Spring 2019

Latent Choice Models to Account for Misclassification Errors in Discrete Transportation Data

Lacramioara Elena Balan
Old Dominion University, lbala001@odu.edu

LATENT CHOICE MODELS TO ACCOUNT FOR MISCLASSIFICATION ERRORS IN DISCRETE TRANSPORTATION DATA

by

Lacramioara Elena Balan

B. Sc. June 2013, Technical University of Civil Engineering of Bucharest, Romania

M.S.E, June 2015, Technical University of Civil Engineering of Bucharest, Romania

A Dissertation Submitted to the Faculty of

Old Dominion University in Partial Fulfilment of the

Requirements for the degree of

DOCTOR OF PHILOSOPHY

CIVIL & ENVIRONMENTAL ENGINEERING

OLD DOMINION UNIVERSITY

May 2019

Approved by:

Rajesh Paleti (Chair)

Mecit Cetin (Chair)

Hong Yang (Member)

ABSTRACT

LATENT CHOICE MODELS TO ACCOUNT FOR MISCLASSIFICATION ERRORS IN DISCRETE TRANSPORTATION DATA

Lacramioara Elena Balan
Old Dominion University, 2019
Director: Dr. Rajesh Paleti

One of the most fundamental tasks when it comes to analyzing data using statistical methods is to understand the relationship between the explanatory variables and the outcome. Misclassification of explanatory variables is a common risk when using statistical modeling techniques. In this dissertation, we define 'misclassification,' as a response that is reported or recorded in the wrong category; for example, a variable is registered as a one when it should have the value zero. Misclassification can easily happen in any data; for example, in an interview setting where the respondent misunderstands the question or the interviewer checks the wrong box.

The results uncovered significant misclassification rates ranging from 1% to 40% for different auto ownership alternatives, in the first part of the dissertation. Also, the results from latent class models provide evidence for variation in misclassification probabilities across different population segments. The second part of the dissertation uses traditional crash databases that record police-reported injury severity data, which are prone to misclassification errors. In addition, we developed a mixed generalized ordered response model that quantifies misclassification rates in the injury severity variable and adjusts the bias in parameter estimates due to misclassification. The model uncovered a 32% misclassification rate in the non-incapacitating severity category. As another case study, the misclassification extent in the telecommuting frequency data is also investigated. Telecommuting frequency is a response variable collected in travel surveys; therefore, it is prone to errors leading to mismeasurements or misclassification. The objective of this investigation of the dissertation is to develop a statistical model to analyze telecommuting data while accounting for potential misclassification errors.

Models that ignore misclassification were not only found to have lower statistical fit but also significantly different elasticity effects, particularly for choice alternatives with high

misclassification probabilities. Overall, the simulation analysis, along with the other models developed, suggests that the models that consider misclassification in the data perform better than the ones that ignore the misclassification. The methods developed in this study can be extended to analyze misclassification in other transportation disciplines.

DEDICATION

To my father Ioan Balan, and my mother, Luminita Balan

ACKNOWLEDGMENTS

The bachelor's degree was - until not too long ago - the proudest accomplishment of my life. Always passionate for mathematics and everything connected with it, I decided to apply for Geodesy engineering program with my family's support. Enrolling at the university with the highest rank for this major in the country, I was at the same time excited and terrified of what was about to happen. I experienced a new world opening to me; stimulated by a new encouraging environment, I excelled academically. I learned that if I try it hard, I could succeed, if I wanted something badly enough. Based on this principle, I continued with my education and immediately after I finished my Master's degree I decided to enroll in a Ph.D. program in the United States. I want to acknowledge all the people who have supported and inspired me during these days. More specifically, I would like to thank a small group of people without whom this dissertation would not have been possible: my dissertation committee members, my office colleagues, my family, and my friends.

First of all, I am indebted to my dissertation advisor, Dr. Rajesh Paleti. Since the first day I started graduate school, he believed in me and gave me endless support. I met Dr. Paleti in the fall of 2015 when he challenged me to learn how to be a researcher. On the academic level, Dr. Paleti taught me fundamentals of conducting scientific research in the transportation area. Under his directions, I learned how to define a research problem, find a solution to it, and finally publish the results. On a personal level, Dr. Paleti inspired me by his hardworking and passionate attitude. To summarize, I would like to give Dr. Paleti most of the credit for helping me to become the kind of scientist I am today.

Also, I would like to thank the rest of the committee members (Dr. Mecit Cetin and Dr. Hong Yong) for their great support, prompt responses, and invaluable advice. I would also like to give my gratitude for my office colleagues for their continued support and for making my experience in the TRI and graduate school exciting and fun.

Last but not least, I would like to show my appreciation to my family and friends. All the work for this dissertation would not have been possible without their warm love, continued patience, and endless support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Research Background

The most fundamental tasks when it comes to analyzing data using statistical methods is to understand the correlation between the explanatory variables and the outcome. Misclassification of explanatory variables is a common risk when using statistical modeling techniques. Misclassification occurs when a subject is falsely classified into a category in which the subject does not belong. It may result from misreporting by study subjects, from the use of less than optimal measurement devices, or by random error. For example, a variable could be recorded with a value of one when it should have the value zero.

The subject of mismeasurement has been extensively studied in econometrics. Most of the studies have focused on analyzing the misclassification of the discrete dependent variable more that mismeasurement of continuous dependent variable. From a statistical standpoint, when the variable analyzed is continuous, this event is referred to as containing errors in variables or simply as measurement errors; when the explanatory variable is discrete, the term used is a misclassification. The observed responses in the data can be viewed as realizations from random variables that depend on true latent responses that are unobservable to the analyst. Misclassification of continuous dependent variables does not result in biased parameter estimates in linear regression specification but less precise statistical estimates [1]. However, in cases of limited dependent variables such as discrete choices, the standard maximum likelihood estimates are biased and inconsistent. For instance, using simulation analysis, Hausman et al. found that misclassification rates as low as 2% can lead to 15-25% bias in the parameter estimates and lower standard errors (or over-precise estimates) in the binary choice scenario [2]. This study also analyzed the decision of workers to change jobs in the past one year and found that 6% of non-job changers reported job change whereas 31% of persons who changed their job did not report job change in the survey data. So, their study found not only evidence for significant misclassification but also different misclassification rates for different responses (job change versus no job change).

1.2 The Importance of Travel Survey

Travel surveys are one of the most important ways of obtaining the critical information needed for transportation planning and decision making. These surveys are not only used to gather

current information about the demographic, socioeconomic, and trip-making characteristics of individuals and households, but they are also used to further understanding of travel about the next choice, location, and also, scheduling of daily activities. Such information will permit us to enlarge travel forecasting methods and improve the capacity to predict changes in daily travel patterns in response to current social and economic trends and new investments in transportation systems and services.

Travel surveys also play a role in evaluating changes in transportation supply and regulation as they occur. In the last half-century, travel survey methods have experienced tremendous change. Originally, travel surveys were done primarily through face-to-face interviews, typically conducted in respondents' homes or at intercept points along major roadways and transit routes or major transportation nodes [3-6]. These changes, coupled with technological advancements such as GPS, have improved the quality of survey data considerably. However, it is still very likely that there are several errors in the data recorded. The response variables collected in these surveys are prone to errors leading to mismeasurement or misclassification. Standard modeling methods that ignore these errors while modeling travel choices can lead to biased parameter estimates. Models that ignore misclassification were not only found to have lower statistical fit but also significantly different elasticity effects, particularly for choice alternatives with high misclassification probabilities.

1.3 The Importance of Safety Analysis

A key crash attribute used for safety analysis is the injury severity of the crash. The crash injury severity level is recorded as the severity level of the most severely injured person in the crash. Reducing the severity of injuries from traffic accidents is one of the most effective means to improve highway safety. Many studies have been conducted to reduce the number of people killed and injured in traffic accidents and to identify the risk factors that can significantly influence the injury outcomes of traffic accidents. Injury severity analysis is a significant topic to investigate for improving motorized vehicles and roadway design, improving control strategies at conflict locations, designing good pedestrian and bicycle facilities and building driver and non-motorized user education programs. Typically, injury severity is recorded based on crash-assigned or hospital-assigned ordinal scoring systems [7]. The crash-assigned injury severity reported in police accident reports is typically recorded on a five-point ordinal scale – fatal, incapacitating, non-

incapacitating, possible, and no injury (KABCO scale). However, the definitions of these ordinal categories in the KABCO scale vary across state jurisdictions. Traditional crash databases that record police-reported injury severity data are prone to misclassification errors. Ignoring these errors in discrete ordered response models used for analyzing injury severity can lead to biased and inconsistent parameter estimates.

1.4 The Importance of Telecommuting Frequency

Telecommuting usually refers to working from home or telecenters using computers or telephones. Other scholars are referring to telecommuting as teleworking, and recently more advanced communication facilities are used to maintain a connection with the office and with central management and administration. Some of the concerns of the mobility management strategies that affect the public and private sector (pollution, traffic congestion, energy consumption, labor shortage, office spaces, and family commitments) can be enhanced using telecommuting. From a cost standpoint towards the users and the time it takes to be implemented, scholars have shown that, among all the strategies, telecommuting is easy to implement, and it has a lower cost [8-10].

Telecommuting has several benefits for both employers and employees. It can improve telecommuter's family-work balance by providing more time to be spent with the family members [11, 12], and it can bring a more efficient way of planning the activity-travel arrangements during the working hours [13]. In addition to the saved commuting time, it has been concluded that telecommuters spend more time on work activities than they would do in the workplace [14], and based on the flexible work schedule, they are working during the time that they are productive [15]. In Finland, for example, it was proven that home-based telecommuting could reduce the total commute distance by 0.7% (almost 0.84 million miles saved every week) [16]. Recently, using data from the Canadian General Security Survey, it was found that as telecommuters have more flexible activity schedules, they mostly take trips during off-peak periods. Also, it was showed that telecommuting could reduce daily travel time by an average of 13 minutes [17]. Although scholars focused on the effects of telecommuting on travel demand and network operation, none of the studies considered errors in the data modeled. Ignoring these errors in discrete count models analysis when using telecommuting frequency data may lead to bias and inconsistent parameter estimates. The number of days a person telecommutes in a month is usually obtained using

household travel surveys [18]. It was previously demonstrated that household travel surveys are prone to misclassification errors. Ignoring this misclassification errors can lead to bias and inconsistent parameter estimates.

1.5 An Overview of Research Objectives

This dissertation addresses this gap in the past literature by developing a framework for analyzing misclassification errors in discrete choice responses, by developing statistical models to analyze different discrete transportation data while accounting for potential misclassification errors using the existing literature in econometrics. The different datasets investigated are household travel surveys: application to auto ownership, safety analysis: application of the injury severity level of the driver, and household travel surveys: application of monthly telecommuting frequency, and for all of them, the model that accounts for misclassification performed better. For these analyses misclassification rates were modeled as high as 40% and 25% for the "three cars" and "two cars" respectively in the case of the auto ownership levels. In the injury severity level and the telecommuting frequency, misclassification rates as high as 32.2%, and 14.4% respectively, were found in the non-incapacitating injuries. It was also proven that the model that accounts for misclassification has a better statistical fit when compared with the model that ignores misclassification for the three datasets investigated.

1.6 The Structure of the Dissertation

The rest of the dissertation is structured as follows. The next section, Chapter 2, provides an overview of the available methods in the econometric literature for handling misclassification and methods previously used to acknowledge the presence of misclassification in the cases of injury severity data, auto ownership data, and telecommuting frequency data. Chapter 3 presents a simulation study to evaluate the performance of the misclassification models using synthetic data. Chapter 4 presents the methodological framework and describes the data followed by the empirical results and post-estimation analysis of misclassification in travel surveys and implications to choice modeling: application to household auto ownership decisions. Chapter 5 presents the methodological framework and describes the data followed by the empirical results and post-estimation analysis using modified mixed generalized ordered response model to handle misclassification in injury severity data. Chapter 6 presents the methodological framework and describes the data followed by the empirical results and post-estimation analysis using a

generalized extreme value model to handle misclassification rates in the telecommuting frequency data. Finally, I will outline the research, and give information on fulfillment of the research objectives and future research directions.

CHAPTER 2

EARLIER RESEARCH AND THE CURRENT STUDY IN CONTEXT

This chapter provides a detailed review of earlier work relevant to the two main objectives of the dissertation The literature review is grouped under the following headings: 1) Misclassification in Travel Surveys and Implications to Choice Modeling: Application to Household Auto Ownership Decisions, 2) A modified Mixed Generalized Ordered Response Model to Handle Misclassification in Injury Severity Data, and 3) A Generalized Extreme Value Model to Handle Misclassification in Telecommuting Frequency Choices Data.

2.1 Misclassification in Travel Surveys and Implications to Choice Modeling: Application to Household Auto Ownership Decisions

Household Travel Survey (HTS) data that records information regarding activity and travel patterns along with detailed socio-demographic details of a representative population in the study area is the primary component underlying all transportation planning and policy analysis. The travel survey methods have evolved over the past two decades both in the format (e.g., travel diary versus activity diary) and the medium (e.g., face-to-face or phone-based interviews versus web-based survey questionnaires) of data collection [3-6]. These changes, coupled with technological advancements such as GPS, have improved the quality of survey data considerably. However, it is still very likely that there are several errors in the data recorded. These errors may be traced back to too many different sources. For instance, the respondent can intentionally provide misinformation. For example, past studies found that self-employed individuals can under-report their income by up to 25% in household travel surveys [19]. In some cases, the errors might be systematically associated with the survey instrument used for collecting data. For instance, respondents in Computer Assisted Telephone Interviewing (CATI) based surveys were found to under-report trip rates, under-report travel distances, and over-report travel times compared to GPS-based studies [20]. Moreover, these errors were also found to vary based on the demographic characteristics of the survey respondents [21, 22]. It is also possible that the respondent provided wrong responses unintentionally, either due to the problem in comprehending the survey questions or miscommunication on the part of the surveyor. In some cases, random errors (e.g., mistakes while recording) can also be the source of errors. Such errors may occur in all types of survey

responses - continuous (e.g., trip duration) and discrete including nominal (mode choice), ordinal (trip/tour frequency), and count (monthly telecommuting frequency).

In the transportation context, this is a critical problem because most of the activity-travel choices are discrete and models that ignore misclassification can provide incorrect travel sensitivities leading to misleading or even wrong policy inferences. Moreover, recent travel demand models take the form of large-scale activity travel simulators that encompass a chain of several discrete choice models tied together sequentially. So, the misclassification errors in an upstream discrete choice model can accumulate and propagate to all downstream models in the activity-travel simulator, affecting several new model forecasts and not just the choice being modeled. However, it is surprising that while there have been studies that attempted to quantify the extent of misclassification [20, 22], most studies have entirely ignored these errors while modeling the travel choice itself.

2.2 A Modified Mixed Generalized Ordered Response Model to Handle Misclassification in Injury Severity Data.

The World Health Organization (WHO) considers road safety a significant public health problem given that almost 1.25 million people lose their lives and another 50 million people sustain non-fatal injuries every year globally [23]. Also, traffic crashes remain the leading cause of death within the 15 to 29 years age group. In the United States alone, 35,000 people lost their lives and 2.44 million people were injured in 2015 [24]. To address this problem, safety engineers undertake data analysis to identify policy measures to enhance roadway safety. In the United States, within each state, traffic accidents are usually investigated by police officers who complete a standard form, usually soon after a crash has occurred, named the police accident report (PAR). The report contains information regarding driver characteristics, vehicle attributes, traffic conditions, environmental conditions, and crash characteristics [25]. Typically, all accidents that are above a specified severity level and threshold for the property damage dollar value are recorded by police [26]. These PARs constitute the primary data component used for safety analysis. For instance, the General Estimates System (GES) data of the National Automotive Sampling System (NASS) is a representative sample of police-reported crashes across the nation. While the PARs are mostly reliable, several factors determine the quality of data recorded. The consistency of coverage and interpretation, missing data, response errors, entry procedure, and level of detail are some of the factors identified in the literature [27].

A key crash attribute used for safety analysis is the injury severity of the crash. The crash injury severity level is recorded as the severity level of the most severely injured person in the accident. In some cases, the PARs record the severity level of injuries sustained by all the people involved in the crash. Typically, injury severity is recorded based on crash-assigned or hospital-assigned ordinal scoring system [7]. The crash-assigned injury severity reported in PARs generally is recorded on a five-point ordinal scale – fatal, incapacitating, non-incapacitating, possible, and no injury. However, the definitions of these ordinal categories in the KABCO scale vary across state jurisdictions. The hospital-assigned injury severity, on the other hand, is an anatomically based Abbreviated Injury Scale (AIS) that rates an injury on a six-point scale (minor, moderate, serious, severe, critical, and maximum) based on the threat to life and is correlated with mortality, morbidity, and hospital stay duration [26].

While some studies found significant differences between the KABCO and AIS scoring systems [26], few others found that these two measures are reasonably consistent [7]. A comparison of police-reported crashes and hospital records in New Zealand found that police recorded only two-thirds of fatal accidents. Furthermore, the reporting rates were found to vary by demographic, accident, seasonal, and geographic factors [28]. Along similar lines, non-fatal pedestrian accidents were found to be under-reported in police records [29]. A similar comparison between police records and trauma registry in France found misclassification in all injury severity categories [30]. Similar under-reporting and misclassification of injury severity of crashes leading to medical care were found in other recent studies [31-34]. On the contrary, police records were found to considerably over-estimate injury severity, and the degree of over-estimation was found to vary by the injury severity score (ISS) and the victim's age and position inside the vehicle [25]. Overall, past research suggests that hospital-based AIS recordings are more precise compared to the KABCO scale police recordings. However, there is no consensus on the exact level of discordance, the nature of discordance, and the factors that lead to discordance between these two measures of injury severity [29, 30, 35, 36]. Even small errors in the injury severity data can have enormous implications for the accuracy of predictions and policy sensitivity analysis. Irrespective of the type of injury severity scoring system, there may be errors in crash databases [37, 38]. For instance, multiple injury severity patterns can have the same score in AIS depending on the weights associated with different body parts. Furthermore, the error rates in the police injury severity recordings that form the basis for most safety research are expected to be higher compared to

hospital-based severity scores. However, while there have been attempts to measure the errors in crash databases by comparing them with alternate data sources (such as hospital and ambulance records), none of these studies attempted to account for these errors in statistical models used for analyzing injury severity

2.3 Generalized Extreme Value Model to Handle Misclassification in Telecommuting Frequency Choices Data

Telecommuting or teleworking, mainly, refers to working from home or telecenters using telephones, computers, or other advanced communications facilities to maintain a connection with the office and with central management and administration. One of the mobility management strategies that address the public and private sector concerns such as pollution, traffic congestion, energy consumption, labor shortage, office space, and family commitments is telecommuting [39-41]. Previous scholars have shown that among these policies, telecommuting has a lower cost for the users and a shorter time to be implemented [8-10]. Telecommuting has several benefits for both employers and employees. It can improve telecommuter's family-work balance by providing more time to be spent with the family members [11, 12], and can bring a more efficient way of planning the activity-travel arrangements during the working hours [13]. In addition to the saved commuting time, it has been concluded that telecommuters spend more time on work activities than they would do in the workplace [14], and based on the flexible work schedule, they are working during the time that they are productive [15].

Several studies have considered two dimensions of telecommuting decisions and analyzed accordingly. First one focused on whether the employer provides telecommuting options for the employees and secondly, how many days the employee is using this option [42, 43]. From a monthly telecommuting frequency, earlier studies have had modeled the actual number of days an employee works from home using different models. Count models [43], ordered response models [44], or even breaking the frequency information into different categories (e.g., infrequent, medium, and high frequency) and using discrete choice models such as MNL are just a few of the models used to model the telecommuting frequency [45].

Telecommuting was mainly previously investigated from two major perspectives. On the one hand, some scholars examined telecommuting, focusing on the worker's adoption behavior and aimed to identify the factors associated with their propensity to adopt this policy [46-49].

Relying on the statistical analysis of workers' decisions about choice and frequency and trying to recognize the connection between their choices and various types of personal, household, job-related and built-environment attributes, as well as their activity planning and scheduling behavior it was of high importance to account for their propensity to adopt telecommuting [50].

On the other hand, some of scholars investigated the potential consequences of telecommuting implementation, the impact of this policy on telecommuter's trip rates and miles driven. There would be conflicting viewpoints about the effect on worker's daily activity-travel behavior, even if it were accepted that telecommuting reduces commute travel. Many studies have shown results supporting the hypothesis that telecommuting can reduce daily trip rates. It was confirmed that the telecommuter's peak period trips could be reduced by 60% and the total distance traveled by 75% on telecommuting days, based on spatial and temporal analysis of travel diaries from California [13]. It was also shown that telecommuting could reduce annual vehicle-miles traveled by up to 0.8% [8]. When comparing the results with approximate vehicle-miles traveled caused by public transit, it was shown that telecommuting is a far more cost-efficient congestion mitigation policy. In Finland, for example, it was proven that home-based telecommuting could reduce the total commute distance by 0.7% (almost 0.84 million miles saved every week) [16]. Recently, it was found that as telecommuters have more flexible activity schedules, they mostly take trips during off-peak periods, according to data from the Canadian General Security Survey. Also, it was shown that telecommuting could reduce daily travel time by an average of 13 minutes [17].

While focusing on the complementary effects of telecommuting, some studies showed an increase in travel measures [10, 51, 52]; overall, the impact of telecommuting on both travel demand and network operation still need to be studied for more empirical evidence on this issue [9, 53]. Although scholars focused on the effects of telecommuting on travel demand and network operation, none of the studies considered errors in the data modeled. Such errors can occur in all types of survey responses (continuous, and discrete including nominal, ordinal, and count data). Statistically, these errors are referred to as 'mismeasurements' errors, more specifically as 'misclassification' errors in the case of discrete responses. It was shown that misclassification rates as low as 2% could lead to 15-25% bias in the parameter estimates and lower standard errors, using simulation analysis [2]. For one study, the decision of workers to change jobs in the past year was

analyzed, and researchers found that 6% of non-job changers reported job change, whereas 31% job changers did not report job change in the survey data. Not only misclassification was found but also different misclassification rates for different responses. On the same topic, while investigating the auto ownership choice data, significant misclassification rates were revealed, ranging from 1% to 40% for different auto ownership alternatives [18]. It was shown that only 68.23% and 62.75% of possible and non-incapacitating injuries were correctly recorded in the 2014 General Estimates System (GES) data, using a mixed generalized ordered response model for quantifying the misclassification rates in the injury severity variables. Also, when compared with the mixed generalized order model that ignores misclassification, it was shown that the model that considers misclassification has better data fit [54]. The objective of this dissertation is to develop a statistical model to analyze telecommuting data while accounting for potential misclassification errors by building upon existing literature in econometrics. The empirical analysis was undertaken using the 2017 National Household Travel Surveys (NHTS).

CHAPTER 3

SIMULATION ANALYSIS

This chapter considers a simulation study to evaluate the extent of misclassification errors using synthetic data. The details of the simulation settings and the results of this analysis are discussed next. To analyze the extent of misclassification errors, the Multinomial Logit model that accounts for misclassification was developed.

3.1 Methodological Framework

Previous econometrics studies developed statistical models (parametric and semi-parametric) that estimate discrete choice models under misclassification. The sufficiency condition needed for consistency is that the probability of being misclassified is smaller than the probability of being correct classified [2]. In their analyses, the estimates from the parametric estimates were similar to those obtained using the semiparametric method, indicating that the parameter approach is reasonable for several applications. For this part of the dissertation, the modified maximum likelihood estimation method was adopted, which is described below. For better understanding let us consider as example, the mode choice.

Let *q,* and *i* be the indices for household and alternatives, respectively. Let *J* denote the total number of alternatives in the choice set (in the current empirical context, say $J = 3$ for example: car, transit, car-pool). For this part of the dissertation, we will assume that the alternatives are to be outcome of the utility maximization principle in the unordered modeling framework. Let $U_{q,i}, V_{q,i}, and\ \varepsilon_{q,i}$ denote the total, observed, and unobserved components of utility associated with alternative *i* for household *q*. In the utility framework, the probability that a person q chooses alternative *i* is given by:

$$\begin{aligned} P_q(i) &= P\left(U_{q,i} > U_{q,j}\right) \forall\ i \neq j \\ &= P\left(V_{q,i} + \varepsilon_{q,i} > V_{q,j} + \varepsilon_{q,j}\right) \forall\ i \neq j \\ &= \boldsymbol{P\left(\varepsilon_{q,j} - \varepsilon_{q,i} < V_{q,i} - V_{q,j}\right) \forall\ i \neq j} \end{aligned} \qquad \textbf{(3. 1)}$$

Assuming the stochastic utility components to be realizations from standard Gumbel distributions that are independent and identically distributed across alternatives and households will result in the standard multinomial logit (MNL) model. In the absence of misclassification, the probability of alternative *i* is given by:

$$\boldsymbol{P_q(i) = \frac{e^{V_{q,i}}}{\sum_{j=1}^{J} e^{V_{q,j}}}} \quad \textbf{(3. 2)}$$

Let $\alpha_{s,t}$ denote the probability that alternative *s* is misclassified as alternative *t.* Any given alternative *s* can be classified as one of the *J* alternatives, so $\sum_{t=1}^{J} \alpha_{s,t} = 1$. Now, if *i* is the observed dependent variable, then the true latent response can be any of the *J* alternatives. So, the probability of observed dependent variable *i* under misclassification is given by:

$$\boldsymbol{P_q(i) = \sum_{t=1}^{J} \alpha_{t,i} \times \frac{e^{V_{q,t}}}{\sum_{j=1}^{J} e^{V_{q,j}}}} \quad \textbf{(3. 3)}$$

In the current simulation context with three alternatives, the misclassification matrix is given by:

$$\begin{vmatrix} \textbf{Best Estimated} \downarrow \textbf{|| Observed} \rightarrow & \textbf{One} & \textbf{Two} & \textbf{Three} \\ \textbf{One} & \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} \\ \textbf{Two} & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} \\ \textbf{Three} & \alpha_{3,1} & \alpha_{3,2} & \alpha_{3,3} \end{vmatrix} \quad \textbf{(3. 4)}$$

The diagonal elements in the above matrix indicate the probability that the observed and the correct response variable are the same or the probability of correct classification. Any observed response *s* in the survey data may be because of misclassification (i.e., the chosen alternative was some other alternative *t* but was misclassified as *s*) or due to correct classification. The intuitive meaning of the sufficiency conditions for consistency is that the probability of observed data being correct must be larger than the probability of being misclassified. If these sufficiency conditions fail, the parameter estimates in the model can have opposite signs from a model that ignores misclassification, and there is little hope in recovering the true parameters consistently [2]. Mathematically, the sufficiency condition translates into the following equation for alternative *s*:

$$\boldsymbol{\sum_{\substack{t=1 \\ t \neq s}}^{J} \alpha_{t,s} < \alpha_{s,s} \forall\, s \in [1, J]} \quad \textbf{(3. 5)}$$

Adding $\sum_{\substack{t=1 \\ t \neq s}}^{J} \alpha_{s,t}$ to both sides of Equation (5) gives the following result:

$$\boldsymbol{\sum_{\substack{t=1 \\ t \neq s}}^{J} \alpha_{t,s} + \sum_{\substack{t=1 \\ t \neq s}}^{J} \alpha_{s,t} < \alpha_{s,s} + \sum_{\substack{t=1 \\ t \neq s}}^{J} \alpha_{s,t} \; \forall\, s \in [1, J]} \quad \textbf{(3. 6)}$$

But, $\alpha_{s,s} + \sum_{\substack{t=1 \\ t \neq s}}^{J} \alpha_{s,t} = \sum_{t=1}^{J} \alpha_{s,t} = 1$. So, the sufficiency condition is equivalent to:

$$\boldsymbol{\sum_{\substack{t=1 \\ t \neq s}}^{J} \alpha_{t,s} + \sum_{\substack{t=1 \\ t \neq s}}^{J} \alpha_{s,t} < 1 \;\; \forall\, s \in [1, J]} \quad \textbf{(3. 7)}$$

For the current empirical application, these sufficiency conditions are:

$$\boldsymbol{(\alpha_{2,1} + \alpha_{3,1}) + (\alpha_{1,2} + \alpha_{1,3}) < 1} \quad \textbf{(3. 8 a)}$$

$$(\alpha_{1,2} + \alpha_{3,2}) + (\alpha_{2,1} + \alpha_{2,3}) < 1 \quad \textit{(3.8 b)}$$

$$(\alpha_{1,3} + \alpha_{2,3}) + (\alpha_{3,1} + \alpha_{3,2}) < 1 \quad \textit{(3.8 c)}$$

3.2 Synthetic Data Generation

Going further with the demonstrations provided in the following chapters, we undertook a simulation study to evaluate the performance of the misclassification models using synthetic data. More details of the simulation set up and the results of this analysis are outlined in this section. The choice conditions considered in the analysis have three alternatives, among which one of the alternatives is considered probabilistically during decision making. The number of independent variables is three, and all these variables were drawn from linear functions of independent normal distributions. The data generation process is designed to create synthetic data that is close to real-world mode choice data with three independent variables and three alternatives: travel time, travel cost, and headway, and car, transit, and carpool, respectively.

Table 1. Synthetic data generation for our three alternatives scenario

Alternate	Travel Time (min)	Travel Cost (in $)	Headway (in min)
Car	7+30+UNIFORM (0,1)	2.0+15+UNIFORM (0,1)	0
Transit	12+30+UNIFORM (0,1)	1.0+10+UNIFORM (0,1)	30+120+UNIFORM (0,1)
Car-pool	9+30+UNIFORM (0,1)	1.5+10+UNIFORM (0,1)	60+120+UNIFORM (0,1)

The mean parameter vector considered for the three independent variables is b = (1, -1, -0.5); additionally the observed parts of the utilities include alternate specific constants (ASCs) given by ASC = (0, -0.4, -0.3). The ASC corresponding to the first alternative is fixed to zero (also during model estimation) because only utility differences matter, and we chose the first alternative as the base alternative. Also, along with the mean parameter vector and the alternate specific constant, the misclassification matrix was also given:

$$\begin{vmatrix} \textit{Best Estimated} \downarrow || \textit{Observed} \rightarrow & \textit{One} & \textit{Two} & \textit{Three} \\ \textit{One} & 0.000 & -1.000 & -2.000 \\ \textit{Two} & -1.000 & 0.000 & -2.000 \\ \textit{Three} & -1.000 & -2.000 & 0.000 \end{vmatrix} \quad (3.9)$$

Synthetic data were generated assuming that the correct data generation is a Multinomial Logit Model while accounting for misclassification. Any given alternative *s* can be classified as one of

the J alternatives, so $\sum_{t=1}^{J} \alpha_{s,t} = 1$. Now, if i is the observed dependent variable, then the true latent response can be any of the J alternatives.

The consideration probability of each variable was obtained using equation 3.4. For each observation record, the observed outcome was generated and compared to $\boldsymbol{P_q(i)}$,in order to determine if the first alternative was considered during decision making. To be specific, if UNIFORM (0, 1) <$\boldsymbol{P_q(i)}$, then the choice set does not include car, if not all of the three alternatives are considered.

3.2 Model Estimation

We consider the Multinomial Logit Model while accounting for misclassification and Multinomial Logit Model that does not account for misclassification in this comparative analysis. Using the maximum likelihood (ML) inference approach, all parameters were estimated, and all model estimation work was undertaken using Gauss programming language. The mean $\bar{\beta}$ and the standard deviation σ of each parameter were computed using the estimation results of the 100 synthetic data sets. The performance of models were evaluated using two metrics: absolute percentage bias (APB) obtained by taking the absolute value of ($\frac{\beta-\bar{\beta}}{\beta}$) x 100 and the t-statistic calculated as ($\frac{\beta-\bar{\beta}}{\sigma}$), where β is the true value of the parameter used during data generation. While APB values indicate the extent of bias, the t-statistic indicates whether the bias is statistically significant (i.e., whether parameter estimates are significantly different from their corresponding true values in the MNL model).

3.3 Simulation Results

Table 3. present the results of the simulation analysis for the MNL model accounting for misclassification and without misclassification scenarios, respectively. In Table 2, it can be seen that the mean APB value of the model and t-statistic are confirming that the MNL model, which accounts for misclassification, and that this is a correct method for modeling latent choice sets. Table 3 presents the estimation results of the model that doesn't account for misclassification. It can be observed that the mean APB value of the model parameters is quite high. Also, the t-statistics of comparison between the correct values and mean estimates of over 100 synthetic datasets show that the settings estimates are significantly different from their corresponding correct values.

Table 2.Simulation results MNL considering misclassification

Variable	True parameter	Mean	Absolute Percentage Bias	SE	T Stat
Constant	-0.400	-0.389	2.875	0.320	0.036
Constant	-0.300	-0.169	43.567	0.208	0.628
First Variable	1.000	0.924	7.650	0.238	-0.322
Second Variable	-1.000	-0.914	8.560	0.209	0.409
Third Variable	-0.500	-0.664	32.720	0.170	-0.963
Additional misclassification parameters					
	0.000	0.000	0.000	0.000	0.000
	-1.000	-1.037	0.000	0.167	-0.219
	-2.000	-2.359	0.000	0.601	-0.598
	-1.000	-1.117	0.000	0.233	-0.504
	0.000	0.000	0.000	0.000	0.000
	-1.000	-0.888	0.000	0.156	0.722
	-2.000	-1.678	0.000	0.389	0.829
	-1.000	-0.994	0.000	0.221	0.028
	0.000	0.000	0.000	0.000	0.000
Mean APB			19.074	0.582	0.009

Table 3.Simulation Results MNL without misclassification

Variable	True parameter	Mean	Absolute Percentage Bias	SE	T Stat
Constant	-0.400	-0.050	87.625	0.046	7.670
Constant	-0.300	-0.155	48.467	0.045	3.203
First Variable	1.000	0.292	70.790	0.025	-28.660
Second Variable	-1.000	-0.295	70.490	0.025	28.771
Third Variable	-0.500	-0.232	53.680	0.024	11.421
Mean APB			331.052	0.164	22.405

CHAPTER 4

MISCLASSIFICATION IN TRAVEL SURVEYS AND IMPLICATIONS TO CHOICE MODELING: APPLICATION TO HOUSEHOLD AUTO OWNERSHIP

In this chapter methods available in the econometrics literature were used to quantify and assess the impact of misclassification errors in auto ownership choice data. To demonstrate that, the modified maximum likelihood estimation and latent class models were adopted.

4.1 Methodological Framework

Hausman and his colleagues developed both parametric (maximum likelihood estimator (MLE)) and semi-parametric (monotone rank estimator) methods to consistently estimate discrete choice models under misclassification. The sufficiency condition needed for consistency is that the probability of misclassification is less than the probability of correct classification [2, 55, 56]. While the semi-parametric estimator is quite robust, the MLE will provide consistent estimates only if the misclassification probabilities are modeled correctly [55]. However, in their analysis, the estimates from the parametric method were quite similar to those obtained using the semi-parametric method, indicating that the parametric approach is reasonable for several practical applications [1, 55]. For this part of the study, we adopted the modified maximum likelihood estimation method that is described below.

Let q and i be the indices for households and alternatives, respectively. Let J denote the total number of alternatives in the choice set (in the current empirical context, $J = 5$ - zero, one, two, three, and four or more cars). Researchers used both ordered and unordered modeling frameworks for analyzing auto ownership choices and found that data fit in both the modeling frameworks is reasonably close [57-59]. In this study, auto ownership choices are assumed to be the outcome of the utility maximization principle in the unordered modeling framework. However, the methodology presented below can be easily extended to the ordered modeling framework. Let $U_{q,i}, V_{q,i}, and\ \varepsilon_{q,i}$ denote the total, observed, and unobserved components of utility associated with alternative i for household q. In the utility framework, the probability that household q chooses alternative i is given by:

$$P_q(i) = P\left(U_{q,i} > U_{q,j}\right) \forall\ i \neq j$$

$$= P\left(V_{q,i} + \varepsilon_{q,i} > V_{q,j} + \varepsilon_{q,j}\right) \forall\, i \neq j$$

$$= \boldsymbol{P\left(\varepsilon_{q,j} - \varepsilon_{q,i} < V_{q,i} - V_{q,j}\right) \forall\, i \neq j} \qquad \textbf{(4. 1)}$$

Assuming the stochastic utility components ($\boldsymbol{\varepsilon_{q,j} - \varepsilon_{q,i}}$) to be realizations from standard Gumbel distributions that are independent and identically distributed across alternatives and households will result in the standard multinomial logit (MNL) model. In the absence of misclassification, the probability of alternative *i* is given by:

$$\boldsymbol{P_q(i) = \frac{e^{V_{q,i}}}{\sum_{j=1}^{J} e^{V_{q,j}}}} \qquad \textbf{(4. 2)}$$

Let $\alpha_{s,t}$ denote the probability that alternative *s* is misclassified as alternative *t.* Any given alternative *s* can be classified as one of the *J* alternatives, so $\sum_{t=1}^{J} \alpha_{s,t} = 1$. Now, if *i* is the observed dependent variable, then the true latent response can be any of the *J* alternatives. So, the probability of observed dependent variable *i* under misclassification is given by:

$$\boldsymbol{P_q(i) = \sum_{t=1}^{J} \alpha_{t,i} \times \frac{e^{V_{q,t}}}{\sum_{j=1}^{J} e^{V_{q,j}}}} \qquad \textbf{(4. 3)}$$

In the current empirical context with five alternatives, the misclassification matrix is given by:

$$\begin{bmatrix} \textbf{Best Estimated} \downarrow \textbf{||Observed} \rightarrow & \textbf{Zero} & \textbf{One} & \textbf{Two} & \textbf{Three} & \textbf{Four +} \\ \textbf{Zero} & \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} & \alpha_{1,5} \\ \textbf{One} & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} & \alpha_{2,4} & \alpha_{2,5} \\ \textbf{Two} & \alpha_{3,1} & \alpha_{3,2} & \alpha_{3,3} & \alpha_{3,4} & \alpha_{3,5} \\ \textbf{Three} & \alpha_{4,1} & \alpha_{4,2} & \alpha_{4,3} & \alpha_{4,4} & \alpha_{4,5} \\ \textbf{Four +} & \alpha_{5,1} & \alpha_{5,2} & \alpha_{5,3} & \alpha_{5,4} & \alpha_{5,5} \end{bmatrix} \qquad \textbf{(4. 4)}$$

The diagonal elements in the above matrix indicate the probability that the observed and the correct response variable are the same or the probability of correct classification. Any observed response *s* in the survey data may be because of misclassification (*i.e.,* the chosen alternative was some other alternative *t* but was misclassified as *s* or due to correct classification. The intuitive meaning of the sufficiency conditions for consistency is that the probability of observed data being correct must be larger than the probability of being misclassified. If these sufficiency conditions fail, the parameter estimates in the model can have opposite signs from a model that ignores misclassification, and there is little hope in recovering the true parameters consistently [2]. Mathematically, the sufficiency condition translates into the following equation for alternative *s*:

$$\boldsymbol{\sum_{\substack{t=1 \\ t \neq s}}^{J} \alpha_{t,s} < \alpha_{s,s} \forall\, s \in [1, J]} \qquad \textbf{(4. 5)}$$

Adding $\sum_{\substack{t=1 \\ t\neq s}}^{J} \alpha_{s,t}$ to both sides of Equation (4.5) gives the following result:

$$\sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{t,s}} + \sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{s,t}} < \boldsymbol{\alpha_{s,s}} + \sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{s,t}} \; \forall \; \boldsymbol{s} \in [\mathbf{1}, \boldsymbol{J}] \qquad \textbf{\textit{(4. 6)}}$$

But, $\alpha_{s,s} + \sum_{\substack{t=1 \\ t\neq s}}^{J} \alpha_{s,t} = \sum_{t=1}^{J} \alpha_{s,t} = 1$. So, the sufficiency condition is equivalent to:

$$\sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{t,s}} + \sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{s,t}} < \mathbf{1} \forall \; \boldsymbol{s} \in [\mathbf{1}, \boldsymbol{J}] \qquad \textbf{\textit{(4. 7)}}$$

For the current empirical application, these sufficiency conditions are:

$$\left(\boldsymbol{\alpha_{2,1}} + \boldsymbol{\alpha_{3,1}} + \boldsymbol{\alpha_{4,1}} + \boldsymbol{\alpha_{5,1}}\right) + \left(\boldsymbol{\alpha_{1,2}} + \boldsymbol{\alpha_{1,3}} + \boldsymbol{\alpha_{1,4}} + \boldsymbol{\alpha_{1,5}}\right) < \mathbf{1} \qquad \textbf{\textit{(4.8 a)}}$$

$$\left(\boldsymbol{\alpha_{1,2}} + \boldsymbol{\alpha_{3,2}} + \boldsymbol{\alpha_{4,2}} + \boldsymbol{\alpha_{5,2}}\right) + \left(\boldsymbol{\alpha_{2,1}} + \boldsymbol{\alpha_{2,3}} + \boldsymbol{\alpha_{2,4}} + \boldsymbol{\alpha_{2,5}}\right) < \mathbf{1} \qquad \textbf{\textit{(4.8 b)}}$$

$$\left(\boldsymbol{\alpha_{1,3}} + \boldsymbol{\alpha_{2,3}} + \boldsymbol{\alpha_{4,3}} + \boldsymbol{\alpha_{5,3}}\right) + \left(\boldsymbol{\alpha_{3,1}} + \boldsymbol{\alpha_{3,2}} + \boldsymbol{\alpha_{3,4}} + \boldsymbol{\alpha_{3,5}}\right) < \mathbf{1} \qquad \textbf{\textit{(4.8 c)}}$$

$$\left(\boldsymbol{\alpha_{1,4}} + \boldsymbol{\alpha_{2,4}} + \boldsymbol{\alpha_{3,4}} + \boldsymbol{\alpha_{5,4}}\right) + \left(\boldsymbol{\alpha_{4,1}} + \boldsymbol{\alpha_{4,2}} + \boldsymbol{\alpha_{4,3}} + \boldsymbol{\alpha_{4,5}}\right) < \mathbf{1} \qquad \textbf{\textit{(4.8 d)}}$$

$$\left(\boldsymbol{\alpha_{1,5}} + \boldsymbol{\alpha_{2,5}} + \boldsymbol{\alpha_{3,5}} + \boldsymbol{\alpha_{4,5}}\right) + \left(\boldsymbol{\alpha_{5,1}} + \boldsymbol{\alpha_{5,2}} + \boldsymbol{\alpha_{5,3}} + \boldsymbol{\alpha_{5,4}}\right) < \mathbf{1} \qquad \textbf{\textit{(4.8 e)}}$$

The analyst must estimate $J \times (J-1)$ additional parameters to account for misclassification. However, in the current empirical context, all the auto ownership alternatives are ordered. So, it is expected that the misclassification probability $\alpha_{s,t}$decreases considerably as $|s-t|$ increases. So, several entries of the misclassification matrix are expected to be zero.

It can be seen that the misclassification probabilities in matrix 4.4 do not vary across household *q*. The model can be generalized to allow misclassification probabilities to differ across different demographic population segments. For instance, two sets of misclassification probabilities can be estimated separately for low and high-income households. The sufficiency conditions in Equation (3.7) must hold within low and high-income households separately (but not necessarily across the entire population). However, the number of misclassification parameters can explode easily as the number of segments increases. To avoid this problem, we used the latent class modeling approach that probabilistically assigns each household to latent segments each with its own set of misclassification probabilities. The probability of belonging to each latent segment can be specified as a function of several household socio-demographics. Dustmann and van Soest [60] and Sullivan [61] used similar methods to allow misclassification probabilities to differ across population groups.

Let l denote the index for latent class, and L indicate the total number of latent classes in the population. The conditional probability of observed dependent variable i for household q with observed utility $V_{q,i}^{l}$ for segment l i,s given by:

$$\boldsymbol{P_q^l(i) = \sum_{t=1}^{J} \alpha_{t,i}^{l} \times \frac{e^{V_{q,t}^{l}}}{\sum_{j=1}^{J} e^{V_{q,j}^{l}}}} \quad \textbf{(4.9)}$$

The sufficiency conditions for MLE to provide consistent estimates are as follows:

$$\boldsymbol{\sum_{\substack{t=1 \\ t\neq s}}^{J} \alpha_{t,s}^{l} + \sum_{\substack{t=1 \\ t\neq s}}^{J} \alpha_{s,t}^{l} < 1 \forall\ s \in [1, J]\ and\ \forall\ l \in [1, L]} \quad \textbf{(4. 10)}$$

where $\alpha_{s,t}^{l}$ is the probability that alternative s is misclassified as alternative t in segment l.

Lastly, the unconditional probability of observed dependent variable i is given by:

$$\boldsymbol{P_q(i) = \sum_{l=1}^{L} P_q^l(i) \times w_q(l)\ where \sum_{l=1}^{L} w_q(l) = 1} \quad \textbf{(4.11)}$$

During model estimation, $\alpha_{t,s}^{l}$ and $w_q(l)$ were parameterized to ensure that $\sum_{t=1}^{J} \alpha_{s,t}^{l} = 1$ and $\sum_{l=1}^{L} w_q(l) = 1$. For instance, the diagonal elements of the misclassification matrix were not estimated but calculated as: $1 - \sum_{\substack{t=1 \\ t\neq s}}^{J} \alpha_{s,t}^{l}$. Similarly, $w_q(l)$ was parameterized as a function of household socio-demographics $\boldsymbol{Z_q^l}$ using a multinomial logit formulation as follows:

$$\boldsymbol{w_q(l) = \frac{e^{\gamma_l' z_q^l}}{\sum_{r=1}^{L} e^{\gamma_r' z_q^r}},} \quad \textbf{(4. 12)}$$

where: $\boldsymbol{\gamma_l'}$ is the parameter vector corresponding to $\boldsymbol{Z_q^l}$ and all elements of $\boldsymbol{\gamma_l'}$ for one of the segments are normalized to zero for identification.

Several metrics can be computed to characterize the population belonging to different segments [62]. For instance, the mean value $\bar{z}^{l}$ (within each segment) of each attribute z_q^l that determine segment membership can be computed as:

$$\boldsymbol{\bar{z}^l = \frac{\sum_q w_q(l) \times z_q^l}{\sum_q w_q(l)}} \quad \textbf{(4.13)}$$

Also, the size of each segment l can be obtained by summing the latent class probabilities $w_q(l)$ across all households as $\sum_q w_q(l)$. Next, the share R_l of the segment, l was computed by dividing $\sum_q w_q(l)$ by the total number of households in the sample.

Lastly, the shares of different auto ownership alternatives can be calculated as:

$$\boldsymbol{S_l(i) = \frac{\sum_q P_q^l(i) \times w_q(l)}{\sum_q \sum_{l=1}^{L} w_q(l)}\ and\ S(i) = \sum_{l=1}^{L} R_l \times S_l(i)} \quad \textbf{(4. 14)}$$

where $S_l(i)$ and $S(i)$ are shares of alternative i in segment l and the entire sample, respectively.

4.2 Empirical Application

Household auto ownership decisions are critical determinants of several short-term travel choices that household members make on a day-to-day basis. Understandably, most travel demand models have an explicit model to predict the household auto ownership levels that are subsequently used as an explanatory variable in several downstream models. In some cases, the auto ownership variable is also used to constrain the choice set of downstream choices instead of being used as a more explanatory variable. For instance, mode choice models of non-working household members who do not have a car (because the household vehicle was taken by the working member in the household) exclude drive alone auto mode completely. So, any errors in the auto ownership forecasts can propagate downstream through the entire modeling system.

Given this particular importance associated with auto ownership decisions, the latent class modeling framework described in the methodology section was used to explore, quantify, and assess the impact of misclassification errors of auto ownership responses in household travel surveys. The data for the analysis was obtained from the Southern California Household Travel Survey (HTS) that collected detailed activity and travel diary information from a representative sample of 35,000 households. This dataset was recently used to analyze auto ownership decisions using latent choice set Manski model [57]. After excluding records with missing information on explanatory variables considered in this study, the data size was reduced to about 30,000 households. The frequency distribution of the dependent variable of analysis in this study, auto ownership, was zero cars (7.7%), one car (31.3%), two cars (40.3%), three cars (14.5%), and four or more cars (6.2%). The relatively lower percentages for the extreme alternatives – zero cars and four or more cars – may be due to misclassification. For instance, low shares of zero cars may be due to households with one car underreporting or households with zero cars over-reporting.

4.2.1 Statistical Fit Comparison

Several models were developed for this dissertation including MNL, MNL with misclassification (‘MNL MC’), latent class MNL (‘LC MNL’), and latent class MNL with two sets of misclassification probabilities (‘LC MNL MC’). For brevity, only the results of the “LC MNL MC” models are presented in the Table 45 along with the misclassification probabilities in ‘MNL MC’ model for comparison purposes. Although the ‘MNL MC’ model has nine misclassification probabilities, only six of them were estimated because diagonal elements were

obtained using the constraint that each row in the misclassification matrix must add up to 1. By the same logic, only nine additional parameters were estimated in the 'LC MNL MC' model compared to the 'LC MNL' model. The likelihood ratio (LR) test of comparison between the MNL (log-likelihood, LL =-28,235) and 'MNL_MC' (-28,141) models was 188.89, which is greater than the critical chi-squared statistic of 12.59 corresponding to six degrees of freedom at 95% confidence level. This suggests that the model that accounts for misclassification is statistically better than the standard MNL model. While the latent class model with two segments could be estimated, our attempts to estimate models with more than two segments were not successful due to convergence problems even with the expectation maximization (EM) algorithm. So, the latent class model with two classes ('LC MNL') was adopted for subsequent analysis. Given that the MNL and 'LC MNL' models are non-nested, they cannot be compared using the LR test. So, Bayesian Information Criterion (BIC) computed as $-2 \times LL + k \times LN(N)$, where k is the number of parameters and N is the sample size, was used to compare the two models. Between two non-nested models, a model with lower BIC value is preferred over the other model. The BIC values for the MNL and 'LC MNL' models are 57,110 and 56,836, respectively suggesting superior data fit in the 'LC MNL' model. Lastly, the 'LC MNL MC' that accounts for different misclassification errors in the two latent segments has nine additional parameters compared to the 'LC MNL' model. The LR test statistic of comparison between the 'LC MNL' (log-likelihood, LL = -27,887) and 'LC MNL MC' (-27,837) models is 100.67, which is greater than the critical chi-squared statistic of 16.92 corresponding to nine degrees of freedom at a 95% confidence level. Overall, the 'LC MNL MC' model that allows misclassification rates to vary between the two segments was found to the best model in this study.

4.2.2 Misclassification Errors in Un-Segmented Model

Table 4 presents the results of the misclassification components for the 'MNL MC' and 'LC MNL MC' models. All non-zero non-diagonal elements in the three matrices were statistically different from 0 at a 95% confidence level. The results in Table 4.a for the 'MNL MC' model show significant misclassification of extreme alternatives compared to intermediate alternatives. Specifically, 25.2% of 'zero cars' responses were wrongly classified as 'one car', whereas 29.2% and 7.7% of 'four or more cars' responses were wrongly classified as 'three cars' and 'two cars,' respectively. There was no evidence for misclassification for the 'two cars' alternative, which is

also the most common auto ownership level in the data. Lastly, 18.4% of 'three cars' responses were wrongly classified as 'four or more cars' alternative.

4.2.3 Latent Class Model and Misclassification Errors

The 'MNL MC' model restricts that misclassification probabilities are the same for all households. To relax this assumption, the 'LC MNL MC' model was estimated. The results corresponding to the latent class membership component in 5 indicate that high-income households, households with more workers, owner-occupied households, single-family detached households, and households with fewer senior adults aged 80 years and above are more likely to belong to the first segment. The mean values of attributes within each segment were computed using Equation 4.13 and shown in Table 45. These mean values are consistent with the parameter signs and the earlier interpretations of the latent class membership component. Also, 43% of households were found to belong to the first segment, whereas the remaining 57% of households belonged to the second segment. The shares of different auto ownership levels in 5 indicate that auto ownership levels tend to be higher in the first segment, whereas they are skewed towards the lower end in the second segment. For instance, the shares of extreme auto ownership levels ('zero cars' and 'four or more cars') are almost flipped in the two segments: (2.4%, 11.4%) in the first segment and (11.6%, 2.2%) in the second segment.

Table 4.Misclassification Probabilities

Table 4a Misclassification in Un-Segmented Model					
Best Estimated ↓ ||Observed →	Zero	One	Two	Three	Four +
Zero	0.7477	0.2523	0.0000	0.0000	0.0000
One	0.0200	0.9471	0.0329	0.0000	0.0000
Two	0.0000	0.0000	1.0000	0.0000	0.0000
Three	0.0000	0.0000	0.0000	0.8157	0.1843
Four or more	0.0000	0.0000	0.0766	0.2928	0.6306
Table 4b Misclassification in Segment 1					
Best Estimated ↓ ||Observed →	Zero	One	Two	Three	Four +
Zero	1.0000	0.0000	0.0000	0.0000	0.0000
One	0.0000	0.9448	0.0552	0.0000	0.0000
Two	0.0000	0.0000	1.0000	0.0000	0.0000

Table 4. Continued

Best Estimated↓||Observed→	Zero	One	Two	Three	Four +
Three	0.0000	0.0000	0.0000	0.8951	0.1049
Four or more	0.0000	0.0000	0.0000	0.0000	1.0000
Table 4c Misclassification in Segment 2					
Best Estimated ↓ ||Observed →	Zero	One	Two	Three	Four +
Zero	0.8703	0.1297	0.0000	0.0000	0.0000
One	0.0140	0.9308	0.0552	0.0000	0.0000
Two	0.0000	0.0000	0.9651	0.0349	0.0000
Three	0.0000	0.0000	0.0000	0.7967	0.2033
Four or more	0.0000	0.0000	0.2487	0.4019	0.3494

Table 5. Latent Segmentation Component

Explanatory Variable	**Segment 2 (Base: Segment 1)**		**Mean Attribute Value**	
	Parameter	**T-Stat**	**Segment 1**	**Segment 2**
Constant	3.9543	15.920		
Household Income (Base: $35,000 or less)				
$35,001-$50,000	-0.7066	-4.743	0.10	0.15
$50,001-$100,000	-1.3177	-8.979	0.38	0.26
>$100,000	-2.0158	-11.415	0.41	0.12
Ratio of workers to driving age adults	-4.0542	-14.195	0.78	0.32
Ratio of adults 80 years or older to driving age adults	1.7288	6.398	0.03	0.13
Housing Type (Base category: Mobile & Other)				
Single family detached household	-0.3706	-3.018	0.74	0.61
Owner-occupied household	-0.2191	-1.496	0.77	0.64
Size of Segment			43%	57%
Mode Shares within Segment				
Zero Cars			2.37	11.64
One Car			18.89	40.51
Two Cars			46.06	36.22
Three Cars			21.29	9.44
Four or More Cars			11.38	2.20

Moreover, the misclassification probabilities were found to be different for the two latent population segments (see 4b and 4c). Interestingly, there were no misclassification errors for extreme alternatives in the first population segment. Also, respondents in the first segment were found to over-report auto ownership levels. Specifically, 5.5% of 'one car' and 10.5% of 'three cars' responses were wrongly classified as 'two cars' and 'four or more cars', respectively. On the contrary, the misclassification errors of extreme alternatives were significant in the second segment. For instance, only 34.9% of 'four or more cars' responses were correctly classified with more than 40%, and 25% responses wrongly classified as 'three cars' and 'two cars,' respectively. These misclassification errors in the second segment are significantly higher than the errors in the un-segmented 'MNL MC' model. Other misclassification probabilities for intermediate alternatives were also higher compared to corresponding misclassification probabilities in 'MNL MC' model except 'one car' responses.

Interestingly, while the un-segmented model found no evidence for misclassification error in the 'two cars' alternative, 3.5% of 'two cars' responses were found to be wrongly classified as 'three cars' in the second segment. These results suggest that not only is misclassification significant, but that it also varies across different population segments. Lastly, it can be seen that all the misclassification probabilities in both the un-segmented and segmented models satisfy the sufficiency conditions in Equation 8. For instance, for the fifth alternative, 'four or more cars' in the second segment, the sufficiency condition is $\left(\alpha_{1,5} + \alpha_{2,5} + \alpha_{3,5} + \alpha_{4,5}\right) + \left(\alpha_{5,1} + \alpha_{5,2} + \alpha_{5,3} + \alpha_{5,4}\right) < 1$ which is equivalent to (0.0000+0.0000+0.0000+0.2033) + (0.0000+0.0000+0.2487+0.4019) < 1 or 0.8539 < 1 which is true.

4.2.4 Utility ComponentTable 5, presents the results of the utility specification for the two latent segments. The 'two cars' alternative was chosen as the base alternative in the utility specification of both segments. In some cases, an alternate specification based on 'household auto-sufficiency' was used. Auto sufficiency is an alternate-specific variable with three categories (excluding the zero cars alternative) –low (fewer cars than driving age (16 years) adults), equal (same number of cars as driving age adults), and high (more cars than driving age adults). For each variable, both the standard way (where the 'two cars' alternative was chosen as the reference alternative) and as an interaction with the auto-sufficiency variable were tested and the specification that provided better data fit was chosen. Also, the constants in the utility specification

were segmented by the number of driving age adults in the household. Given that there are several continuous variables in the utility specification, there is no clear behavioral interpretation for the constants.

4.2.5 First Latent Segment Utility Component

Households with more senior adults aged 80 years and above are more likely to choose high auto-sufficiency levels. Higher income levels were associated with higher auto ownership levels beyond two, whereas lower income levels were associated with lower inclination to own less than two cars. Households with higher educational attainment (associate's degree and higher) are less likely to own less than two cars compared to households with an education attainment of high school degree and lower. Single-family detached and non-rental households are less likely to own fewer cars, whereas non-rental households tend to own more than two cars. Households residing in residential neighborhoods with high household density are more likely not to own a car as well as less likely to own more than two cars. Also, a higher percentage of residence zones in high-quality transit areas (HQTA) and transit priority areas (TPA) are associated with lower auto ownership levels. Lastly, households in high transit accessibility neighborhoods are less inclined to choose higher auto ownership levels. Interestingly, auto ownership preferences of households in the first segment were not related to the average commute distance within the household.

4.2.6 Second Latent Segment Utility Component

Households with more workers, pre-driving age children (<16 years), and senior adults (aged 65-79 years) are more inclined to own at least one car. Also, households with more senior adults aged 65-79 years are less likely to choose high auto-sufficiency levels. Households with more senior adults 80 years and above are less inclined to choose high sufficiency alternatives and more likely to choose zero and low sufficiency levels compared to equal sufficiency alternatives. Similar to the results in the first segment, higher (lower) income levels were associated with higher (lower) auto ownership levels. Households with higher educational attainment are less inclined to own fewer than two cars. While households in single-family detached households are less inclined to own fewer cars, households in apartments have different auto ownership preferences. Also, owner-occupied households tend to own three cars and less inclined to own fewer than two cars compared to rental households. High residential density, high bus stop density, and a high percentage of HQTA and TPA were associated with lower auto-sufficiency levels. Interestingly, transit accessibility was not found to influence auto ownership choices for households in the

second segment. Lastly, higher average commute distance was found to be associated with a lower likelihood of owning fewer than two cars. This does not imply causality because it is possible that households with more cars chose to reside in suburban neighborhoods with longer commutes.

4.3 Elasticity Effects Analysis

To quantify the impact of ignoring misclassification errors on parameter estimates and model forecasts, elasticity effects that indicate the percentage change in the shares of different auto ownership levels for a unit change in an explanatory variable were computed. First, market shares of different auto ownership alternatives were computed in the base scenario using Equation 4.14. Next, the market shares were recomputed in the policy scenario using the same equation but with a unit increase in the variable for which the elasticity is being calculated. The unit change is 0 to 1 in the case of dummy variables such as high-income indicator variable and one unit increment in case of ordinal variables such as workers indicator variable. Table 7 present the results of the elasticity analysis for the 'LC MNL' and 'LC MNL MC' models. The last column presents the absolute difference between the elasticity effects of the two models. The first number under the column 'LC MNL MC' indicates that households with income between $35,001 and $50,000 are 57.3% less likely not to own cars compared households with income less than $35,000. Other numbers in the table can be interpreted similarly. For instance, households with one additional worker are, on average, 27.6% more likely to own four or more cars. The elasticity estimates of the two models are quite similar for the 'one car' and 'two cars' alternatives. This is consistent with the fact that the misclassification probabilities of these intermediate alternatives are relatively low. Alternatively, the elasticity effects of extreme alternatives (four or more cars, three cars, and zero cars) can differ significantly, which is again consistent with higher misclassification errors associated with these alternatives. For instance, the elasticity effects of the variable representing the household income with 'four or more cars' alternative differ by up to 75 percentage points. These results suggest that misclassification errors can result in biased parameter estimates, leading to incorrect model forecasts and policy sensitivity.

Table 6. Utility Component

Explanatory Variables	**Generic Parameters**		**Parameters Specific to Choice Alternatives**							
			0		**1**		**3**		**4+**	
	Seg 1	**Seg 2**	**Seg 1**	**Seg 2**	**Seg 1**	**Seg 2**	**Seg 1**	**Seg 2**	**Seg 1**	**Seg 2**
Household Demographic & Socio-Economic Variables										
Number of Driving Age Adults in the Household										
One			3.304	3.446	3.919	2.595	-1.406	-3.468	-4.004	-0.986
Two					-1.614	0.910	-1.862	-3.510	-5.597	-2.493
Three							1.531	-2.508	-2.267	
Four or more							3.009	-1.388		
Household Income (Base: Household income $35,000 or less)										
$35,001-$50,000			-2.052	-1.702	-1.046	-0.611				
$50,001-$100,000			-2.052	-2.213	-1.046	-0.907			0.599	
>$100,000			-3.105	-3.016	-1.443	-1.415	0.299	0.982	1.161	
Housing Type (Base category: Mobile & Other)										
Single family detached household			-1.093	-0.227	-0.799					
Single family attached household										
Multi-family household				0.518		0.272				
Highest Educational Attainment (Base: Less than high school)										
High school				-0.660						
Associate degree			-1.526	-1.095	-0.422	-0.182		0.377		
Bachelor degree			-1.561	-1.552		-0.423				
Graduate degree			-1.344	-1.909		-0.358				

Table 6. Continued

Explanatory Variables	**Generic Parameters**		**Parameters Specific to Choice Alternatives**							
			0		**1**		**3**		**4+**	
	Seg 1	**Seg 2**	**Seg 1**	**Seg 2**	**Seg 1**	**Seg 2**	**Seg 1**	**Seg 2**	**Seg 1**	**Seg 2**
Owner-occupied household			-1.423	-1.731	-0.347	-0.468	0.469	1.124	1.587	
Average commute distance (in miles/100)				-6.278		-0.828				
Ratio of workers to driving age adults				-1.563						
Low sufficiency	0.691									
Ratio of pre-driving age children to driving age adults				-0.393						
Ratio of adults 65-79 years or older to driving age adults				-0.455						
High sufficiency		-0.406								
Ratio of adults 80 years or older to driving age adults				0.387						

Table 7. Elasticity Effects

Explanatory Variable	**Alternative**	**LC MNL MC**	**LC MNL**	**Absolute Difference**
Household Income $35,001-$50,000	Zero	-57.3%	-57.6%	0.3%
	One	-11.4%	-13.6%	2.1%
	Two	14.4%	15.1%	0.7%
	Three	13.5%	15.1%	1.6%
	Four or more	26.8%	19.9%	6.9%
Household Income $50,001-$100,000	Zero	-67.3%	-66.4%	0.8%
	One	-25.2%	-26.8%	1.6%
	Two	17.2%	19.0%	1.7%
	Three	41.4%	39.6%	1.8%
	Four or more	102.1%	70.2%	32.0%
Household Income >$100,000	Zero	-76.2%	-76.0%	0.3%
	One	-42.4%	-44.4%	1.9%
	Two	16.3%	19.9%	3.6%
	Three	77.1%	69.9%	7.2%
	Four or more	217.0%	141.9%	75.1%
Single family detached household	Zero	-7.3%	-8.2%	0.9%
	One	-12.3%	-13.2%	1.0%
	Two	7.4%	6.8%	0.7%
	Three	8.5%	11.6%	3.1%
	Four or more	16.4%	18.4%	1.9%
Owner-occupied household	Zero	-59.6%	-59.0%	0.6%
	One	-4.9%	-7.1%	2.2%
	Two	1.7%	2.4%	0.8%
	Three	46.3%	35.9%	10.3%
	Four or more	104.6%	108.7%	4.1%
Number of workers	Zero	-38.3%	-34.1%	4.2%

Table 7. Continued

Explanatory Variable	Alternative	LC MNL MC	LC MNL	Absolute Difference
	One	-1.6%	-4.7%	3.1%
	Two	0.8%	2.2%	1.4%
	Three	13.9%	13.3%	0.6%
	Four or more	27.6%	20.4%	7.2%
Number of senior adults aged 80 years and above	Zero	-5.1%	6.5%	11.7%
	One	13.7%	11.8%	1.9%
	Two	-5.4%	-5.8%	0.4%
	Three	-6.9%	-8.9%	2.0%
	Four or more	-12.5%	-9.0%	3.4%

4.4 Conclusion

Household Travel Survey (HTS) data is prone to several errors either due to intentional or unintentional misinformation provided by the respondents. Ignoring these errors while modeling travel decisions using standard discrete choice models can result in biased parameter estimates. In this study, methods available in the econometrics literature for handling misclassification were used to quantify and assess the impact of misclassification in travel survey data. Individually, misclassification in household auto ownership choices was analyzed using Southern California HTS. The auto ownership survey response was recorded into five categories – zero, one, two, three, and four or more cars and was modeled as an unordered discrete response variable. The results indicate that misclassification errors can be as high as 40%, particularly for the extreme auto ownership levels. Comparatively, the misclassification in 'one car' and 'two cars' alternatives was lower. However, un-segmented models restrict that misclassification rates are the same for the entire population. To relax this assumption, latent class auto ownership model that allows the misclassification probabilities to vary across different latent segments was developed. The empirical analysis uncovered two latent classes in the population about auto ownership preferences and also significant differences in the misclassification rates between the two segments. Statistical fit comparisons and elasticity analysis illustrate that models that ignore misclassification have not only worse data fit but also biased parameter estimates with significant policy implications.

The modern suite of advanced travel demand models including tour-based and activity-based models encompass several discrete choice models to predict daily activity and travel preferences of people. The underlying idea of these models is that people travel to participate in different types of events at locations dispersed in space and time. So, the critical response variables that form the basis of these models are activity purpose, activity duration, mode, departure time, and destination. All these responses in HTS data are prone to measurement errors and must be analyzed using similar modeling methods to those used in this study to quantify and assess the impact of misclassification on parameter estimates of respective choice models. Also, it is a common practice for researchers to collect their data to analyze new empirical contexts with limited revealed preference data. For example, several studies used web-based surveys that elicit preferences for new vehicle technologies including connected and autonomous vehicles and electric vehicles. It is a useful exercise to explore the quality of these survey responses by quantifying misclassification to demonstrate the validity and confidence of these study findings. Lastly, the models developed in this study can be applied to other transportation disciplines. For instance, in the transportation safety arena, police reported injury severity in crash databases is a key dependent variable that safety engineers analyze to explore the factors that determine the severity of a crash conditionally at the crash occurrence. These injury severity recordings are prone to errors either due to the subjectivity of classification or the stress that police are subjected to during crash incidents. Ignoring these errors can potentially result in over or under-estimation of critical variables such as seat belt effectiveness and alcohol involvement.

CHAPTER 5

A MODIFIED GENERALIZED ORDERED RESPONSE MODEL TO HANDLE MISCLASSIFICATION IN INJURY SEVERITY

The objective of this part of the study is to develop a statistical model to analyze police-reported injury severity while accounting for potential misclassification errors by building upon the existing literature in econometrics.

5.1 Methodological Framework

The ordered response (OR) framework assumes a single latent propensity function that is mapped into one of *J* ordered outcomes by *J-1* threshold parameters that are strictly ordered. The latent propensity function is specified as the sum of linear-in-parameters deterministic component (which is a function of observed attributes) and a random component (that represents all the unobserved factors that influence the ordered outcome). The specification of the OR model is completed by assuming a continuous probability density function for the random component. The two most commonly used assumptions for the density function of the unobserved part are the standard normal distribution (leading to the ordered response probit (ORP) model) and the standard logistic distribution (pointing to the ordered response logit (ORL) model) [63, 64]. Earlier applications of the OR models assumed constant threshold parameters that do not vary across observations. However, for the same reasons that the latent propensity function varies across observations, the threshold parameters can vary systematically across observations. This idea led to the formulation of the Generalized Ordered Response (GOR) framework that parameterized thresholds as a function of observation-specific attributes [65]. The next significant extension of the OR framework is capturing the unobserved heterogeneity of parameters in the propensity and threshold components, i.e., the effects of different observed attributes can vary across observations because of the moderating influence of unobserved factors not considered in the model [66]. Researchers developed the mixed GOR (MGOR) model that assumes the parameters in the propensity and threshold components to be stochastic realizations from multivariate probability density functions to address this problem [65, 67]. This study adopted the mixed generalized ordered probit (MGORP) framework for modeling injury severity outcomes conditional on crash occurrence.

Let q $(1,2,\ldots.Q)$ be the index for crash and $j(1,2,\ldots J)$ be the index for injury severity outcome. In the current context, $J = 4$ with the four ordered alternatives being no injury $(j = 1)$, possible injury $(j = 2)$, non-incapacitating injury $(j = 3)$, and incapacitating or fatal injury $(j = 4)$. In the OR framework, the latent propensity y_q^* is related to the $K \times 1$ vector of observed attributes $\boldsymbol{x}_q$ (including constant) as:

$$\boldsymbol{y_q^* = \beta_q' x_q + \varepsilon_q} \qquad \textit{(5. 1)}$$

where $\boldsymbol{\beta}_q$is the vector of parameters corresponding to the observed attributes $\boldsymbol{x}_q$ and ε_q is the stochastic component of propensity assumed to be a realization from the standard normal distribution, *i.e.* $\varepsilon_q \sim N(0,1)$. The subscript q to the parameter vector indicates unobserved heterogeneity across observations. The $\boldsymbol{\beta}_q$ vector is assumed to be a realization from a multivariate normal distribution with mean vector **b** and $K \times K$ covariance matrix**Σ,** *i.e.* $\boldsymbol{\beta}_q \sim N(\boldsymbol{b}, \boldsymbol{\Sigma})$. Equation (4.1) can now be re-written as follows:

$$\boldsymbol{y_q^* = b' x_q + \widetilde{\beta}_q' x_q + \varepsilon_q = \ b' x_q + \eta_q} \qquad \textit{(5. 2)}$$

where $\widetilde{\boldsymbol{\beta}}_q \sim N(\mathbf{0}_K, \boldsymbol{\Sigma})$ and $\mathbf{0}_K$ is a $K \times 1$ vector of zeros. The variance of the effective error term $\eta_q$is equal to $\boldsymbol{x}_q' \boldsymbol{\Sigma} \boldsymbol{x}_q + 1$.

The latent propensity y_q^* is mapped into ordinal outcomes by threshold parameters ψ_q^k as follows:

$$\boldsymbol{y_q = j \quad if \quad \psi_q^{j-1} < y_q^* < \psi_q^j} \qquad \textit{(5. 3)}$$

The strict monotonicity of thresholds is ensured by using the following parameterization:

$$\boldsymbol{\psi_q^k = \psi_q^{k-1} + exp(\gamma_{k,q}' z_{k,q}), \psi_q^0 = -\infty, \psi_q^J = \infty, and\ \psi_q^1 = exp(\alpha_1)} \qquad \textit{(5. 4)}$$

where $\boldsymbol{z}_{k,q}$ is $L_k \times 1$ vector of observed attributes affecting the k^{th} threshold and $\boldsymbol{\gamma}_{k,q}$ is the corresponding vector of coefficients, which is assumed to be a stochastic realization from a multivariate normal distribution with mean $\boldsymbol{c}_k$ and $L_k \times L_k$ covariance matrix $\boldsymbol{\Omega}_k$, i.e., $\boldsymbol{\gamma}_{k,q}' \sim N(\boldsymbol{c}_k, \boldsymbol{\Omega}_k)$. Please note that the observed attributes $\boldsymbol{x}_q$ can include a constant because in Equation 5.4 all the thresholds are constrained to be positive using the exponential parameterization.

Let $\boldsymbol{\theta}_q = \left(\boldsymbol{\gamma}'_{1,q}, \boldsymbol{\gamma}'_{2,q}, \ldots \boldsymbol{\gamma}'_{J-1,q}\right)'$ and $\boldsymbol{c} = \left(\boldsymbol{c}'_1, \boldsymbol{c}'_2, \ldots . \boldsymbol{c}'_{J-1}\right)'$ denote $\left(\sum_{k=1}^{J-1} L_k\right) \times 1$ vectors of vertically stacked parameters $\boldsymbol{\gamma}_{k,q}$ and $\boldsymbol{c}_k$. The probability of ordinal outcome *j* conditional on random parameter vectors $\boldsymbol{\gamma}_{k,q}$ in thresholds can be obtained as follows:

$$P(y_q = j|\boldsymbol{\theta}_q) = P\left(\psi_q^{j-1} < y_q^* < \psi_q^j\right) = P\left(\psi_q^{j-1} < \boldsymbol{b'x}_q + \eta_q < \psi_q^j\right)$$

$$= P\left(\psi_q^{j-1} - \boldsymbol{b'x}_q < \eta_q < \psi_q^j - \boldsymbol{b'x}_q\right)$$

$$= \boldsymbol{\Phi}\left(\frac{\psi_q^j - \boldsymbol{b'x}_q}{\sqrt{\boldsymbol{x}'_q \boldsymbol{\Sigma} \boldsymbol{x}_q + 1}}\right) - \boldsymbol{\Phi}\left(\frac{\psi_q^{j-1} - \boldsymbol{b'x}_q}{\sqrt{\boldsymbol{x}'_q \boldsymbol{\Sigma} \boldsymbol{x}_q + 1}}\right) \qquad \textbf{(5. 5)}$$

The unconditional probability of ordinal outcome *j* is obtained by integrating the random components $\boldsymbol{\gamma}_{k,q}$ as follows:

$$\boldsymbol{P}(\boldsymbol{y_q = j}) = \int_{\boldsymbol{\theta_q}} \left[\boldsymbol{\Phi}\left(\frac{\psi_q^j - \boldsymbol{b'x}_q}{\sqrt{\boldsymbol{x}'_q \boldsymbol{\Sigma} \boldsymbol{x}_q + 1}}\right) - \boldsymbol{\Phi}\left(\frac{\psi_q^{j-1} - \boldsymbol{b'x}_q}{\sqrt{\boldsymbol{x}'_q \boldsymbol{\Sigma} \boldsymbol{x}_q + 1}}\right)\right] \boldsymbol{f}(\boldsymbol{\theta_q}) \boldsymbol{d\theta_q} \qquad \textbf{(5. 6)}$$

where $f\left(\boldsymbol{\theta}_q\right)$ is the multivariate normal probability density function of $\boldsymbol{\theta}_q \sim N(\boldsymbol{c}, \boldsymbol{\Xi})$ and $\boldsymbol{\Xi}$ is a $\left(\sum_{k=1}^{J-1} L_k\right) \times \left(\sum_{k=1}^{J-1} L_k\right)$ block diagonal matrix with $\boldsymbol{\Omega}_k$ as the k^{th} diagonal matrix.

5.1.1 Modified Likelihood Function to Handle Misclassification

Let $\alpha_{s,t}$ denote the probability that ordinal alternative *s* is misclassified as ordinal alternative *t*. Any given alternative *s* can be classified as one of the *J* alternatives, so $\sum_{t=1}^{J} \alpha_{s,t} = 1$. Now, if *j* is the observed ordinal outcome, then the true latent response can be any of the *J* alternatives. So, the probability of observed ordinal outcome *j* under misclassification is given by:

$$\widetilde{\boldsymbol{P}}_{\boldsymbol{q}}(\boldsymbol{y_q = j}) = \sum_{\boldsymbol{t=1}}^{\boldsymbol{J}} \boldsymbol{\alpha_{t,i}} \times \boldsymbol{P}(\boldsymbol{y_q = j}) \qquad \textbf{(5. 7)}$$

In the current empirical context with four injury severity alternatives, the misclassification matrix is given by:

$$\begin{bmatrix} \boldsymbol{Best\ Estimated} \downarrow \boldsymbol{||Observed} \rightarrow & \boldsymbol{j=1} & \boldsymbol{j=2} & \boldsymbol{j=3} & \boldsymbol{j=4} \\ \boldsymbol{j=1, No\ injury} & \boldsymbol{\alpha_{1,1}} & \boldsymbol{\alpha_{1,2}} & \boldsymbol{\alpha_{1,3}} & \boldsymbol{\alpha_{1,4}} \\ \boldsymbol{j=2, Possible\ injury} & \boldsymbol{\alpha_{2,1}} & \boldsymbol{\alpha_{2,2}} & \boldsymbol{\alpha_{2,3}} & \boldsymbol{\alpha_{2,4}} \\ \boldsymbol{j=3, Non-Incapacitating\ Injury} & \boldsymbol{\alpha_{3,1}} & \boldsymbol{\alpha_{3,2}} & \boldsymbol{\alpha_{3,3}} & \boldsymbol{\alpha_{3,4}} \\ \boldsymbol{j=4, Incapacitating\ Injury} & \boldsymbol{\alpha_{4,1}} & \boldsymbol{\alpha_{4,2}} & \boldsymbol{\alpha_{4,3}} & \boldsymbol{\alpha_{4,4}} \end{bmatrix} \qquad \textbf{(5. 8)}$$

The diagonal elements in the above matrix indicate the probability that the observed and the true response variable are the same or the probability of correct classification. Any observed ordinal outcome *s* in the crash database may be because of misclassification (*i.e.,* the true severity outcome was some other alternative *t* but was misclassified as *s*) or due to correct classification. The intuitive meaning of the sufficiency conditions for consistency is that the probability of observed data being correct must be larger than the likelihood of being misclassified. If these sufficiency conditions fail, the parameter estimates in the model can have opposite signs from a model that ignores misclassification, and there is little hope in recovering the true parameters consistently [2]. Mathematically, the sufficiency condition translates into the following equation for alternative *s*:

$$\sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{t,s}} < \boldsymbol{\alpha_{s,s}} \forall\ \boldsymbol{s} \in [\mathbf{1}, \boldsymbol{J}] \qquad \textbf{(5. 9)}$$

Adding $\sum_{\substack{t=1 \\ t\neq s}}^{J} \alpha_{s,t}$ to both sides of Equation 9 gives the following result:

$$\sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{t,s}} + \sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{s,t}} < \boldsymbol{\alpha_{s,s}} + \sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{s,t}} \forall\ \boldsymbol{s} \in [\mathbf{1}, \boldsymbol{J}] \qquad \textbf{(5. 10)}$$

But, $\alpha_{s,s} + \sum_{\substack{t=1 \\ t\neq s}}^{J} \alpha_{s,t} = \sum_{t=1}^{J} \alpha_{s,t} = 1$. So, the sufficiency condition is equivalent to:

$$\sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{t,s}} + \sum_{\substack{t=1 \\ t\neq s}}^{J} \boldsymbol{\alpha_{s,t}} < \mathbf{1} \forall\ \boldsymbol{s} \in [\mathbf{1}, \boldsymbol{J}] \qquad \textbf{(5. 11)}$$

For the current empirical application, these sufficiency conditions are:

$$\left(\boldsymbol{\alpha_{2,1}} + \boldsymbol{\alpha_{3,1}} + \boldsymbol{\alpha_{4,1}}\right) + \left(\boldsymbol{\alpha_{1,2}} + \boldsymbol{\alpha_{1,3}} + \boldsymbol{\alpha_{1,4}}\right) < \mathbf{1} \qquad \textbf{(5.12 a)}$$

$$\left(\boldsymbol{\alpha_{1,2}} + \boldsymbol{\alpha_{3,2}} + \boldsymbol{\alpha_{4,2}}\right) + \left(\boldsymbol{\alpha_{2,1}} + \boldsymbol{\alpha_{2,3}} + \boldsymbol{\alpha_{2,4}}\right) < \mathbf{1} \qquad \textbf{(5.12 b)}$$

$$\left(\boldsymbol{\alpha_{1,3}} + \boldsymbol{\alpha_{2,3}} + \boldsymbol{\alpha_{4,3}}\right) + \left(\boldsymbol{\alpha_{3,1}} + \boldsymbol{\alpha_{3,2}} + \boldsymbol{\alpha_{3,4}}\right) < \mathbf{1} \qquad \textbf{(5.12 c)}$$

$$\left(\boldsymbol{\alpha_{1,4}} + \boldsymbol{\alpha_{2,4}} + \boldsymbol{\alpha_{3,4}}\right) + \left(\boldsymbol{\alpha_{4,1}} + \boldsymbol{\alpha_{4,2}} + \boldsymbol{\alpha_{4,3}}\right) < \mathbf{1} \qquad \textbf{(5.12 d)}$$

The analyst must estimate $J \times (J-1)$ additional parameters to account for misclassification. However, in the current empirical context, all the alternatives are ordered. So, it is expected that the misclassification probability $\alpha_{s,t}$decreases considerably as $|s-t|$ increases. So, several entries of the misclassification matrix are expected to be zero. The parameters of the MGORP model were estimated using the Maximum Simulated Likelihood (MSL) estimation approach using 150 Halton draws.

5.2 Empirical Application

The data used for undertaking the analysis was obtained from the 2014 General Estimates System (GES) database maintained by the National Highway Traffic Safety Administration (NHTSA)'s National Center for Statistics and Analysis. The GES database is a nationally representative sample of police-recorded accidents that involved at least one motor vehicle traveling on a traffic-way and resulted in property damage, injury, or death. The database provided detailed information on about 53,000 accidents involving 93,000 vehicles. Including (a) details of all people involved in the crash (age, gender, seating position, seat belt use, alcohol involvement, whether the occupant was ejected, and injury severity level sustained), (b) attributes of all vehicles involved in the crash (body type of the car and whether the vehicle rolled-over), roadway geometric attributes (details regarding the regulatory signs/control at the accident location, number of lanes, roadway type, and speed limit), environment factors (lighting and weather conditions), and crash characteristics (type of collision, whether the collision occurred at an intersection, and number of vehicles involved). The injury severity of each occupant was recorded on a five-point KABCO ordinal scale: (1) No injury, (2) Possible injury, (3) Non-incapacitating injury, (4) Incapacitating injury, and (5) Fatal injury. Given that the focus of this analysis is on accidents involving colliding motor-vehicles, all non-collision crash records, motorcycle crashes, and crashes involving bicyclists and pedestrians were excluded from the analysis. Also, the analysis was limited to driver injury severity. So, all records corresponding to passengers were removed. Lastly, after cleaning the data and eliminating the records with missing information on key explanatory variables and the injury severity variable, the size of the data reduced to about 42,100 driver records from 25,708 crashes. In this final estimation sample, the percentage of drivers who sustained fatal injury was less than 1%. Because of this low percentage of the fatal injury records, the fatal and incapacitating injury categories were combined and labeled as '*incapacitating or fatal'* injury. The distribution of the dependent variable in the final estimation sample was as follows: no injury (69%), possible injury (12.3%), non-incapacitating injury (12.2%), and incapacitating or fatal injury (6.5%).

Past research findings, statistical significance, and parameter intuitiveness guided the model estimation. Only parameters that were statistically significant at a 95% confidence level were retained in the final model specification. Two models – misclassification-adjusted MGORP (MMGORP) and standard MGORP model that ignores misclassification – were estimated for comparison purposes.

5.2.1 Misclassification Rates

Table 8. Misclassification Probabilities in MMGORP Model8 presents the estimated misclassification matrix in the best specification of the MMGORP model. The identification conditions in Equations (5.12a-5.12d) are satisfied by the estimated misclassification rates. Interestingly, the misclassification rates of all injury severity categories were found to be zero except for the non-incapacitating injury. Specifically, 32.2% of non-incapacitating injuries were wrongly classified as possible injuries. Alternatively, only 67.8% of non-capacitating injuries were correctly classified. This is consistent with the expectation that there may be considerable subjectivity while classifying less severe crashes without significant bodily harm into possible and non-incapacitating categories.

Table* 8. *Misclassification Probabilities in MMGORP Model

$Best\ Estimated \downarrow \|\|Observed \rightarrow$	No Injury	Possible Injury	Non-Incapacitating Injury	Incapacitating or Fatal Injury
No Injury	1.0000	0.0000	0.0000	0.0000
Possible Injury	0.0000	1.0000	0.0000	0.0000
Non-Incapacitating Injury	0.0000	0.3218	0.6782	0.0000
Incapacitating or Fatal Injury	0.0000	0.0000	0.0000	1.0000

5.2.2 Parameters Interpretation

Table 7. Elasticity Effects 9 presents the parameter estimates of the MMGORP model and the bias between the parameter estimates of MMGORP (β_{MMGORP}) and MGORP (β_{MGORP}) models computed as:

$$\boldsymbol{Bias = \frac{(\beta_{MGORP} - \beta_{MMGORP})}{\beta_{MMGORP}} \times 100} \qquad \textit{(5. 12)}$$

All the parameter estimates have the same sign in the two models except for the constant parameter in the last threshold. However, there was considerable bias in the parameter estimates of MGORP models that ignore misclassification as shown in the last column of Table 7. Elasticity Effects 9. The MGORP model seems to under-estimate the parameters in the propensity and second threshold and over-estimate the parameters in the third threshold relative to the MMGORP model. This finding is consistent with the over-representation of the less severe possible injury category

due to misclassification in the more severe non-incapacitating injury category. So, the MGORP model skews the propensity to the left and the second and third thresholds to the right to account for over-representation of the possible injury category and under-representation of the non-incapacitating injury category. Also, the bias in the threshold parameters was higher than bias in the propensity parameters. Specifically, the average absolute percentage bias values were 11.1%, 18.0%, 43.8%, and 22.7% for the propensity, first, second, and third threshold parameters, respectively. So, the bias in the second threshold parameters between possible and non-incapacitating injury categories and the third threshold parameters between incapacitating and non-incapacitating injury categories were higher compared to other parameters. This is consistent with the significant misclassification rate in the non-incapacitating injury category. However, the bias, although relatively lower in magnitude, propagates to the propensity and other threshold parameters.

From an interpretation standpoint, higher propensity values translate into higher probabilities of more severe injury outcomes. Also, higher values for the second threshold will lead to a higher probability for possible injury and lower probability for non-incapacitating injury. Similarly, higher values of the third threshold will lead to a higher probability of non-incapacitating injury and a lower probability of incapacitating or fatal injury outcome. Everything else being the same, men tend to sustain less severe injuries compared to women. Older drivers are more likely to sustain severe injuries compared to younger drivers. As expected, driving under the influence of alcohol and driver ejection increase whereas seat belt use lowers the propensity to sustain severe injuries. Drivers in SUVs, vans, light trucks, and heavy trucks have lower risk propensity compared to passenger car drivers. Drivers in rolled-over vehicles tend to sustain severe injuries. Crashes along two-way divided roadways and ramps tend to be more severe compared to those occurring on one-way and two-way undivided roadways. Drivers in crashes along multi-lane roadways have higher risk propensity than those who have accidents along single lane roads. Interestingly, the speed limit was not found to influence the risk propensity of drivers. This is probably because of the correlation between roadway geometry variables (roadway type and some lanes) and speed limit. Intersection crashes have marginally higher risk propensity compared to accidents elsewhere. The kind of traffic control at the intersection also had a significant impact on severity. Specifically, accidents at controlled intersections including traffic signals, stop, and yield signs have higher risk propensity compared to crashes at uncontrolled intersections. There is no

base category for collision type variable because the constant in the propensity was completely segmented by the type of crash. Front-front, angled, and opposite direction sideswipe collisions tend to be more severe compared to front-rear and same direction sideswipe collisions. Drivers involved in accidents with fixed objects have a higher risk of propensity compared to drivers involved in vehicular collisions. Dawn and cloudy conditions have marginally higher risk propensity compared to dark and daylighting conditions. The higher the number of vehicles involved in a crash, the higher the risk propensity becomes. Significant unobserved heterogeneity was observed in the effect of vehicle body type, vehicle rollover indicator, roadway type, surface condition, traffic control type, and lighting conditions, as indicated by the standard deviation parameters in Table 7. Elasticity Effects 9. Also, all the three thresholds were found to be stochastic, as noted in the standard deviation parameters on the constants in these thresholds. These results underscore the importance of unobserved heterogeneity in crash severity modeling.

5.2.3 Statistical Fit Comparison

The final log-likelihood values of the MGORP and MMGORP models were -36402.40 and -36350.29, respectively. The MMGORP model that handles misclassification has only one additional parameter (*i.e.*, the misclassification rate corresponding to non-incapacitating injury) and nests the MGORP model as a particular case. So, these two models can be compared using the log-likelihood ratio (LR) test. The LR statistic of comparison between the two models was equal to $-2 \times (-36402.40 + 36350.29) = 104.22$, which is considerably higher than the critical chi-squared value of 3.96 corresponding to one degree of freedom. This suggests a superior data fit in the MMGORP model.

Table 9.The MMGORP Model Results

	Propensity		**Second Threshold**		**Third Threshold**	
	Coeff.	Bias	Coeff.	Bias	Coeff.	Bias
Explanatory Variables						
Constant			-1.459	-48.39	0.171	-161.93
Standard Deviation			0.329	-16.09	0.248	-49.47

Table 9. Continued

Explanatory Variables						
Gender (Base: Female)						
Male	-0.327	-7.25	-0.220	-8.68	-0.188	4.20
Age (Base: <=15 years)						
16 to 19 years	-0.316	-11.38				
20 to 25 years	-0.132	-10.79				
46 to 60 years	0.110	-2.90				
61 to 75 years	0.125	-1.69			-0.076	18.87
>=76 years					-0.147	38.16
DUI (Base: No)						
Yes	0.311	-5.37				
Wearing a seat belt (Base: No)						
Yes	1.367	-13.12				
Ejected						
Yes	1.778	-21.90				
Body Type (Base: Passenger car)						
SUV	-0.203	-10.75				
Van	-0.155	-11.84				
Light Truck	-0.274	-8.64				
Heavy Truck	-1.291	-14.38				
Standard deviation	0.509	-20.83				
Vehicle rolled over (Base: No)						

Explanatory Variables						
Yes	1.723	-40.93			0.817	-61.12
Standard Deviation	0.810	-66.70				
Roadway type (Base: One-way Two-way Undivided)						
Two-way Divided Unprotected	0.206	-3.78				
Two-way Divided Protected	0.247	-3.24	0.275	-50.36		
Ramp	0.196	-18.87				
Standard deviation	0.289	0.38				
# of lanes (Base: One lane)						
Two Lanes	0.195	-12.62				
Three Lanes	0.178	-15.19				
Four or More Lanes	0.213	-14.85				
Surface condition (Base: Normal)						
Wet					0.071	17.26
Snow	-0.258	-7.75				
Standard deviation	0.409	-8.18				
Ice	-0.321	-6.72				
Standard deviation	0.371	-6.14				
Traffic control (Base: No control)						
Traffic Signal	-0.075	-4.16				
Stop Sign	-0.323	-4.02				
Standard deviation	0.362	-3.93				

Explanatory Variables						
Yield Sign	-0.611	2.06				
Standard deviation	0.378	8.10				
Type of Collision						
Front End Collision	1.066	-8.89	0.384	-50.29	-0.172	74.56
Rear End Collision	-0.499	-10.11				
Angled Collision	0.436	-7.57	0.293	-44.44	-0.083	76.08
Side Swipe Same Direction	-0.437	-10.48				
Side Swipe Opposite Direction	0.284	-1.73			-0.260	36.59
Crash occurred at Intersection						
Yes	0.071	-5.90				
Lighting conditions (Base: Day/Dark lighting)						
Dark no lighting			-0.570	-65.17		
Dawn	0.147	-8.98				
Standard deviation	0.328	-5.55				
Cloudy	0.061	-16.34	-0.235	-65.25	0.066	-4.98
Fixed object crash						
Yes	0.749	-13.28			-0.107	114.33
Number of vehicles (Base: <3)						
3 or more	0.155	-10.56	0.311	-46.01		
First Threshold						
Constant	-0.257 (% Difference = 8.98%)					

Explanatory Variables	
Standard Deviation	-0.976 (% Difference = -27.00%)
Log-likelihood at convergence	
MGORP Model	-36,350.29
MMGORP Model	-36,402.40

5.3 Elasticity Effects Analysis

The bias in a parameter estimate of a variable does not necessarily mean significantly different policy implications. This is because the parameter estimates in Table 7. Elasticity Effects 10 do not directly indicate the magnitude of variable effects. To gain understanding into the relative effects of different variables and the policy implications of misclassification, the aggregated elasticity effects were calculated as the percentage change in the shares of varying injury severity levels for a unit change in an explanatory variable. First, the shares of different injury severity categories were computed in the base scenario by summing the probabilities obtained using Equation 5.6 across all observations. Next, the shares were recomputed in the policy scenario using the same equation but with a unit increase in the variable for which the elasticity is being calculated. Given that all variables in the final model specification are indicator variables, the unit change is 0 to 1. Table 10 presents the results of the elasticity analysis for the MMGORP model and the percentage bias in the MGORP model elasticity effects. The elasticity corresponding to the 'Gender' variable for the 'incapacitating or fatal' injury was -11.7%, indicating that male drivers are 11.7% less likely to sustain an 'incapacitating or fatal' injury compared to women in the event of a crash. The next set of four numbers in the first row indicate the bias of elasticity effects in the MGORP model that ignores misclassification. For instance, the MGORP model over-estimates the elasticity effect for 'incapacitating or fatal' injury by 7%. Other numbers in the table can be interpreted similarly. From the relative magnitude of elasticity effects, it can be seen that ejection from the vehicle, front-front vehicle collisions, vehicle rollover, and fixed object collisions are most likely scenarios to result in 'incapacitating or fatal' injury. Similarly, seat belt use, drivers in heavy trucks, rear-rear collisions, yield sign traffic control, and younger drivers are least likely scenarios to result in 'incapacitating or fatal' injury. The results indicate considerable bias in the

elasticity estimates of the MGORP model. Furthermore, the bias is much higher for the possible injury and non-incapacitating injury categories compared to 'no injury' and 'incapacitating or fatal' injury categories. For instance, drivers under the influence of alcohol are 21% more likely to sustain non-incapacitating injury. However, the MGORP model underestimates this elasticity effect by 30%. The average absolute bias values for the four injury severity levels are 6%, 58%, 24%, and 4%, respectively.[1] These findings are consistent with the statistically significant misclassification rate corresponding to the possible and non-incapacitating injury category in 8. Overall, the results indicate that misclassification in injury severity data can result in biased parameter estimates leading to incorrect policy sensitivity results.

Table 10. Elasticity Effects of MMGORP Model[2]

	GORP- Misclassification				**% Bias of MGORP Model**			
Explanatory Variables/ Alternative	NI	PI	NII	IFI	NI	PI	NII	IFI
Gender (Base: Female)								
Male	10.5	-27.1	-27.3	-11.7	7.0	3.9	4.4	7.1
Age (Base: Less than <=15 years)								
16 to 19 years	9.2	-11.6	-21.6	-32.8	1.2	26.8	11.0	-3.0
20 to 25 years	3.9	-4.3	-9.1	-15.0	2.4	37.5	14.4	-2.8
46 to 60 years	-3.3	3.2	7.8	14.3	11.9	62.0	28.1	5.9
61 to 75 years	-3.8	3.5	3.2	25.6	13.6	67.1	33.8	-2.3
>=76 years	-6.2	5.1	2.4	46.3	12.0	72.9	-56.8	0.4
DUI								
Yes	-9.6	6.9	21.3	44.5	10.4	84.3	29.8	3.6
Wearing a seat belt								

[1] The unusually high bias for the 'Ramp' traffic control for possible injury was excluded from the average bias calculation. This high bias value is due to extremely low elasticity effect (0.1%) in the MMGORP model.
[2] NI: No injury; PI: Possible injury; NII: Non-Incapacitating injury, IFI: Incapacitating or Fatal injury

Table 10. Continued

	GORP- Misclassification				% Bias of MGORP Model			
Explanatory Variables/ Alternative	NI	PI	NII	IFI	NI	PI	NII	IFI
No	-42.7	-9.8	58.3	305.4	6.5	-276.5	60.1	-5.1
Ejected								
Yes	-54.0	-25.3	49.4	402.1	-1.8	-139.0	90.2	-16.7
Vehicle Type (Base: Passenger car)								
SUV	6.0	-6.8	-13.8	-22.2	2.3	35.2	13.6	-2.7
Van	4.5	-5.3	-10.8	-17.5	1.1	32.8	12.4	-3.8
Light Truck	8.1	-9.7	-18.7	-29.0	4.3	34.0	15.0	-0.6
Heavy Truck	25.3	-55.5	-63.7	-69.9	2.0	1.4	3.9	2.1
Vehicle rolled over								
Yes	-47.5	-30.4	160.4	186.9	-18.8	-152.7	-14.7	-4.9

	GORP- Misclassification				% Bias of MGORP Model			
Explanatory Variables/ Alternative	NI	PI	NII	IFI	NI	PI	NII	IFI
Roadway type (Base: One-way or two-way undivided)								
Two Way Divided Unprotected Roadway	-6.2	38.2	9.6	17.3	11.3	-37.6	32.3	11.2
Two Way Divided Protected Roadway	-7.4	6.7	17.8	34.4	11.7	69.3	29.8	5.9
Ramp	-6.5	0.1	11.5	37.1	1.3	3410.3	29.4	-3.1
Number of lanes (Base: One Lane)								
Two Lanes	-5.7	5.8	14.3	26.8	0.6	46.8	15.7	-5.4
Three Lanes	-5.3	5.0	12.6	23.8	-2.1	44.6	12.4	-8.4
Four or More Lanes	-6.3	6.4	15.6	29.2	-1.6	41.8	12.3	-8.2
Surface conditions (Base: Normal)								
Snow	5.4	-15.7	-16.5	-9.4	0.6	0.2	-2.8	4.8
Ice	7.4	-16.9	-20.3	-18.9	2.7	7.9	3.8	0.9
Traffic control type (Base: No control)								
Traffic Signal	2.2	-2.3	-5.1	-8.7	10.2	52.2	23.8	4.1
Stop Sign	7.6	-16.7	-20.6	-20.0	5.0	12.1	6.8	2.0
Yield Sign	14.2	-28.4	-36.8	-42.6	8.0	20.5	12.1	1.2
Type of collision								

	GORP- Misclassification				**% Bias of MGORP Model**			
Explanatory Variables/ Alternative	NI	PI	NII	IFI	NI	PI	NII	IFI
Front End Collision	-33.6	53.1	35.1	223.0	9.6	-4.9	11.7	4.4
Rear End Collision	13.6	-20.8	-34.0	-47.4	2.1	19.8	10.0	-1.4
Angled Collision	-12.9	49.7	20.2	66.1	6.8	-21.5	-2.8	3.0
Side Swipe Same Direction	12.5	-17.4	-29.8	-43.0	1.7	22.5	10.4	-2.0
Side Swipe Opposite Direction	-8.8	6.5	-2.0	75.3	14.6	86.2	146.2	0.2
Crash occurred at intersection								
Yes	-2.1	2.2	5.1	9.1	8.2	51.8	23.0	2.4
Lighting Conditions (Base: Day)								
Dawn	-5.2	-2.2	8.0	33.1	12.3	-172.5	59.6	2.2
Cloudy	-1.8	-18.9	12.9	7.0	-3.7	-75.2	1.2	0.8
Fixed object collision (Base: No)								
Yes	-23.2	8.7	37.7	166.6	2.1	174.6	-3.4	1.0
Number of vehicles involved (Base: <3 vehicles)								
3 or more	-4.7	41.0	5.2	8.7	3.3	-41.7	6.1	-8.4

5.4 Conclusions

The police-reported injury severity recordings in crash databases are prone to errors. Previous research that measured the discordance between police-reported injury severity data and hospital/ambulance records confirmed the presence of misclassification errors in traditional crash databases. However, these databases remain the primary data sources for safety analysis including aggregate crash frequency and disaggregate injury severity analysis conditional on crash occurrence. Ignoring the errors in the injury severity data during modeling can lead to biased and inconsistent parameter estimates. However, it is surprising that none of the earlier studies attempted to quantify and adjust the bias caused by misclassification in injury severity models. In this study, the misclassification-adjusted mixed generalized ordered response probit (MMGORP) model was developed to analyze driver injury severity using the 2014 General Estimates System (GES) data. The results indicate that more than 30% of non-incapacitating injuries were wrongly classified as possible injuries. Also, the MGORP model that ignores misclassification has not only lower data fit but also considerable bias in the parameter and elasticity effects leading to incorrect policy implications. The model developed in this study can be used to investigate misclassification errors in ordinal response variables in other empirical contexts beyond transportation safety.

However, there are several possible avenues for future research. For instance, the misclassification rates in the model developed do not vary across observations. However, earlier studies found that the discordance rates between police and hospital records vary as a function of different factors including the driver, crash, and geographic factors. The model developed in this study can be extended to allow the misclassification rates to vary across different segments. For instance, two sets of misclassification rates can be estimated separately for crashes that occur in urban and rural neighborhoods. The sufficiency conditions in Equation 11 must hold within urban and rural neighborhoods separately (but not necessarily in the two regions together). However, the number of misclassification parameters can explode easily as the number of segments increases. To avoid this problem, latent class models that probabilistically assigns each driver/crash record to latent segments each with its own set of misclassification rates can be developed [60, 61]. The probability of belonging to each latent segment can be specified as a function of the driver, vehicle, and crash variables. Recently, this latent modeling approach was used for analyzing misclassification rates in household auto-ownership responses in travel surveys [18]. Next, not only injury severity recordings but also other variables in crash databases are prone to

misclassification. For instance, police tend to over-estimate seat-belt use in road casualties [68]. Seat-belt use also has the endogeneity problem whereby there can be common unobserved factors that influence the decision to wear a seat-belt and the injury severity outcome [69]. In this context, future research that develops an integrated modeling framework to account for misclassification in critical explanatory variables in addition to the injury severity response variable is warranted.

CHAPTER 6

GENERALIZED EXTREME VALUE MODEL TO HANDLE MISCLASSIFICATION IN TELECOMMUTING FREQUENCY CHOICES DATA

In this chapter, we are investigating the misclassification extended in the telecommuting frequency data using a form of the General Extreme Value Model, recently developed by other scholars. Specifically, the Negative Binomial Model re-casted as the Multinomial Logit Model with maximum count set to 31 days while accounting for misclassification errors.

6.1 Methodological Framework

The number of days that a person telecommutes in a month are count responses variables. Count data are typically analyzed using parametric count models including Poisson and Negative Binomial (NB) models. While the Poisson model is suited for count data with equidispersion property (i.e., mean is equal to variance), the NB model is apt for modeling over-dispersed data (i.e., mean is less than variation) [70]. Another common feature of count data is the 'excess zeroes' problem, i.e., zero count outcome is over-represented, making it difficult for standard count models to account for additional probability mass associated with the zero count outcomes. In the past, researchers used the two-step hurdle or zero-inflated models with an explicit modeling step to account for probability mass associated with zero count outcome [71, 72]. In the current empirical context, the response variables of interest are the misclassification errors for the number of days a person telecommutes in a month. It is very likely that this data is skewed to the right, leading to over-representation of multiple non-zero count outcomes. However, it is difficult to extend the hurdle and zero-inflated models to account for the additional probability mass associated with various count outcomes (i.e., excess ones, excess twos, etc.). Recently, Generalized Extreme Value (GEV) count models that can easily handle the probability mass deviations of multiple count outcomes were developed [73]. Each worker has several or zero days that he/she telecommutes in a corresponding month. So, it is likely that there are workers-specific unobserved factors that influence the number of days chosen to telecommute across the month. These common unobserved factors that remain the same across all time-periods can be captured by introducing worker-specific random effects into the mean parameter of the count model. The current study adopted the

Generalized Extreme Value (GEV) count modeling framework for jointly modeling the number of days a worker telecommutes while accounting for misclassification errors and the characteristics of telecommuting frequency data. More specifically, the Negative Binomial Model re-casted as the Multinomial Logit Model with the maximum count set to 31 days that account for misclassification.

For the Negative Binomial model, the probability of observing count outcome *y* conditional on the expected value parameter λ and dispersion parameter $r > 0$ is given by:

$$\boldsymbol{P(Y = y)} = \left(\frac{r}{r+\lambda}\right)^{r} \frac{\Gamma(r+y)}{\Gamma(y+1)\Gamma(r)} + \left(\frac{\lambda}{r+\lambda}\right)^{y} \qquad \textbf{(6. 1)}$$

Where Γ is the gamma function defined as follows:

$$\boldsymbol{\Gamma(t)} = \begin{cases} \int_{x=0}^{\infty} x^{t-1}e^{-x}\,dx \;\; \textbf{for positive non} - \textbf{integer } t \\ (t-1)! \;\; \textbf{for positive integer } t \end{cases} \qquad \textbf{(6. 2)}$$

Also, the Gamma function, Γ has the following property:

$$\boldsymbol{\Gamma(t+1) = t\Gamma(t) \;\; for\ any\ positive\ real\ number\ t} \qquad \textbf{(6. 3)}$$

The variance of the negative binomial distribution $\lambda + \frac{\lambda^2}{r}$ which is always greater than the expected value parameter λ, the Negative Binomial Model is better suited for handling over-dispersion as mentioned earlier. The NB model collapses to the Poison Model for large values of the dispersion parameter *r* and to the Geometric Model when $r = 1$. It is important that the parameter *r* will not take integer values.

It was previously shown by other scholars [73] that the Negative Binomial Model could be recast as special cases of the simplest GEV model, the multinomial logit. Considering that, the probability that an outcome *k* with observed utility $\tilde{V}_k$ is chosen from a set of *K* mutually exhaustive and exclusive outcomes is given by:

$$\boldsymbol{P(Y = y) = \frac{e^{\tilde{V}_y}}{\sum_{k=0}^{31} e^{\tilde{V}_k}}, \; where \;\; \tilde{V}_k = \; LN\left[\frac{\Gamma(r+k)}{\Gamma(r)\Gamma(k+1)} + \left(\frac{\lambda}{r+\lambda}\right)^{k}\right]} \qquad \textbf{(6. 4)}$$

Eq. 6.4 can be viewed as the probability expression of an MNL model with infinite ordinal outcomes (starting from 0) in the choice set and the observed utility of count outcome *k* given by $\tilde{V}_k = LN\left[\frac{\Gamma(r+k)}{\Gamma(r)\Gamma(k+1)} + \left(\frac{\lambda}{r+\lambda}\right)^{k}\right]$.

$$\boldsymbol{P(Y=y)=\frac{\frac{\Gamma(r+y)}{\Gamma(y+1)\Gamma(r)}+\left(\frac{\lambda}{r+\lambda}\right)^{y}}{\sum_{k=0}^{\infty}\frac{\Gamma(r+k)}{\Gamma(r)\Gamma(k+1)}+\left(\frac{\lambda}{r+\lambda}\right)^{k}}} \quad \textit{(6. 5)}$$

It can be seen that more r $\rightarrow \infty$, this utility expression collapses to LN $\left[\frac{\lambda^k}{k!}\right]$, which is the utility expression in the Poisson model.

Let $k(0,1,2,\ldots K)$ be the index for the number of days the worker chooses to telecommute outcome. In the current context, $K = 31$ with the outcome being zero days $(k = 0)$, one day $(k = 1)$, two days, $(k = 2)$and so on until last the day of the month $(k = 31)$.

Modified Likelihood Function to Handle Misclassification

Let $\boldsymbol{\rho_{s,t}}$ be the probability that count outcome *s* is misclassified as count outcome *t*, where $0 \leq \rho \leq 1$. Any given outcome *s* can be classified as one of the *K* outcomes, so $\sum_{t=1}^{K}\boldsymbol{\rho_{s,t}} = 1$. Now, if *k* is the observed ordinal outcome, then the true response can be any of the *K* outcomes. So, the probability of observed ordinal outcome *k* under misclassification is given by:

$$\boldsymbol{P\,(Y=k)=\sum_{t=1}^{K}\rho_{s,t}\times P(y=t)} \quad \textit{(6. 6)}$$

In the current empirical context with 31 days to telecommute outcomes, the misclassification matrix is given by:

$$\begin{bmatrix} \boldsymbol{Best\ Estimated \downarrow || Observed \rightarrow} & \boldsymbol{k=0} & \boldsymbol{k=1} & \boldsymbol{k=2} & \boldsymbol{k=3} & . & \boldsymbol{k=31} \\ \boldsymbol{k=0} & - & \boldsymbol{\rho_2^1} & \boldsymbol{\rho_2^2} & \boldsymbol{\rho_2^3} & . & \boldsymbol{\rho_2^{31}} \\ \boldsymbol{k=1} & \boldsymbol{\rho_1^1} & - & \boldsymbol{\rho_2^1} & \boldsymbol{\rho_2^2} & . & \boldsymbol{\rho_2^{30}} \\ \boldsymbol{k=2} & \boldsymbol{\rho_1^2} & \boldsymbol{\rho_1^1} & - & \boldsymbol{\rho_2^1} & . & \boldsymbol{\rho_2^{29}} \\ \boldsymbol{k=3} & \boldsymbol{\rho_1^3} & \boldsymbol{\rho_1^2} & \boldsymbol{\rho_1^1} & - & . & \boldsymbol{\rho_2^{28}} \\ \ldots & . & . & . & . & . & . \\ \boldsymbol{k=30} & \boldsymbol{\rho_1^{30}} & \boldsymbol{\rho_1^{29}} & \boldsymbol{\rho_1^{28}} & \boldsymbol{\rho_1^{27}} & . & \boldsymbol{\rho_2^1} \\ \boldsymbol{k=31} & \boldsymbol{\rho_1^{31}} & \boldsymbol{\rho_1^{30}} & \boldsymbol{\rho_1^{29}} & \boldsymbol{\rho_1^{28}} & . & - \end{bmatrix}$$

(6. 7)

The diagonal elements in the above matrix indicate the probability that the observed and the true response variable are the same or the probability of correct classification. Any observed ordinal outcome *s* in the data may be because of misclassification (*i.e.,* the true outcome representing the number of days telecommuting is some other outcome *t* but was misclassified as *s*) or due to correct classification. The intuitive meaning of the sufficiency conditions for consistency is that the probability of observed data being correct must be larger than the probability of being

misclassified. If these sufficiency conditions fail, the parameter estimates in the model can have opposite signs from a model that ignores misclassification, and there is little hope in recovering the true parameters consistently [2]. Mathematically, the sufficiency condition translates into the following equation for alternative *s*:

$$\sum_{\substack{t=1 \\ t\neq s}}^{K} \boldsymbol{\rho_{t,s}} < \boldsymbol{\rho_{s,s}} \forall\ \boldsymbol{s} \in [\mathbf{1}, \boldsymbol{K}] \qquad \textbf{\textit{(6. 8)}}$$

Adding $\sum_{\substack{t=1 \\ t\neq s}}^{K} \rho_{s,t}$ to both sides of Equation (6.8) gives the following result:

$$\sum_{\substack{t=1 \\ t\neq s}}^{k} \boldsymbol{\rho_{t,s}} + \sum_{\substack{t=1 \\ t\neq s}}^{K} \boldsymbol{\rho_{s,t}} < \boldsymbol{\rho_{s,s}} + \sum_{\substack{t=1 \\ t\neq s}}^{K} \boldsymbol{\rho_{s,t}} \ \forall\ \boldsymbol{s} \in [\mathbf{1}, \boldsymbol{K}] \qquad \textbf{\textit{(6. 9)}}$$

But, $\rho_{s,s} + \sum_{\substack{t=1 \\ t\neq s}}^{K} \rho_{s,t} = \sum_{t=1}^{K} \rho_{s,t} = 1$. So, the sufficiency condition is equivalent to:

$$\sum_{\substack{t=1 \\ t\neq s}}^{K} \boldsymbol{\rho_{t,s}} + \sum_{\substack{t=1 \\ t\neq s}}^{K} \boldsymbol{\rho_{s,t}} < \mathbf{1} \forall\ \boldsymbol{s} \in [\mathbf{1}, \boldsymbol{K}] \qquad \textbf{\textit{(6. 10)}}$$

For the current empirical application, these sufficiency conditions are:

$$\left(\boldsymbol{\rho_2^1} + \boldsymbol{\rho_2^2} + \dots + \boldsymbol{\rho_2^{31}}\right) + \left(\boldsymbol{\rho_1^1} + \boldsymbol{\rho_1^2} + \dots + \boldsymbol{\rho_1^{31}}\right) < \mathbf{1} \qquad \textbf{\textit{(6.11. a)}}$$

$$\left(\boldsymbol{\rho_1^1} + \boldsymbol{\rho_2^1} + \dots + \boldsymbol{\rho_2^{30}}\right) + \left(\boldsymbol{\rho_2^1} + \boldsymbol{\rho_1^1} + \dots + \boldsymbol{\rho_1^{30}}\right) < \mathbf{1} \qquad \textbf{\textit{(6.11. b)}}$$

$$\left(\boldsymbol{\rho_1^2} + \boldsymbol{\rho_1^1} + \dots + \boldsymbol{\rho_2^{29}}\right) + \left(\boldsymbol{\rho_2^2} + \boldsymbol{\rho_2^1} + \dots + \boldsymbol{\rho_2^{29}}\right) < \mathbf{1} \qquad \textbf{\textit{(6.11. c)}}$$

……………………………………………………………..

$$\left(\boldsymbol{\rho_1^{31}} + \boldsymbol{\rho_1^{30}} + \dots + \boldsymbol{\rho_1^{30}}\right) + \left(\boldsymbol{\rho_2^{31}} + \boldsymbol{\rho_2^{30}} + \dots + \boldsymbol{\rho_2^1}\right) < \mathbf{1} \qquad \textbf{\textit{(6.11. d)}}$$

The analyst must estimate $K \times (K-1)$ additional parameters to account for misclassification. In our case the outcome is count in nature, so most of the misclassification terms are likely to be zero as the distance between two outcomes increases. In the current empirical context, all the number of days a person telecommutes alternatives are ordered. So, it is expected that the misclassification probability $\boldsymbol{\rho}$ decreases considerably as $|s-t|$ increases. So, several entries of the misclassification matrix are expected to be zero.

The probability of observed outcome in Equation 6.6 is a function of all the misclassifications of parameter $\rho_{t,k}$. Please note that the observed number of days a person telecommutes outcome, *k,* varies across workers in addition to all the parameters of $P(y=t)$.

6.2 Empirical Analysis

The data for this analysis were obtained from the 2017 National Household Travel Surveys (NHTS) that collected detailed socio-demographics, employment characteristics, and travel diary information from a representative sample of the US population. In the complete sample, there are about 106,580 workers, and 82.6 % of these workers have their primary work location outside the home. Among these workers with an out-of-home workplace, only 17.21% (about 18,341) had the option to telecommute. These 18,341 workers constitute the target sample for the analysis in this study. After excluding records with missing information about key explanatory variables used during model estimation, the final estimation sample reduces to 18,306 workers.

Figure 1. Monthly Telecommuting Frequency

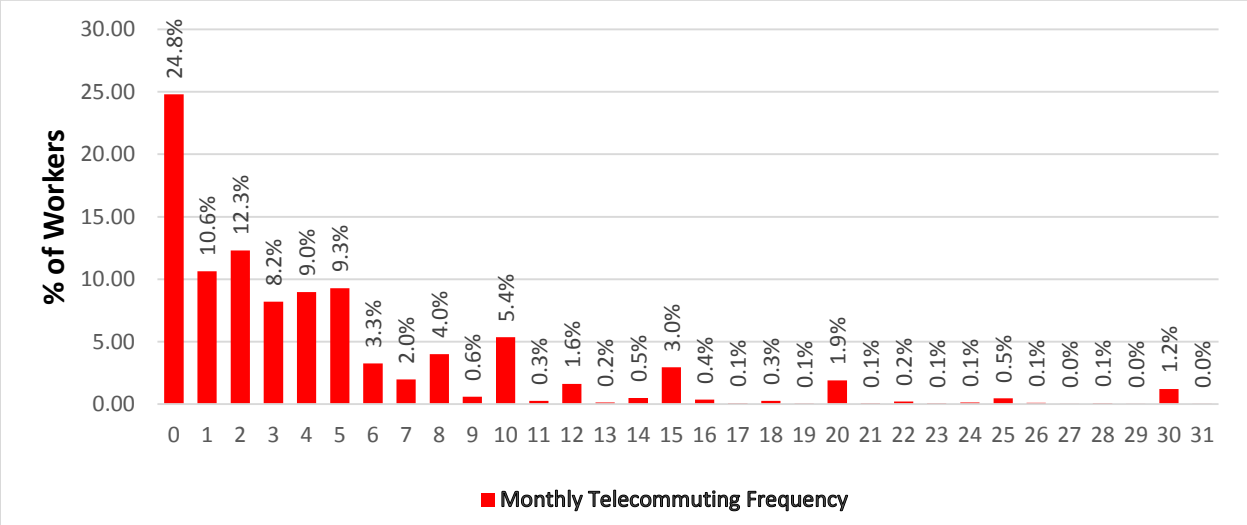

Figure 1 present the frequency distribution of telecommuting frequency in the final sample. It can be seen from the picture that 24.8% of workers do not telecommute although they have the option. There are a significant number of workers who have high telecommuting frequency. For instance, 3.0% and 1.9% of workers telecommute 15 and 20 days a month, respectively. The mean frequency is 4.61 days, and the variance is 33.5. So, the preliminary descriptive analysis suggests over-dispersion in telecommuting frequency data. Based on the descriptive statistics presented in Figure 1, we observe the presence of rounding in the responses. Rounding is a type of misclassification in the dependent variable that can lead to inconsistent parameter estimates [2].

Several models were developed in this study including NB, NB MNL, a model that accounts for rounding, NB MNL with one misclassification parameter, and NB MNL with two

misclassification parameters. For brevity, only the results of the "NB MNL with two misclassification parameters" model and the standard NB MNL, along with the misclassification parameters were presented in this study. The likelihood ratio (LR) test of comparison between the NB MNL model (log-likelihood, LL= -46,998.5) and the NB MNL with one misclassification parameters (-46,994.8) models was 7.32, which is greater than the critical chi-squared statistic of 3.84 corresponding to one degree of freedom at a 95% confidence level. This suggests that the model that accounts for one misclassification parameter for the entire dataset is statistically better than the standard NB MNL model. The likelihood ratio test of comparison between the NB MNL with one misclassification parameter (-46,994.8) and the NB MNL with two misclassification parameters (-46,994.1) is 1.46, which is smaller than the critical chis-squared statistic of 3.84. Based on that we cannot reject the null hypothesis, and the difference between the two models is not statistically significant. Because we choose to represent only the results from the NB MNL and the NB MNL with two misclassification parameters, we also computed the log-likelihood ratio test. The value of the log-likelihood ratio test between these two models (-46,994.8 and -46.994.1) is 8.79, which is greater than the critical chi-squared statistic of 5.99 corresponding to two degrees of freedom at av95% confidence interval. The magnitude of the dispersion parameter *r* in the NB MNL tow parameters was -0.0898 which is lower than -0.6157 for the NB MNL model. The constant in Table 9.The MMGORP Model Results0 cannot be interpreted directly because they are controlling for the range of the continuous commute distance variables. Considering that, the constants in Table 9.The MMGORP Model Results1, can be interpreted as a measure of over-representation (because all constants are positive) of the corresponding count outcomes in the telecommuting frequency data.

Table* 11*.Estimation Results

	NB MNL		NB MNL TWO PARAM	
Explanatory Variables	Parameter	t-stat	Parameter	t-stat
Person Attributes				
Constant	1.4308	24.447	1.3075	16.296
Gender (base: Female)				
Male	0.0465	1.701	0.0736	2.375
Immigration Status (base: Immigrant)				

Table 11. Continued

	NB MNL		NB MNL TWO PARAM	
Explanatory Variables	Parameter	t-stat	Parameter	t-stat
US Born Citizen	0.0632	1.825	0.0652	1.750
Job category (base case: 'sales and service,' 'other jobs.'				
Clerical and administration	-0.8474	-13.940	-0.9032	-12.901
Manufacturing, construction, maintenance, and farming	0.4857	5.278	0.5806	4.963
Professional. Managerial, or technical	-0.5352	-13.030	-0.5813	-11.844
Commute distance/100 (miles)	0.6028	8.006	0.6646	7.756
Works part-time	0.4395	9.465	0.4809	9.051
Uses Internet frequently	0.1512	6.047	0.1595	5.887
Household socio-demographics				
Auto Ownership (base case: fewer or same cars as driving age adults)				
More cars than driving age adults	0.1352	4.956	0.1508	5.057
Household income(base case: ≥ 35,000)				
Low income (≤ 34,999)	0.3787	7.512	0.4417	7.282
Dispersion Parameter, r	-0.6157		-0.8098	
Misclassification Parameter				
ρ_1			-2.1917	-4.537
ρ_2			-1.7861	-6.992

Model results, shown above, as expected and consistent with previous studies, that employed men are less likely to telecommute compared with working women, as indicated by the positive coefficient for the "Male" variable. Immigrant employees tend to telecommute less frequently compared with native citizens. Workers in the 'clerical and administration' positions, along with those in 'professional, managerial, or technical' positions tend to telecommute more compared with the ones in 'sales and service' or 'other jobs.' On the other hand, as expected, workers in 'manufacturing, construction, maintenance, and farming' positions are not using the telecommuting option when compared with the ones in 'sales and service' or 'other jobs'. Also, as expected, part-time workers tend to telecommute less frequently compared with full-time

employees. The parameter estimates on the commute distance variables indicate a robust positive relationship between commute distance and telecommuting frequency. Specifically, employees with longer commute distances are inclined to telecommute more frequently compared with employees with relatively shorter commuting distance. People who use the internet regularly are less inclined to telecommute. Workers in households with more cars than the driving age are less likely to telecommute compared to households with fewer cars then the driving age adults. From this data analysis, it is not possible to understand if workers that own fewer cars telecommute frequently or they telecommute because they have fewer cars. The decisive parameter on the low-income variable suggests that low-income workers are less inclined to telecommute. As expected, and based on the results, misclassification errors exist in the telecommuting frequency data reporting.

Table* 12*.Misclassification Probabilities

Best Estimated || Observed (days)	**0**	**1**	**2**	**.**	**30**	**31**
0	0.8324	0.1436	0.0206	.	0.0000	0.0000
1	0.1005	0.7319	0.1436	.	0.0000	0.0000
2	0.0101	0.1005	0.7218	.	0.0000	0.0000
…	.	.	.	.	.	.
30	0.0000	0.0000	0.0000	.	0.7447	0.1436
31	0.0000	0.0000	0.0000	.	0.1005	0.8883

Error! Reference source not found.2, presents the results of the misclassification components or the NB MNL with two misclassification parameters. The results in **Error! Reference source not found.**2, show significant misclassification of extreme alternatives compared with intermediate alternatives. Specifically, only, 83.24% of the responses of the workers that do not telecommute were correctly classified, 14.36% were wrongly classified as telecommuting ‘one day’, whereas 2.06% as telecommuting ‘two days’. Same in the case of ‘one day’ outcome, only 73.19% of the responses were correctly classified whereas, 10.05% and 14.36% of the responses were wrongly classified as ‘zero days’ and ‘two days’ respectively. Based on the results and consistent with previous studies, the respondents tend to over-estimate more than under-estimate the numbers of days that they telecommute. The misclassification parameters show that there is a tendency of misclassifying of the number of days a worker telecommutes.

6.4 Conclusion

By nature, and consistent with other studies, Household Travel Surveys (NHTS) data is prone to errors that can be grouped in intentional or unintentional misinformation provided by the person being interviewed. Ignoring these errors while modeling telecommuting frequencies using standard discrete count models can result in biased parameter estimates. In this part of the dissertation, the General Extreme Value models available in the literature for handling misclassification were used to quantify the impact of misclassification in telecommuting frequency data. Correctly, the frequency of telecommuting days was analyzed using the Negative Binomial re-casted as the Multinomial Logit Model. The misclassification parameter was calculated for both over-reporting and under-reporting. The misclassification errors can be as high as 14% over-reported and as high as 10% under-reported, particularly for the neighboring values. Statistical fit comparison between the models used shows that models that ignore misclassification have not only worse data fit but also biased parameter estimates with significant policy implications.

CHAPTER 7

CONCLUSION

This chapter gives a brief overview of this dissertation's findings, key implications of misclassification errors and future research directions. In all the three different discrete transportation datasets investigated, some evidence of misclassification errors was found based on the proposed model. The extent of these errors is different for each dataset depending on the variables considered. It was shown that ignoring misclassification errors can potentially result in over or under-estimation of critical variables that are important for future policy implications. The sufficiency condition considered in the three models developed is the same: the probability of an outcome being misclassified is smaller than the likelihood of being correctly observed. Based on this, the models developed showed a better statistical fit of the parameter estimates than the models that do not account for misclassification. The police-reported injury severity recordings in crash databases are prone to errors. Past research that measured the discordance between police-reported injury severity data and hospital/ambulance records confirmed the presence of misclassification errors in traditional crash databases. However, these databases remain the primary data sources for safety analysis including aggregate crash frequency and disaggregate injury severity analysis conditional on crash occurrence. Ignoring the errors in the injury severity data during modeling can lead to biased and inconsistent parameter estimates.

Misclassification errors are varying for different data sets. In the auto ownership investigation, it was shown that the misclassification errors could be as high as 40%, particularly for the extreme auto ownership levels. Comparatively, the misclassification in 'one car' and 'two cars' alternatives was lower. The un-segmented models used in this part of the dissertation restrict that misclassification rates are the same for the entire population. To relax this assumption, a latent class auto ownership model that allows the misclassification probabilities to vary across different latent segments was developed. The empirical analysis uncovered two latent classes in the population with regards to auto ownership preferences and also significant differences in the misclassification rates between the two segments. The misclassification-adjusted mixed generalized ordered response probit (MMGORP) model was developed to analyze driver injury severity using the 2014 General Estimates System (GES) data.

The research methodology adopted in this dissertation treats the observed injury severity outcomes as realizations from discrete random variables that depend on true latent injury severities that are unobservable to the analyst. The results indicate that 31.77% of possible injuries were wrongly recorded as no injuries; 29.80% of non-incapacitating injuries were wrongly classified as possible injuries: and 7.45% of non-incapacitating injuries were wrongly recorded as incapacitating or fatal injuries. Also, the MGORP model that ignores misclassification has not only lower data fit but also considerable bias in the parameter and elasticity effects, leading to incorrect policy implications. Ignoring these errors while modeling telecommuting frequencies using standard discrete count models can result in biased parameter estimates.

In another part of the dissertation, the General Extreme Value models available in the literature for handling misclassification were used to quantify the impact of misclassification in telecommuting frequency data. Specifically, the telecommuting frequency was analyzed using the Negative Binomial re-casted as the Multinomial Logit Model. The misclassification parameter was calculated for both over-reporting and under-reporting. The misclassification errors can be as high as 14% over-reported and as high as 10% under-reported, particularly for the neighboring values. Statistical fit comparison between the models used shows that models that ignored misclassifications have not only worse data fit but also biased parameter estimates with significant policy implications. The model developed in this study can be used to investigate misclassification errors in ordinal response variables in other empirical contexts beyond discrete transportation data.

However, there are several possible avenues for future research. For instance, the results indicate that there is no significant misclassification in the incapacitating/fatal injury category, i.e., all true incapacitating and fatal injuries are recorded correctly. However, it is possible that the misclassification rates in incapacitating injuries are zero because of merging the incapacitating and fatal injury categories as one alternative. This is because fatal crashes are rarely misclassified and the non-zero misclassification in incapacitating injuries are weighed down by the zero-misclassification rate in fatal crashes. Future studies must examine the impact of aggregation of injury severity categories on misclassification rates as an avenue for future research. Also, the misclassification rates in the model developed do not vary across observations. However, earlier studies found that the discordance rates between police and hospital records vary as a function of different factors including the driver, crash, and geographic factors. The model developed in this

chapter can be extended to allow the misclassification rates to vary across different segments. For instance, two sets of misclassification rates can be estimated separately for crashes that occur in urban and rural neighborhoods. The sufficiency conditions must hold within urban and rural areas separately (but not necessarily in the two regions together). However, the number of misclassification parameters can explode easily as the number of segments increases. To avoid this problem, latent class models that probabilistically assign each driver/crash record to latent segments each with its own set of misclassification rates can be developed [60, 61]. The probability of belonging to each latent segment can be specified as a function of the driver, vehicle, and crash variables. Recently, this latent modeling approach was used for analyzing misclassification rates in household auto-ownership responses in travel surveys [18].

Next, not only injury severity recordings but also other variables in crash databases are prone to misclassification. For instance, police tend to over-estimate seat-belt use in road casualties [68]. Seat-belt use also has the endogeneity problem whereby there can be common unobserved factors that influence the decision to wear a seat-belt and the injury severity outcome [69]. In this context, future research that develops an integrated modeling framework to account for misclassification in key explanatory variables in addition to the injury severity response variable is warranted. The modern suite of advanced travel demand models including tour-based and activity-based models that encompass several discrete choice models to predict daily activity and travel preferences of people. The underlying idea of these models is that people travel to participate in different types of activities at locations dispersed in space and time. So, the key response variables that form the basis of these models are activity purpose, activity duration, mode, departure time, and destination. All these responses in HTS data are prone to measurement errors and must be analyzed using similar modeling methods used in this study to quantify and assess the impact of misclassification on parameter estimates of respective choice models. Also, it is a common practice for researchers to collect their data to analyze new empirical contexts with limited revealed preference data. For example, several studies used web-based surveys that elicit preferences for new vehicle technologies including connected and autonomous vehicles and electric vehicles. It is a useful exercise to explore the quality of these survey responses by quantifying misclassification to demonstrate the validity and confidence of these study findings. Lastly, the models developed in this study can be applied to other transportation disciplines.

REFERENCES

1. Hausman, J., *Mismeasured variables in econometric analysis: problems from the right and problems from the left.* The Journal of Economic Perspectives, 2001. **15**(4): p. 57-67.
2. Hausman, J.A., J. Abrevaya, and F.M. Scott-Morton, *Misclassification of the dependent variable in a discrete-response setting.* Journal of Econometrics, 1998. **87**(2): p. 239-269.
3. Stopher, P.R., *Use of an activity-based diary to collect household travel data.* Transportation, 1992. **19**(2): p. 159-176.
4. Bonnel, P. and M. Le Nir, *The quality of survey data: telephone versus face-to-face interviews.* Transportation, 1998. **25**(2): p. 147-167.
5. Wolf, J., et al., *A Case Study: Multiple Data Collection Methods and The NY/NJ/CT regional travel survey.* Transport Survey Methods: Best Practice for Decision Making, 2013: p. 71.
6. Cottrill, C., et al., *Future mobility survey: Experience in developing a smartphone-based travel survey in Singapore.* Transportation Research Record: Journal of the Transportation Research Board, 2013(2354): p. 59-67.
7. Burch, C., L. Cook, and P. Dischinger, *A comparison of KABCO and AIS injury severity metrics using CODES linked data.* Traffic injury prevention, 2014. **15**(6): p. 627-630.
8. Choo, S., P.L. Mokhtarian, and I.J.T. Salomon, *Does telecommuting reduce vehicle-miles traveled? An aggregate time series analysis for the US.* 2005. **32**(1): p. 37-64.
9. Kim, S.-N.J.I.J.o.S.T., *Is telecommuting sustainable? An alternative approach to estimating the impact of home-based telecommuting on household travel.* 2017. **11**(2): p. 72-85.
10. Zhu, P., S.G.J.I.J.o.E.S. Mason, and Technology, *The impact of telecommuting on personal vehicle usage and environmental sustainability.* 2014. **11**(8): p. 2185-2200.
11. Mokhtarian, P.L., et al., *Telecommuting, residential location, and commute-distance traveled: evidence from State of California employees.* 2004. **36**(10): p. 1877-1897.
12. Hilbrecht, M., et al., *'I'm home for the kids': contradictory implications for work–life balance of teleworking mothers.* 2008. **15**(5): p. 454-476.
13. Pendyala, R.M., K.G. Goulias, and R.J.T. Kitamura, *Impact of telecommuting on spatial and temporal patterns of household travel.* 1991. **18**(4): p. 383-409.
14. Grawitch, M.J., L.K.J.C.P.J.P. Barber, and Research, *Work flexibility or nonwork support? Theoretical and empirical distinctions for work–life initiatives.* 2010. **62**(3): p. 169.
15. Kirk, J. and R.J.J.o.E.C. Belovics, *Making e-working work.* 2006. **43**(1): p. 39-46.

16. Helminen, V. and M.J.J.o.T.G. Ristimäki, *Relationships between commuting distance, frequency and telework in Finland.* 2007. **15**(5): p. 331-342.
17. Lachapelle, U., G. A Tanguay, and L.J.U.S. Neumark-Gaudet, *Telecommuting and sustainable travel: Reduction of overall travel time, increases in non-motorised travel and congestion relief?* 2017: p. 0042098017708985.
18. Paleti, R. and L. Balan, *Misclassification in Travel Surveys and Implications to Choice Modeling: Application to Household Auto Ownership Decisions*. 2017.
19. Hurst, E., G. Li, and B. Pugsley, *Are household surveys like tax forms? Evidence from income underreporting of the self-employed.* Review of economics and statistics, 2014. **96**(1): p. 19-33.
20. Stopher, P., C. FitzGerald, and M. Xu, *Assessing the accuracy of the Sydney Household Travel Survey with GPS.* Transportation, 2007. **34**(6): p. 723-741.
21. Bricka, S. and C. Bhat, *Comparative analysis of global positioning system-based and travel survey-based data.* Transportation Research Record: Journal of the Transportation Research Board, 2006(1972): p. 9-20.
22. Bricka, S.G., et al., *An analysis of the factors influencing differences in survey-reported and GPS-recorded trips.* Transportation research part C: emerging technologies, 2012. **21**(1): p. 67-88.
23. WHO, *Global status report on road safety 2015*. 2015: World Health Organization.
24. NHTSA, *2015 motor vehicle crashes: overview, Traffic safety facts* National Highway Traffic Safety Administration, 2016. **2016**: p. 1-9.
25. Tsui, K., et al., *Misclassification of injury severity among road casualties in police reports.* Accident Analysis & Prevention, 2009. **41**(1): p. 84-89.
26. Farmer, C.M., *Reliability of police-reported information for determining crash and injury severity.* 2003.
27. O'Day, J., *Accident data quality*. Vol. 192. 1993: Transportation Research Board.
28. Alsop, J. and J. Langley, *Under-reporting of motor vehicle traffic crash victims in New Zealand.* Accident Analysis & Prevention, 2001. **33**(3): p. 353-359.
29. Sciortino, S., et al., *San Francisco pedestrian injury surveillance: mapping, under-reporting, and injury severity in police and hospital records.* Accident Analysis & Prevention, 2005. **37**(6): p. 1102-1113.
30. Amoros, E., et al., *Road crash casualties: characteristics of police injury severity misclassification.* Journal of Trauma and Acute Care Surgery, 2007. **62**(2): p. 482-490.
31. Loo, B.P. and K. Tsui, *Factors affecting the likelihood of reporting road crashes resulting in medical treatment to the police.* Injury prevention, 2007. **13**(3): p. 186-189.

32. McDonald, G., G. Davie, and J. Langley, *Validity of police-reported information on injury severity for those hospitalized from motor vehicle traffic crashes.* Traffic injury prevention, 2009. **10**(2): p. 184-190.
33. Watson, A., B. Watson, and K. Vallmuur, *Estimating under-reporting of road crash injuries to police using multiple linked data collections.* Accident Analysis & Prevention, 2015. **83**: p. 18-25.
34. Couto, A., M. Amorim, and S. Ferreira, *Reporting road victims: assessing and correcting data issues through distinct injury scales.* Journal of safety research, 2016. **57**: p. 39-45.
35. Aptel, I., et al., *Road accident statistics: discrepancies between police and hospital data in a French island.* Accident Analysis & Prevention, 1999. **31**(1): p. 101-108.
36. Lopez, D.G., et al., *Complementing police road-crash records with trauma registry data—an initial evaluation.* Accident Analysis & Prevention, 2000. **32**(6): p. 771-777.
37. Austin, K., *The identification of mistakes in road accident records: part 2, casualty variables.* Accident Analysis & Prevention, 1995. **27**(2): p. 277-282.
38. Osler, T., S.P. Baker, and W. Long, *A modification of the injury severity score that both improves accuracy and simplifies scoring.* Journal of Trauma and Acute Care Surgery, 1997. **43**(6): p. 922-926.
39. DOT, U.J.U.D.o.T.U.G.P.O., Washington, DC, *Transportation implications of telecommuting.* 1993.
40. Mokhtarian, P.L.J.T., *Telecommuting and travel: state of the practice, state of the art.* 1991. **18**(4): p. 319-342.
41. Kugelmass, J., *Telecommuting: A manager's guide to flexible work arrangements*. 1995: Jossey-Bass.
42. Bernardino, A. and M.J.T.R.R.J.o.t.T.R.B. Ben-Akiva, *Modeling the process of adoption of telecommuting: Comprehensive framework.* 1996(1552): p. 161-170.
43. Singh, P., et al., *On modeling telecommuting behavior: option, choice, and frequency.* 2013. **40**(2): p. 373-396.
44. Asgari, H., X. Jin, and A.J.T.R.R.J.o.t.T.R.B. Mohseni, *Choice, Frequency, and Engagement: Framework for Telecommuting Behavior Analysis and Modeling.* 2014(2413): p. 101-109.
45. Walls, M., E. Safirova, and Y. Jiang, *What Drives Telecommuting?:Relative Impact of Worker Demographics, Employer Characteristics, and Job Types.* 2007. **2010**(1): p. 111-120.
46. Alexander, B., M. Dijst, and D.J.T. Ettema, *Working from 9 to 6? An analysis of in-home and out-of-home working schedules.* 2010. **37**(3): p. 505-523.
47. Drucker, J. and A. Khattak, *Propensity to Work from Home: Modeling Results from the 1995 Nationwide Personal Transportation Survey*. 2000. **1706**: p. 108-117.

48. Mannering, J.S. and P.L. Mokhtarian, *Modeling the choice of telecommuting frequency in California: An exploratory analysis.* Technological Forecasting and Social Change, 1995. **49**(1): p. 49-73.
49. Sener, I.N. and C.R. Bhat, *A Copula-Based Sample Selection Model of Telecommuting Choice and Frequency.* Environment and Planning A: Economy and Space, 2011. **43**(1): p. 126-145.
50. Paleti, R. and I. Vukovic, *Telecommuting and Its Impact on Activity–Time Use Patterns of Dual-Earner Households.* 2017. **2658**: p. 17-25.
51. Koenig, B., D. Henderson, and P. Mohktarian, *The travel and emissions impacts of telecommuting for the State of California Telecommuting Pilot Project.* 1996.
52. Nilles, J.M.J.T., *Telecommuting and urban sprawl: mitigator or inciter?* 1991. **18**(4): p. 411-432.
53. Zong, F., Z. Juan, and H.J.T. Jia, *Examination of staggered shifts impacts on travel behavior: a case study of Beijing, China.* 2013. **28**(2): p. 175-185.
54. Balan, L. and R.J.T.R.R. Paleti, *Modified Mixed Generalized Ordered Response Model to Handle Misclassification in Injury Severity Data.* 2018: p. 0361198118796352.
55. Abrevaya, J. and J.A. Hausman, *Semiparametric estimation with mismeasured dependent variables: an application to duration models for unemployment spells.* Annales d'Economie et de Statistique, 1999: p. 243-275.
56. Ramalho, E.A., *Regression models for choice-based samples with misclassification in the response variable.* Journal of Econometrics, 2002. **106**(1): p. 171-201.
57. Paleti, R., *Implicit choice set generation in discrete choice models: Application to household auto ownership decisions.* Transportation Research Part B: Methodological, 2015. **80**: p. 132-149.
58. Bhat, C.R. and V. Pulugurta, *A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions.* Transportation Research Part B: Methodological, 1998. **32**(1): p. 61-75.
59. Anowar, S., et al., *Analyzing car ownership in Quebec City: a comparison of traditional and latent class ordered and unordered models.* Transportation, 2014. **41**(5): p. 1013-1039.
60. Dustmann, C. and A. Van Soest, *An analysis of speaking fluency of immigrants using ordered response models with classification errors.* Journal of Business & Economic Statistics, 2004. **22**(3): p. 312-321.
61. Sullivan, P., *Estimation of an occupational choice model when occupations are misclassified.* Journal of Human Resources, 2009. **44**(2): p. 495-535.
62. Bhat, C.R., *An endogenous segmentation mode choice model with an application to intercity travel.* Transportation science, 1997. **31**(1): p. 34-48.

63. Boes, S. and R. Winkelmann, *Ordered response models.* Allgemeines Statistisches Archiv, 2006. **90**(1): p. 167-181.

64. Greene, W.H. and D.A. Hensher, *Modeling ordered choices: A primer*. 2010: Cambridge University Press.

65. Eluru, N., C.R. Bhat, and D.A. Hensher, *A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes.* Accident Analysis & Prevention, 2008. **40**(3): p. 1033-1054.

66. Mannering, F.L., V. Shankar, and C.R. Bhat, *Unobserved heterogeneity and the statistical analysis of highway accident data.* Analytic Methods in Accident Research, 2016. **11**: p. 1-16.

67. Fountas, G. and P.C. Anastasopoulos, *A random thresholds random parameters hierarchical ordered probit analysis of highway accident injury-severities.* Analytic methods in accident research, 2017. **15**: p. 1-16.

68. Schiff, M.A. and P. Cummings, *Comparison of reporting of seat belt use by police and crash investigators: variation in agreement by injury severity.* Accident Analysis & Prevention, 2004. **36**(6): p. 961-965.

69. Eluru, N. and C.R. Bhat, *A joint econometric analysis of seat belt use and crash-related injury severity.* Accident Analysis & Prevention, 2007. **39**(5): p. 1037-1049.

70. Greene, W., *Functional forms for the negative binomial model for count data.* Economics Letters, 2008. **99**(3): p. 585-590.

71. Gurmu, S., *Generalized hurdle count data regression models.* Economics Letters, 1998. **58**(3): p. 263-268.

72. Lord, D., S.P. Washington, and J.N. Ivan, *Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory.* Accident Analysis & Prevention, 2005. **37**(1): p. 35-46.

73. Paleti, R., *Generalized Extreme Value models for count data: Application to worker telecommuting frequency choices.* Transportation Research Part B: Methodological, 2016. **83**: p. 104-120.

VITA

Lacramioara Balan was born to his parents, Luminita and Ioan Balan, in Bacau, Romania. She did most of her schooling in two cities: Bacau and Bucharest. After completing high school in the Technical College "Gheorge Asachi", Bacau, she joined Technical University of Civil Engineering, Bucharest in the year 2009 to pursue undergraduate studies in Civil Engineering. She successfully graduated with a Bachelor's degree in Civil Engineering in the year 2013. In fall 2013, she continued with the graduate studies at the same university, in Geomatics under the guidance of Dr. Valentin Danciu. She received her M.S. degree in Civil Engineering from the Technical University of Civil Engineering at Bucharest in June 2015, and later continued for her doctoral degree under the guidance of Dr. Rajesh Paleti.

Permanent Address: 85 Matei Basarab,bloc L119, Sector 3,
Bucharest 030573, Romania

This Dissertation was typed by the author.