Paper*, 5:0400, 2005.

[62] NHTSA. Motor vehicle traffic crashes as a leading cause of death in the united states, 2015. Report DOT HS 812 499, 2018.

[63] NHTSA. Traffic safety facts 2016: A compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system. Report DOT HS 812 554, 2018.

[64] Cynthia Owsley, Beth T Stalvey, Jennifer Wells, Michael E Sloane, and Gerald McGwin. Visual risk factors for crash involvement in older drivers with cataract. *Archives of ophthalmology*, 119(6):881–887, 2001.

[65] Byung-Jung Park and Dominique Lord. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention*, 41(4):683–691, 2009.

[66] X. Qin, J. N. Ivan, and N. Ravishanker. Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accident Analysis and Prevention*, 36(2): 183–191, 2004. ISSN 0001-4575. doi: 10.1016/S0001-4575(02)00148-3.

[67] Romesh Ranawana and Vasile Palade. Optimized precision-a new measure for classifier performance evaluation. In *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*, pages 2254–2261. IEEE, 2006.

[68] Fridulv Sagberg, Selpi, Giulio Francesco Bianchi Piccinini, and Johan Engström. A review of research on driving styles and road safety. *Human factors*, 57(7):1248–1275, 2015.

[69] K. Saha and S. Paul. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, 61(1):179–185, 2005. ISSN 0006-341x. doi: DOI10.1111/j.0006-341X.2005.030833.x.

[70] V. N. Shankar, G. F. Ulfarsson, R. M. Pendyala, and M. B. Nebergall. Modeling crashes involving pedestrians and motorized traffic. *Safety Science*, 41(7):627–640, 2003.

[71] Venky Shankar, John Milton, and F Mannering. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis & Prevention*, 29 (6):829–837, 1997.

[72] David Shinar. *Traffic safety and human behavior*. Emerald Publishing Limited, 2017.

[73] Bruce G Simons-Morton, Marie Claude Ouimet, Jing Wang, Sheila G Klauer, Suzanne E Lee, and Thomas A Dingus. Hard braking events among novice teenage drivers by passenger characteristics. In *Proceedings of the... International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, volume 2009, page 236. NIH Public Access, 2009.

[74] Bruce G Simons-Morton, Zhiwei Zhang, John C Jackson, and Paul S Albert. Do elevated gravitational-force events while driving predict crashes and near crashes? *American journal of epidemiology*, 175(10):1075–1079, 2012.

[75] Bruce G Simons-Morton, Kyeongmi Cheon, Feng Guo, and Paul Albert. Trajectories of kinematic risky driving among novice teenagers. *Accident Analysis & Prevention*, 51: 27–32, 2013.

[76] Susan A Soccolich, Jeffrey S Hickman, Richard J Hanowski, et al. Identifying high-risk commercial truck drivers using a naturalistic approach. Technical report, Virginia Tech. Virginia Tech Transportation Institute, 2011.

[77] Paweenuch Songpatanasilp, Harutoshi Yamada, Teerayut Horanont, and Ryosuke Shibasaki. Traffic accidents risk analysis based on road and land use factors using glms and zero-inflated models. In *Proceedings of 14th International Conference on Computers in Urban Planning and Urban Management (CUPUM 2015)*, pages 7–10, 2015.

[78] Jane C Stutts, Jean W Wilkins, J Scott Osberg, and Bradley V Vaughn. Driver risk factors for sleep-related crashes. *Accident Analysis & Prevention*, 35(3):321–331, 2003.

[79] Qian Chayn Sun, Robert Odolinski, Jianhong Cecilia Xia, Jonathan Foster, Torbjörn Falkmer, and Hoe Lee. Validating the efficacy of gps tracking vehicle movement for driving behaviour assessment. *Travel Behaviour and Society*, 6:32–43, 2017.

[80] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[81] Pål Ulleberg. Personality subtypes of young drivers. relationship to risk-taking preferences, accident involvement, and response to a traffic safety campaign. *Transportation Research Part F: Traffic Psychology and Behaviour*, 4(4):279–297, 2001.

[82] Prathyusha Vangala, Dominique Lord, and Srinivas Reddy Geedipally. Exploring the

application of the negative binomial–generalized exponential model for analyzing traffic crash data with excess zeros. *Analytic methods in accident research*, 7:29–36, 2015.

[83] Trent Victor, Marco Dozza, Jonas Bärgman, Christian-Nils Boda, Johan Engström, Carol Flannagan, John D Lee, and Gustav Markkula. Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk. Technical report, 2015.

[84] GR Wood. Generalised linear accident models and goodness of fit testing. *Accident Analysis & Prevention*, 34(4):417–427, 2002.

[85] CF Jeff Wu and Michael S Hamada. *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons, 2011.

[86] Kun-Feng Wu, Jonathan Aguero-Valverde, and Paul P Jovanis. Using naturalistic driving data to explore the association between traffic safety-related events and crash risk at driver level. *Accident Analysis & Prevention*, 72:210–218, 2014.

[87] Yuanchang Xie and Yunlong Zhang. Crash frequency analysis with generalized additive models. *Transportation Research Record: Journal of the Transportation Research Board*, (2061):39–45, 2008.

[88] Shouyi Yin, Jinjin Duan, Peng Ouyang, Leibo Liu, and Shaojun Wei. Multi-cnn and decision tree based driving behavior evaluation. In *Proceedings of the Symposium on Applied Computing*, pages 1424–1429. ACM, 2017.

[89] Xiaoyu Zhu, Yifei Yuan, Xianbiao Hu, Yi-Chang Chiu, and Yu-Luen Ma. A bayesian network model for contextual versus non-contextual driving behavior assessment. *Transportation research part C: emerging technologies*, 81:172–187, 2017.

[90] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net.

Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.

[91] Yajie Zou, Dominique Lord, Yunlong Zhang, and Yichuan Peng. Comparison of sichel and negative binomial models in estimating empirical bayes estimates. *Transportation Research Record: Journal of the Transportation Research Board*, (2392):11–21, 2013.

[92] Yajie Zou, Lingtao Wu, and Dominique Lord. Modeling over-dispersed crash data with a long tail: examining the accuracy of the dispersion parameter in negative binomial models. *Analytic Methods in Accident Research*, 5:1–16, 2015.