Penalized Regression Methods for Modelling Rare Events Data with Application to Occupational Injury Study

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Biostatistics Program of School of Public Health
University of Saskatchewan
Saskatoon

By
Roya Gavanji

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Director of School of Public Health
Health Sciences Building E-Wing, 104 Clinic Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 2Z4 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

Occupational injuries are a serious public health concern for workers around the world. Among all occupational injuries reported to the Workers' Compensation Board of Saskatchewan (WCB-SK) from 2007-2016, 177 (0.06%) out of 280,704 injury claims were fatal. Although work-related injuries are relatively rare, they have tremendous impact on the workers, their family, as well as a company's overall productivity, hiring/training costs, and insurance premiums. To help inform prevention of fatal claims, this study identified factors that increase the probability of fatal injury claims in Saskatchewan.

WCB Saskatchewan's administrative occupational injury claims data from 2007-2016 was used to extract fatal and non-fatal occupational events. Potential covariates included worker characteristics (age, gender, occupation) and incident characteristics (source of injury, cause of injury, part of body). Given the fatality being rare in this study, conventional logistic regression including multiple categorical covariates with over 40 parameters yielded biased parameter estimates. Penalized logistic regression methods, such as bias-correction method, i.e. Firth's method as well as the model selection methods, i.e., lasso and elastic net were compared to identify an optimal modelling strategy for calculating the odds ratio (OR) and 95% confidence intervals (CI) for probability of a WCB claim being fatal (vs. non-fatal).

Based on the best-fitting model, i.e., Firth's logistic regression of the selected variables under the elastic net method, odds of a claim being fatal was 5.5 (95% CI: 2.77,12.46) times higher among men than women and was 6.59 (95% CI: 3.59,12.20) times higher for seniors aged 65-85 as compared with those who are aged 14-24. Odds of a claim being fatal among those who work in primary industry is 2.85 (95% CI: 1.07,9.39) higher than those working in social sciences. The odds of injury being fatal for machinery sources is 51 (95% CI: 10.38,505.38) times higher than chemical products as the source.

Men workers are at higher risk of a claim being fatal (vs non-fatal). With respect to age, result of analysis showed that the middle-aged workers are at a lower risk, and the young workers are at a higher risk than middle aged workers. The risk of a claim being fatal increased sharply as age increased from 45 to 85. Primary industry sector and machinery have a disproportionate share of fatal claims. This knowledge can improve workplace safety by

learning from past incidents, identifying significant risk factors, and implementing targeted prevention strategies. Through development of effective interventions, we hope to prevent fatal injuries in Saskatchewan.

Acknowledgements

This research was undertaken with support from the Workers' Compensation Board of Saskatchewan (WCB-SK) as well as the Mathematics of Information Technology and Complex Systems (MITACS) Accelerate Grant. In the course of my studies, I also received funding support from the Canadian Centre for Health and Safety in Agriculture (CCHSA) through the Founding Chairs Fellowship and Natural Sciences and Engineering Research Council (NSERC) Discovery Grant. This work was supported in part by the Canada Research Chairs Program.

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr. Catherine Trask, and Dr. Cindy Xin Feng. I am grateful for their continuous support, immense knowledge, and motivation during this chapter of my life. Thank you for your support, for your trust and for believing in me and my abilities. I was very lucky to have such great mentors guiding me through this period, and I truly appreciate and value everything I have learned from you.

Besides my supervisors, I am very grateful to my advisory committee Dr. Sean Tucker and Dr. Punam Pahwa. I would like to thank Dr. Tucker for his insightful comments and invaluable suggestions which broaden my perspectives in this research. My sincere thanks also go to Dr. Punam Pahwa, for her constant advice and guidance through my studies as the Biostatistics program chair. I would also like to thank Dr. Sinden for serving as my external examiner and for all of your timely feedback.

Last but not least, a special gratitude goes to my beloved family for their unconditional love throughout my life. I would like to thank my wonderful parents for their continuous support, guidance, and encouragement. This journey would not have been possible without them, and I dedicate this thesis to them.

Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike's Information Criterion
AUC	Area Under the Curve
AWCBC	Association of Workers Compensation Board of Canada
BIC	Bayesian Information Criterion
BLS	Bureau Labor Statistic
CV	Cross Validation
FTE	Full Time Equivalent
GCV	Generalized Cross Validation
GLM	Generalized Linear Model
LASSO	Least Absolute Shrinkage and Selection Operator
MCP	Minimax Concave Penalty
MSE	Mean Square Error
NTOF	National Traumatic Occupational Fatalities
OLS	Ordinary Least Square
ROC	Receiver Operating Characteristic
SCAD	Smoothly Clipped Absolute Deviation
VIF	Variance Inflation Factor

Chapter 1

Introduction

1.1 Background and Motivation

Occupational injuries continue to present a serious public health concern for workers all around the world. Work-related injuries not only impact the worker and their family, but they also affect a company's overall productivity, hiring and training costs, and insurance premium costs [2]. In Canada, an average of about one million occupational injury claims have been reported each year by provincial and territorial Workers' Compensation Boards (WCBs) [3]. The total direct annual costs of occupational injuries and fatalities to the Canadian economy were approximately $9.7 billion in 2008 [3].

Fatalities represent the most serious type of WCB claims; in 2017, the number of workplace fatalities in Canada was 951 [4]. Between 2013 and 2017, Saskatchewan's five-year average acute injury fatality rate ranked highest (4.9 per 100,000 workers) among provinces with over 100,000 workers [5]. Saskatchewan also showed the greatest percentage increase (63%) in occupational disease fatality rates during this period. The number of work-related fatalities in SK in 2018 was 28 [6]. These numbers include only the claims reported to and accepted by the compensation boards, so the total number of workers affected by occupational injuries and illness may be even higher.

The Saskatchewan Workers' Compensation Board has identified several gaps and problems in their most recent strategic review, which resulted in a "fatalities problem statement": **"We have too many fatalities in this province and we do not know enough about them in order to develop a strategy to eliminate/mitigate them."** The negative effects of fatal occupational claims on workers, their families, and the economy of Saskatchewan demonstrate the need for further research to identify the risk factors associated with fatal occupational injuries, and this investigation forms the primary goal of this study.

A few challenges arise in modelling WCB data which may not be adequately handled by typical regression modelling strategies. For example, in WCB data, not only are the events rare, but also the covariates are mostly categorical variables with many levels and the distribution of the covariates are highly imbalanced. All these characteristics of the data may lead to quasi-complete separation problem. A quasi-complete separation happens when a logistic regression model perfectly or nearly perfectly predicts the response. In this case, as unique maximum likelihood estimates do not exist, the model fails to converge [7–9]. Often this happens when there is a categorical predictor with no variability in the response, which means all cases in one category of the predictor have the same response, which is the case in our WCB data. Even if there is no quasi-complete separation, separation may be nearly complete, so the standard error for a parameter estimate can become very large. Perfect prediction or complete separation can occur for many reasons. One of the possible scenarios for quasi separation to arise is when the event of interest in rare. The likelihood of separation is higher for categorical predictors with rare categories compared to continuous predictors [1]. In the presence of separation, maximum likelihood-based logistic regression faces problems including lack of convergence of maximum likelihood; even if it converges it produces biased (sometimes infinite) estimates of the regression coefficients [9–11].

One common strategy to address quasi-complete separation problem is to use Firth's method [12], which is a bias-preventive approach in which the parameter is not corrected after estimation, but a systematic corrective procedure is used to the score function which the parameter estimate is calculated from. This method provides consistent estimates of logistic regression parameters in the presence of separation [9]. Another solution is to reduce the number of covariates through model selection. The traditional model selection methods are forward, backwards or stepwise selection [13, 14], which find a subset of covariates to fit a regression model. These methods are useful when there are many potential covariates, and they can search for the presence of interactions, but the problem is that the traditional model selection methods are prone to overfitting and have been shown to yield models with low prediction accuracy [15]. To reduce the problem of overfitting, penalization or regularization [16, 17] method, such as lasso (least absolute shrinkage and selection operator) [15] or elastic net [16] can be used, which impose penalty to the log likelihood function to reduce (shrink) the coefficient values toward zero [18]. Penalized regression methods will be discussed in details in Chapter 3.

1.2 Objectives

In our analysis of fatal injury claim based on the WCB claim data, the events are rare ($< 1\%$) with many potential categorical covariates (listed in Appendix C), which leads to the problem of quasi-complete separation. Penalized regression methods, such as Firth's method or the model selection methods can help to find a parsimonious model for identifying risk factors associated with fatal occupational injuries.

To the best of our knowledge, except one study that has been conducted to identify factors associated with fatal occupational accidents among Mexican workers using Firth's method [19], other penalized regression methods have not yet been applied in occupational health studies and it is also not clear that which of these methods would be the best for analysis of WCB claims data with several challenging characteristics. Therefore, we aim to examine each of these methods on the data to evaluate their estimation performance to get a new perspective on this problem by applying these methods. We are particularly interested in examining whether model selection methods such as traditional backward regression, lasso or elastic net can fully help to solve the quasi-complete separation problem and in doing so if they lead to an inferior fit to the data; moreover, we propose to examine whether applying Firth's method after model selection can further improve the model fit.

We aim to answer two primary research questions by doing this study. The first research question is: what is the best-performing penalized logistic regression method within this context of quasi-complete separation and a rare event? The second research question is: what, if any, are the statistically-significant relationships between worker and incident characteristics and the likelihood of a workers' compensation claim being fatal?

1.3 Outline

The remainder of the thesis is organized as follows: In Chapter 2, the problem of fatal occupational health claims, gaps in understanding, and a summary of challenges in working with this data will be demonstrated. The principals of Firth's logistic regression method, and some model selection methods including lasso and elastic net will be provided in Chapter 3. In Chapter 4, administrative Saskatchewan WCB data set will be introduced and model comparison and result interpretation will be provided. Discussion and concluding remarks will be given in Chapter 5.

Chapter 2

Motivating Study

This chapter will discuss the problem of fatal occupational health claims, gaps in understanding, and highlight the necessity of learning more about this problem. Section 2.1 gives an overview of Saskatchewan Workers' Compensation Board data set. Literature review on the analysis of WCB data in Canada will be presented in Section 2.2. The analytic challenges in analysis of WCB-SK administrative data set will be discussed in Section 2.3, and finally a summary of the problem and challenges will be discussed in Section 2.4.

2.1 Overview of Workers' Compensation Board

Workplace mortality and morbidity result in suffering and hardship for the worker and their family, but they also result in loss of time at work, reduction of overall productivity for the enterprise and economy, and increased additional hiring and training costs due to staff replacement [20]. The Workers' Compensation Board (WCB) is an insurance system for workplace injuries and illnesses that delivers financial help, medical treatment, and rehabilitation to injured workers, and they also do prevention [21]. It is a no-fault system, which means that neither an employer's nor a worker's fault has to be proven for workers to get financial help and health benefits in case of occupational injury [22]. The WCB of Saskatchewan [21] is an independent agency that manages Saskatchewans workers' compensation system and operates under a provincial law known as The Workers' Compensation Act [21].

2.2 Literature Review

Occupational health studies including WCB claims data have been conducted in Canada [23–29], some of which studied serious and/or fatal claims. A number of studies analyzed WCB claims data from the provinces of Manitoba, British Columbia, and Ontario to understand the characteristics of the high risk groups of occupational injury across Canada. Tucker et al. [23] conducted descriptive analysis on WCB claims data using data from Association of Workers' Compensation Board of Canada (AWCBC) and estimated full time equivalent (FTEs); they derived the fatality rate and injury and illness rates for different provinces and compared them [23]. Fan et al. [24] analyzed the WCB serious claims data from British Columbia using negative binomial regression analysis to examine the rate and distribution of serious work-related injuries by demographic, work, and injury characteristics. McLeod et al. [27] conducted detailed analysis of work disability duration across jurisdictions including Manitoba, British Columbia, and Ontario using Cox proportional hazard model. Table A.1 in Appendix A summarizes the Canadian WCB studies conducted in British Columbia, Manitoba, and Ontario. The study of work-related injury claims has also been conducted in other countries such as Italy, United States, Australia, and Mexico [19, 30–33]. For example, In 1998, Chen et al. [31] applied the National Traumatic Occupational Fatalities (NTOF) surveillance system to assess risks of occupational fatal injuries related to cause and occupation among U.S. construction workers. They derived fatality injury rate and working lifetime risk. This study was the first to provide a comprehensive national profile of occupational injury risk for construction workers in United States [31]. In another study using extracted data from NTOF surveillance system for a 12 year period, Kisner et al. [34] calculated fatality rates and risk ratios using annual average employment data from Bureau Labor Statistic (BLS). In this study, rate ratios were reported for cause of death and industry divison combinations and cause of death and occupation divison combinations. In another study, Gonzalez et al. [19] utilized information from National Occupational Risk Information System to identify risk factors associated with fatality using logistic regression with Firth's approach. They considered sociodemographics (including age, sex and occupation), the work environment and workplace characteristics in their study.

Several modelling methods were used in these studies including negative binomial and Cox PH models, but the methods are for modelling different outcomes, so these studies are not only different because of the study time periods, their outcomes are also different. Although the researchers analyzed workplace claims data during different time periods, few studies explored penalized regression methods for improving the effect estimation. This might be due to the fact that they did not encounter quasi-complete separation problem (for binary outcome) in their data set. To our knowledge, the only study that used penalization method is a study by Gonzalez et al. [19] that used Firth's approach for identifying factors associated with fatal occupational accident, in which the number of fatalities is 1,140 out of 406,222 with almost 60 parameters (EPV=19). In the current study, our event of interest (fatality) is even more rare than their study. With respect to method, in the current study, we have not considered using negative binomial for modelling counts as aggregating data into count outcome may result in loss of information and may limit our ability to explore the impact of many categorical variables. The reason why we could not consider time to event outcome is because we did not have access to that kind of information in our data.

To our knowledge, most research in Canada have investigated serious occupational injuries, and there have not been any studies conducted on the fatal occupational injuries in Canada and specially in Saskatchewan. Although there are a few studies conducted on occupational fatalities in other countries, which mostly focused on fatality rate, those studies have not used penalized regression method. This remaining gap in using penalized regression methods forms one of the main objectives of this study. In modelling the rare events data with many parameters, model selection is another strategy to resolve the overfitting problem. Besides the traditional model selection methods (backward, forward, stepwise), regularization techniques, such as lasso and elastic net, that reduce the size of the coefficient estimates (shrinking them towards zero) have gained increasing popularity recently. The key strength of this thesis is to move beyond the conventional logistic regression method to investigate the penalized regression modelling methods for addressing the challenges arising from analyzing rare events data.

2.3 Methodological Problem Statement: Challenges

In this section, the analytic challenges that were encountered in modelling the risk of occupational fatality based on the WCB data are discussed as follows.

2.3.1 Rare Event

Rare events are dichotomos dependent variables with dozens to thousands of times fewer "ones" (i.e. events, such as wars, vetoes, or epidemiological infections) than "zeros" (i.e. non-events) [35]. Many studies have shown that rare-event variables are difficult to explain and predict; common statistical methods like logistic regression can underestimate the probability of rare events [35]. When the number of event of interest is small in comparison with estimated regression coefficients, overfitting is likely to occur [36]. Overfitting happens when a model can accurately classify data that is very closely related to the training data, but it performs poorly when using it for data point that are not closely related to training data, which means that random fluctuation and the noise in the training data is learned and negatively affect the model's ability to gerneralize. This problem may arise in the studies of rare events or rare diseases in health research [36].

Preliminary analysis of the WCB claim data shows that, out of 280,704 WCB traumatic injury claims between 2007 and 2016, only 0.06% (177) of WCB claims were fatalities, which indicates that EPV is less than 10 as the number of coefficients to be estimated is around 40. EPV can be calculated by dividing the number of events by the number of covariates used in developing a prediction model, or equivalently the number of EPV is the number of events divided by the number of degrees of freedom needs to present all of the variables in the model [37]. Roughly 10 EPVs are required for true estimation of regression coefficients in logistic regression model [38].

2.3.2 Multiple Covariates with Many Levels

The problem of many potential predictors is a concern with the WCB dataset, since not only the number of variables but also the number of levels in the categorical covariates is high. Rare events and multiple covariates with many levels cause low EPV, which leads to unstable parameter estimates. Under these circumstances, an alternative to standard regression techniques is needed to address this problem. The most common selection methods are forwards or backwards stepwise selection [13, 14], which find a subset of covariates to fit a regression model. This is useful when there are many potential covariates, and they can search for the presence of interactions, but the problem is that stepwise methods are prone to overfitting and have been shown to yield models with low prediction accuracy [15].

In WCB claims data, we consider 6 covariates including age, gender, occupation, part of body, source of injury, and cause of injury. Many of those variable have multiple levels; for example, for the source of injury, there is 10 levels. In total, there are around 40 dummy variables.

2.3.3 Quasi-Complete Separation

Another problem that arises in the analysis of WCB claims data is quasi-complete separation, which happens when one or some of covariates can perfectly or nearly perfectly predict the response variable. Table 2.1 shows an example of separation. We can see that in presence of complete separation, observations with Y=A all have values of X=0, and observations with Y=B all have values of X=1. In other words, Y separates X perfectly or X predicts Y perfectly because X=1 corresponds to Y=B and X=0 corresponds to Y=A.

Table 2.1: Example of complete and quasi-complete separation for binary covariate X against outcome variable Y, based on Rahman et al [1]

Complete separation				Quasi-complete separation			
		Y				**Y**	
		A	**B**			**A**	**B**
X	0	177	0	**X**	0	177	0
	1	0	177		1	2	175

As shown in the Table 2.2, the problem of quasi-complete separation is present in this data. In some of the levels, the number of fatal claim injuries is equal to zero. For example, there are not any fatalities in the occupations in art and science category in Table 2.2. Some of other characteristics also have the problem of zero cells, which will be discussed further in Chapter 4, Section 4.2.1. These zero cells here are an indicator of the presence of quasi-complete separation, which can be problematic while analyzing WCB-SK data set with traditional regression methods such as logistic regression.

Table 2.2: Distribution of the occupation of injured workers from Saskatchewan WCB who had fatalities vs. those who did not, SK, Canada, 2007-2016

Occupation	**Injury being fatal**	
	Yes (%)	**No (%)**
social sciences	1(0.00)	6160 (2.19)
business/advertising	4 (0.01)	12785 (4.55)
health	1(0.00)	27617 (9.84)
natural/applied sciences	9(0.00)	3592 (1.28)
primary industry	24 (0.01)	12480 (4.45)
art/culture	0(0.00)	1228(0.44)
sale/services	12(0.01)	59437 (21.17)
trade/transport	93(0.03)	95142 (33.89)
processing/manufacturing	9(0.00)	22028 (7.85)
not stated	24 (0.01)	40058 (14.27)

2.3.4 Multicollinearity

Another challenge in working with Saskatchewan WCB claims data is the presence of multicollinearity. Multicollinearity is "a situation in which two or more independent variables are perfectly or nearly perfectly correlated" [39]. Using multiple regression models, multicollinearity can lead to several problems including: increased variance of estimated regression coefficients and unstable parameter estimates [40, 41]. Variance Inflation Factor (VIF) is a statistic that measures the level of multicollinearity [39]. The VIF is defined as follows

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{2.1}$$

where R_i^2 is the square of the multiple correlation coefficient from the regression of the j-th explanatory variable on the remaining explanatory variables. Ringle et al [42] suggested that the maximum acceptable level of VIF has to be smaller than 5. However, a rough rule of thumb is that variance inflation factors greater than 10 can be problematic in multiple linear regression. In this study, we will use the more traditional maximum level of VIF for logistic regression which is 2.5 [43, 44].

Table 2.3 represents the VIF for the dummy variables of all the categorical variables. As shown in the Table 2.3, some of dummy variables are highly correlated with VIFs greater than the recommended maximum of 2.5. For example, VIF for bodily reaction in the cause of injury is 132.87 and VIF for upper extremities in part of body is 14.42, both of which are higher than the recommended value of 2.5.

One of the concerns under high multicollinearity is the interpretation of regression coefficients [45]. The predictor variables in the model will largely affect the same portion of variance as none of them can make a unique contribution, so one must be cautious interpreting the partial coefficients of a set of variables [46]. Another problem with high multicollinearity is that the parameter estimates might show sample to sample variation, which means they are not reliable [47, 48].

Table 2.3: Variance inflation factor (VIF) for examining the multicollinearity among the dummy variables of the categorical covariates in the analysis of WCB injury claim data

Variable	VIF	Variable	VIF	Variable	VIF
men	1.56	primary industry	1.24	other sources	11.28
age 25-34	1.69	processing	1.4	parts/materials	8.85
age 35-44	1.65	body system	2.4	tools/instruments	8.9
age 45-54	1.71	head	7.6	vehicles	5.6
age 55-64	1.46	lower extremities	9.63	assaults	16.49
age 65-85	1.07	multiple parts	5.04	bodily reaction	132.87
occupation business	1.26	other parts	1.13	contact with objects	132.58
applied sciences	1.07	trunk	13.64	harmful substances	34.5
health	1.75	upper extremities	14.42	falls	69.8
art/culture	1.03	containers	7.53	other events	26.26
sales/services	1.93	furniture	3.23	transportation accidents	15.56
trades/transport	2.22	machinery	4.5		

2.4 Summary of Problem and Challenges

In the WCB fatal claims data, the event of interest (fatal injury) is rare 177 over 280,704 (<1%) claims, with multiple categorical covariates containing many levels as shown in Appendix C. During our analysis, we found that the estimated regression coefficients tend to be unstable with wide confidence intervals, which is undesirable for estimation. We also encountered the problem of a rare event and presence of quasi-complete separation problem that we need to address by using some statistical methods. This motivates current study seeking for a more appropriate analytic strategy to address those challenges including the Firth's method to correct the bias in the parameter estimates and using model selection methods (such as lasso and elastic net) to build a more parsimonious model and deliver better-estimated coefficients for modelling rare events. Given that few published reports have evaluated the performance of different methods in this context, this study has the potential to advance knowledge of identifying more appropriate analytic tools for identifying risk factors for fatal WCB claims.

Chapter 3

Statistical Methods Overview

To address the analytic challenges presented in Chapter 2, several statistical methods are presented in this Chapter as the potential solutions for overcoming these challenges.

3.1 Conventional Logistic Regression Model

Logistic regression models are commonly used to estimate the relationship between a binary response variable and one or more covariates. The popularity of logistic regression mainly comes from its mathematical convenience and the easy interpretation of results in terms of odds ratios. Let y_i be the outcome variable for the i-th subject, and it is Bernoulli distributed and takes on the value 1 with probability $\pi_i = P(y_i = 1|x_i)$, where $\mathbf{x}_i = (x_1, ..., x_p)^T$ is the i-th subject's covariate vector, and value 0 with probability $1 - \pi_i$. The logistic regression model with the logit link function can be written as:

$$\pi_i = \frac{exp(\beta_0 + x_i^T \beta)}{1 + exp(\beta_0 + x_i^T \beta)} \tag{3.1}$$

where β_0 is an intercept term, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a $p \times 1$ vector of estimated regression coefficients on the logit scale.

Equation 3.1 is a generalized linear model. If parameter $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta})^T$, then the corresponding log-likelihood function is given by the following equation as it was also shown by [49]:

$$\ell_{\boldsymbol{\theta}} = \sum_{i=1}^{n} [y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)] \tag{3.2}$$

By replacing π_i from Equation 3.1 in Equation 3.2, we have:

$$\ell_{\boldsymbol{\theta}} = \sum_{i=1}^{n} \left[y_i(\beta_0 + \mathbf{x}_i{}^T \boldsymbol{\beta}) - log(1 + exp(\beta_0 + \mathbf{x}_i{}^T \boldsymbol{\beta})) \right] . \tag{3.3}$$

In the maximum likelihood method, the goal is finding a set of values for $\boldsymbol{\theta}$ that can maximize Equation 3.3. One of the most common ways of doing this is differentiating this equation with respect to $\boldsymbol{\theta}$, set the derivative to 0, and then solve the equation to find estimated regression coefficients using MLE [7]. However, for most data and models, there is not a closed form or explicit solution for this equation. In these cases, numerical methods such as Newton-Raphson algorithm will be used [7]. For more information about this method we refer the reader to Anderson [7]. With respect to this thesis, the question is what would happen with this algorithm when we have quasi-complete separation problem. Using this algorithm when we have the problem of quasi-complete separation, "at each iteration, the parameter estimate for the variable (or variables) with separation gets larger in magnitude. Iterations continue until the fixed iteration limit is exceeded. At whatever limit is reached, the parameter estimate is large and the estimated standard error is extremely large" [7], which in turn leads to lack of convergence of ML.

Although maximum likelihood (ML) estimation is one of the most common methods to estimate unknown regression coefficients, ML is also known to have finite sample properties [36]. For example, when the event per variable (EPV) is low and the quasi-complete separation or complete separation may occur, ML estimation could lead to infinite estimates of coefficients [12].

3.2 Firth's Logistic Regression

One of the possible solutions for the problem of separation in WCB data is using Firth's logistic regression. Heinze and Schemper showed that Firth's method is an ideal solution to the issue of separation [9].

The expectation of the estimate is always larger in absolute value than the true parameter, so maximum likelihood estimates of $\boldsymbol{\theta}$ are biased away from 0 [50]. As shown by Firth [12], the bias of the ML estimates of $\boldsymbol{\theta}$ can be expanded asymptotically as

$$Bias(\boldsymbol{\theta}) = E(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta} = \frac{\beta_1(\boldsymbol{\theta})}{n} + \frac{\beta_2(\boldsymbol{\theta})}{n^2} + ... \tag{3.4}$$

Most bias-corrective methods remove the first asymptotic order bias from $\hat{\boldsymbol{\theta}}$ by using $\hat{\boldsymbol{\theta}}_{BC} = \hat{\boldsymbol{\theta}} - \frac{\beta_1(\hat{\boldsymbol{\theta}})}{n}$ [12]. These methods rely on calculating the MLE and correcting MLE by subtracting the first-order bias $\frac{\beta_1(\boldsymbol{\theta})}{n}$ [51]. In presence of complete or quasi-complete separation, it is not feasible because MLEs do not exist. To tackle this problem, Firth [12] introduced a bias-preventive approach in which the parameter is not corrected after estimation, but a systematic corrective procedure is used to the score function from which the parameter estimate is calculated. This method provides consistent estimates of logistic regression parameters in the presence of separation [9]. The detail of Firth's logistic regression method can be found in Appendix B, and we refer the reader to these papers [12, 52–55] for more information on this method.

As it was also shown in [1], taking the natural logarithm of the Equation B.6 gives us the corresponding log likelihood function

$$\ell^*(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \frac{1}{2} log|I(\boldsymbol{\theta})| \tag{3.5}$$

If Firth's method is used in binary logistic regression model as defined in Equation 3.1, where $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta})^T$ this is known as Firth's logistic regression. The penalized log likelihood function in this case is

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} [y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)] + \frac{1}{2} log|I(\boldsymbol{\theta})|, \tag{3.6}$$

Which the information matrix is $I(\boldsymbol{\theta}) = \mathbf{X^T W X}$, with $\mathbf{W} = diag[\pi_i(1 - \pi_i)]$ and $\pi_i = P(y = 1|x_i, \boldsymbol{\theta})$. The second term on right hand side of the above equation is maximized at $\pi_i = 0.5$ for $i = 1, 2, ..., n$ which occurs in $\boldsymbol{\theta} = 0$. So the parameters are shrunk towards zero. The penalized-likelihood estimates will be smaller in absolute value than standard MLEs [1, 12].

Heinze and Schemper [9] applied Firth's logistic regression to data sets that have separation. The results of their study showed that Firth's penalized likelihood estimator is an ideal solution in case of separation problem in logistic regression, which is the case in our data set. By comparing the estimates derived by Firth's method with those derived by ordinary MLE, they concluded that in presence of small samples Firth's method is superior to ordinary MLE as point estimates have lower variability and confidence intervals are more reliable.

3.3 Lasso Penalized Logistic Regression

The Least Absolute Shrinkage and Selection Operator (lasso) is a penalization (regularization) method introduced by Tibshirani in 1996 [15], which can be used for regression coefficient estimation and variable selection when the number of covariates (regression coefficients) p is larger than the number of sample size n. This method performs both regularization through penalizing and shrinking parameter estimates, and variable selection as it is able to shrink parameter estimates to exactly zero.

Lasso is an alternative to ridge regression with a different penalty term, and it is able to overcome the disadvantage of ridge regression. Ridge regression shrinks the regression coefficients towards zero by imposing constraint, but it does not shrink the regression coefficients to exactly zero, which is why it can not be used as a variable selection method [56, 57]. Therefore, lasso will be used as an alternative method to do variable selection.

In conventional logistic regression the parameter estimates are derived by maximizing the log-likelihood function:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{y_i log(\pi_i) + (1-y_i) log(1-\pi_i)\} = \sum_{i=1}^{n} \{y_i(\beta_0 + \mathbf{x}_i{}^T\boldsymbol{\beta}) - log\left[1 + exp(\beta_0 + \mathbf{x}_i{}^T\boldsymbol{\beta})\right]\} \tag{3.7}$$

The lasso logistic regression estimator depends on the choice of tuning (shrinking) parameter $\lambda \geq 0$, that can be chosen by cross validation or generalized cross validation [15]. As shown in [58], by Maximizing the penalized log-likelihood function shown in Equation 3.8, the regression coefficients estimates will be derived [17].

$$\ell_\lambda^L(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \lambda\sum_{j=1}^{p}|\beta_j| = \sum_{i=1}^{n} \{y_i(\beta_0 + \mathbf{x}_i{}^T\boldsymbol{\beta}) - log\left[1 + exp(\beta_0 + \mathbf{x}_i{}^T\boldsymbol{\beta})\right]\} - \lambda\sum_{j=1}^{p}|\beta_j| \tag{3.8}$$

The ℓ_1 penalty in lasso sets some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is large enough. Models generated from lasso are generally easier to interpret than models produced by ridge regression, and lasso yields sparse models as it excludes ‘unnecessary’ predictors by shrinking their coefficients to exactly zero, yielding a more parsimonious model [59]. Selection of a good value of λ for lasso is critical and it can be driven by cross validation.

3.3.1 Choice of Regularization Parameter

Choosing the suitable regularization parameter λ is an important thing that needs to be considered in penalized logistic regression. When λ becomes larger, the bias increases and the variance decreases, and this is where we need to decide how much bias we take to decrease the variance which leads to the choice of optimal tuning parameter λ. We are interested in finding λ that gives us the model with the lowest mean square error (MSE). In cross validation, test sets and training sets are made by splitting the data set to K groups. Common choices of K is 5 or 10, where one group is chosen as a test set and the remaining K-1 groups form the training set [60].

For K-fold cross validation, first we split the data into K equal size parts. Then for each part (k-th), we fit the model to the other K-1 parts of the data and calculate the MSE of fitted model when predicting the k-th part of the data [60]. At next stage, we repeat the procedure for k=1,2,...,K and average the K estimates of mean square error, which gives us a cross validation error curve [60].

For instance, 10 fold cross validation consists of splitting the data into 10 sub samples with the same size, before fitting the considered model on 9 sub samples (in this case 90% of the data is in the training set) and evaluating the model's performance on the remaining one sub sample (10% of the data is in the validation set) [61]. Then, this process is repeated for all 10 cases, where each of the 10 sub samples would be used one time as validation set. The value of λ that results in the lowest MSE rate is then chosen.

When the cross validation error curve achieves the minimum, the estimate of λ is chosen. This choice of tuning parameter often results in insufficient regularization, which means that too many variables stay in the model [62]. Hastie et al. [17] reports that the model based on the one standard error rule is the best cross validated model; this means that selected model will be the most regularized model with error within one standard error of the minimal error.

The simplest model whose accuracy is comparable with the best model will be chosen by this rule [17].

3.4 Elastic Net Logistic Regression

The elastic net was introduced by Zou and Hastie [16] as another regularization and variable selection method which is capable of outperforming lasso, especially where the number of predictors is significantly larger than the sample size (i.e. $p >> n$), while this method retains a similar sparsity. Elastic net eliminates the problems that occurs when lasso is used in the presence of highly correlated variables [63].

Elastic net includes the tuning parameter $\alpha \geq 0$, and the penalty term in this method is a combination of ridge and lasso as it was also shown by [16, 58]:

$$\alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j| \tag{3.9}$$

Elastic net is the combination of ℓ_1 and ℓ_2 penalties that conveys the desirable properties of both ridge and lasso [16]. The method can effectively shrink the coefficients of non-informative features to exactly 0, and it is also able to control the group of correlated features. For more detailed features about the elastic net method we refer the reader to Zou et al. [16] and Tibshirani et al. [60].

3.5 Other Penalized Regression Methods

Many other penalties have been introduced after introducing lasso by Tibshirani [15]. Fan and Li [64] introduced the Smoothly Clipped Absolute Deviation (SCAD) penalty. They showed how the penalized estimator in SCAD is optimal in the sense that it performs as if the active variables are known [65]. Later on, Zhang introduced the Minimax Concave Penalty (MCP) which is a similar method to SCAD [66]. In 2006, Zou [67] introduced another penalized estimator called adaptive lasso that has the oracle property. Adaptive lasso is much simpler than the SCAD and MCP penalties. This thesis focuses on exploring the two mostly commonly used penalized model selection methods, namely, lasso and elastic net.

3.6 Method Comparison Criteria

3.6.1 Akaike's Information Criterion (AIC)

For a given model, the AIC is a measure of the loss of information resulted by the use of model to explain a specific pattern or variable [68].

$$AIC = -2logL + 2k \tag{3.10}$$

Where k is the number of estimated parameters in the model. The log-likelihood of the model given the data shows the overall fit of the model. AIC penalized for addition of the parameters, which means it selects the model that fits the data well with a minimum number of parameters [68]. The smaller the AIC, the more accurate the model.

3.6.2 Area Under the Curve (AUC)

One of the ways to rate the predictive performance of a model is Area Under the Curve (AUC), which measures the area under the Receiver Operating Characteristic (ROC) curve [69, 70]. AUC shows a trade off between specificity and sensitivity [71]. Sensitivity is the proportion of events that are correctly predicted while specificity is the proportion of non-events that are correctly predicted [72]. The ideal is for both of these proportions to be high. "ROC is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied" [73]. By plotting the true positive rate (TPR) vs the fraction of false positive rate (FPR) at various threshold settings ROC will be made [69, 73]. The AUC=0.5 is the baseline, and AUC=1 shows perfect prediction. For more information about ROC, we refer the reader to Tom Fawcett [70].

Chapter 4
Application to WCB Data

In this Chapter, the performance of different penalized regression methods will be compared to see which one fits the data better and then the results of analysis for WCB data based on the best method will be provided. At first, a brief background about the study will be given, then the data set and the variables will be described and finally, method comparison, the analysis results, and the interpretation will be provided.

4.1 Data Sources and Descriptions

4.1.1 Study Population

The data used for this study is the administrative occupational injury claim data from 2007 to 2016 for workers in Saskatchewan which was provided by Saskatchewan's WCB. A summary of all explanatory and outcome variables is reported in Appendix C. There are 280,704 observations and near 40 features with fatality as the response variable or outcome of interest. Illness-related fatalities (such as occupational cancers) are not included in this analysis. The minimum age considered in this study is 14 because the minimum legal working age in Saskatchewan is 14, and the maximum age considered in this study for workers is 85 as it is close to the life expectancy in Canada.

4.1.2 Outcome Variable

The outcome variable that will be considered in this analysis of WCB data is fatal injury claim. The percentages of fatal claims in WCB data are reported in Table 4.1.

Table 4.1: Frequency and percentage of fatal injury vs. non-fatal injury claims

Fatal claim indicator	Frequency	Percentage
Non-fatal	280,527	99.94
Fatal	177	0.06

4.1.3 Potential Covariates

All the potential covariates and their categories are given in Appendix C. Table C.1 presents the descriptive statistics/frequency for a subset of covariates. For the remainder of the thesis, all explanatory variables will be explored.

Demographic Characteristics

- Gender: Men vs. women, with women as the reference group since our preliminary results showed that women have a lower chance of a claim being fatal.

- Age: In our study, age was stratified into six categories: 14-24, 25-34, 35-44, 45-54, 55-64, 65-85 with 14-24 as the youngest group and 65-85 as the oldest group.

- Occupation of the worker at the time of the occupational injury: There are ten categories of occupations including: (1) business and finance; (2) health; (3) natural and applied sciences; (4) primary industry; (5) art, culture, recreation and sport; (6) social sciences and education; (7) sale and services; (8) trade and transport; (9) processing, and manufacturing; and (10) not stated.

Characteristic of injury

- Cause of injury: Why injury happened (7 categories; please see Appendix C for more details)

- Part of body injured: Part of body that injured at time of injury (7 categories; please see Table C.1 in Appendix C for more details)

- Source of injury: (10 categories; please see Table C.1 for more details)

- Year: Year when the injury happened.
- Month: Month when the injury happened.

4.2 Saskatchewan WCB Data Analysis Results

Based on the nature of data and challenges we encountered some of which explained in Chapter 2, this analysis will use several methods, including: conventional logistic regression, conventional logistic regression after variable selection, Firth's logistic regression, and Firth's logistic regression after lasso, and elastic net variable selection methods to analyze WCB fatal claims data.

4.2.1 Descriptive Statistics for Covariates

Selected descriptive statistics for covariates of interest are described within three categories: personal characteristics Table 4.2, incident characteristics Table 4.3, and temporal characteristics Table 4.4. Characteristics of the study population by fatality status and the distribution of injured workers in Saskatchewan across different categories of the covariates considered in our analysis are presented in these tables.

As shown in the descriptive statistics tables, the problem of quasi-complete separation is present while working with WCB-SK data. For example, in personal characteristics there are not any fatalities in the occupations in art and science category, and for incident characteristics, there are not any fatalities in tools and equipments and lower extremities in source of injury and part of body characteristics respectively.

As shown in Table 4.3 and based on our preliminary analysis, part of body and cause of injury are two main characteristics that cause the problem of quasi-complete separation.

Table 4.2: Distribution of the personal characteristics of injured workers from Saskatchewan WCB who had fatalities vs. those who did not, SK, Canada, 2007-2016 (p-values are based on bivariate analysis)

Characteristics	**Frequency** (%)	**Fatal injury**			**Type3**
		Yes (%)	**No (%)**	**P-value**	**P-value**
Gender					<.0001
women (ref)	92432 (32.93)	8 (0.01)	92424 (99.99)		
men	188272 (67.07)	169 (0.09)	188103 (99.91)	<.0001	
Age					<.0001
14 to 24 (ref)	56576 (20.16)	23 (0.04)	56553 (99.96)		
25 to 34	67495 (24.04)	34 (0.05)	67461 (99.95)	.4270	
35 to 44	57903 (20.63)	16 (0.03)	57887 (99.97)	.2355	
45 to 54	61297 (21.84)	43 (0.07)	61254 (99.93)	.0346	
55 to 64	33106 (11.79)	33 (0.10)	33073 (99.90)	.0010	
65 to 85	4327 (1.54)	28 (0.65)	4299 (99.35)	<.0001	
Occupation					<.0001
social sciences (ref)	6161 (2.19)	1 (0.02)	6160 (99.98)		
business/advertising	12789 (4.56)	4 (0.03)	12785 (99.97)	.0005	
health	27618 (9.84)	1 (0.00)	27617 (100)	.56	
natural/applied sciences	3601 (1.28)	9 (0.25)	3592 (99.75)	.054	
primary industry	12504 (4.46)	24 (0.19)	12480 (99.81)	.0008	
processing/manufacturing	22037 (7.85)	9 (0.04)	22028 (99.96)	.66	
art/culture	1228 (0.44)	0 (0.00)	1228 (100)	.97	
sale and services	59449 (21.18)	12 (0.02)	59437 (99.98)	.45	
trade/transport	95235 (33.93)	93 (0.10)	95142 (99.90)	.026	
not stated	40082 (14.28)	24 (0.06)	40058 (99.94)	.23	

Table 4.3: Distribution of the incident characteristics of the injury from Saskatchewan WCB, SK, Canada, 2007-2016 (p-values are based on bivariate analysis)

Characteristics	**Frequency** (%)	**Fatal injury**			**Type3**
		Yes (%)	**No (%)**	**P-value**	**P-value**
Source of injury					<.0001
chemical products (ref)	5109 (1.82)	1 (0.02)	5108 (99.98)		
furniture/fixture	7977 (2.84)	1 (0.01)	7976 (99.99)	.75	
parts/materials	30835 (10.98)	13 (0.04)	30822 (99.96)	.46	
structure/surfaces	39435 (14.05)	20 (0.05)	39415 (99.95)	.35	
vehicles	14660 (5.22)	71 (0.48)	14589 (99.52)	.0014	
containers	24759 (8.82)	2 (0.01)	24757 (99.99)	.47	
machinery	12809 (4.56)	14 (0.11)	12795 (99.89)	.096	
persons/animals	68549 (24.42)	28 (0.04)	68521 (99.96)	.47	
tools/equipments	30000 (10.69)	0 (0.00)	30000 (100)	.96	
other sources	46571 (16.59)	27 (0.06)	46544 (99.94)	.29	
Part of body					<.0001
other (ref)	5283 (1.88)	46 (0.01)	5237 (0.99)		
body systems	4569 (1.63)	30 (0.66)	4539 (99.34)	.0002	
head	31786 (11.32)	9 (0.03)	31777 (99.97)	.96	
lower extremities	47910 (17.07)	0 (0.00)	47910 (100)	<.0001	
upper extremities	87424 (31.14)	1 (0.00)	87423 (100)	<.0001	
multiple	20390 (7.26)	85 (0.42)	20305 (99.58)	<.0001	
trunk	83342 (29.69)	6 (0.01)	83336 (99.99)	<.0001	
Cause of injury					<.0001
reference category [1]	8124 (2.90)	3 (0.00)	8121 (100)		
contact with objects	99675 (35.51)	22 (0.02)	99653 (99.98)	.40	
bodily reaction/exertion	97801 (34.84)	2 (0.00)	97799 (100)	.002	
transportation accidents	6941 (2.47)	74 (1.07)	6867 (98.93)	<.0001	

[1] assaults/violent acts and fires and explosions

falls	37847 (13.48)	21 (0.06)	37826 (99.94)	.51
other events	12916 (4.60)	20 (0.15)	12896 (99.85)	.021
harmful substances	17400 (6.20)	35 (0.2)	17365 (99.80)	.005

Table 4.4: Distribution of the temporal characteristics of the injury event from Saskatchewan WCB, SK, Canada, 2007-2016 (p-values are based on bivariate analysis)

Characteristics	Frequency (%)	Fatal injury			Type3
		Yes (%)	No (%)	P-value	P-value
Year					0.3447
2007	30437 (10.84)	13 (0.04)	30424 (99.96)	0.9429	
2008	30360 (10.82)	15 (0.05)	30345 (99.95)	0.7695	
2009	26755 (9.53)	19 (0.07)	26736 (99.93)	0.2059	
2010	26887 (9.58)	20 (0.07)	26867 (99.93)	0.1614	
2011	28574 (10.18)	19 (0.07)	28555 (99.93)	0.2752	
2012	29287 (10.43)	28 (0.10)	29259 (99.90)	0.0290	
2013	29170 (10.39)	17 (0.06)	29153 (99.94)	0.4669	
2014	28211 (10.05)	19 (0.07)	28192 (99.93)	0.2606	
2015	26013 (9.27)	16 (0.06)	25997 (99.94)	0.3917	
2016 (ref)	25010 (8.91)	11 (0.04)	24999 (99.96)		
Month					.7519
January	23603 (8.41)	16 (0.07)	23587 (99.93)	.4408	
February	21372 (7.61)	8 (0.08)	21364 (99.99)	.5116	
March	23822 (8.49)	17 (0.01)	23805 (99.92)	.3591	
April	21559 (7.68)	13 (0.07)	21546 (99.93)	.6578	
May	23332 (8.31)	13 (0.06)	23319 (99.94)	.8061	
July	24760 (8.82)	12 (0.05)	24748 (99.95)	.9195	
August	25191 (8.97)	21 (0.09)	25170 (99.91)	.1660	
September	24404 (8.69)	14 (0.06)	24390 (99.94)	.7460	
October	24960 (8.89)	20 (0.09)	24940 (99.91)	.2061	
November	23938 (8.53)	18 (0.08)	23920 (99.92)	.2854	
December	20003 (7.13)	13 (0.07)	19990 (99.93)	.5286	
June (ref)	23760 (8.46)	12 (0.06)	23748 (99.94)		

4.2.2 Results of Model Fitting

In this section, we present the results of analysis based on different methods including conventional logistic regression, lasso logistic regression, and elastic net logistic regression, Firth's logistic regression, and Firth's logistic regression after doing variable selection via lasso and elastic net. First, we demonstrate model selection procedure from bivariate analysis to investigate interactions, and then we present the results based on different methods after doing multivariable analysis.

In the last two columns of tables 4.2, 4.3, and 4.4, the p-values for bivariate analysis and type 3 analysis were shown. In this study, we kept those variables with p-value less than 0.25 in the bivariate analysis for the first multivariable model. Based on p-values from tables 4.2, 4.3, 4.4, variables included in the model are gender, age, occupation, source of injury, cause of injury, and part of the body. Year and month will not be considered to be in the model for multivariable analysis as the p-value for year and month is 0.3447 and 0.7519 respectively, which are greater than 0.25.

The next step in model selection is fitting the multivariable model with all covariates identified for inclusion in bivariate analysis, then we do backward model selection based on p-values.

In presence of quasi-complete separation, *SAS* gives the results at the last iteration in case of using logistic regression [74]. After fitting conventional logistic regression method, age, gender, cause of injury, source of injury, and part of body were kept in the model.

The next step in model selection is investigating the assumption of the presence of interaction. To our knowledge, there have not been any studies on WCB claims data in Canada investigating the presence of interactions.

After using conventional logistic regression, *SAS* and *R* gave an error indicating that there exist the quasi-complete separation problem in this analysis. Then lasso and elastic net logistic regression methods were used to see whether applying variable selection methods can address the problem of quasi-complete separation or not. Based on the error from *R* and the results provided in tables 4.6 and 4.7, the problem of quasi-complete separation was solved after using lasso logistic regression with $\lambda = \lambda.1se$, but the separation problem is still there after using elastic net logistic regression.

Now that we presented the analysis results based on conventional logistic regression, lasso logistic, and elastic net logistic regression, we are going to apply Firth's method to address the quasi-complete separation problem and compare the results with other penalized methods. The *logistf* package [75] in *R* was used to run the Firth's method analysis. After fitting Firth's logistic regression methods, age, gender, cause of injury, source of injury, and part of body were kept in the model. The results of this analysis can be found in tables 4.8 and 4.9.

As mentioned earlier, there are some challenges in working with Saskatchewan WCB data including separation, rare events and multiple covariates with many levels (low EPV), and multicollinearity, all of which can be addressed by penalized regression. Now that we addressed the separation problem of the data by using Firth's method, we are going to try some variable selection methods including lasso and elastic net as regularization or penalization methods to address the remaining problems.

To determine if the variable selection improves the Firth's method, we will compare the Firth's method after variable selection with Firth's method before variable selection in terms of AIC. The results of analysis after variable selection using some of penalized logistic regression methods (lasso and elastic net) will be presented in this section to see what characteristics will be selected by these methods to stay in the final model. First, the results after doing variable selection by lasso method will be presented.

Firth's Logistic Regression Results after Lasso Variable Selection

Several implementations of lasso are offered in *R* like in packages liblinear [76], *glmnet* [77], *lars* [78], and *genlasso* [79]. We chose to use the *glmnet* package as model fitting is easy by using this package, which provides easy transition between lasso and elastic net models.

For fitting the lasso method, *cv.glmnet()* from *glmnet* package will be used. It performs 10 fold cross validations to find the best value of the tuning parameter λ.

First, the tuning parameter λ in the lasso penalty will be chosen using cross validation procedure, which choose λ_{opt} to be the one that minimize the deviance with respect to logistic regression. Figure 4.1 shows the cross validation plot generated by package *glmnet*. Figure 4.1 includes cross validation curve (red dotted line), and upper and lower standard deviation curves along the λ sequence (error bars). The larger the value of λ, the more variables will be eliminated from the model.

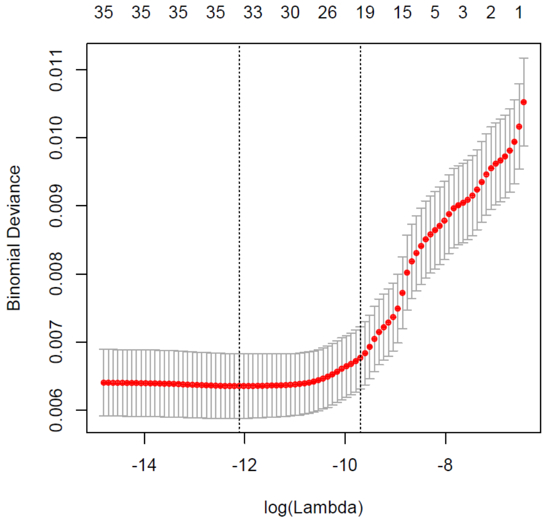

Figure 4.1: Cross validation plots for λ_{opt} for lasso method with WCB fatality as an outcome

There is two vertical lines in Figure 4.1. The one at the minimum is the one that minimize out of sample CV ($\lambda.min$) and the other vertical line is for $\lambda.1se$ which is the largest λ value within 1 standard error of $\lambda.min$. The numbers on the top of the Figure 4.1 give the number of non-zero coefficients, which means that for our data, we would be using 19 dummy variables instead of using 36 dummy variables for selected model if we would choose the one standard error estimate.

After doing variable selection via lasso, we refit the Firth's logistic regression again to compare the derived model with the results of Firth's method without doing variable selection. The results of fitted model is presented in Table 4.5. Another choice for tuning parameter λ is $\lambda.min$. If we select this λ as the final tuning parameter, the number of dummy variables remaining in the model will be 34, which means lasso removed only 1 dummy variable from the model with this choice of λ. A list of these variables can be found in Table C.2, Appendix C.

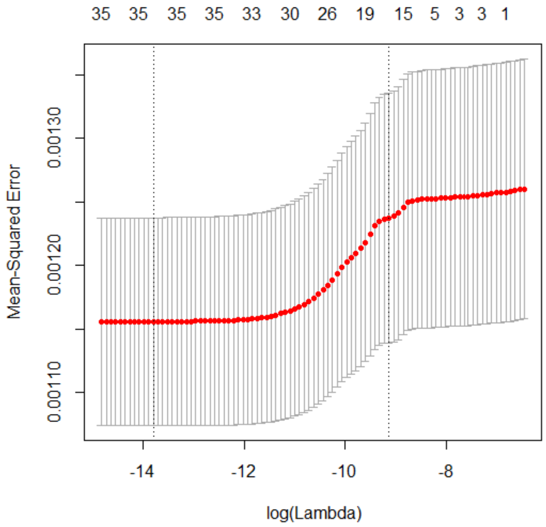

Figure 4.2: Cross validation plot for log(λ) vs mean square error for lasso method with WCB fatality as an outcome

Firth's Logistic Regression Results after Elastic Net Variable Selection

In this section, the results of variable selection conducted by elastic net will be presented. For fitting elastic net, *cv.glmnet* will be used. Just as with the lasso method, in elastic net we also have two choices for choosing tuning parameter λ, so we present the result of analysis based on these two quantities. Figure 4.3 shows the cross validation plot generated by package *glmnet*, which shows larger the value of λ, the more variables will be eliminated from the model. As shown in Table 4.5, if we choose $\lambda = \lambda.min$, then we would have 35 dummy variables in the model while we would have 26 dummy variables in the model if we choose $\lambda = \lambda.1se$. The result of Firth's method after elastic net variable selection can be found in tables 4.8 and 4.9.

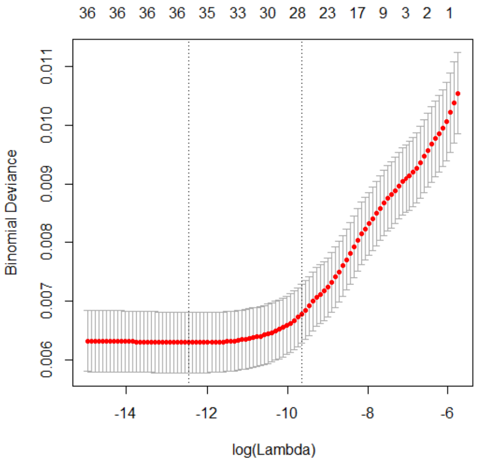

Figure 4.3: Cross validation plots for λ_{opt} for elastic net method with WCB fatality as an outcome

4.2.3 Method Comparison

In this section, a comparison between different methods that were used will be presented to find the best method to analyze WCB fatal claims data and interpret findings. Several model performance criteria have been used to compare conventional logistic, lasso logistic, elastic net logistic, Firth's, and Firth's after lasso, and Firth's after elastic net together, some of which mentioned in Section 3.6.

Table 4.5 reports method comparison score AIC and AUC for these methods, and it also reports the number of dummy variables selected to stay in the model in final analysis (i.e. model parsimony) plus the intercept. Based on AIC performance criteria shown in Table 4.5, Firth's after elastic net with the choice of $\lambda = \lambda.min$ has the lowest AIC, and it can be a good candidate to analyze this data set. Another thing that needs to be considered in model selection is model parsimony. Compared to the Firth's method, Firth's method after doing variable selection by elastic net.min has a lower AIC, and it has one dummy variable less than the Firth's method. Firth's method after elastic net variable selection also can address the problem of separation with WCB data. The AUC is also high for this method compared to other methods. All of these show that this model performs better for WCB claims data set. Therefore, all interpretation for the purpose of WCB applications will be done using the Firth's method after elastic net variable selection.

It is good to mention that, the performance of Firth' logistic, Firth' logistic after lasso.min, and Firth's logistic after elastic net.min are very similar and they are just slightly different, but for the case of this study based on our model performance criteria, we decided to choose Firth's method after elastic net.min as the best method to interpret the results based on. In the current study, bias-correction or Firth's method is doing most of the work, and the positive effect of variable selection methods is only marginal, but for other cases of the data the other two methods might perform better. With respect to lasso logistic regression, although this method could address the problem of separation, we did not present the final results based on because its AIC is higher than other methods.

Table 4.5: Method comparison scores for the logistic, Firth's, and Firth's after lasso and elastic net variable selection methods. The bolded numbers in the table indicates the model with the optimized performance metric

Method	AIC	AUC	number of dummy variables	-logL
logistic	1766.72	0.98	37	846.36
Lasso.1se logistic	1807.64	0.97	20	883.82
Lasso.min logistic	1762.78	0.98	35	846.39
Elastic net.1se logistic	1767.22	0.98	27	856.61
Elastic net.min logistic	1764.73	0.98	36	846.36
Firth's	1703.30	0.98	37	814.65
Firth's after lasso.1se	1761.77	0.97	20	860.88
Firth's after lasso.min	1702.67	0.97	35	816.33
Firth's after elastic net.1se	1713.11	0.97	27	829.55
Firth's after elastic net.min	**1701.96**	0.975	36	814.98

In addition to the method performance comparison criteria presented in Table 4.5, the width of confidence intervals was also considered to compare these methods. The tables of estimated OR and 95% CIs are provided for all included methods that we used to analyze the WCB Saskatchewan data. Table 4.8 shows the OR and 95% CIs for personal characteristics for conventional logistic, Firth's logistic, and Firth's after lasso and elastic net variable selection with different tuning parameters. Table 4.9 represents the OR and 95% CIs for incident characteristics as calculated by conventional logistic, Firth's, and Firth's after lasso and elastic net variable selection for two different tuning parameters.

As shown in odds ratio tables, 95% confidence intervals are wide for some of the levels in some characteristic using conventional logistic regression, although using the Firth's method gives us shorter CIs compared to conventional logistic regression in addition to address the separation problem. For example, in Table 4.9, 95% confidence interval for point estimate of machinery in source of injury based on conventional logistic regression is (11.92,1406.23) while this CI based on Firth's method is (10.31,515.65). For furniture and fixture in source of injury, the 95% CI is (0.83,603.30) based on logistic regression although the 95% CI from Firth's method is (1.67,300.65).

Based on our literature review, using conventional logistic regression is not ideal in the presence of quasi-complete separation as it gives very wide CIs, which is consistent with the result from our analysis, but it seems that the Firth's method could solve this problem. It is also good to mention that as the results of conventional logistic regression are based on the last maximum likelihood iteration, validity of the model fit is questionable for this method; note that we presented the results of conventional logistic regression only to make a comparison with other methods.

Based on the results from OR tables 4.8 and 4.9, the length of 95% CI is shorter for Firth's method after using lasso and elastic net variable selection methods especially for $\lambda = \lambda.1se$ as this λ prevents overfitting. As mentioned earlier compared to Firth's logistic regression, AIC for Firth's after elastic net.min is lower.

The plots of odds ratio (OR) and 95% CIs for OR are provided in Figure 4.4 and Figure 4.5 for personal and incident characteristics for conventional logistic, Firth's logistic, and Firth's logistic after elastic net.min methods. Red dots are ORs, and the green lines are 95% CIs for the ORs. Blue lines show very wide CIs for ORs, and $\circ$ shows tiny CIs in these two OR plots. These tables show that the length of 95% CIs for Firth's method after variable selection (elastic net.1se) is shorter.

Table 4.6: Odds Ratios and 95% Confidence Intervals for personal characteristics from conventional logistic regression, logistic regression after using lasso and elastic net as variable selection methods for the relation between the covariates and a claim being fatal, Saskatchewan WCB, 2007-2016

Characteristics[2]	**Method**				
	logistic (95% CI)	LA.lse (95% CI)[3]	LA.min (95% CI) [4]	EN.1se (95% CI) [5]	EN.min (95% CI) [6]
Gender					
women					
men	5.82 (2.88,13.5)	7.2 (4.4,11.63)	5.8 (2.88,13.5)	6.1 (3,13.98)	5.82 (2.88,13.5)
Age					
14 to 24					
25 to 34	1.06 (0.62,1.86)	-	-	-	1.064 (0.62,1.867)
35 to 44	0.48 (0.25,0.93)	0.49 (0.27,0.84)	0.47 (0.26,0.80)	0.48 (0.261,0.82)	0.484 (0.25,0.93)
45 to 54	1.52 (0.90,2.62)	1.46 (0.97,2.18)	1.46 (0.96,2.2)	1.45 (0.96,2.18)	1.52 (0.90,2.62)
55 to 64	1.78 (1.02,3.17)	1.77 (1.12,2.75)	1.7 (1.1,2.70)	1.74 (1.1,2.72)	1.79 (1.020,3.17)
65 to 85	6.70 (3.62,12.49)	7.2 (4.4,11.6)	6.5 (3.8,10.67)	6.71 (4,11.01)	6.70 (3.63,12.48)

[2] The first category for each characteristic is the reference category for logistic method, but the reference category for other methods is different and consists of the first category of each variable (reference category in logistic) plus those dummy variable kicked out from the model shown by $-$ in each method in the table

[3] Logistic regression after lasso variable selection when $\lambda = \lambda.1se$

[4] Logistic regression after lasso variable selection when $\lambda = \lambda.min$

[5] Logistic regression after elastic net variable selection when $\lambda = \lambda.1se$

[6] Logistic regression after elastic net variable selection when $\lambda = \lambda.min$

Occupations [7]					
social sciences					
business	3.01 (0.93,11.62)	2.81 (1.25,5.67)	3 (0.92,11.6)	2.49 (1.1,5.2)	3.01 (0.93,11.62)
health	0.45 (0.02,3.15)	-	0.45 (0.023,3.14)	0.33 (0.02,1.59)	0.45 (0.02,3.14)
applied sciences	0.29 (0.01,2.08)	-	0.29 (0.015,2.07)	0.22 (0.01,1.13)	0.29 (0.02,2.08)
primary industry	3.14 (1.14,11.13)	2.91 (1.70,4.85)	3.13 (1.13,11.1)	2.53 (1.43,4.4)	3.14 (1.14,11.13)
art/culture	0.00* [8] (0.00*, 0.00*)	1 (0.01,9.74)	0.00* (0.00*,0.00*)	-	0.00* (0.00*,0.00*)
sale/services	0.90 (0.3,3.28)	-	0.90 (0.3,3.28)	1.74 (1.1,2.7)	0.90 (0.30,3.28)
trade/transport	1.68 (0.67,5.65)	1.59 (1.11,2.31)	1.67 (0.67,5.63)	1.37 (0.92,2.08)	1.68 (0.67,5.65)
manufacturing	1.59 (0.49,6.11)	-	1.58 (0.49,6.1)	-	1.59 (0.49.6.11)
not stated	1.20 (0.44,4.23)	-	1.2 (0.44,4.2)	-	1.20 (0.44,4.23)

[7] Reference category for logistic is the first category in the table (social sciences), and for other methods the reference consists of social sciences in addition to those dummy variables kicked out from the model shown by $-$ in the table

[8] Reports very low numbers less than $< \times 10^{-3}$

Table 4.7: Odds Ratios and 95% Confidence Intervals for incident characteristics from conventional logistic regression, logistic regression after using lasso and elastic net as variable selection methods for the relation between the covariates and a claim being fatal, Saskatchewan WCB, 2007-2016

Characteristics[9]	**Method**				
	logistic (95% CI)	LA.lse (95% CI)[10]	LA.min (95% CI) [11]	EN.1se (95% CI) [12]	EN.min (95% CI) [13]
Source of injury					
chemical products					
furniture/fixture	22.36 (0.83,603.30)	-	22.27 (0.83,596.56)	-	22.35 (0.83,598.40)
parts/materials	37.60 (6.80,703.75)	-	37.58 (6.86,702)	3.41 (1.70,6.44)	37.59 (6.86,701.8)
structure/surfaces	29.92 (4.46,605.74)	-	29.91 (5,580.47)	-	29.89 (5.01,580.16)
vehicles	43.77 (7.01,857.84)	3.87 (1.74,8)	43.78 (7.2,851.6)	3.39 (1.5,7.2)	43.75 (7.16,851.01)
containers	19.21 (1.67,439.65)	-	19.23 (1.7,437.3)	-	19.21 (1.68,436.68)
machinery	72.57 (11.92,1406.23)	6.3 (2.9,12.37)	72.67 (12,1399.48)	5.9 (2.64,12.19)	72.55 (12.13,1397.03)
persons/plants	22.19 (4.66,397.79)	-	22.24 (4.7,398.5)	2.60 (1.50,4.6)	22.19 (4.66,397.66)
tools/equipments	0.00* [14] (0.00*,0.00*)	-	0.00* (0.00*,0.00*)	0.00* (0.00*,0.00*)	0.00* (0.00*,0.00*)
other sources	9.29 (1.74,171.80)	-	9.3 (1.7,171.7)	-	9.29 (1.75,171.76)

[9] The first category for each characteristic is the reference category for logistic method, but the reference category for other methods is different and consists of the first category of each variable (reference category in logistic) plus those dummy variable kicked out from the model shown by $-$ in each method in the table

[10] Logistic regression after lasso variable selection when $\lambda = \lambda.1se$

[11] Logistic regression after lasso variable selection when $\lambda = \lambda.min$

[12] Logistic regression after elastic net variable selection when $\lambda = \lambda.1se$

[13] Logistic regression after elastic net variable selection when $\lambda = \lambda.min$

[14] Reports very low numbers less than $< \times 10^{-3}$

Part of body					
other					
body systems	0.01 (0.00*,0.01)	0.01 (0.007,0.03)	0.06 (0.002,0.013)	0.006 (0.003,0.012)	0.006 (0.002,0.01)
lower extremities	0.06 (0.04,0.10)	0.14 (0.09,0.20)	0.061 (0.04,0.098)	0.067 (0.04,0.12)	0.060 (0.04,0.098)
upper extremities	0.00* (0.00*,0.00*)	0.001(0.002,0.012)	0.00* (0.00*,0.001)	0.00* (0.00*,0.00*)	0.00* (0.00*,0.00*)
multiple parts	0.00* (0.00*,0.02)	0.01 (0.00,0.04)	0.005 (0.001,0.016)	0.005 (0.001,0.018)	0.005 (0.001,0.02)
head	0.00* (0.00*,0.00*)	0.00* (0.00*,0.00*)	0.00* (0.00,0.01)	0.00* (0.00*,0.00*)	0.00* (0.00*,0.00*)
trunk	0.00* (0.00*,0.01)	0.006 (0.002,0.012)	0.02 (0.001,0.006)	0.003 (0.001,0.006)	0.002 (0.001,0.006)
Cause of injury					
violent acts					
bodily reaction	0.13 (0.02,0.85)	0.16 (0.03,0.55)	0.134 (0.02,0.54)	0.068 (0.011,0.23)	0.134 (0.02,0.54)
transportation	5.25 (1.35,27.32)	3.60 (1.76,7.81)	5.25 (2.14,14.1)	3.10 (1.55,6.69)	5.26 (2.14,14.09)
falls	1.00 (0.23,5.40)	-	-	-	-
contact (objects)	1.77 (0.52,8.36)	-	1.78 (0.8,3.98)	-	1.78 (0.8,3.98)
other events	7.17 (2.12,33.68)	6.67 (3.87,11)	7.17 (2.73,19.86)	3.33 (1.723,6.278)	7.18 (2.736,19.87)
harmful substances	0.43 (0.15,1.86)	-	0.43 (0.18,1.1)	0.162 (0.086,0.30)	0.43 (0.18,1.1)

Table 4.8: Odds Ratios and 95% Confidence Intervals for personal characteristics from conventional logistic regression, Firth's logistic regression, and Firth's after using lasso and elastic net as variable selection methods for the relation between the covariates and a claim being fatal, Saskatchewan WCB, 2007-2016

Characteristics[15]	Method					
		Firth's Method				
	logistic (95% CI)	Firth's (95% CI)	LA.lse (95% CI)[16]	LA.min (95% CI) [17]	EN.1se (95% CI) [18]	EN.min (95% CI) [19]
Gender						
women						
men	5.82 (2.88,13.5)	5.49 (2.76,12.45)	7.68 (3.96,16.99)	5.5 (2.77,12.47)	5.73 (2.91,12.91)	5.50 (2.77,12.46)
Age						
14 to 24						
25 to 34	1.06 (0.62,1.86)	1.06 (0.62,1.84)	-	-	-	1.06 (0.62,1.84)
35 to 44	0.48 (0.25,0.93)	0.49 (0.25,0.93)	0.50 (0.28,0.85)	0.48 (0.27,0.82)	0.49 (0.27,0.83)	0.49 (0.25,0.93)
45 to 54	1.52 (0.90,2.62)	1.5 (0.90,2.58)	1.46 (0.97,2.18)	1.46 (0.97,2.20)	1.45 (0.96,2.18)	1.50 (0.90,2.58)
55 to 64	1.78 (1.02,3.17)	1.77 (1.02,3.13)	1.78 (1.14,2.75)	1.73 (1.09,2.69)	1.75 (1.11,2.72)	1.77 (1.02,3.12)
65 to 85	6.70 (3.62,12.49)	6.6 (3.6,12.23)	7.23 (4.41,11.61)	6.42 (3.83,10.58)	6.7 (4.01,10.94)	6.59 (3.59,12.20)

[15] The first category for each characteristic is the reference category for logistic and Firth method, but the reference category for other methods is different and consists of the first category of each variable (reference category in logistic and Firth's) plus those dummy variable kicked out from the model shown by $-$ in each method in the table

[16] Firth's method after lasso variable selection when $\lambda = \lambda.1se$

[17] Firth's method after lasso variable selection when $\lambda = \lambda.min$

[18] Firth's method after elastic net variable selection when $\lambda = \lambda.1se$

[19] Firth's method after elastic net variable selection when $\lambda = \lambda.min$

Occupations [20]						
social sciences						
business	3.01 (0.93,11.62)	2.83 (0.91,10.11)	2.93 (1.33,5.82)	2.81 (0.91,10.07)	2.57 (1.13,5.33)	2.82 (0.91,10.11)
health	0.45 (0.02,3.15)	0.6 (0.06,3.34)	-	0.60 (0.06,3.35)	0.49 (0.05,1.90)	0.60 (0.06,3.35)
applied sciences	0.29 (0.01,2.08)	0.38 (0.04,2.17)	-	0.38 (0.04,2.17)	0.32 (0.04,1.33)	0.38 (0.04,2.17)
primary industry	3.14 (1.14,11.13)	2.85 (1.07,9.39)	2.92 (1.72,4.86)	2.84 (1.07,9.36)	2.53 (1.44,4.38)	2.85 (1.07,9.39)
art/culture	0.00* [21] (0.00*, 0.00*)	1 (0.01,9.74)	-	1 (0.01,9.72)	-	1.00 (0.01,9.74)
sale/services	0.90 (0.3,3.28)	0.83 (0.29,2.83)	-	0.83 (0.29,2.83)	0.75 (0.37,1.42)	0.84 (0.29,9.74)
trade/transport	1.68 (0.67,5.65)	1.51 (0.63,4.7)	1.58 (1.1,2.3)	1.5 (0.62,4.69)	1.36 (0.91,2.06)	1.51 (0.63,4.70)
manufacturing	1.59 (0.49,6.11)	1.49 (0.48,5.32)	-	1.48 (0.48,5.29)	-	1.49 (0.48,5.32)
not stated	1.20 (0.44,4.23)	1.10 (0.42,3.59)	-	1.10 (0.42,3.58)	-	1.10 (0.42,3.59)

[20] Reference category for logistic and Firth is the first category in the table (social sciences), and for other methods the reference consists of social sciences in addition to those dummy variables kicked out from the model shown by $-$ in the table

[21] Reports very low numbers less than $< \times 10^{-3}$

Table 4.9: Odds Ratios and 95% Confidence Intervals for incident characteristics from conventional logistic regression, Firth's logistic regression, and Firth's after using lasso and elastic net as variable selection methods for the relation between the covariates and a claim being fatal, Saskatchewan WCB, 2007-2016

Characteristics	Method					
		Firth's Method				
	logistic (95% CI)	Firth's (95% CI)	LA.lse (95% CI)	LA.min (95% CI)	EN.1se (95% CI)	EN.min (95% CI)
Source of injury						
chemical products						
furniture/fixture	22.36 (0.83,603.30)	22.40 (1.67,300.65)	-	22.02 (1.65,292.77)	-	22.07 (1.66,293.19)
parts/materials	37.60 (6.80,703.75)	26.19 (5.77,249.66)	-	25.96 (5.78,246.75)	3.47 (1.75,6.50)	25.96 (5.78,246.66)
structure/surfaces	29.92 (4.46,605.74)	20.26 (3.60,216.45)	-	19.82 (4.01,199.20)	-	19.79 (4.01,198.95)
vehicles	43.77 (7.01,857.84)	30.57 (5.91,310.47)	3.98 (1.80,8.17)	30.21 (5.97,303.75)	3.45 (1.54,7.27)	30.16 (5.96,303.21)
containers	19.21 (1.67,439.65)	15.96 (1.90,189.55)	-	15.79 (1.90,185.93)	-	15.76 (1.89,185.56)
machinery	72.57 (11.92,1406.23)	51.68 (10.31,515.65)	6.59 (3.11,12.81)	51.20 (10.40,506.78)	6.14 (2.80,12.51)	51.06 (10.38,505.38)
persons/plants	22.19 (4.66,397.79)	15.02 (3.89,134.97)	-	15.10 (3.91,135.61)	2.60 (1.46,4.58)	15.06 (3.90,135.28)
tools/equipments	0.00*[22] (0.00*,0.00*)	1.90 (0.01,38.56)	-	1.88 (0.01,37.91)	0.25 (0.00,1.83)	1.88 (0.01,37.93)
other sources	9.29 (1.74,171.80)	6.51 (1.49,61.09)	-	6.49 (1.49,60.87)	-	6.49 (1.49,60.88)
Part of body						
neck (throat)						
body systems	0.01 (0.00*,0.01)	0.01 (0.00,0.01)	0.02 (0.01,0.03)	0.01 (0.00,0.01)	0.01 (0.00,0.01)	0.01 (0.0026,0.01)
lower extremities	0.06 (0.04,0.10)	0.06 (0.04,0.10)	0.14 (0.09,0.21)	0.06 (0.04,0.10)	0.07 (0.04,0.11)	0.06 (0.04,0.1)

[22] Reports very low numbers less than $< \times 10^{-3}$

upper extremities	0.00* (0.00*,0.00*)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00* (0.00*,0.00*)
multiple parts	0.00* (0.00*,0.02)	0.01 (0.00,0.02)	0.01 (0.00,0.04)	0.01 (0.00,0.02)	0.01 (0.00,0.02)	0.01 (0.00,0.02)
head	0.00* (0.00*,0.00*)	0.00* (0.00*,0.00*)	0.00 (0.00,0.01)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.006 (0.00*,0.00*)
trunk	0.00* (0.00*,0.01)	0.00 (0.00,0.01)	0.01 (0.00,0.01)	0.00 (0.00,0.01)	0.00 (0.00,0.01)	0.0027 (0.00*,0.01)
Cause of injury						
violent acts						
bodily reaction	0.13 (0.02,0.85)	0.14 (0.02,0.78)	0.20 (0.04,0.61)	0.16 (0.03,0.60)	0.08 (0.02,0.26)	0.16 (0.03,0.60)
transportation	5.25 (1.35,27.32)	4.30 (1.17,20.21)	3.44 (1.71,7.44)	4.89 (2.01,13.06)	2.98 (1.50,6.37)	4.90 (2.01,13.08)
falls	1.00 (0.23,5.40)	0.87 (0.21,4.29)	-	-	-	-
contact (objects)	1.77 (0.52,8.36)	1.54 (0.47,6.50)	-	1.74 (0.79,3.91)	-	1.75 (0.79,3.92)
other events	7.17 (2.12,33.68)	6.03 (1.89,25.30)	6.74 (3.94,11.07)	6.79 (2.61,18.70)	3.34 (1.74,6.27)	6.79 (2.62,18.71)
harmful substances	0.43 (0.15,1.86)	0.38 (0.14,1.45)	-	0.43 (0.18,1.08)	0.16 (0.09,0.31)	0.43 (0.18,1.08)

Reflecting on all of these findings, the best performing model is Firth's after doing elastic net variable selection with the choice of $\lambda = \lambda.min$, and in order to answer the application question stated in the second objective of the study, the next section will focus solely on the results of the this method.

4.2.4 Estimating Risk of a Claim Being Fatal (Results)

In this section, we discuss variables that are associated with higher risk of a claim being fatal. Based on what we discussed in Section 4.2.3, we interpret the results of the Firth's method after elastic net.min to see what variables increase the risk of a claim being fatal. Tables 4.10 and 4.11 show the results of the analysis based on this method.

Interpretation

The results of the multivariate analysis for personal characteristics (Table 4.10) indicate that the risk of a claim being fatal for the seniors aged 65-85 years of age is 6.59 (95% CI: 3.59-12.20) times higher as compared with those who are 14-24. Similarly, the odds of a claim being fatal for those aged 55-64 years of age is 1.77 (95% CI: 1.02-3.12) times higher as compared with those who are aged 14-24. The odds of a claim being fatal for those aged 35-44 years of age is 0.49 (95% CI: 0.25-0.93) less than those who are aged 14-24. Comparing workers 14 to 24 years old and workers aged 45 to 54 reveals no significant difference in claims being fatal. In addition, odds of a claim being fatal among men is 5.5 (95% CI: 2.77-12.46) times higher than women. Odds of a claim being fatal among those who work in primary industry vs. those who work in social sciences is 2.85 (95% CI: 1.07-9.39). Comparing other occupations and occupations in social sciences does not show any significant differences.

The results of the multivariate analysis for incident characteristics (Table 4.11) indicates that odds of a claim being fatal for machinery in source of injury is 51 (95% CI: 10.38-505.38) times higher than odds of a claim being fatal in chemical products. For part of body, the only significant OR is related to lower extremities, and as other levels of this variable have very tiny CIs and are very close to zero, we did not consider them statistically significant relevant to the application. For cause of injury, the odds of claims being fatal for other events and exposures is 6.79 (95% CI: 2.62-18.71) times higher as compared with 'reference category'

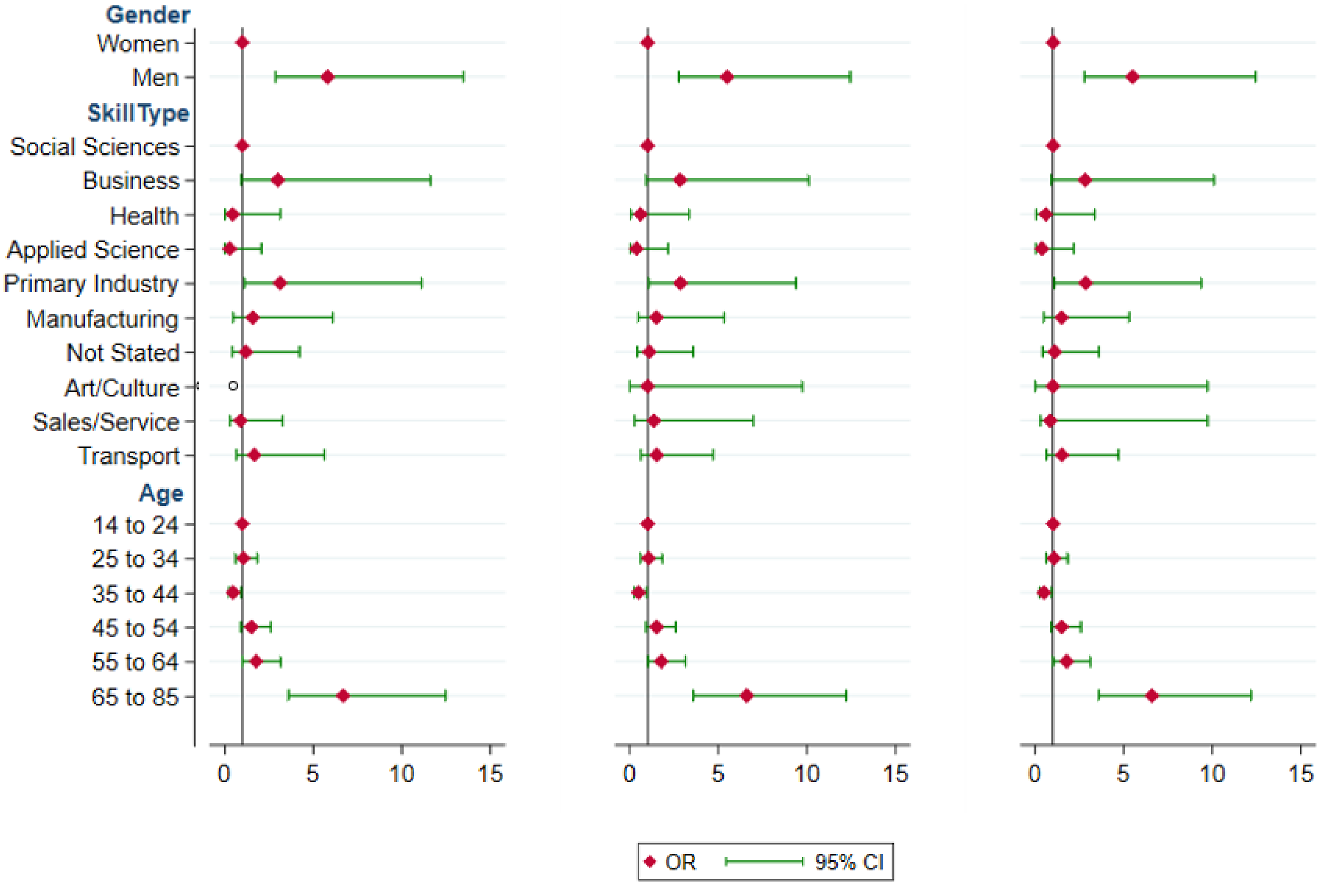

Figure 4.4: Odds Ratios and 95% Confidence Intervals for personal characteristics from conventional logistic regression, Firth's logistic regression, and Firth's logistic after using elastic net.min as variable selection methods from left to right respectively for the relation between the covariates and a claim being fatal, WCB Saskatchewan, 2007-2016

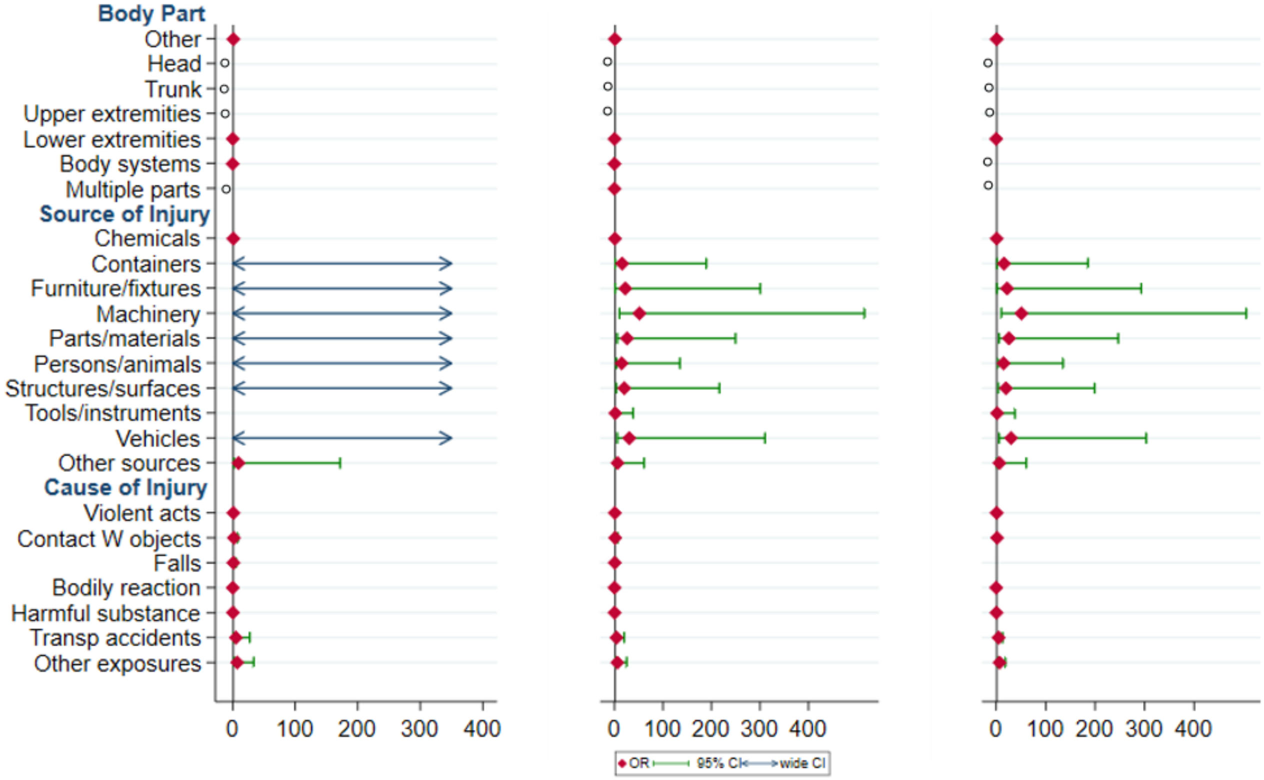

Figure 4.5: Odds Ratios and 95% Confidence Intervals for incident characteristics from conventional logistic regression, Firth's logistic regression, and Firth's logistic after using elastic net.min as variable selection methods from left to right respectively for the relation between the covariates and a claim being fatal, Saskatchewan WCB, 2007-2016

including fires and explosions, violent acts and falls. Moreover, the odds of claims being fatal for cause of injury in transportation vs. ‘reference category’ (fires and explosions ; violent acts; and falls) is 4.90 (95% CI: 2.01-13.08) times higher. The odds of a claim being fatal for bodily reaction is 0.16 times less likely than for ‘reference category’ (OR: 0.16, 95% CI: 0.03-0.6). Comparing contact with harmful substances and objects with ‘reference category’ shows no significant difference in odds of a claim being fatal.

Results of the previous section shows that men, occupations in primary industry, ‘machinery’ source of injury, and lower extremities are significant predictors of injury claims being fatal in Saskatchewan.

Table 4.10: The multivariate analysis reporting the estimated odds ratios (OR), the corresponding 95% confidence interval (CI) and p-values for all the potential personal risk factors in the WCB fatality data analysis using Firth's after elastic net.min variable selection method

Characteristic	**OR**	**95% CI**	**P-value**
Gender			
men vs women	5.50	(2.77,12.46)	<.0001
Age			
25-34 vs 14-24	1.06	(0.62,1.84)	0.8
35-44 vs 14-24	0.49	(0.25,0.93)	0.03
45-54 vs 14-24	1.50	(0.90,2.58)	0.1
55-64 vs 14-24	1.77	(1.02,3.12)	0.04
65-85 vs 14-24	6.59	(3.59,12.20)	<.0001
Occupation			
business/advertising vs social sciences	2.82	(0.91,10.11)	0.07
health vs social sciences	0.60	(0.06,3.35)	0.58
natural/applied sciences vs social sciences	0.38	(0.04,2.17)	0.29
primary indusrty vs social sciences	2.85	(1.07,9.39)	0.03
art and culture vs social sciences	1	(0.01,9.74)	0.99
sale and services vs social sciences	0.84	(0.29,2.84)	0.75
trade and transport vs social sciences	1.51	(0.63,4.70)	0.38
processing/manufacturing vs social sciences	1.49	(0.48,5.32)	0.49
not stated vs social sciences	1.10	(0.42,3.59)	0.86

Table 4.11: The multivariate analysis reporting the estimated odds ratios (OR), the corresponding 95% confidence interval (CI) and p-values for all the potential incident risk factors in the WCB fatality data analysis using Firth's method after elastic net.min

Characteristic	**OR**	**95% CI**	**P-value**
Source of injury			
furniture/fixtures vs chemical products	22.07	(1.66,293.19)	0.02
parts/materials vs chemical products	25.96	(5.78,246.66)	<.0001
structure/surfaces vs chemical products	19.79	(4.01,198.95)	<.0001
vehicles vs chemical products	30.16	(5.96,303.21)	<.0001
containers vs chemical products	15.76	(1.89,185.56)	0.01
machinery vs chemical products	51.06	(10.38,505.38)	<.0001
persons, plants/animals vs chemical products	15.06	(3.90,135.28)	<.0001
tools/equipments vs chemical products	1.88	(0.01,37.93)	0.7
other sources vs chemical products	6.49	(1.49,60.88)	<.0001
Part of body			
body systems vs 'other body part'	0.01	(0.0026,0.01)	<.0001
lower extremities vs 'other body part'	0.06	(0.04,1)	<.0001
upper extremities vs 'other body part'	0.00033	(0.00004,0.001)	<.0001
multiple body part vs 'other body part'	0.01	(0.001,0.02)	<.0001
head vs neck and 'other body part'	0.0002	(0.000002,0.001)	<.0001
trunk vs 'other body part'	0.0027	(0.001,0.01)	<.0001
Cause of injury			
bodily reaction/exertion vs 'ref category' [a]	0.16	(0.03,0.60)	0.005
transportation accidents vs 'ref category'	4.90	(2.01,13.08)	<.0001
contact with objects vs 'ref category'	1.75	(0.79,3.92)	0.17
harmful substances vs 'ref category'	0.43	(0.18,1.08)	0.07
other events/exposures vs 'ref category'	6.79	(2.62,18.71)	<.0001

[a] fires and explosions, violent acts, and falls

Chapter 5
Final Remarks

5.1 Summary

Motivated by the statistical challenges encountered in modelling rare event data in a real application with Saskatchewan Workers' Compensation Board Data, this thesis went beyond basic descriptive analysis by applying penalized logistic regression and different variable selection methods to identify the characteristics of the vulnerable population with a high risk of a claim being fatal. We used administrative WCB-SK claims data, which is representative of all fatal claims obtained from population-based data at the individual level, along with workers and incident characteristics. To our knowledge, no studies have applied penalized regression methods in the context of occupational injury data in order to address the analytic challenges except one study that used logistic regression with Firth's approach in identifying facors associated with fatal accidents among Mexican workers [19].

The analytic challenges are mostly due to the strong imbalance of the outcome variable as well as the categorical covariates. The outcome of interest in our study, i.e., fatal injury claim was very rare (177 out of 280,704) and about 40 regression coefficients (multiple covariates with many levels) were estimated, which resulted in low EPV, i.e. $177/40 \approx 4.4$. In many epidemiological and medical studies, an EPV of ≥ 10 is widely used as a rule-of-thumb to determine the reliability of the statistical analysis. Variable selection is often used as a strategy to reduce the number of variables in the model to overcome the problem of low EPV. In addition to the issue of low EPV, several categorical explanatory variables in our analysis, such as source of injury or part of the body have many levels and are highly imbalanced. The low EPV in combinations with highly imbalanced multi-categorical covariates caused the quasi-complete separation problem. Under the combinations of these problems, the maximum likelihood estimation under the conventional logistic regression model failed to converge and yielded very unstable parameter estimates. Firth's logistic regression is a standard tool for solving the problem of quasi-complete separation. However, Firth's method does not perform model selection and there has been very limited research investigated if model selection methods can fully circumvent the quasi-complete separation problem.

Therefore, various model selection methods were applied, such as traditional backward model selection method, lasso and elastic net penalized regression methods to examine if quasi-complete separation problem can be resolved at increased EPV. Unsurprisingly, model selection methods reduced the number of parameters in the model and therefore increased EPV; however, quasi-complete separation problem still exists especially in using elastic net logistic regression method. As a result, Firth's method was used after the model selection methods as a bias-correction method. Our results showed that Firth's method after model selection based on elastic net with $\lambda = \lambda.min$ outperformed other methods, which gave lower AIC, higher AUC, and shorter CIs. Previous studies showed that elastic net can outperform lasso while encouraging a grouping effect and enjoying the same sparsity [16]. In the presence of highly correlated variables, empirical studies have shown that elastic net outperforms lasso [16], which was the case in the current study using WCB-SK data.

This is not to say that the Firth's method after variable selection by elastic net.min should be preferred over all other methods in all scenarios. Indeed, when analyzing rare event data with many categorical covariates, which may lead to separation problem and in presence of multicollinearity, additional care needs to be given to choosing the best method that fits the data well. Depending on the nature of data and the objectives of the study, other methods can be preferable.

Results of the Section 4.2.4 showed that men, 'primary industry' occupation, 'machinery' source of injury, 'other events/exposures' and 'transportation' cause of injury are significant covariates in higher odds of a claim being fatal for workers in Saskatchewan. These results more or less confirm the findings of other researchers for analysis of occupational claims data.

With respect to age, the relationship between age and the risk of a claim being fatal is not simply linear, and the middle-aged workers are at a lower risk. The young workers are at a higher risk than middle-aged workers, and the risk of having fatal injury increased sharply as age increased from 45 to 85. Our analysis showed that age group 65-85 years old is most prone to having a claim being fatal compared to 14-24. The majority of studies that examined serious occupational injuries showed that older age workers suffer from a higher number of severe or fatal injuries in comparison with younger workers [26, 80–86]. In addition, previous studies using different occupational data have reported different age patterns for different injury types [87]. For example, using workers' compensation claims data from Ontario, Canada, Choi et al. [88] reported that workers aged 30-59 were more likely to have strain and sprain occupational injuries. However, in the current study, we did not have access to such a variable to investigate this relationship.

Our results showed that men have higher odds of a claim being fatal compared to women. With regards to the effect of gender on making fatal claim injuries, our study revealed higher odds of fatal claims among men. Similar findings were reported in other studies, for instance, Fan et al. [24] reported lower overall serious injury rate for women compared to men in British Columbia. However, for some studies, the rate of fracture (injury type) was similar across age groups for men but increased with age for women [24]. Lots of studies have shown an increased risk of fatal accidents related to gender [89, 90], some of which show a higher risk for men, some higher risk to women. For example, Ward et al. showed that from 1990 to 1996, there were 11 times as many agriculture-related fatalities for men compared to women [91]. Although these studies give insight into fatal accidents, we could not find any reports specifically investigating the outcome of a claim being fatal.

An interesting result from our analysis is the highest odds of a claim being fatal are for occupations in primary industry (such as mining, oil and gas drilling and service, fishing vessel deckhands, etc.) and sales and services. No studies were found which compare WCB claims resulting from primary industry and sale and services industry. However, there is ample evidence that injury risk is higher when there is a risk of falling [92], or working with fire and explosive materials [93, 94].

With respect to the part of body, as shown in Table 4.11, the length of CIs are very short and the CIs for different levels of part of body variable are very close to zero. Therefore, the only level of this variable that we can have a conclusion for is lower extremities. The odds of a claim being fatal for lower extremities is only 6% that of the reference category (other body part). The results of other levels of part of body variable were not considered statistically significant as the CIs were very short and close to zero. However, part of body can provide useful information about the nature of the incident and the mechanism of a fatality. For example, in the current study, the analysis of occupational injuries were considered and we did not take occupational illnesses into consideration. As most occupational diseases are related to lung illnesses and lung cancers, the results derived from part of body after adding those workplace injuries to the data set would probably be different.

5.2 Limitations and Future Work

Although our data captured the majority of Saskatchewan's employed workforce, some workers such as self-employed workers and farmers are excluded from the data, which may bias the results. It is also possible that workplace injuries are under-reported for compensation, but this study likely included most of them as fatalities are most likely to be reported [95]. Therefore, locally representative survey sample could be collected in the future to have a more representative sample of the study population. Motor-vehicle collisions are particularly hazardous in Saskatchewan [96]. To learn more about the nature of industries in Saskatchewan and for prevention efforts, WCB-SK data could be possibly linked with other data sources such as Saskatchewan Government Insurance (SGI) data and coroner death data for incorporating more and different kinds of valuable information, such as environmental factors for identifying risk factors associated with occupational traffic crashes more accurately [97]. As a potential next step to this study, several characteristics such as length of employment for workers can be collected by WCB-SK to investigate whether there is any statistically significant relationship between this covariate and occupational fatalities or not. Some of other variables which might be useful to collect include where occupational injury fatality happened and what the weather was like (especially for traffic events, which form a large proportion of the fatalities.). As most studies in different countries used fatality rate and fatality risk to analyze occupational claims data, using a slightly different outcome in the current study makes it difficult to compare the results of our analysis with those from other studies. In future work, fatality rate can be calculated to get the results and compare them with other studies.

Other regression shrinkage and variable selection methods can be considered (e.g., SCAD [65], adaptive lasso [67]) to analyze WCB-SK data. Simulation studies will be conducted in the future to compare various penalized likelihood methods to provide recommendations of optimal strategies in conditions of rare events data with low EPV and quasi-complete separation problems. In this thesis, we primarily focused on binary logistic regression (fatal. vs. non-fatal injuries). Injury severity level (fatal vs. serious vs. non-serious injuries) as a three-level outcome variable may be also of interest to be modelled. Ordinal regression is often used for modelling outcome variable with 'ordered' multiple levels. The analytic challenges such as quasi-complete separation and low EPV can also occur with such a discrete outcome. Penalized regression methods can be applied to investigate if combinations of strategies, such as Firth's penalization after lasso or elastic net variable selections can yield better model performances.

The present study found statistically significant relationships between personal characteristics such as gender, age, and occupation, and some incident characteristics and the possibility of death in case of occupational injury. The findings from our study enable us to identify the most vulnerable groups who are at higher risk of a claim being fatal. Based on the results of the current analysis, we propose some strategies for WCB-SK to prevent occupational injuries. Improving occupational injury prevention programs by monitoring and promoting safe work area, implementing more rigorous legal control measures, and improving enforcement activities such as focused inspection and training could be useful interventions to consider and evaluate.

Bibliography

[1] M Shafiqur Rahman and Mahbuba Sultana. Performance of firth-and logf-type penalized methods in risk prediction for small or sparse binary data. *BMC medical research methodology*, 17(1):33, 2017.

[2] Maryam Sadat Hajakbari and Behrouz Minaei-Bidgoli. A new scoring system for assessing the risk of occupational accidents: A case study using data mining techniques with iran's ministry of labor data. *Journal of Loss Prevention in the Process Industries*, 32:443–453, 2014.

[3] Jaclyn Gilks and Ron Logan. Occupational injuries and diseases in canada, 1996–2008: injury rates and cost to the economy. *Ottawa, Canada: Human Resources and Skills Development Canada*, 2010.

[4] Canadian Centre for Occupational Health and Safety. *Beyond the Statistics*, 2018. https://www.ccohs.ca/events/mourning/ [Accessed: August 2019].

[5] Sean Tucker. *2019 Report on Work Fatality and Injury Rates in Canada*, 2019. https://www.uregina.ca/business/faculty-staff/faculty/file_download/2019-Report-on-Workplace-Fatalities-and-Injuries.pdf [Accessed: August 2019].

[6] Saskatchewan WCB. *2018 Annual Report Saskatchewan Workers' Compensation Board*, 2018. https://www.wcbsask.com/wp-content/uploads/2019/04/2018-Annual-Report.pdf [Accessed: August 2019].

[7] Paul D Allison. Convergence failures in logistic regression. In *SAS Global Forum*, volume 360, pages 1–11, 2008.

[8] Paul Allison. Convergence problems in logistic regression. *Numerical issues in statistical computing for the social scientist*, pages 238–252, 2003.

[9] Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16):2409–2419, 2002.

[10] Adelin Albert and John A Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

[11] Robert L Schaefer. Bias correction in maximum likelihood logistic regression. *Statistics in Medicine*, 2(1):71–78, 1983.

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

[13] Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[14] MA Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203, 1960.

[15] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[16] 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.

[17] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[18] Michele Usuelli. *R Machine Learning Essentials*. Packt Publishing Ltd, 2014.

[19] Mery Gonzalez-Delgado, Héctor Gómez-Dantés, Julián Alfredo Fernández-Niño, Eduardo Robles, Víctor H Borja, and Miriam Aguilar. Factors associated with fatal

occupational accidents among mexican workers: a national analysis. *PloS one*, 10(3): e0121490, 2015.

[20] Alexander-Stamatios G Antoniou. *Handbook of managerial behavior and occupational health*. Edward Elgar Publishing, 2009.

[21] Saskatchewan Workers' Compensation Board. *Saskatchewan WCB*, 2019. http://www.wcbsask.com/about-wcb/who-we-are/vision-mission-values/ [Accessed: September 2019].

[22] Workers' Compensation Board of British Columbia. *WCB of British Columbia*, 2017. https://www.worksafebc.com/en [Accessed: September 2017].

[23] Sean Tucker. *Workplace Fatalities in Saskatchewan*, 2018. https://www.uregina.ca/business/faculty-staff/faculty/file_download/Workplace%20fatalities%20in%20SK%20October%2011%202018.pptx.pdf [Accessed: July 2019].

[24] Jonathan Fan, Christopher B McLeod, and Mieke Koehoorn. Descriptive epidemiology of serious work-related injuries in british columbia, canada. *PloS one*, 7(6):e38750, 2012.

[25] Andrew Sharpe and Jill Hardt. Five deaths a day: Workplace fatalities in canada. *Centre for the Study of Living Standards*, 2006.

[26] Curtis Breslin, Mieke Koehoorn, Peter Smith, and M Manno. Age related differences in work injuries and permanent impairment: a comparison of workers compensation claims among adolescents, young adults, and adults. *Occupational and environmental medicine*, 60(9):e10–e10, 2003.

[27] Christopher B. McLeod, Robert A. Macpherson, William Quirke, Jonathan Fan, Benjamin C Amick, Cameron A Mustard, Sheilah Hogg-Johnson, Allen Kraut, and Mieke Koehoorn. *Work Disability Duration: A Comparative Analysis of Three Canadian Provinces*, 2017. https://www.wcb.mb.ca/sites/default/files/files/Koehorn%20et_al%20Work%20Disability%20Duration%20PWHS%20July%202017%20Final%20%20Report.pdf [Accessed: August 2019].

[28] BMPH Pratt, J Cheesman, C Breslin, and MT Do. Occupational injuries in canadian youth: an analysis of 22 years of surveillance data collected from the canadian hospitals injury reporting and prevention program. *Health promotion and chronic disease prevention in Canada: research, policy and practice*, 36(5):89, 2016.

[29] Allen Kraut. Estimates of the extent of morbidity and mortality due to occupational diseases in canada. *American journal of industrial medicine*, 25(2):267–278, 1994.

[30] B Fabiano, F Curro, and R Pastorino. Occupational injuries in italy: risk factors and long term trend (1951–98). *Occupational and environmental medicine*, 58(5):330–338, 2001.

[31] Guang-Xiang Chen and David E Fosbroke. Work-related fatal-injury risk of construction workers by occupation and cause of death. *Human and Ecological Risk Assessment*, 4(6):1371–1390, 1998.

[32] Glenn S Pransky, Katy L Benjamin, Judith A Savageau, Douglas Currivan, and Kenneth Fletcher. Outcomes in work-related injuries: A comparison of older and younger workers. *American Journal of Industrial Medicine*, 47(2):104–112, 2005.

[33] Peter Matthew Smith and Janneke Berecki-Gisolf. Age, occupational demands and the risk of serious work injury. *Occupational medicine*, 64(8):571–576, 2014.

[34] Suzanne M Kisner and Stephanie G Pratt. Occupational fatalities among older workers in the united states: 1980-1991. *Journal of occupational and environmental medicine*, 39(8):715–721, 1997.

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

[36] Menelaos Pavlou, Gareth Ambler, Shaun R Seaman, Oliver Guttmann, Perry Elliott, Michael King, and Rumana Z Omar. How to develop a more accurate risk prediction model when there are few events. *Bmj*, 351:h3868, 2015.

[37] Peter C Austin and Ewout W Steyerberg. Events per variable (epv) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical methods in medical research*, 26(2):796–808, 2017.

[38] Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R Holford, and Alvan R Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12):1373–1379, 1996.

[39] Charles W Champ and Deborah K Shepherd. Encyclopedia of statistics in quality and reliability. 2007.

[40] Donald E Farrar and Robert R Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pages 92–107, 1967.

[41] Wonsuk Yoo, Robert Mayberry, Sejong Bae, Karan Singh, Qinghua Peter He, and James W Lillard Jr. A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology*, 4(5):9, 2014.

[42] Christian M Ringle, Sven Wende, and Jan-Michael Becker. Smartpls 3. *Boenningstedt: SmartPLS GmbH*, 2015.

[43] Paul Allison. *When Can You Safely Ignore Multicollinearity?*, 2012. https://statisticalhorizons.com/multicollinearity [Accessed: August 2019].

[44] Fushing Y Hsieh, Daniel A Bloch, and Michael D Larsen. A simple method of sample size calculation for linear and logistic regression. *Statistics in medicine*, 17(14):1623–1634, 1998.

[45] Yann-Yann Shieh and Rachel T Fouladi. The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. *Educational and psychological measurement*, 63(6):951–985, 2003.

[46] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. Applied multiple regression. *Correlation Analysis for the Behavioral Sciences*, 2, 1983.

[47] WilliamL Hays. Statistics . new york: Holt, rinehart, winston. *HaysStatistics1981*, 1981.

[48] John Neter, William Wasserman, and George A Whitmore. *Applied statistics*. Number 519.5 N469. Allyn and Bacon, 1978.

[49] Scott A Czepiel. Maximum likelihood estimation of logistic regression models: theory and implementation. *Available at czep. net/stat/mlelr. pdf*, 2002.

[50] Szilard Nemes, Junmei Miao Jonasson, Anna Genell, and Gunnar Steineck. Bias in odds ratios by logistic regression modelling and sample size. *BMC medical research methodology*, 9(1):56, 2009.

[51] Georg Heinze and Rainer Puhr. Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Statistics in medicine*, 29(7-8):770–777, 2010.

[52] Sander Greenland and Mohammad Ali Mansournia. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in medicine*, 34(23):3133–3143, 2015.

[53] Georg Heinze and Michael Schemper. A solution to the problem of monotone likelihood in cox regression. *Biometrics*, 57(1):114–119, 2001.

[54] Georg Heinze and Meinhard Ploner. Fixing the nonconvergence bug in logistic regression with splus and sas. *Computer Methods and Programs in Biomedicine*, 71(2): 181–187, 2003.

[55] Georg Heinze, Meinhard Ploner, D Dunkler, and H Southworth. Firth's bias reduced logistic regression. *R package version*, 1, 2013.

[56] Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

[57] Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1): 191–201, 1992.

[58] Jose Manuel Pereira, Mario Basto, and Amelia Ferreira da Silva. The logistic lasso and ridge regression in predicting corporate failure. *Procedia Economics and Finance*, 39: 634–641, 2016.

[59] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[60] Robert Tibshirani, Martin Wainwright, and Trevor Hastie. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[61] Brownlee Jason. *A Gentle Introduction to k-fold Cross-Validation*, 2018. https://machinelearningmastery.com/k-fold-cross-validation/ [Accessed: August 2019].

[62] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[63] Warren M Persons. The correlation of economic statistics. *Publications of the American Statistical Association*, 12(92):287–322, 1910.

[64] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348–1360, 2001.

[65] Jian Huang, Huiliang Xie, et al. Asymptotic oracle properties of scad-penalized least squares estimators. In *Asymptotics: Particles, processes and inverse problems*, pages 149–166. Institute of Mathematical Statistics, 2007.

[66] Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[67] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[68] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[69] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[70] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[71] John A Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857): 1285–1293, 1988.

[72] Robert Trevethan. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Frontiers in public health*, 5:307, 2017.

[73] Janpu Hou. *ROC on Classifiers*, 2018. http://rstudio-pubs-static.s3.amazonaws.com/359286_c8e26df825464105aff4db12e9da32d7.html [Accessed: August 2019].

[74] Y. So. A tutorial on logistic regression. *Proceedings of the Eighteenth Annual SAS Users Group International Conference*, 1993.

[75] Georg Heinze, Meinhard Ploner, Daniela Dunkler, and Harry Southworth. logistf: Firths bias reduced logistic regression. r package version 1.21. *R Foundation for Statistical Computing, Vienna, Austria. Hu G, Dai Z, Long L, Han Y, Hou S, and Wu L (2002) Bioequivalence of clavulanate potassium and amoxicillin (1: 7) dispersible tablets in healthy volunteers. J Huazhong Univ Sci Technolog Med Sci*, 22:224–227, 2013.

[76] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[77] J Friedman, T Hastie, N Simon, and R Tibshirani. Lasso and elastic-net regularized generalized linear models. r-package version 2.0-5. 2016, 2016.

[78] Trevor Hastie, Brad Efron, and Maintainer Trevor Hastie. Package lars. 2013.

[79] Taylor B Arnold, Ryan J Tibshirani, Maintainer Taylor Arnold, and TRUE ByteCompile. Package genlasso. *Statistics*, 39(3):1335–1371, 2019.

[80] Joanne O Crawford, Richard A Graveling, HA Cowie, and Ken Dixon. The health safety and health promotion needs of older workers. *Occupational medicine*, 60(3): 184–192, 2010.

[81] Cynthia K Grandjean, Patricia C McMullen, Kenneth P Miller, William O Howie, Kevin Ryan, Alice Myers, and Richard Dutton. Severe occupational injuries among older workers: demographic factors, time of injury, place and mechanism of injury, length of stay, and cost data. *Nursing & health sciences*, 8(2):103–107, 2006.

[82] Sara F Jacoby, Theimann H Ackerson, and Therese S Richmond. Outcome from serious injury in older adults. *Journal of Nursing Scholarship*, 38(2):133–140, 2006.

[83] Larry A Layne and Deborah D Landen. A descriptive analysis of nonfatal occupational injuries to older workers, using a national probability sample of hospital emergency departments. *Journal of occupational and environmental medicine*, 39(9):855–865, 1997.

[84] Larry A Layne and Keshia M Pollack. Nonfatal occupational injuries from slips, trips, and falls among older workers treated in hospital emergency departments, united states 1998. *American journal of industrial medicine*, 46(1):32–41, 2004.

[85] Olivia S Mitchell. The relation of age to workplace injuries. *Monthly Lab. Rev.*, 111:8, 1988.

[86] Simo Salminen. Have young workers more injuries than older ones? an international literature review. *Journal of safety research*, 35(5):513–521, 2004.

[87] Centers for Disease Control, Prevention (CDC, et al. Nonfatal occupational injuries and illnesses among older workers—united states, 2009. *MMWR. Morbidity and mortality weekly report*, 60(16):503, 2011.

[88] Bernard CK Choi, Marianne Levitsky, Roxanne D Lloyd, and Ilene M Stones. Patterns and risk factors for sprains and strains in ontario, canada 1990: an analysis of the work-

place health and safety agency data base. *Journal of Occupational and environmental Medicine*, 38(4):379–389, 1996.

[89] Union européenne. Commission européenne and EUROSTAT. *Work and Health in the EU: A Statistical Portrait.* Office for official publications of the European communities, 2004.

[90] Anne Marie Feyer, AM Williamson, N Stout, T Driscoll, H Usher, and John Desmond Langley. Comparison of work related fatal injuries in the united states, australia, and new zealand: method and overall findings. *Injury Prevention*, 7(1):22–28, 2001.

[91] H Dimich-Ward, JR Guernsey, William Pickett, D Rennie, L Hartling, and RJ Brison. Gender differences in the occurrence of farm related injuries. *Occupational and environmental medicine*, 61(1):52–56, 2004.

[92] Evan Nadhim, Carol Hon, Bo Xia, Ian Stewart, and Dongping Fang. Falls from height in the construction industry: a critical review of the scientific literature. *International journal of environmental research and public health*, 13(7):638, 2016.

[93] Nitin Roy, Fadwa Eljack, Arturo Jiménez-Gutiérrez, Bin Zhang, Preetha Thiruvenkataswamy, Mahmoud El-Halwagi, and M Sam Mannan. A review of safety indices for process design. *Current opinion in chemical engineering*, 14:42–48, 2016.

[94] Chia-Fen Chi, Tin-Chang Chang, and Hsin-I Ting. Accident patterns and prevention measures for fatal occupational falls in the construction industry. *Applied ergonomics*, 36(4):391–400, 2005.

[95] Ken D Rosenman, Joseph C Gardiner, J Wang, J Biddle, A Hogan, MJ Reilly, K Roberts, and Ed Welch. Why most workers with occupational repetitive trauma do not file for workers compensation. *Journal of Occupational and Environmental Medicine*, 42(1):25, 2000.

[96] Saskatchewan Government Insurance. *Major Contributing Factors of our Traffic Accident Information System (TAIS) report.*, 2019. https://www.sgi.sk.ca/documents/

625510/627017/TAIS_2015_03.pdf/4582d742-0919-4d40-a145-c6f09cfea797 [Accessed: August 2019].

[97] Government of Canada. *Canadian Motor Vehicle Traffic Collision Statistics: 2017*, 2017. https://www.tc.gc.ca/eng/motorvehiclesafety/canadian-motor-vehicle-traffic-collision-statistics-2017.html [Accessed: August 2019].

[98] Benjamin A Goldstein, Ann Marie Navar, and Rickey E Carter. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal*, 38(23):1805–1814, 2016.

[99] David Firth. Bias reduction, the jeffreys prior and glim. In *Advances in GLIM and Statistical Modelling*, pages 91–100. Springer, 1992.

[100] P McCullough and JA Nelder. Generalized linear models chapman and hall. *New York*, 1989.

[101] Michelle Botes et al. *Comparing logistic regression methods for completely separated and quasi-separated data.* PhD thesis, University of Pretoria, 2013.

[102] David Collett. *Modelling binary data.* Chapman and Hall/CRC, 2002.

[103] Mohammad Ali Mansournia, Angelika Geroldinger, Sander Greenland, and Georg Heinze. Separation in logistic regression: causes, consequences, and control. *American journal of epidemiology*, 187(4):864–870, 2017.

[104] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

Appendix. A

Summary of Literature Review of Published WCB Claims Analysis

Table A.1: Summary of literature review of published WCB claims analysis

Author (Year)	**Data**	**Goal**	**Method**	**Results**
McLeod et al (2017) [27]	MB, BC, ON Workers Compensation Board	Conduct detailed analysis of work disability duration across jurisdictions and analyze long duration injury claims among 3 Canadian provinces (MB, BC, ON) to investigate trends and variations in work disability duration across these provinces.	Cox proportional hazard model	Reducing long work disability duration claim is a key policy objective of Canadian WCB. Large differences in the average number of disability days paid were observed across province and industry sector. Jurisdiction has a marked effect on duration of work disability.
Fan et al (2012)[24]	Workers Compensation Board of BC	Examine the rate and distribution of serious work-related injuries by demographic, work and injury characteristics in British Colombia from 2002 to 2008.	Negative binomial regression	Women had a lower overall serious injury rate compared to men. The 35-44 age group had the highest overall rate compared to youngest age group. The rate for severe strain was similarly high in both men and women group in the 35-44 age group. Although there is a differential pattern by gender for other types of injuries.

Pratt et al (2016) [28]	Youth aged 10 to 17 years, inclusive, who had completed a CHIRPP form between January 1, 1991, and December 31, 2012)	Describes features of work-related injuries in young Canadians to identify areas for potential occupational injury prevention strategies.	none	"Of the 6046 injuries (0.72% of events in this age group) that occurred during work, 63.9% were among males. Youth in food and beverage occupations (54.6% males) made up 35.4% of work-related ED visits and 10.2% of work-related hospital admissions, while primary industry workers (76.4% males) made up 4.8% of work-related ED visits and 24.6% of work-related hospital admissions [28]"
Tucker (2016-2018)	WCB-SK AWCBC	To compare fatality rate and occupational injury rate among different provinces and suggest some advice to WCB and policy makers	Descriptive statistics and calculation of fatality and occupational injury rates	No statistical comparison of risk groups. Descriptive is too lengthy to summarize, please refer to the report

Table A.2: Summary of literature review of using penalized regression methods for risk prediction and model selection

Author (Year)	Data	Goal	Method	Results
Pavlou et al [36] (2015)	Penile cancer	Predicting low dimensional binary outcomes when the number of events is small compared to the number of regression coefficient using penalized regression methods	Penalized regression methods including LASSO, ridge regression, elastic net, adaptive lasso and SCAD	Ridge regression performs well except when we have noise predictors. LASSO performs better than ridge in case of noise predictor and worse in case of correlated predictors. Elastic net performs well in all scenarios, and adaptive lasso and SCAD perform best in all scenarios with many noise predictors.
Goldstien et al [98] (2016)	Data derived from their institution's electronic health record (13 regularly measured laboratory markers)	Use of machine learning methods for development of risk prediction models	Regression based including lasso, ridge regression, principal component regression, random forests	Machine learning algorithms can be advantageous over traditional regression methods because they can be used to solve the problem of multiple and correlated predictors, non-linear relationships, interaction between predictors and endpoints and most importantly large datasets
Lu et al (2017)	Bangladesh cohort	To show application of penalized linear regression methods including SCAD, adaptive lasso to the selection of environmental biomarkers.	SCAD, MCP	Simulation studies show that SCAD, adaptive lasso and MCP are better variable selection methods compared to traditional stepwise regression methods.
Rahman et al [1] (2017)	Stress echocardiography data and simulation study	Evaluation of the performance of Firth-and log F-type penalized methods in risk prediction for small or sparse binary data	Firth, log F-type penalized method	All penalized methods offered some improvements in calibration, discrimination, and overall predictive performance. Although the Firth and log-F type methods showed almost equal amount of improvement, Firth type penalization produces some bias in the average predictive probability and the amount of bias is even larger than that produced by MLE.

Appendix. B

Firth's Method

Reduction of bias in maximum likelihood estimates is one of the popular ways to address the problem of separation [99]. The maximum likelihood estimates are unbiased with asymptotic variance which is equal to $I(\boldsymbol{\theta}) = \mathbf{X}^T\mathbf{W}\mathbf{X}$, the inverse of Fisher information matrix, in which $\mathbf{X}$ is the model matrix, and $\mathbf{W}$, an $n \times n$ matrix, when $\mathbf{W} = diag(\pi_i(1-\pi_i))$. McCullagh and Nelder [100] showed that for a large sample size,

$$E(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = O(n^{-1}). \tag{B.1}$$

Then Firth [12] showed that for an m dimensional model, the asymptotic bias of a single ML estimate $\hat{\theta}$ of parameter θ can be written in the following form as it was also shown in [101]:

$$b(\theta) = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + ... \tag{B.2}$$

Reducing the bias of parameter estimates by removing the $O(n^{-1})$ term is the main goal of Firth's method [99].

The maximum likelihood estimate is a solution to the score equation

$$\triangledown \ell(\theta) = U(\theta) = 0, \tag{B.3}$$

Where $\ell(\theta) = logL(\theta)$ is log likelihood function [12]. An exponential family model can be written as $\ell(\theta) = t\theta - K(\theta)$, in which θ is scalar [12]. Then we have

$$U(\theta) = \ell^{'}(\theta) = t - K^{'}(\theta), \tag{B.4}$$

and as shown in Equation B.4, the sufficient statistic t only affects the location of $U(\theta)$, and it would not have any effects on its shape [12]. As discussed in Firth [12], the bias in $\hat{\theta}$ comes from two factors including: unbiasedness of the score function at the true value of θ $(E(U(\theta)) = 0$ and curvature of the score function $(U^{'''}(\theta) \neq 0)$ [12].

The main focus of Firth's method is that the bias in $\hat{\theta}$ can be reduced by introducing a small bias in score function [12]. The best modification to $U(\theta)$ is given by simple triangle geometry, shown in Figure B.1.

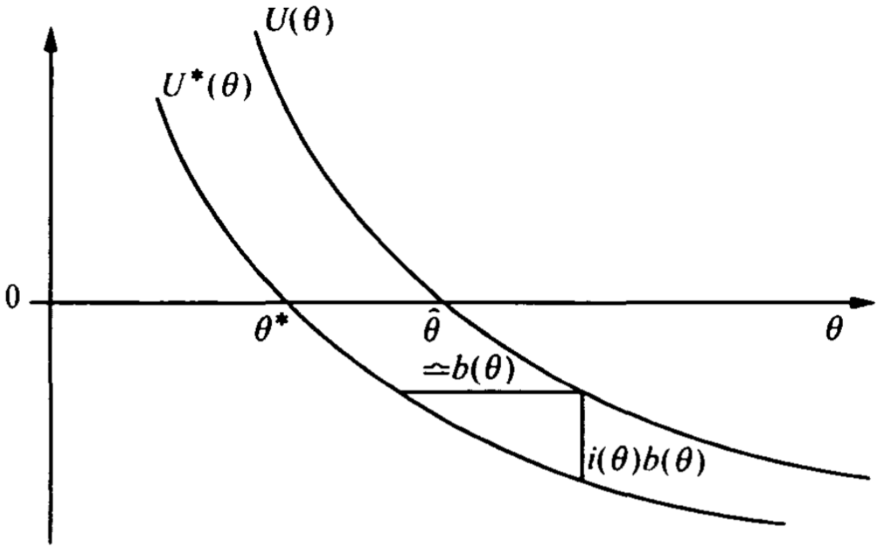

Figure B.1: Modified score function reproduced from DAVID FIRTH, Bias reduction of maximum likelihood estimates, Biometrika 1993; 80 (1): 2738, doi:10.1093/biomet/80.1.27. Reprinted by permission of Oxford University Press on behalf of the Biometrika Trust.

"If $\hat{\theta}$ is subject to a positive bias $b(\theta)$, the score fucntion is shifted downward at each point θ by an amount $i(\theta)b(\theta)$, where $-i(\theta) = U''(\theta)$ is the local gradient" [12]. The modified score function will be defined by

$$U^*(\theta) = U(\theta) - i(\theta)b(\theta), \tag{B.5}$$

Where θ^* is a solution to $U^*(\theta) = 0$. When we have a vector parameter, Equation B.5 will be read as a vector equation, and $i(\theta)$ will be the Fisher information matrix. We refer the reader to Firth [12] for more information.

The log likelihood function can be penalized by Jeffrey's invariant prior [104] to obtain the modified score function above [55, 101]. The Jeffrey's invariant prior density is $|I(\boldsymbol{\theta})|^{1/2}=$ $|\mathbf{X}^T\mathbf{W}(\boldsymbol{\theta})\mathbf{X}|^{1/2}$, when $\boldsymbol{\theta}$ is the vector of unknown parameters, and $I(\boldsymbol{\theta}) = \mathbf{X}^T\mathbf{W}\mathbf{X}$ is Fisher information matrix. The penalised likelihood function in Firth's method can be written as

$$l^*(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) \times |I(\boldsymbol{\theta})|^{1/2} \tag{B.6}$$

Taking natural log of Equation B.6, we have

$$logl^*(\boldsymbol{\theta}) = logl(\boldsymbol{\theta}) + (1/2)log|I(\boldsymbol{\theta})|. \tag{B.7}$$

For more detail about the method we refer readers to the original paper on Firth's method by Firth [12].

Appendix. C

List of Covariates

Table C.1: Potential explanatory variables and their categories from claims data

Variable	**levels**	
Cause of injury	Contact with objects and equipment	Falls
	Bodily reaction and exertion	Exposure to harmful substances or environments
	Transportation accidents	Other events or exposures
	Assaults, violent acts, attacks, harassment	
Source of injury	Chemicals and chemical products	Containers
	Furniture and fixture	Machinery
	Parts and materials	Persons, plants, animals, and minerals
	Structures and surfaces	Tools, instruments, and equipment
	Vehicles	Other sources
Occupations	Art, culture, recreation and sport	Business and finance
	Health	Natural and applied sciences
	Primary industry	Social sciences and education
	Sale and services	Trade and transport
	Processing and manufacturing	Not stated
Part of body injured	Other body parts	Head
	Trunk	Body system
	Lower extremities	Upper extremities
	Multiple body parts	
Year	2007, 2008, ...,2016	
Age	14,15,...,85	
Month	January, February,..., December	

Table C.2: List of covariates selected by lasso (for λ=lambda.min and lambda.1se)

Variable	**Selected levels**	
	Lasso with $\lambda.min$	**Lasso with** $\lambda.1se$
Age	25-34 , 35-44 ,45-54, 55-64, 65-85	25-34 , 35-44 ,45-54, 55-64, 65-85
Gender	Male	Male
Source of injury	machinery, vehicles containers, furniture other sources, parts and materials persons/plants, structures tools, instruments	machinery, vehicles
Occupations	natural and applied sciences primary industry, trade and transport business, health, social sciences art/culture, sale/services, processing, not stated	natural and applied sciences primary industry, trade and transport
Part of body	neck including throat, head trunk, multiple body parts lower extremities, upper extremities	neck including throat, head trunk, multiple body parts lower extremities, upper extremities
Cause of injury	bodily reaction and exertion, other events transportation accidents, objects and equipment harmful substances	bodily reaction and exertion, other events transportation accidents