WASHINGTON UNIVERSITY
THE HENRY EDWIN SEVER GRADUATE SCHOOL
DEPARTMENT OF CIVIL ENGINEERING

THE RELATIONSHIP OF FACTORS CONTRIBUTING TO THE FAULT OF PARTIES INVOLVED IN CYCLIST-MOTORIST COLLISIONS

by

Craig Wademan Schiller

Prepared under the direction of Professor Gudmundur Freyr Ulfarsson

Thesis presented to the Henry Edwin Sever Graduate School of
Washington University in partial fulfillment of the
requirements of the degree of
MASTER OF SCIENCE
December 2005
Saint Louis, Missouri

WASHINGTON UNIVERSITY

THE HENRY EDWIN SEVER GRADUATE SCHOOL

DEPARTMENT OF CIVIL ENGINEERING

ABSTRACT

THE RELATIONSHIP OF FACTORS CONTRIBUTING TO THE FAULT OF PARTIES INVOLVED IN CYCLIST-MOTORIST COLLISIONS

by

Craig Wademan Schiller

ADVISOR: Professor Gudmundur Freyr Ulfarsson

December 2005

Saint Louis, Missouri

The determination of fault, not in the legal sense but rather of the act of committing a violation, in cyclist-motorist collisions may be considered in terms of a "civilian's perspective" and an "officer's perspective". These perspectives may be represented by an ordered probit, and either two independent binary probit models or a single bivariate probit respectively. Simulation enables comparison between the three forms. The ordered probit predicts fault with at least a 62.5% accuracy 99% of the time provided fault must be assigned to either party. Without such provision, accuracy drops to 60.9% for the dataset upon which the other models are simulated. In estimation of the two binary probit models, it was found best to use only observations where a party was solely at fault. The result was a prediction with at least a 54.8% accuracy 99% of the time. For the purposes of statistical efficiency, the models should not be estimated separately. Thus a bivariate probit is used to simultaneously estimate the models. Like the binary probit, it was found best to estimate the model from only sole fault. The result was a prediction with at least a 55.6% accuracy 99% of the time. In comparison of the models the difference in perspectives accounts for 92% of the variation – suggesting the "civilian's perspective" may better represent the decision process.

to my parents who made all of this possible

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank the committee members who took their time to review this thesis: Dr. Gudmundur Freyr Ulfarsson, Dr. Phillip L. Gould, Dr. Kevin Z. Truman, and Dr. Sungyop Kim. Furthermore, I would like to thank Dr. Ulfarsson, my graduate advisor, for his assistance and regular reviews of my master's work. I would also like to thank Dr. Kim for recommending the topic and preparing the dataset.

Special thanks to Dr. Truman for persuading me to come to Washington University in St. Louis. I would also like to thank Prof. Bill Kankolenski and Prof. Jerry Craig for their confidence in me, particularly during my first year. I would like to thank Dr. Luis Porrello for convincing me to enter transportation engineering and Prof. John Brancaglione for encouraging me to continue particularly in its planning aspects.

Thanks to Prof. Michael L. Nobs, a brilliant professor, who taught me many of the concepts that appear here and will appear in my future work. Also, thanks to Dr. Muthanna Al-Dahhan whose instruction on numerical methods and simulation was invaluable to this study. I would also like to thank Kimberly Jones & Linda Buckingham for putting up with me.

Finally, I would like to thank Dr. Phillip L. Gould, my undergraduate advisor, for his advice, without which I would certainly not have made it here today.

Chapter 1

Introduction

The relationship of factors contributing to the fault of parties involved in cyclist-motorist collisions may be revealed through the use of various econometric models. For each model, such relationships may be justified with inferences regarding the physical meaning of the phenomena. The potential to make valid inferences of physical meaning goes to motivation. To further this motivation, the models are themselves selected for their hypothesized physical meaning. The chapter begins with a discussion of motivation, followed by separately detailing the approaches of data oriented modeling and decision oriented modeling.

1.1 Motivation

The study is motivated by the potential to make valid inferences of physical meaning behind the factors contributing to the fault of parties involved in a cyclist-motorist collision. Furthermore, the form of the model utilized may itself have physical meaning. Even outside the context of cyclist-motorist collisions, the methods have broad application in both the public and private sector of society.

1.1.1 Application within the Public Sector

Governments and public organizations look to models in order to evaluate existing and implement new programs, policies, etc. Their interests are in evaluating both factors that may influence outcomes and also processes that may influence outcomes.

In consideration of evaluating the factors that influence the outcome, if the study were to report that, say, the cyclist was more likely to be at fault when it is dark, programs may be put in place to provide cyclists with warning lights or policies could be made to prohibit cyclists out at night. And, while it should be noted that, without further information as to why the phenomena occurs, such interpretations are essentially speculation as to a cause, they also are potentially viable explanations in a set of many viable explanations upon which public figures may choose to act. As a matter of maintaining the science, the public figures acting in their role as decision makers may implement alternatives as hypotheses to be tested. For example, here, they hypothesize that warning lights reduce the likelihood for a cyclist to be at fault, they implement a program to provide cyclists with such gadgets, and they wait to see whether there was a significant reduction in cyclist at fault collisions.

In consideration of evaluating the processes that influence the outcome, consider the selection of the model. Each model may be hypothesized to capture a certain process, such as the processes influencing the decision to assign fault to the individual parties involved in a collision. Say it were hypothesized that the individual assigning fault just twirls the spinner from “The Game of Life®” modulus two, then the strength of a model using a uniform random distribution to represent such a decision making process may suggest whether this hypothesis may be accepted. Policy makers could then modify the process if need be. For the example of randomly assigning fault to the parties involved, a strong model may suggest that they should investigate and potentially revamp their fault assignment methods.

1.1.2 Application within the Private Sector

Corporations, individuals and other private sector entities often look to models for the same reasons as found in the public sector: to evaluate existing and implement new programs, policies, etc. The difference lies within the criteria with which they evaluate each. Consider their interests in evaluating both factors that may influence outcomes and also processes that may influence outcomes.

In consideration of factors that influence outcomes, private sector entities see how it may further their interests. Should the study report that a cyclist was more likely to be at fault in when it is dark, a corporation may devise a new warning light system to sell to cyclists or an individual cyclist may choose to wear an illuminated vest. Again, it is noted that these are plausible explanations in a set of many plausible explanations to describe the phenomena of a cyclist's increased likelihood of fault given that the accident occurred while it was dark; it is up to the discretion of the entity to further study the relationships before proceeding with a plan.

Another tool given the factors that influence outcomes is the ability to predict an outcome given a set of factors. For example, an actuary on behalf of an insurance company may determine the risk associated with an individual cyclist or motorist of being at fault in a collision. Another application in insurance is allocating blame when deciding claims since a strong model may be extrapolated to predict the fault of each party.

In consideration of the processes that influence the outcome, private entities, again, seek to further their interests. Certain processes may describe certain outcomes. If the determinations of fault were made by a judge, one might go about selecting a court based upon the model hypothesized to describe the decision process that favors a particular outcome. For example, a motorist who committed various violations may seek a judge who decides cases by tossing a coin.

1.2 Data Oriented Modeling

Data oriented modeling examines correlations within the data to select the model to describe it. An example of such modeling is creating a histogram, noting the bell-shape and selecting a model with a standard normal or a Student's *t* distribution. Not that the approach is incorrect, it is just that it does not aim to describe why such distribution makes sense in the real-world.

To its credit, the data oriented approach is an easy method to quickly capture the relationship of the variables in the dataset. So, when the motivation is to create a model to determine the relationship of factors in cyclist-motorist collisions in terms of the available data, this is a quick and easily justifiable method. The problem is that it does not describe the process by which the outcome is decided.

1.3 Decision Oriented Modeling

Decision oriented modeling selects the model based on how the data should relate rather than how it may in the dataset. An example of such modeling is selecting a model with a standard normal distribution to describe error in a chemistry experiment. Granted error is by no means a "decision" per se, but it does reflect the formulation of an outcome. Indeed, the selection of a model requires an advanced understanding of the relationship of interest; however, positing one plausible relationship to compare with another plausible relationship may be used in gaining understanding of the relationship.

Selecting a model requires qualitative insights and forethought. For example, if it were known that cyclists were at least partially at fault 71% of the time (which they are in the dataset), one might simply create a model to reflect this proportion and do quite well. Clearly, or at least reasonably, a model based on this information, while strong, is meaningless unless qualified by such limitations as those which bound the art of forecasting.

For the case of how fault is assigned to a party involved in a cyclist-motorist collision, there are also many plausible processes describing that determination. Here, the processes behind the decision are described in terms of hypothetical decision makers embodying the “civilian’s perspective” and “officer’s perspective” as defined below.

1.3.1 A “Civilian’s Perspective”

For the purposes of describing a “civilian’s perspective”, a “civilian” refers to a stereotype of the sort of generic, reasonable person who, when presented a hypothetical scenario of a speeding, intoxicated, heavily-armed motorist that overtakes and ultimately collides with a cyclist who failed to signal a right-hand turn, would be so disgusted by the severity of the motorist’s transgressions that the cyclist’s contribution would be disregarded. To “civilians”, fault is relative. *Really bad* conceals *somewhat bad*. Thus, this blame game has an associated ordering from the driver at fault to the cyclist at fault with a category there between for, when assignment is indistinguishable, both at fault. The “civilian’s perspective” may describe the assignment of fault by insurance companies or courts interested in weighing the contribution of the parties to the situation at hand.

1.3.2 An "Officer's Perspective"

For the purposes of describing an "officer's perspective", an "officer" refers to a stereotype of the sort high-caliber, impersonal individual who, when confronted with the same hypothetical scenario of a speeding, intoxicated, heavily-armed motorist that overtakes and ultimately collides with a cyclist who failed to signal a right-hand turn, would look at the motorist and mark the columns for speeding, intoxication and overtaking, and then look at the cyclist and mark the column for failure to signal a turn – checks in both rows, both are at fault. To the "officer", fault is absolute. *Lesser wrongs* do not equate to *rights*. Thus fault is determined for the motorist, and fault is determined for the cyclist. The determination of the fault of one party is made independent of the other. The "officer's perspective" may describe the assignment of fault by individuals entrusted by society to reserve judgment for the courts and operate fairly to enforce the law as written.

Chapter 2

Literature Review

In 2003, there was an estimated 100000 reported accidents involving cyclists and motorists in the United States of America of which 28000 were injury accidents and 700 were fatal injury accidents (National Safety Council, 2004). Many studies have been done to capture the relationship of variables present in vehicular accidents; some studies have gone further to discuss the affects of fault; and a few studies have even explicitly modeled the assignment of fault. The contribution this study attempts to make is the consideration of the form of the model as a relation to the decision-making process behind the assignment of fault; this consideration hence controls selection of an appropriate model for which to model the assignment of fault. The literature reviewed breaks down into the three areas of factors affecting accidents, fault affecting accidents, and factors affecting fault.

2.1 Factors Affecting Accidents

The factors affecting accidents are discussed in many previous studies. These types of factors recognized shaped the variables used in the model as discussed in the next chapter. One factor is the effects as that of stress as discussed in Legree, Heffner, Psotka, Martin and Medsker (2003). Van der Flier and Schoonman (1988) discuss affects of time of day, change of shifts, age of driver, hours on duty, length of service and experience in terms of railway stop-signal abuse. Hanowski, Wierwille and Dingus (2003) investigate fatigue in the context of local trucking. The effects of various sleep related issues are shown in a study by Sagberg (1999).

Regarding accident propensity, Stamatiadis and Deacon (1995) make various observations particularly with respect to aging. They found middle-aged drivers to be safer than younger drivers and younger drivers to be safer than older drivers. They assert that female drivers are on average safer than male drivers, the same relationship holds for younger drivers, but the relationship inverts for older drivers.

2.2 Fault Affecting Accidents

There are a few studies that include the fault of a party among the other factors affecting accidents. Rerrari and Russell (2001) discuss the affect of fault in an accident as part of their biopsychosocial model of whiplash injuries. Vernon, Diller, Cook, Reading, Suruda and Dean (2002) compared rates of “adverse driving events” including at-fault determinations for licensed Utah drivers with medical conditions. Underwood, Chapman, Wright and Crundall (1999) describe fault’s affect on the provocation of feelings of anger. Stewart (2005) describes the individual assignment of responsibility and the implications on the emotional health of survivors of collisions. Clarke, Ward and Jones (1998) used the driver “most at fault” to describe ten types of overtaking accidents.

How fault affects seatbelt use is used in the context of the drivers’ risk assessments in Calisir and Lehto (1996). Another idea relating to self-regulation is shown in Ball, Owsley, Stalvey, Roenker, Sloane and Graves (1998). Ball et al. (1998) concludes that older drivers who were determined to be at fault in a collision within the past five years are more likely to avoid such potentially challenging driving situations as rain, dark, and heavy traffic.

Quddus, Noland and Chin (2002) analyzed motorcycle injury and vehicle damage severity with ordered probit models to determine that a motorcyclist was more likely to have severe injuries if determined to be at fault for the accident. Dissanayake and Lu (2002) found a similar relationship between fault and severity except in the context of elderly drivers involved in fixed object crashes via two sets of sequential binary logistic regression models.

Solnick and Hemenway (1995) discuss the affect of fault in terms of its affect on the “hit-and-run” in fatal pedestrian accidents. They conclude that drivers are more likely to run when at fault, as suggested by a positive state of intoxication, or simply when escape seems easy. Such interpretations of fault may be built upon as potential factors upon which to base the determination of fault.

Fault was used by Al-Balbissi (2003) to identify reasons for the differences in accidents. In such a manner McCartt, Northrup and Retting (2004) reviewed the diagrams and narrative descriptions found in police reports detailing ramp-related motor vehicle crashes in Northern Virginia to identify the type of accident and determine what factors influenced the at-fault driver. Their study described the facility type, actions taken, and other characteristics. With this they concluded with proposals for countermeasures that may be taken to mitigate the affects of the characteristics suggested to cause the problems, such as increased length of acceleration lanes to reduce sideswipe/cutoff crashes.

A study by Gruder, Romer and Korth (1978) used a female to call randomly selected phone numbers to request for help. Based on such factors as her fault for her troubles, the telephone subscriber's assistance was measured. It was found that the victim's dependency is a determinant for helping; and the greater the dependency the greater the helping. A study by Peltzer and Renner (2004) describes the psychosocial effects of fault; that is the effect of fault with respect to others involved in the traffic accidents studied. They conclude that, in the context of how South Africans cope with trauma, holding another party responsible lowers one's psychological well-being. Gruder et al. (1978) and Peltzer and Renner (2004) contribute to this study by positing aspects that may affect the assignment of fault.

2.3 Factors Affecting Fault

The factors affecting fault are discussed in a few studies. Legree, Heggner, Psotka, Martin and Medsker (2003) study how stress elevates at-fault crash risk. Sjögren, Eriksson and Öström (1996) describe how drivers with "intrinsic medical factors" were often at fault and were even more so for the elderly group. The increased likelihood for an elderly driver to be at fault was briefly acknowledged by McGwin and Brown (1999). Raedt and Ponjaert-Kristoffersen (2001) attempt to predict at-fault auto accidents of older elderly drivers based on tests administered to older drivers. Hing, Stamatiadis and Aultman-Hall (2003) describe the negative impacts of multiple passengers on the probability an elderly driver is found at-fault in an accident. A similar study shown in Preusser, Ferguson and Williams (1998) describing the effect of passengers in the context of younger drivers reveals that the presence of passengers is associated with greater risk of at fault collisions for drivers under twenty-four years of age.

In the context of involvement in fatal vehicle accidents, Garretson and Peck (1982) found drivers that were found to be at fault in a wreck were found to have worse driving records than those drivers in the general driving population. Elliot, Waller, Raghunathan, Shope and Little (2000) describe how the odds of one being at fault is increased by nearly 50% if that one had been at fault in an accident within the previous year. In a later study, Waller et al. (2001) describe how the risk of being at fault in an accident decreases with years of licensure. An earlier set of studies with similar conclusions is available in Chen, Cooper and Pinili (1995) and in Cooper, Pinili and Chen (1995). Chen et al. found a driver previously determined to be at fault in an accident would be more likely to be involved in another such accident in the future. Cooper et al. (1995) examined the effects of restrictions on licensure of novice drivers through the graduated licensing system.

Retting Weinstein and Solomon (2003) find that, for motor vehicle accidents at stop signs, drivers under eighteen and drivers over sixty-five are more likely to be found at fault. Kim, Li, Richardson and Nitz (1998) discuss the influences of age, gender and vehicle type on the risk of being at fault in an accident. Yannis, Golias and Papadimitriou (2005) describe at-fault risk in terms of driver age and motorcycle engine size. The paper most related to determining the fault of parties involved in a cyclist-motorist collision is the paper that does just that by Kim and Li (1996).

Kim and Li (1996) utilized a logistical regression analysis of the likelihood of drivers being at fault as based on data collected by the police for the state of Hawaii from 1986 to 1991. The variables they made use of were motorist age and the square thereof, cyclist age and the square thereof, cyclist use of alcohol, cyclist helmet use, driver turning action, cyclist turning action, and rural area. In their study, the square of driver's age, the cyclist's age, the cyclist's use of helmet, and the driver's turning action all implicated the driver; whereas the driver's age, the square of the cyclist's age, the cyclist's use of alcohol, the cyclist's turning action, and the rural area did not implicate the driver.

This study makes use of many of the same variables as these prior, including such functional relationships as the one utilized by Kim and Li (1996) to square the age. Indeed, while comparison will likely reveal the same relationship of the factors within the models created herein, such is not the purpose of the study. Variables will be included and excluded based on availability and significance rather than for comparison to prior studies. This study compares various models representing different perspectives on the decision to assign on party a violation, and, in that respect, it is a contribution to the field of modeling factors relating to the assignment of fault for a collision.

Chapter 3

Methodology

Estimation of an econometric model finds a mathematical relationship between the data by fitting it to the model's form. However, the selection of the proper form is often a reflection of the relationship itself. For the case of the determination of fault between a cyclist and a motorist, there are multiple plausible relationships. Thus multiple models must be examined and compared in order to properly determine the relationship. Here, models are based on two perspectives: (1) that of a civilian determining fault; and (2) that of an officer of the law determining fault.

3.1 A Civilian's Perspective

Chapter 1 introduces the "civilian's perspective" in terms of a stereotype of a "civilian" and how that "civilian" assigns fault to parties involved in a collision. To summarize the perspective, fault is a relative. That is there is an associated ordering from the driver at fault to the cyclist at fault with a category there between for, when assignment is indistinguishable, both at fault.

3.1.1 An Order

A "civilian" weighs factors: factors implicating the cyclist tip the balance to the right: factors implicating the motorist tip the balance to the left. Illustrated, this becomes the ordered fault line in Figure 3-1.

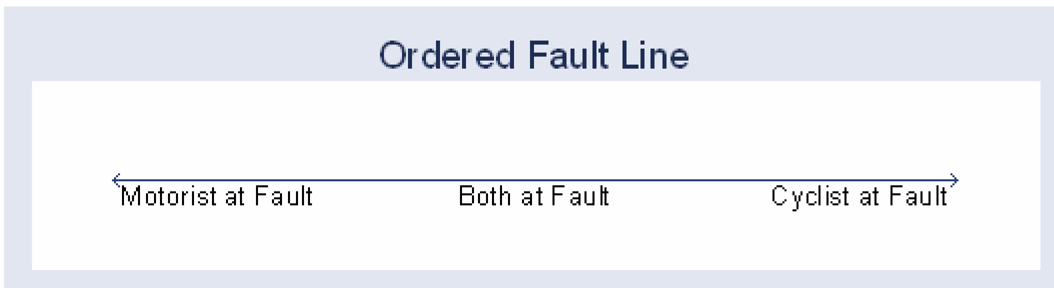


FIGURE 3-1. Ordered Fault Line

The designation of motorist at fault, both at fault, and cyclist at fault are ordered categories. An ordered probit model may describe this relationship.

3.1.2 Ordered Probit

The ordered probit model has long been used to describe ordered alternative selection for such applications as the description of customer satisfaction (in, say, gradations of unsatisfied, somewhat unsatisfied, somewhat satisfied, and satisfied) in terms of such other factors as customer age, gender, location, etc. In more general terms, the ordered probit model describes the ordered outcomes by assigning coefficients to the variables thought to explain the phenomena – the more positive a coefficient, the greater the outcome level. For three outcomes ordered with motorist at fault on the low end, both at fault in the middle, and cyclist at fault in the high end, negative coefficients pull the fault toward the motorist, positive coefficients pull the fault toward the cyclist, and both at fault is the neutral case.

Description

The ordered probit is used as it appears in Intercooled Stata® 8.1; commands to implement the ordered probit in the software are included in Appendix C. The ordered probit associates a utility with each outcome in a form presented by Train (2003) as having observed and unobserved parts:

$$U_n = \boldsymbol{b}\boldsymbol{x}_n + \epsilon_n \tag{3-1}$$

where $\boldsymbol{b}\boldsymbol{x}_n$ is the observed portion of utility known as the probit index comprised of a vector of estimable coefficients, $\boldsymbol{b}$, multiplied by the observation specific variables, $\boldsymbol{x}_n$, and ϵ_n is the unobserved factors, random variation or error. It is assumed that utility is normally distributed about zero with a standard deviation of one. This standard normal distribution is shown as:

$$\phi(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) \tag{3-2}$$

The ordered probit assumes a distribution of standard normal, so, in calculation, it uses a cumulative standard normal distribution:

$$P(t \leq y) = \Phi(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt \tag{3-3}$$

Probabilities are calculated based on the probability that the outcome lies between a set of cutoff points. Thus it takes the form presented by Train (2003) and StataCorp (2003):

$$P_{ni} = \Phi\left(k_i - \boldsymbol{b}\boldsymbol{x}_n\right) - \Phi\left(k_{i-1} - \boldsymbol{b}\boldsymbol{x}_n\right) \tag{3-4}$$

where Φ is the standard cumulative normal distribution described in equation 3-3, $\boldsymbol{b}\boldsymbol{x}_n$ is the observed portion of the utility described in equation 3-1, and k_i and k_{i-1} are the cutoff points above and below the predicted outcome respectively.

The model is estimated by maximum-likelihood. The log-likelihood function to be maximized takes the form:

$$\mathrm{LL} = \sum_{n=1}^{N} \ln\left(P_{ni}\right) \tag{3-5}$$

where P_{ni} is the probability of the experienced outcome i for observation n, and N is the sample size. For further discussion on numerical maximization, consult Train (2003).

One indicator of the strength of the model compared to a naïve constant-only model is the pseudo r-squared, ρ^2:

$$\rho^2 = 1 - \left(\frac{LL_{end}}{LL_{start}}\right) \tag{3-6}$$

where LL_{start} is the log-likelihood at starting with a naïve constant-only model, and LL_{end} is the log-likelihood at convergence.

Simulation

The ability for a model to perform accurately in simulation reflects upon the ability for the model to account for the relationship in the data. It is also a method by which to compare models of varying structure, such as this ordered probit representing the civilian's perspective with those models representing the officer's perspective of the following section.

Before simulation can occur, Stata® is used to predict the probabilities associated with each outcome for each observation. The predicted probabilities are exported to a spreadsheet. A Perl script is used to simulate (see Appendix B). Since each of the three outcomes have a given probability, their sum for a given observation equal one. Thus a prediction may be made by comparing a random value between zero and one to the ranges of an outcome within the summation of the three. The script makes a prediction in such a manner for each observation in order to obtain the percentage of observations whose outcomes were correctly predicted. The script iterates over all observations 200000 times, each time recording the percentage of correctly predicted outcomes.

When simulation is complete, the percentage accurately predicted are sorted and identified by their percentiles. Plotting the percentiles on the percent accuracy creates an s-curve which visually presents the confidence in the accuracy of the model.

3.2 An Officer's Perspective

Chapter 1 introduces the "officer's perspective" in terms of a stereotype of an "officer" and how that "officer" assigns fault to parties involved in a collision. To summarize the perspective, fault is absolute. That is the determination of the fault of one party is made independently of the other.

3.2.1 Independent Estimation

For an "officer" there are two determinations: (1) the fault of the motorist; and (2) the fault of the cyclist. Two choices for each, a binary: at fault, not at fault. A combination of two separate binary probit models, representing the assignment of fault to each party, forms a model that, in simulation, provides all outcome possibilities: motorist at fault; cyclist at fault; both at fault; and neither at fault. Alternatively, the role of a single "officer" making the determination for both may be acknowledged by using a single bivariate probit model.

Taking it a step further, if these determinations are truly independent, then each will be based on the same characteristics. This means that a cyclist solely responsible for a collision would have his or her fault determined based on the same characteristics as a cyclist who shares fault with the driver. So the model may be estimated on either partial-fault or sole-fault designations (represented visually in Figure 3-2 and Figure 3-3 respectively).

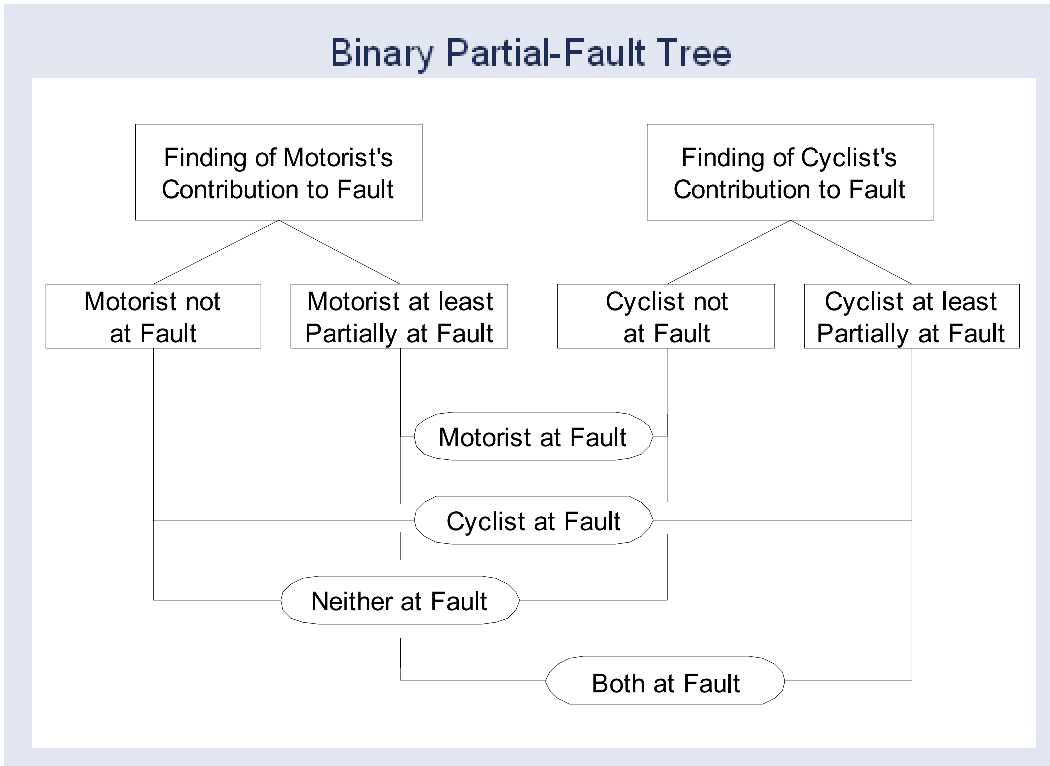


FIGURE 3-2. Binary Partial-Fault Tree

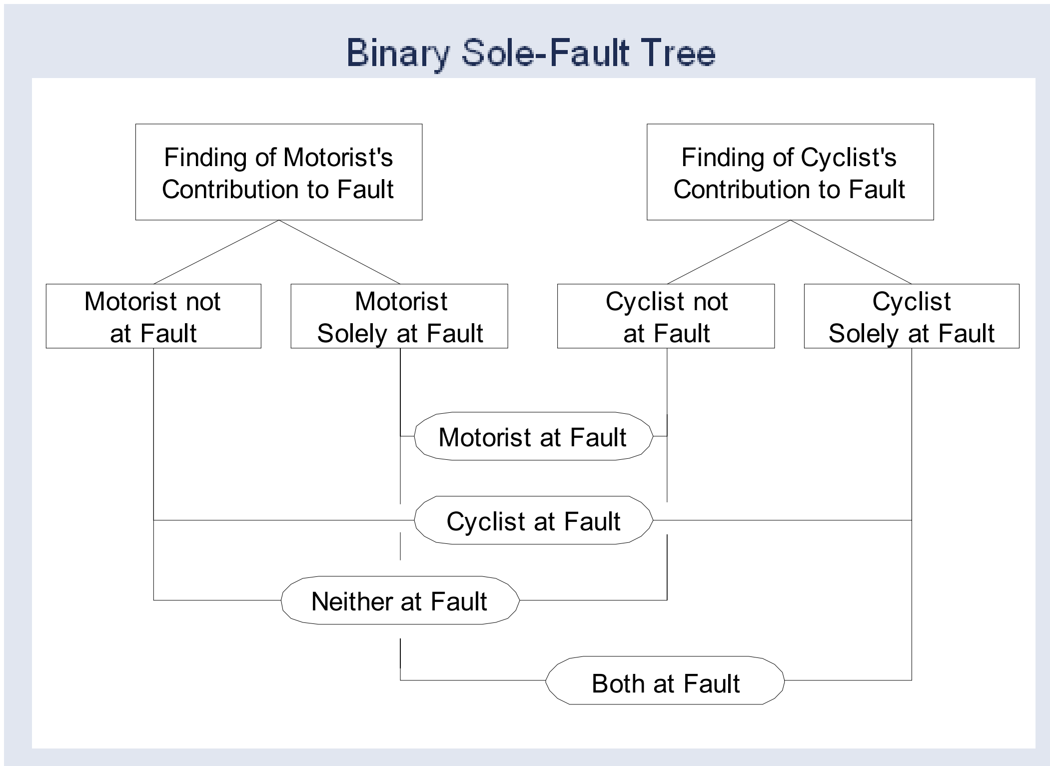


FIGURE 3-3. Binary Sole-Fault Tree

Comparing models estimated with either partial-fault or sole-fault scenarios may provide insight into the strength of the independence of the determinations. For example, the sole-fault case will likely prove a stronger predictor when extrapolated to determining partial fault, because, if the fault is not truly independently assigned in the partial case, the determination of sole fault will be less convoluted, the relationship clearer.

The model will be estimated for both sole-fault and partial-fault designations. But, before getting into the details of the binary probit and bivariate probit, a certain "problem" with estimation should be discussed.

Estimation Paradox

There appears to be a slight problem with the estimation of the model. Since neither at fault also includes fault undetermined, the category is improper to use in estimation. Thus the motorist not at fault in Figure 3-2 and 3-3 equates to cyclist solely at fault – a result that cannot yield the asserted full array of fault determinations!

So there are discrepancies with what the model is estimating and what is meant to be shown. For the partial-fault model, what had been shown as the categories of motorist not at fault and cyclist not at fault which ultimately combine to neither at fault, equates, in estimation, to cyclist solely at fault and motorist solely at fault which combine to a second category of both at fault. For the sole-fault model has the same problem as the partial-fault model, but it also suggests without justification that the combination of a motorist solely at fault and a cyclist solely at fault makes both at fault.

Visually, the partial-fault and sole-fault models break down as shown in Figure 3-4 and Figure 3-5, respectively. The discrepancies with respect to the desired breakdowns in Figure 3-2 and Figure 3-3 are highlighted.

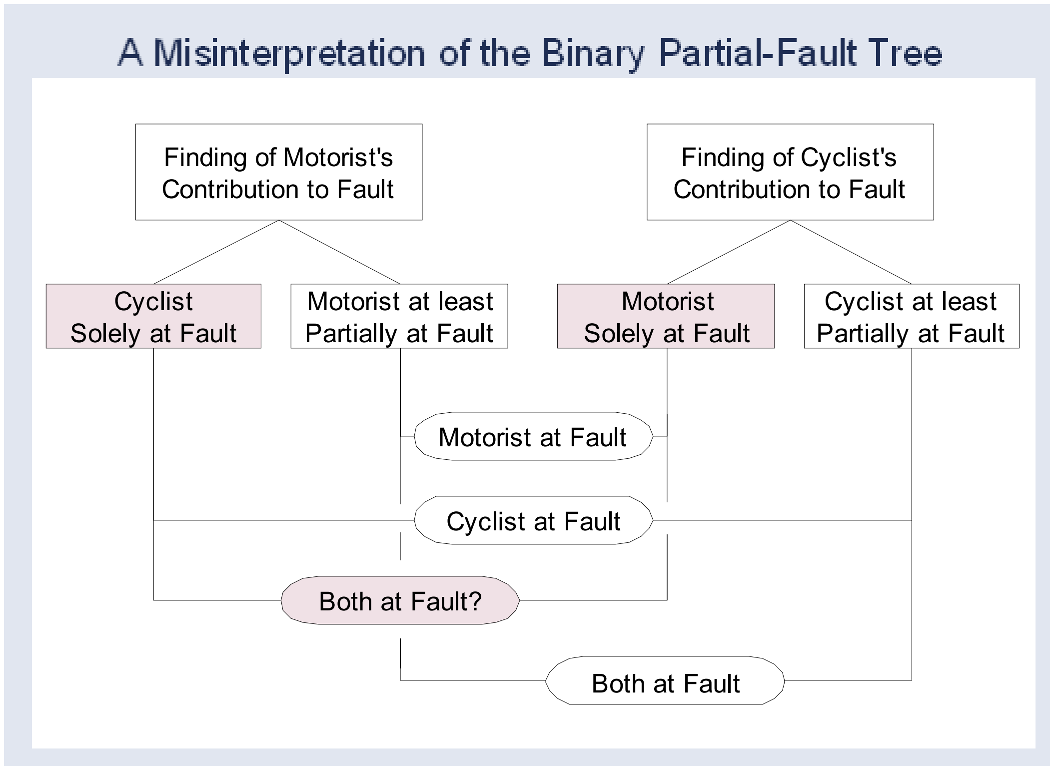


FIGURE 3-4. A Misinterpretation of the Binary Partial-Fault Tree

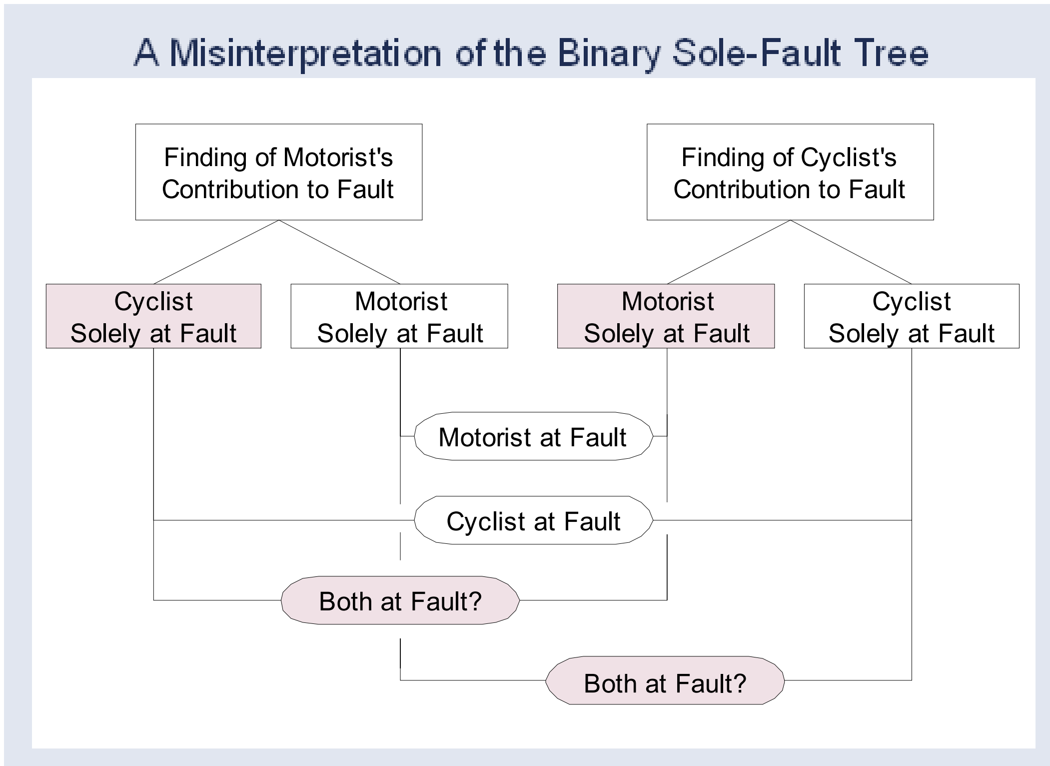


FIGURE 3-5. A Misinterpretation of the Sole-Fault Tree

This is not a paradox. The models are independent by selection of explanatory variables used in estimation of the parties' respective fault.

Explanatory Variables

While a model could literarily be distinguishing between a party at fault and another party at fault, if the variables explain nothing of the other party then they may not be used to determine the other party's fault. For example, determining the fault of the motorist may be done with such variables as driver intoxication, speeding, overtaking the cyclist, or vehicle turning; but without variables unrelated to the driver's actions such as cyclist on the wrong side of the street. The result is a model that implicates the driver but not the cyclist when extrapolated to the entire dataset in which fault is acknowledged. With this restriction of independence upon the explanatory variables, the design is not shown to be inappropriate.

3.2.2 Binary Probit Models

The binary probit model has long been used to describe selections between two alternatives for such applications as the description of going to work or staying home from work in terms of such other factors as snow, ice, temperature, career, etc. A binary probit model describes a phenomena of two outcomes, here fault or the absence thereof, by assigning coefficients to the variables thought to explain the phenomena – the more positive a coefficient, the more it contributes to the likelihood of a party being at fault. Since fault is assumed to be determined independently for the driver and cyclist, each determination is estimated by its own binary probit: the first binary probit describes whether or not the motorist is at fault; the second binary probit describes whether or not the cyclist is at fault. The two are later combined by simulation.

Description

The probit is used as it appears in Intercooled Stata® 8.1; commands to implement the probit in the software are included in Appendix C. The probit is based on observed and unobserved portions of utility as shown in equation 3-1. As in the ordered probit, the model is assumed to be distributed standard normal by equation 3-2. Probabilities are calculated based on the relative utility of outcomes shown by:

$$P(U_{n1} \geq U_{n0}) \tag{3-7}$$

where U_{n1} and U_{n0} are the utilities for the first outcome and the base outcome respectively. Since utilities are relative, the base case may be considered as having zero utility, so equation 3-7 becomes:

$$P(U_{n1} \geq 0) \tag{3-8}$$

where U_{n1} is the utility for the first outcome. By plugging in the observed and unobserved parts for a utility presented in equation 3-1, this becomes:

$$P(\epsilon \geq -\boldsymbol{b}\boldsymbol{x}_n) \tag{3-9}$$

where $\boldsymbol{b}\boldsymbol{x}_n$ is the observed portion of utility known as the probit index comprised of a vector of estimable coefficients, $\boldsymbol{b}$, multiplied by the observation specific variables, $\boldsymbol{x}_n$, and ϵ_n is the unobserved factors, random variation or error. Since the normal distribution is symmetric:

$$P(\epsilon \leq \boldsymbol{b}\boldsymbol{x}_n) \tag{3-10}$$

The probability of the outcome is the area under the normal curve before the probit index, so the cumulative standard normal function in equation 3-3 is used. Thus it takes the form of the binary probit presented by Train (2003) and the probit presented by StataCorp (2003):

$$\begin{aligned} P_{n1} &= \Phi\left(\boldsymbol{b}\boldsymbol{x}_n\right) \\ P_{n0} &= 1 - P_{n1} \end{aligned} \tag{3-11}$$

where Φ is the standard cumulative normal distribution and $\boldsymbol{b}\boldsymbol{x}_n$ is the probit index. As with the ordered probit, the model is estimated via the method of maximum-likelihood. The log-likelihood takes the form of equation 3-5 and the ρ^2 is determined by equation 3-6.

Simulation

As stated before, simulation is an important tool to gauge the accuracy of the model and to compare with models of different forms. But for the case of two separate models, simulation is required to join the two. Simulation uses one binary probit model to predict the fault of the motorist and then the other binary probit model to predict the fault of the cyclist. The combination provides all of the fault determination alternatives via four joint outcomes:

1. motorist at fault and cyclist at fault,
2. motorist at fault and cyclist not at fault,
3. motorist not at fault and cyclist at fault, and
4. motorist not at fault and cyclist not at fault.

Once again, a Perl script is used to simulate (see Appendix B). To begin, Stata® is used to predict the probabilities associated with each outcome for each observation. The predicted probabilities are exported to a spreadsheet. The script is run by its interpreter for the simulation. There are two probabilities for each observation, that of the cyclist being at fault and that of the motorist being at fault. Thus a prediction may be made by comparing one random value between zero and one to the probability that the cyclist is at fault and another random value between zero and one to the probability that the motorist is at fault. When these predictions are considered jointly, they describe all four fault determination alternatives. The script makes a prediction in such a manner for each observation in order to obtain the percentage of observations whose outcomes were correctly predicted. The script iterates over all observations 200000 times, each time recording the percentage of correctly predicted outcomes.

When simulation is complete, the percentage accurately predicted are sorted and identified by their percentiles. Plotting the percentiles on the percent accuracy creates an s-curve which visually presents the confidence in the accuracy of the model.

3.2.3 Bivariate Probit Model

A bivariate probit model does what the binary probit does except it does it simultaneously estimates the two seemingly independent determinations for the parties. In this manner, the knowledge that a single officer makes the determination can be utilized to increase statistical efficiency. While such other popular models as the multinomial probit and the multinomial logit also simultaneously estimate the possible outcomes, these do not maintain the separation between the assignments of fault for each party. Since the bivariate probit model determines fault for each party simultaneously, no special combination is required as done for the two binary probit models.

Description

The bivariate probit is used as it appears in Intercooled Stata® 8.1; commands to implement the bivariate probit in the software are included in Appendix C. Like the models of the previous sections, the bivariate probit is based on the idea of observed and unobserved utility with a distributed standard normal as in equation 3-2. Additionally, bivariate probit takes into account the relation between the two equations. Probabilities are calculated based on the probability that the outcome lies in the area before the probit index so calculation makes use of the cumulative normal distribution in equation 3-3. Thus it takes the form of the bivariate probit presented by StataCorp (2003):

$$\begin{aligned} P_{n11} &= \Phi_2\left(\boldsymbol{b}\,\boldsymbol{x}_n, \boldsymbol{g}\,\boldsymbol{z}_n, \eta\right) \\ P_{n10} &= \Phi_2\left(\boldsymbol{b}\,\boldsymbol{x}_n, -\boldsymbol{g}\,\boldsymbol{z}_n, -\eta\right) \\ P_{n01} &= \Phi_2\left(-\boldsymbol{b}\,\boldsymbol{x}_n, \boldsymbol{g}\,\boldsymbol{z}_n, -\eta\right) \\ P_{n00} &= \Phi_2\left(-\boldsymbol{b}\,\boldsymbol{x}_n, -\boldsymbol{g}\,\boldsymbol{z}_n, \eta\right) \end{aligned} \tag{3-12}$$

where subscripts 11, 10, 01 and 00 reference positive outcome for both equations, the first equation only, the second equation only and neither equation respectively, Φ_2 is the bivariate standard normal distribution, η is the covariance in the errors of the two binary probits comprising this bivariate probit, and $\boldsymbol{b}\,\boldsymbol{x}_n$ and $\boldsymbol{g}\,\boldsymbol{z}_n$ are the probit indexes for the first and second equations respectively. The method of maximum-likelihood is used to estimate the model. The log-likelihood takes the form of equation 3-5.

Simulation

As with the binary probit, simulation remains an important tool to gauge the accuracy of the model and to compare with models of different forms; however, it is no longer utilized for the purposes of creating joint outcomes. Since the bivariate probit estimates both outcomes simultaneously, it will also simultaneously provide predictions for the probability of both at fault, motorist at fault, cyclist at fault, and neither at fault.

As before, Stata® is used to predict the probabilities associated with each outcome for each observation. The predicted probabilities are exported to a spreadsheet which a Perl script uses to simulate (see Appendix B). Since each of the four outcomes have a given probability, their sum for a given observation equal one. Thus a prediction may be made by comparing a random value between zero and one to the ranges of an outcome within the summation of the four. The script makes a prediction in such a manner for each observation in order to obtain the percentage of observations whose outcomes were correctly predicted. The script iterates over all observations 200000 times, each time recording the percentage of correctly predicted outcomes.

When simulation is complete, the percentage accurately predicted are sorted and identified by their percentiles. Plotting the percentiles on the percent accuracy creates an s-curve which visually presents the confidence in the accuracy of the model.

3.3 Adjuvant Tests and Statistics

While the models are the focus of this study, various statistical tests are performed to test the significance of the variables making up the models, to test the strength of models in comparison with other models, and to test for significant changes in proportions accurately predicted. The tests utilized are the t-test, two-sample t-test and the proportion z-test.

3.3.1 t-Test

The t-test is used to test the significance of each variable tested within the model. This significance is in terms of the variables' coefficients' differences from zero. For a non-directional t-test the two hypotheses are:

H_0: $b = 0$: that the coefficient is not significantly different from zero; and

H_A, $b \neq 0$: that the coefficient is significantly different from zero where *b* is the coefficient. Given a certain confidence level, here 62% ($|t^*| \approx 0.87$ for 2500 degrees of freedom), then the null hypothesis may be rejected if the calculated t-statistic is further away from zero than t^*.

Before discussing the mechanics of the test, the selection of the confidence level should be discussed. Experience in modeling the data revealed a natural break in significance that corresponded with the 62% confidence level. 62% is rather low in statistics and is particularly undesirable when modeling because weak variables weaken the model. However, this may be justified in terms of the effects of the tradeoff in Type I and Type II error. Type I error results in false rejection of the null hypothesis; that is the model is estimated with weak variables. Type II error results in failure to reject a false null hypothesis, that is the model does not include an important variable. For the purposes of identifying factors contributing to the fault of parties in cyclist-motorist collisions, it is clearly detrimental to leave out an important factor by committing a Type II error; whereas accidentally providing decision makers with insignificant information may only be inefficient. Thus, a low level of confidence is used.

To test the significance of a variable the t-statistic is computed:

$$t = \frac{b - 0}{Std.\ Error} \tag{3-13}$$

where *b* is the coefficient. The t-distribution has an associated degrees of freedom equal to the number of observations in the estimation sample less one. For large numbers of observations, there is little variation in the shape of the distribution so a single t-statistic, t^*, based on 2500 observations was used as the basis of these tests.
The test proceeds by obtaining the t-statistic for each coefficient with equation 3-13. The absolute value thereof is compared with the absolute value of the t-statistic

corresponding with the 62% confidence level, $|t^*| \approx 0.87$. If $t > t^*$, then the null hypothesis is rejected and the coefficient stays in the model; else, the null hypothesis is not rejected and the coefficient is restricted to zero unless it is of special interest to the model (e.g. a study interested in the affect of weather on fault may choose not remove the weather variable despite its statistical insignificance). For the purposes of noting the strength of each variable in the model, the results for the final model will include the t-statistic for each coefficient.

3.3.2 Two-Sample t-Test

The two-sample t-test is used in this study to compare the results of the models in simulation. While the hypotheses of the test may reflect a preference for a particular model, none has been made for the widely varying perspectives presented in this study. Thus, the hypotheses to be tested are:

H_0: $\nu_1 = \nu_2$; there is not a significant difference; and

H_A: $\nu_2 \neq \nu_2$: there is a significant difference between the models

where ν_1 and ν_2 are the arithmetic average accuracy of the first and second model respectively as found in simulation. Runyon, Coleman and Pittenger (2000) note that the t-test will always be significant for large sample sizes. Since the comparison is made between models of sample size of two hundred thousand iterations each, even a small variation in the models will be significant (and if it were not the computer could iterate infinitely more). To account for this, Runyon et al. (2000) recommend using the omega squared, ω^2, statistic to describe the significance.

The statistic ω^2 is interpreted similarly to the statistic r^2 as a correlation coefficient. Literally, it means the independent variable – here, the form of the model – accounts for ω^2 proportion of the variance in the dependent variable – here, the accuracy there reflected. The statistic is calculated by:

$$\omega^2 = \frac{t^2 - 1}{t^2 + n_1 + n_2 - 1} \tag{3-14}$$

where n_1 and n_2 are the sample sizes of the first and second sets respectively, and t is the t-statistic calculated for two independent samples:

$$t = \frac{v_1 - v_2}{\sqrt{\left[\frac{1}{n_1}\left(\frac{\sum\left(x_1 - v_1\right)^2}{n_1 - 1}\right)\right] + \left[\frac{1}{n_2}\left(\frac{\sum\left(x_2 - v_2\right)^2}{n_2 - 1}\right)\right]}} \tag{3-15}$$

where n_1 and n_2 are the sample sizes of the first and second sets respectively, x_1 and x_2 are values from each sample, and v_1 and v_2 are the mean arithmetic average accuracy of the first and second samples (Runyon et al., 2003).

For the purposes of this study, the threshold of the amount of variation for which to be accounted by a model for the difference between the models to not be considered trivial is arbitrarily set at 50%.

3.3.3 Proportion z-Test

The proportion z-test may be used to determine if the difference between the proportions of observations accurately predicted and of those expected to be so predicted. This test may be directional or non-directional based on whether or not a preference for one outcome is made in the hypothesis being tested. This study makes such a hypothesis in favor of a particular outcome, so the hypotheses to be tested are:

H_0: $p \leq P_0$; there is not a significant increase; and

H_A: $p > P_0$: there is a significant increase in the proportion correctly predicted where p is the proportion correctly predicted, and P_0 is the proportion expected to be correctly predicted. Unlike the t-test, this study does not make use of the z-test during the course of modeling so the risk associated with committing a Type II error is not significant because it does not preclude a factor from inclusion in the study. For this reason a high, 99% confidence level is used and the z-test is executed with the corresponding P of .01.

From the z-distribution, obtain the z-statistic corresponding to the P of .01, $z^* = 2.326$. Then calculate a z-statistic for the sampled proportion:

$$z = \frac{p - P_0}{\sqrt{\dfrac{P_0\left(1 - P_0\right)}{n}}} \qquad (3\text{-}16)$$

where p is the proportion correctly predicted, and P_0 is the proportion expected to be correctly predicted. The null hypothesis is rejected if the z calculated for the sample proportion exceeds z^*.

Chapter 4

Data Description

The dataset upon which this study is based was provided by the North Carolina Highway Safety Research Center on the campus of University of North Carolina – Chapel Hill. The dataset contains details of each cyclist-motorist accident in North Carolina for the period of 1997 to 2002.

4.1 Available Data

There are 5639 observations, i.e. cyclist-motorist accidents, with accompanying details. These details include the designation of fault, temporal factors, weather conditions, road surface conditions, lighting conditions, location features, road classification, posted speed limit, road geometry, traffic direction/division, traffic signs, paved or not, land use, municipal population, accident location with respect to intersections, whether alcohol was a factor, whether speeding was a factor, road defects, type of accident, position of the bicycle with respect to facility, direction of the bicycle with respect to traffic, geographical information, state of intoxication of both motorist and cyclist, demographic information of both motorist and cyclist, whether the cyclist wore a helmet, and the severity of the cyclist's injury.

Additional variables were derived from the originals. As discussed in Chapter 2, it has been recognized that there is a relationship with the square of age (Kim & Li, 1996). Other relationships were also tested including the square and square root of the speed limit and the logarithm of age. Other variables were formed from outcome combinations; such as partial fault from the relationship that if one was at fault or both were at fault, then that one was at least partially at fault. And other variables were formed by the combination of two explanatory conditions; such as weekend from day of week being either Saturday or Sunday. While it is not improper to also create variables via such cross-classifications as gender and helmet use, such interaction terms were not created based on the limited number of observations.

4.1.1 Variables Tested

Many combinations of explanatory variables were used to estimate the models described in Chapter 3. The idea is that the combination that most strongly defines the relationship would be used to estimate the final model[1]. While every combination could eventually be tested, both the plethora of variables makes that impractical and also the random significant combination of potentially irrelevant variables makes that improper. For the purposes here, variables were tested in combination and were pruned based on their significance to the model.

The variables tested need to be themselves viable indicators. For example, while geographic information such as the street name might have been a significant indicator of accident fault, it does not describe the actions that cause one party to be at fault. Not only should the variable be relevant, but it should also best explain the phenomena. For example, while road classification describes the quality of the road,

[1] Here, however, the variables that strongly described one binary probit model were strictly copied into the forms for both bivariate probit models and the other binary probit model. This enables comparison.

the speed limit also describes the quality of the road but does so with greater precision and finer increments.

There is a strong argument that when an important variable is not present in the dataset, irrelevant variables may become significant because they can be indicative of the missing information. Such an argument would encourage the use of variables such as street names in the hopes that it, if significant, could reveal unobserved factors such as the precinct of the officer writing the accident report. For the purposes of this study, that argument is headed only for variables with known correlations to relevant phenomena. For example, ethnicity has a correlation with wealth which is correlated with education and sense of self[2]. Thus ethnicity may be used as an indicator for the explanatory variable for the relevant variables of education and sense of self.

With that in mind, the following categories were not tested: temporal factors except weekend and weekday since weather better describes month, a Tuesday does not differ from a Wednesday, and lighting better describes the time of day; road classification, traffic direction/division, paved or not since speed limit describes the condition of the facility; municipal population since, while it may hint at exposure, it does not itself describe fault; accident location with respect to intersections since the signage also describe the type of intersection; whether alcohol was a factor since the cyclist and motorist are individually noted for drinking; road defects since this does not describe fault; geographical information since the jurisdiction should not affect the assignment of fault; and severity of the cyclist's injury since the determination of fault should be independent of the severity of a party's injury.

[2] While ethnicity does not cause wealth and wealth does not cause ethnicity, historical factors beyond the bounds of this study created a relationship between ethnicity and education that seems to propagate itself into the twenty-first century. Furthermore, it may be shown that race/ethnicity does not itself implicate a party by noting that the coefficients for these terms are not stable when estimating the model on an otherwise homogenous subset of the population.

4.1.2 Variables Used

Of the many variables tested, the following were used in at least one of the models: ages of the motorist and the cyclist, their squares, and their logarithms; whether the driver was intoxicated; whether speed was a factor; speed limit and its square root; estimated vehicle speed; accident type involving a motorist turning or overtaking, and accident type involving the cyclist turning; weather conditions clear or rainy; whether accident occurred in a driveway; bicycle position in a bike lane, on a bike path, or in a side crossing; whether cyclist was traveling against traffic; lighting conditions; land use; ethnicity of the motorist as black, Hispanic, Native-American, or other; the vehicle type of sports utility vehicle or motorcycle; gender of the cyclist; whether the cyclist is black; whether the cyclist wore a helmet; presence of signage or signals; and whether or not it is a weekend.

That was just an overview; the variables are discussed in an easy-to-follow, organized tabular format in the sections detailing the specific data used in estimation and simulation of the models (Table 4-1 thru Table 4-3). With that list of variables in mind, consider the variables not included in estimation.

4.1.3 Variables Excluded

Many variables did not significantly affect the model[3] as determined by the t-test described in section 3.3.1. For example, only the vehicle types of sports utility vehicle and motorcycle were significantly different from car to prompt inclusion. While this was a common factor to prompt exclusion of a variable from the model, it

[3] Some of these insignificant variables were kept in the models because they were significant in a separate model or were representative of some other interest to the study.

does not explain the exclusion of such seemingly significant variables as whether or not the cyclist was intoxicated.

Indeed the state of intoxication of the cyclist is significant to the models. The problem is that the cyclist does not regularly have this determined. In the interest of efficient use of the already small number of observations, the variable is excluded.

4.1.4 Data Limitations

Before proceeding, the limitations of the dataset upon which this study is based should be acknowledged. These limitations qualify the applicability of the models created.

First, the term "fault" is not defined in its legal sense. The dataset created the variable "fault" based on whether a party committed a violation as documented by the police officer at the scene from the evidence and witness testimony. In fact, there is no field for the officer to designate a party to be "at fault" – such determination is reserved for the insurance companies and courts (see the sample police report form in Appendix A). So, for the purposes of this study, the term "fault" is hereby defined as "propensity to commit a violation"; the term "at fault" is hereby defined as "in violation". While it may be argued that it would be more appropriate to describe this study as one of the relationship of factors contributing to the propensity to commit a violation of parties involved in cyclist-motorist collisions, such debate over semantics is unnecessary. A dataset that acknowledged fault in the legal sense would provide more powerful conclusions; however this weaker set precludes neither the use of methods presented in this study nor the use of the inferences there based. The conclusions are simply limited to the acknowledgement that a party was at fault for contributing to a collision by committing a violation; the decision oriented models there based describe the determination of such violations.

The second limitation is that the data refers only to that of North Carolina and may not be applicable outside of this area. A third limitation is that the dataset does not (and could not) include all potentially important variables: there is no field regarding such police officer's notes as the contribution of witnesses in determining the violations of the party; there is no way to know the variation in such things as the ability for a police officer to discern such difficult factors as the "intent" on the part of the motorist to hit the cyclist; and police reports do not have fields for such interesting factors as whether the cyclist was riding alone or in a group. Additional limitations may apply with regard to reporting, uniformity in reporting standards over the time, temporal variations, etc. Always be skeptical when interpreting statistics.

4.2 Ordered Probit

Table 4-1 summarizes the variables used in estimation and in simulation of the ordered probit model.

TABLE 4-1. Summary of Datasets Used for the Ordered Probit Model

Variable	Estimation Dataset			Simulation Dataset		
	Min.	**Max.**	**Mean**	**Min.**	**Max.**	**Mean**
Assignment						
Fault (-1 Motorist, 0 Both, 1 Cyclist)	-1	1	0.413	-1	1	0.411
Actions						
Speed was a Factor	0	1	0.013	0	1	0.013
Est. Vehicle Speed (mph)	0	100	23.906	0	100	24.652
Motorist Overtook Cyclist	0	1	0.089	0	1	0.094
Motorist Turned	0	1	0.086	0	1	0.088
Motorist Intoxicated	0	1	0.022	0	1	0.022
Cyclist Turned	0	1	0.193	0	1	0.196
Helmet Used	0	1	0.051	0	1	0.052
Conditions						
Rainy Weather	0	1	0.038	0	1	0.039
Dawn or Dusk	0	1	0.053	0	1	0.052
Dark, Street Lighting	0	1	0.094	0	1	0.095
Dark, no Street Lighting	0	1	0.084	0	1	0.087
Weekend	0	1	0.259	0	1	0.259
Position						
Accident on Driveway	0	1	0.079	0	1	0.077
Cyclist in Bike Lane	0	1	0.037	0	1	0.038
Cyclist on Bike Path	0	1	0.008	0	1	0.007
Cyclist on Side Crossing	0	1	0.113	0	1	0.105
Cyclist Facing Traffic	0	1	0.262	0	1	0.252
Signage						
Traffic Sign (Stop, Yield, etc.)	0	1	0.271	0	1	0.275
Traffic Signal	0	1	0.111	0	1	0.119
Other Traffic Control (RR Crossing, etc.)	0	1	0.003	0	1	0.004
Land Use						
Commercial Land Use	0	1	0.330	0	1	0.326
Demographics						
Cyclist Age	1	90	24.956	1	90	25.141
Logarithm Cyclist Age	0	4.5	3.010	0	4.5	3.020
Cyclist Black	0	1	0.456	0	1	0.457
Motorist Age	1	91	39.153	15	96	39.255
Logarithm Motorist Age	0	4.51	3.575	2.71	4.56	3.578
Number of Observations	3743			3700		

*unless otherwise specified, variables coded: 1 = yes; 0 = no

The estimation dataset represents the dataset for which each accident had a recorded value for each variable. The simulation dataset is the same except it includes the possibility that neither is at fault and is further restricted to attempt to emulate the dataset of the models to follow. In short, the ordered probit model is simulated on essentially the same dataset as the probit and bivariate probit models.

Running a simulation of the ordered probit on the simulation dataset is unfair because the model can never predict the outcome of neither at fault. This simulation is done merely for sake of comparison to the models that follow. To gauge the effectiveness of the ordered probit, a simulation on the estimation dataset will be made.

4.3 Binary Probit

Table 4-2 and Table 4-3 summarize the variables used in estimation and in simulation of the binary probit models estimated by partial-fault and sole-fault respectively.

TABLE 4-2. Summary of Datasets Used for the Partial-Fault Probit Models

Variable	Estimation Dataset			Simulation Dataset		
	Min.	**Max.**	**Mean**	**Min.**	**Max.**	**Mean**
Assignment						
Motorist at least Partially at Fault	0	1	0.368	0	1	0.35
Cyclist at least Partially at Fault	0	1	0.797	0	1	0.76
Actions						
Speed was a Factor	0	1	0.013	0	1	0.013
Motorist Overtook Cyclist	0	1	0.091	0	1	0.093
Motorist Turned	0	1	0.089	0	1	0.088
Motorist Intoxicated	0	1	0.022	0	1	0.022
Cyclist Turned	0	1	0.202	0	1	0.196
Helmet Used	0	1	0.052	0	1	0.053

(*continued*)

TABLE 4-2. Summary of Datasets Used for the Partial-Fault Probits (cont'd)

Variable	Estimation Dataset			Simulation Dataset		
	Min.	Max.	Mean	Min.	Max.	Mean
Conditions						
Clear Weather	0	1	0.819	0	1	0.818
Dawn or Dusk	0	1	0.052	0	1	0.052
Dark, Street Lighting	0	1	0.093	0	1	0.095
Dark, no Street Lighting	0	1	0.086	0	1	0.087
Weekend	0	1	0.259	0	1	0.259
Position						
Accident on Driveway	0	1	0.079	0	1	0.077
Cyclist in Bike Lane	0	1	0.039	0	1	0.04
Cyclist on Bike Path	0	1	0.007	0	1	0.01
Cyclist Facing Traffic	0	1	0.256	0	1	0.25
Posted Speed Limit (mph)	3	60	38.365	3	60	38.236
Square Root Speed Limit	1.73	7.75	6.127	1.73	7.75	6.12
Signage						
Traffic Sign (Stop, Yield, etc.)	0	1	0.279	0	1	0.275
Traffic Signal	0	1	0.114	0	1	0.119
Other Traffic Control (RR Crossing, etc.)	0	1	0.003	0	1	0.004
Land Use						
Commercial Land Use	0	1	0.322	0	1	0.326
Institutional Land Use	0	1	0.017	0	1	0.017
Agricultural Land Use	0	1	0.156	0	1	0.154
Industrial Land Use	0	1	0.004	0	1	0
Cyclist Demographics						
Cyclist Age	0	90	25.03	0	90	25.1
Cyclist Age Squared	0	8100	873.3	0	8100	876.2
Cyclist Gender (0 Female, 1 Male)	0	1	0.851	0	1	0.85
Cyclist Black	0	1	0.454	0	1	0.46
Motorist Demographics						
Motorist Age	15	91	39.3	15	96	39.3
Motorist Age Squared	225	8281	1822.4	225	9216	1823.8
Vehicle Type: Motorcycle	0	1	0.005	0	1	0.005
Sports Utility Vehicle	0	1	0.036	0	1	0.035
Motorist Black	0	1	0.321	0	1	0.322
Motorist Native-American	0	1	0.008	0	1	0.008
Motorist Hispanic	0	1	0.011	0	1	0.011
Motorist of Other Non-White/Asian Ethnicity	0	1	0.021	0	1	0.021
Number of Observations	3543			3702		

*unless otherwise specified, variables coded: 1 = yes; 0 = no

TABLE 4-3. Summary of Datasets Used for the Sole-Fault Probit Models

Variable	Estimation Dataset Min.	Estimation Dataset Max.	Estimation Dataset Mean	Simulation Dataset Min.	Simulation Dataset Max.	Simulation Dataset Mean
Assignment						
Motorist Solely at Fault	0	1	0.243	0	1	0.19
Cyclist Solely at Fault	0	1	0.757	0	1	0.61
Actions						
Speed was a Factor	0	1	0.012	0	1	0.013
Motorist Overtook Cyclist	0	1	0.094	0	1	0.093
Motorist Turned	0	1	0.081	0	1	0.088
Motorist Intoxicated	0	1	0.019	0	1	0.022
Cyclist Turned	0	1	0.230	0	1	0.196
Helmet Used	0	1	0.055	0	1	0.053
Conditions						
Clear Weather	0	1	0.825	0	1	0.818
Dawn or Dusk	0	1	0.053	0	1	0.052
Dark, Street Lighting	0	1	0.087	0	1	0.095
Dark, no Street Lighting	0	1	0.086	0	1	0.087
Weekend	0	1	0.266	0	1	0.259
Position						
Accident on Driveway	0	1	0.090	0	1	0.077
Cyclist in Bike Lane	0	1	0.036	0	1	0.04
Cyclist on Bike Path	0	1	0.007	0	1	0.01
Cyclist Facing Traffic	0	1	0.177	0	1	0.25
Posted Speed Limit (mph)	3	60	38.931	3	60	38.236
Square Root Speed Limit	1.73	7.75	6.174	1.73	7.75	6.12
Signage						
Traffic Sign (Stop, Yield, etc.)	0	1	0.268	0	1	0.275
Traffic Signal	0	1	0.106	0	1	0.119
Other Traffic Control (RR Crossing, etc.)	0	1	0.003	0	1	0.004
Land Use						
Commercial Land Use	0	1	0.280	0	1	0.326
Institutional Land Use	0	1	0.016	0	1	0.017
Agricultural Land Use	0	1	0.173	0	1	0.154
Industrial Land Use	0	1	0.004	0	1	0
Cyclist Demographics						
Cyclist Age	0	90	24.20	0	90	25.1
Cyclist Age Squared	0	8100	831.14	0	8100	876.2
Cyclist Gender (0 Female, 1 Male)	0	1	0.851	0	1	0.85
Cyclist Black	0	1	0.446	0	1	0.46

(*continued*)

TABLE 4-3. Summary of Datasets Used for the Sole-Fault Probits (cont'd)

Variable	Estimation Dataset Min.	Max.	Mean	Simulation Dataset Min.	Max.	Mean
Motorist Demographics						
Motorist Age	15	91	39.4	15	96	39.3
Motorist Age Squared	225	8281	1838.7	225	9216	1823.8
Vehicle Type: Motorcycle	0	1	0.006	0	1	0.005
Sports Utility Vehicle	0	1	0.034	0	1	0.035
Motorist Black	0	1	0.323	0	1	0.322
Motorist Native-American	0	1	0.008	0	1	0.008
Motorist Hispanic	0	1	0.012	0	1	0.011
Motorist of Other Non-White/Asian Ethnicity	0	1	0.021	0	1	0.021
Number of Observations	2959			3702		

*unless otherwise specified, variables coded: 1 = yes; 0 = no

For both the partial-fault and sole-fault instances, the estimation dataset represents the dataset for which each accident had a recorded value for each variable. The simulation dataset also requires each accident to have a recorded value; the difference is that it includes all designations of fault so that they may be compared with the predicted value.

4.4 Bivariate Probit

Because the bivariate probit is the simultaneous estimation of the two binary probit models of before, the same explanatory variables were used. Thus, Table 4-2 and Table 4-3 also summarize the variables used in estimation and in simulation of the bivariate probit models estimated by partial-fault and sole-fault respectively.

Chapter 5

Results

Estimates and the simulations there based were made per the methodologies of Chapter 3 with the appropriate variables as discussed in Chapter 4. These are the results of the ordered probit model, binary probit models, and bivariate probit models with corresponding simulations for verification and comparison purposes.

5.1 Ordered Probit

The ordered probit model's estimates are shown in Table 5-1. Figure 5-1 shows the results of simulation for the dataset on which the model was estimated.

TABLE 5-1. Ordered Probit Model

Variable	Coeff.	St. Err.	t-score
Constant			
Cutoff 1	-0.9600	0.62	—
Cutoff 2	-0.1531	0.62	—
Actions			
Speed was a Factor	-1.2394	0.19	-6.54
Estimated Vehicle Speed (mph)	0.0276	0.00	15.43
Motorist Overtook Cyclist	-1.9240	0.10	-19.34
Motorist Turned	-1.1644	0.08	-14.65
Motorist Intoxicated	-1.1108	0.15	-7.55
Cyclist Turned	0.9078	0.09	9.74
Helmet Used	-0.4680	0.11	-4.46
Conditions			
Rainy Weather	0.1801	0.12	1.48
Dawn or Dusk	-0.0729	0.10	-0.71
Dark, Street Lighting	0.2269	0.08	2.88
Dark, no Street Lighting	0.5494	0.10	5.59
Weekend	-0.0570	0.05	-1.04
Position			
Accident on Driveway	1.0687	0.14	0.06
Cyclist Positioned in Bike Lane	-0.2696	0.13	-2.14
Cyclist Positioned on Bike Path	-0.1263	0.25	-0.50
Cyclist Positioned on Side Crossing	-0.1460	0.07	-2.07
Cyclist Facing Traffic	0.2274	0.06	3.90
Signage			
Traffic Sign (Stop, Yield, etc.)	0.1066	0.06	1.85
Traffic Signal	0.4566	0.08	6.01
Other Traffic Control (RR Crossing, etc.)	0.5869	0.44	1.35
Land Use			
Commercial Land Use	-0.0053	0.05	-0.10
Demographics			
Cyclist Age	0.0156	0.00	3.20
Logarithm Cyclist Age	-0.8801	0.12	-7.04
Cyclist Black	0.2805	0.05	5.84
Motorist Age	-0.0131	0.01	-2.24
Logarithm Motorist Age	0.6200	0.22	2.77
Number of Observations	3743		
Log-Likelihood at Constants	-3454		
Log-Likelihood at Convergence	-2386		
ρ^2	0.3094		

*unless otherwise specified, variables coded: 1 = yes; 0 = no

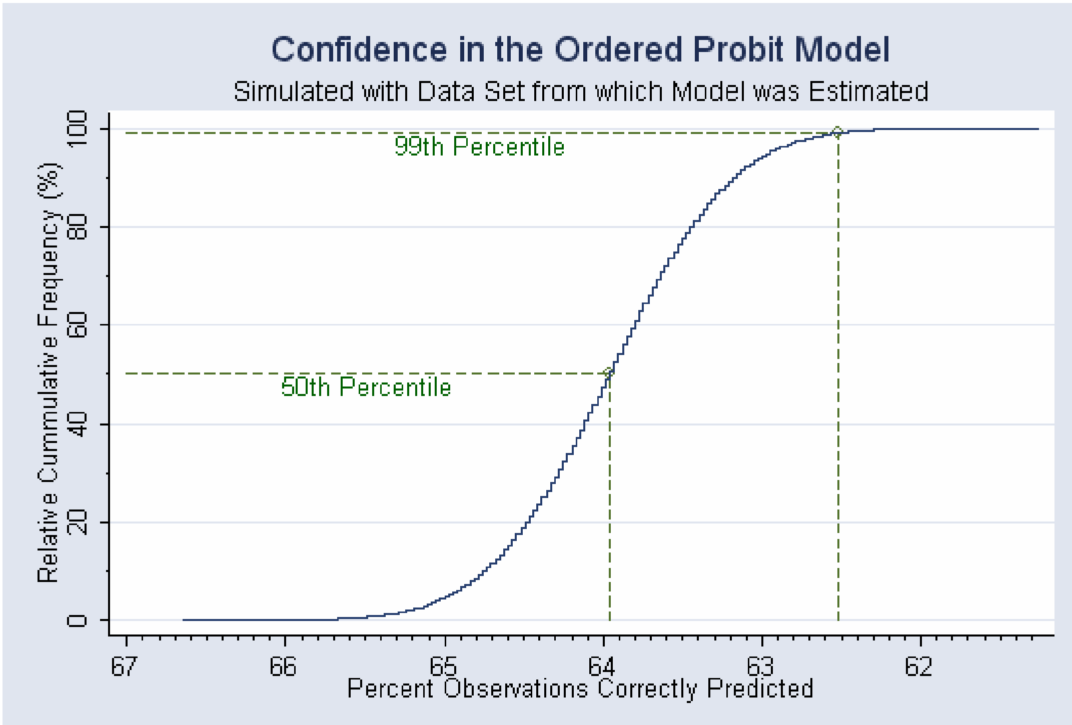


FIGURE 5-1. Ordered Probit on Estimated Dataset

5.1.1 Accuracy

The ordered probit has a pseudo r-squared, ρ^2, of 0.3094, which means it predicts better than a naïve constant-only model. In simulation the model was, on average, about 64.0% accurate; additionally, one could expect, with 99% confidence, the model to be at least 62.5% accurate. For a dataset where both are at fault 16.9%, the motorist is at fault 20.9% and the cyclist is at fault 62.2% of the time, another naïve constant-only model could simply blame the cyclist and be correct 62.2% of the time. While the ordered probit model's high accuracy may not seem quite so impressive when compared to the "blame the bike" model, bear in mind that the model also may explain a relationship between factors.

The stability of the model should also be noted. How the explanatory variables affect the outcome remained similar during model construction. Also, the core model is strong; for example there was less than a percentage point drop in accuracy given the omission of such seemingly important variables as the presence of traffic signs, traffic signals and other controls.

5.1.2 Assumptions

Before interpreting the model, a quick recap of the assumptions on which it is based. The ordered probit model assumes a civilian's perspective by assigning fault based on the severity of the offence. For example, if the motorist was both intoxicated and also overtaking the cyclist, those two large negative coefficients are a difficult combination for the cyclist's offences to outweigh. Thus the model seems to perform as expected.

The problem with this assumption is that it neglects the outcome of neither at fault or fault undetermined. This means if the model were to be extrapolated to the dataset on which fault was acknowledged, it would always predict that outcome incorrectly. Since the neither at fault outcome only affects about four percent of the observations, it still correctly predicts, with 99% confidence, at least 60.9% of the outcomes as shown in Figure 5-2.

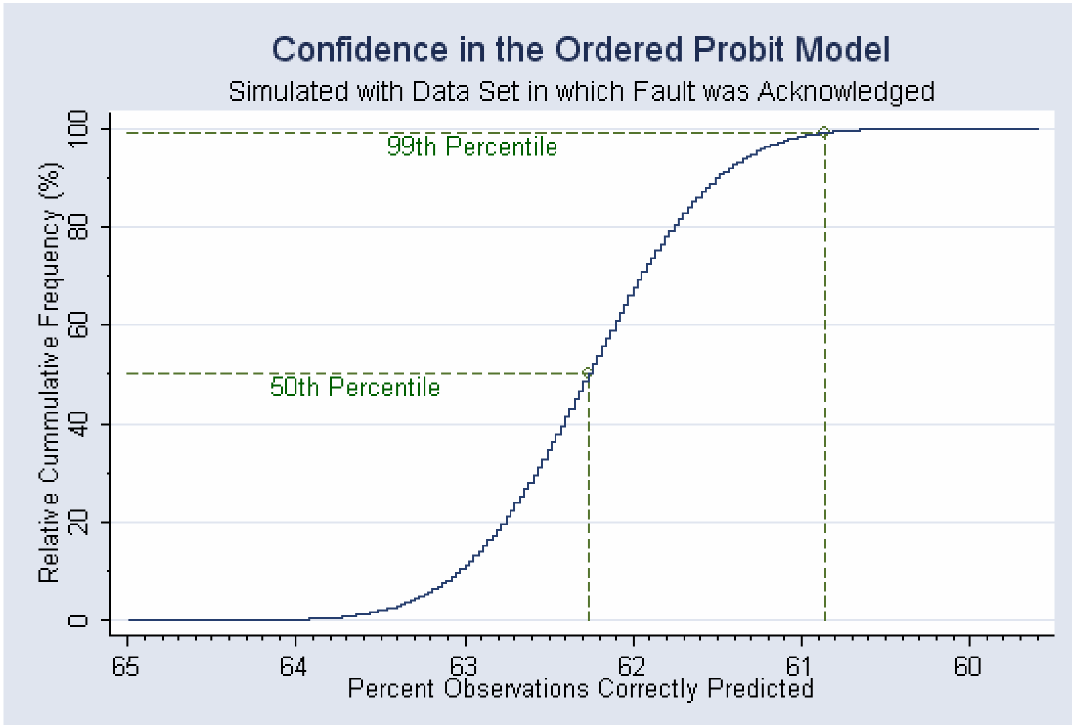


FIGURE 5-2. Ordered Probit on Data where Fault Acknowledged (incorrect)

Since extrapolating a model to a dataset where it cannot correctly predict an outcome is incorrect, interpretations based solely upon it are also incorrect. However, this does have value in making comparisons which will be discussed further along with an application of the model.

5.1.3 Interpretation

There are a few stages to interpreting the model. The first is to see whether it describes the relationship it is intended to well; something the mean accuracy of 64.0% suggests it does. The second is to list variables by their coefficient's sign as in Table 5-2; here, negative weights on the fault of the motorist and positive weights on the fault of the cyclist. The third is to use personal experience to hypothesize and interpret the meaning of the coefficients in support of the purported effect of each variable.

TABLE 5-2. Ordered Probit Model's Variables by Implication

Motorist	***Cyclist***
Helmet Used	Estimated Vehicle Speed
Logarithm Cyclist Age	Cyclist Age
Motorist Age	Logarithm Motorist Age
Motorist Intoxicated	Dark, Street Lighting
Dawn or Dusk	Dark, no Street Lighting
Speed was a Factor	Cyclist Turned
Motorist Overtook Cyclist	Rainy Weather
Motorist Turned	Cyclist Facing Traffic
Commercial Land Use	Cyclist Black
Cyclist in Bike Lane	Accident on Driveway
Cyclist on Bike Path	Traffic Sign
Cyclist on Side Crossing	Traffic Signal
Weekend	Other Traffic Control

Begin with the factors that implicated the motorist. A helmet does not affect the motorist, but it does suggest a more responsible cyclist thus implicating the motorist. The logarithm of the cyclist's age, which suggests the motorist at fault, must be interpreted with the cyclist's age, which suggests the cyclist is at fault, the net effect still places blame with the motorist for increases in cyclist's age but with a plateau at age 40. Clearly a confirmed state of motorist intoxication implicates a driver. Dawn or Dusk suggests that while the cyclist should still be visible to the motorist thus implicating the motorist. When speed is a factor, that is the motorist was speeding, the driver is, as expected, implicated. Such potentially dangerous actions of a motorist as overtaking the cyclist and turning accordingly implicated the motorist.

Commercial land use could suggest either density of curb cuts or a motorist's mental state as a characteristic of trip purpose, and may be justified in implicating the driver. When a cyclist is not located in a traffic lane, it follows that blame may lie with the motorist for potentially entering the cyclist's space. While not clear how a weekend may implicate the motorist, it could be justified by an argument relating to the mentality of the recreational cyclist.

Continue with the factors that implicate the cyclist. While not immediately clear why greater estimated vehicle speeds implicate the cyclist instead of the motorist, it may suggest that the motorist could not have reacted to a cyclist's error. The logarithm of the motorist's age, which suggests the cyclist at fault, must be interpreted with the motorist's age, which suggests the motorist is at fault, the net effect still places blame with the cyclist for increases in motorist's age but with a plateau at age 40. Darkness implicates the cyclist, suggesting that the cyclist does not adequately equip the bike with adequate lighting. Clearly the potentially dangerous action of a cyclist turning may implicate the cyclist. Rainy weather implicating the cyclist suggests that the bike lacks adequate lighting, that a wet cyclist is a reckless cyclist (or at least a more accident-prone cyclist), or a combination of both. Facing traffic, while the law for pedestrians on the roadway, is generally illegal for cyclists thus implicating them. While ethnicity/race in itself does not implicate the cyclist, being black may, via correlations external to this study[4], reflect upon some other characteristics that do implicate the cyclist. An accident on a driveway implicates the cyclist because they are not expected, particularly given the motorist's generally reduced field of vision on such facility. Traffic controls implicate the cyclist probably because of bike's limited stopping and strenuous restarting and, hence, cyclist's general disregard for traffic controls.

[4] For example, ethnicity/race may, for historical reasons external to this study, be correlated with wealth.

This interpretation of the variables is merely one interpretation of many possible interpretations. It should also be noted, that for future interpretation, correlation is not causation. Cyclists will not be absolved of fault by simply removing the traffic controls which implicate them. It should also be noted that absence of evidence is not evidence of absence. There may very well be other more significant factors than those highlighted in the model. As described in Chapter 1, this interpretation should be considered as a start to the process of finding which factors implicate which party, how, and why.

5.1.4 Application

The ordered probit model has many applications. The variables may be interpreted in terms of their respective coefficients in order to determine which factors implicate the driver or cyclist, how they do so, and why the do so. From this, policy decisions may be made to, say, supply lights and reflective vests to increase visibility of the cyclist at night, if this were determined to be the reason that darkness implicates the cyclist.

Another application of the model is to gauge how fault is determined, or for a quick case study here, how it is *not* determined. Based on the assumption that fault is determined by the civilian's perspective, perhaps the determination of neither at fault is based on the same factors. Given the ordered fault line, if neither were at fault, then neither the cyclist nor the motorist would have factors weighing the determination; that is, the neither at fault would overlap the both at fault determination. So, simulating with the assumption that a prediction of both at fault equates to both neither at fault and both at fault, arrive at the s-curve in Figure 5-3.

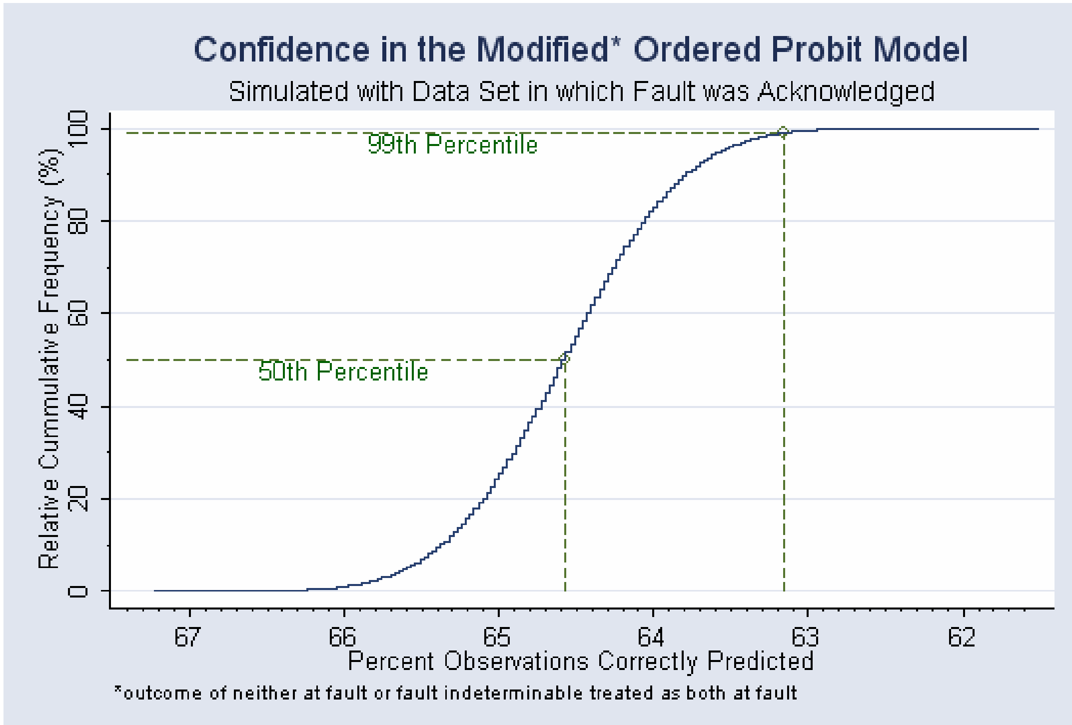


FIGURE 5-3. An Application of the Ordered Probit in a Case Study

The simulation of this modified ordered probit model on the dataset in which fault was acknowledged shows the model is, with 99% confidence, at least about 63.17% accurate. Compared with the original ordered probit model simulated on the same dataset in Figure 5-2 and its approximately 60.87% accuracy at the 99% confidence level, the modified model seems better. However, the original lacked the ability to predict the case of neither at fault when it occurred, so the possibility that the improvement is random chance must be examined.

For the 3700 observations 159 were neither at fault. The difference in accuracy of the models at the 99% confidence level is approximately 2.3%. Thus it predicted an additional 85 observations correctly. As a percentage, it predicts neither at fault as both at fault approximately 53.5% of the time. Using a proportion z-test to compare that to the expected 16.9% correctly predicted by chance, it is shown that this improvement is beyond random variation for the 99% confidence level (z = 12.3).

This means that, for the model to correctly predict the outcome, there is a balance between the characteristics that implicate either party given the neither at fault outcome. The ability to predict the outcome may have been higher, except the category of neither at fault also includes fault undetermined. Further study may be useful in revealing potential patterns, or problems, in the lack of assignment of fault.

5.2 Binary Probit

The two sets of binary probit models were estimated, one on the partial-fault dataset and the other on the sole-fault dataset. Simulation took place on the dataset on which fault was acknowledged. Figure 5-4 and Figure 5-5 show the results of each simulation for the partial-fault estimated model and the sole-fault estimated model, respectively.

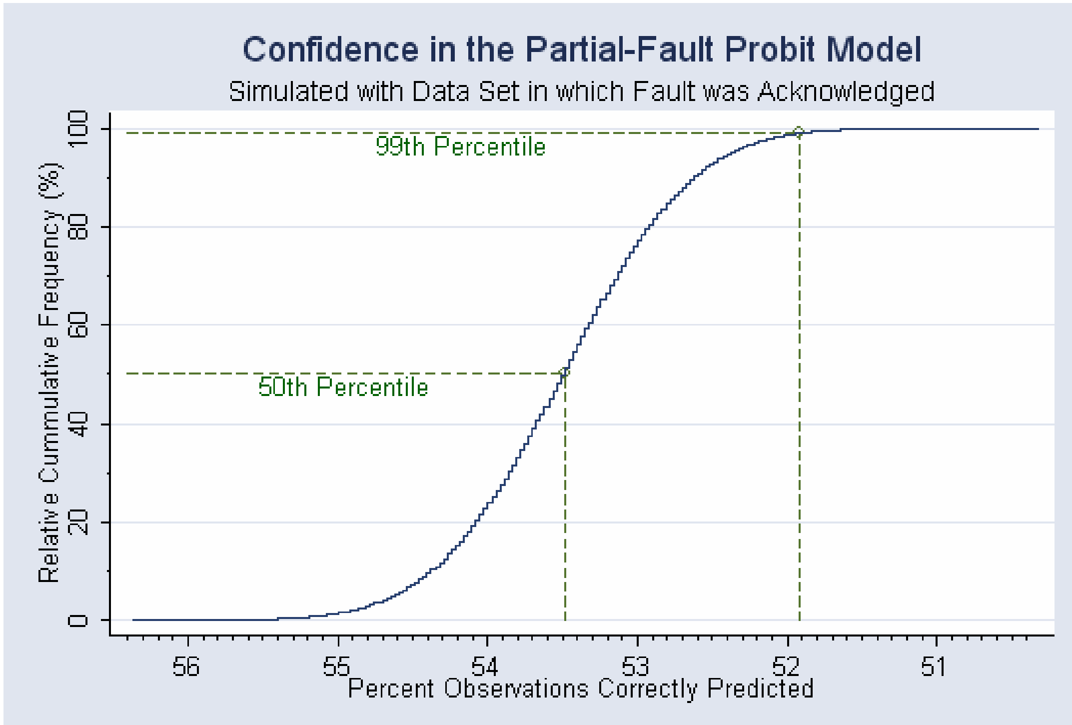


FIGURE 5-4. Binary Probit Partial-Fault on Data where Fault Acknowledged

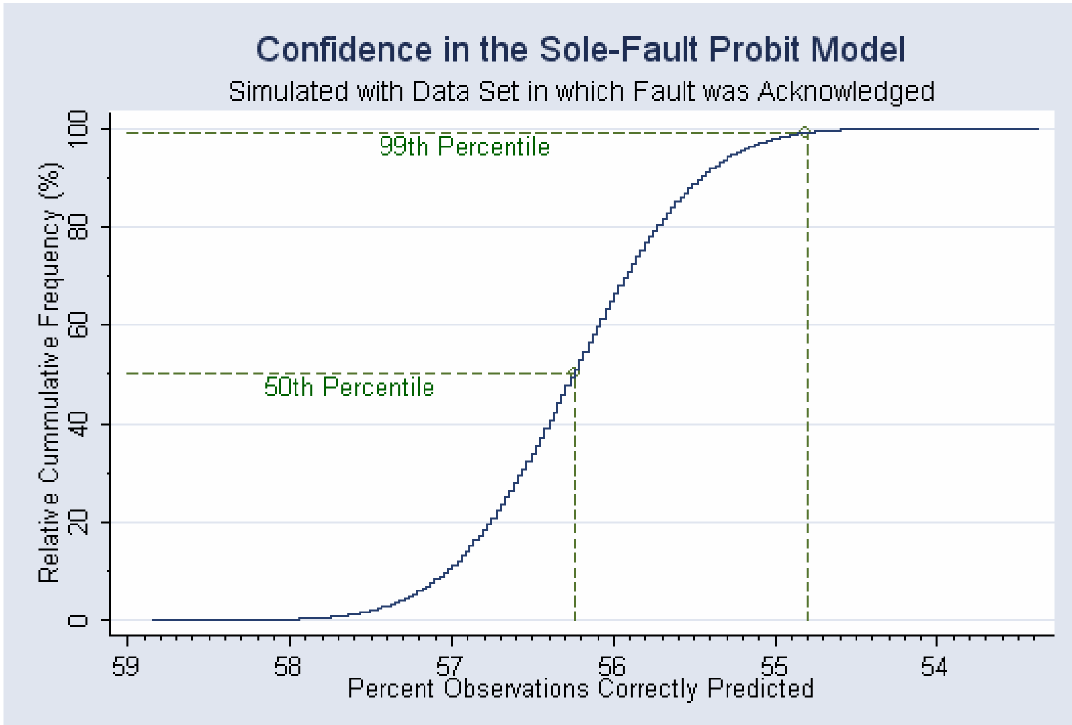


FIGURE 5-5. Binary Probit Sole-Fault on Data where Fault Acknowledged

5.2.1 Accuracy

The partial-fault estimated binary probit models, when combined in simulation, were, on average, 53.5% accurate; additionally one could expect, with 99% confidence, the model to be at least 51.9% accurate. The sole-fault estimated binary probit models, when combined in simulation, were, on average, 56.2% accurate; additionally one could expect, with 99% confidence, the model to be at least 54.8% accurate. Clearly the sole-fault estimated model was better; however there is more to want.

The binary probit model to predict whether the cyclist was at fault has a pseudo r-squared, ρ^2, of 0.3983. The binary probit model to predict whether the motorist was at fault has a pseudo r-squared, ρ^2, of 0.3599. While both models predict better than the naïve constant-only model they are compared to, the joint probability of two independent determinations creates an inefficient model. That is if there is, say, a 90% chance of getting something right and, say, a 60% chance of getting something right then the probability of getting both correct together is their product or a 54% chance. This is a problem of elementary probability and a problem of the assumption of independence of assignment when neither assignment is perfectly identified in terms of the variables. Part of this problem may be captured by acknowledging the variation in fault assignment by a single entity who assigns fault to each party for each observation. The simultaneous estimation used in the bivariate probit that follows this discussion takes that into account.

To the binary probit model's credit, it is stable across estimation sets. How the explanatory variables affect the outcome remained similar during model construction. Also, the core model is strong; for example there was less than a percentage point drop in accuracy given the omission of such seemingly important variables as the presence of traffic signs, traffic signals and other controls.

5.2.2 Assumptions

Before interpreting the model, a quick recap of the assumptions on which it is based. The binary probit models are based on the assumption of an officer's perspective by independently assigning fault to each party, the combination of which encompasses all fault assignment outcomes. For example, if one model determines the motorist to be at fault and the other model determines the cyclist to be at fault, then the combination of the two is the motorist is at fault and the cyclist is at fault – both are at fault. The accuracy of the model seems to suggest that it performs as expected.

5.2.3 Interpretation

The sole-fault estimated binary probit models performed better than the partial-fault estimated binary probit models. This may be interpreted to mean, as hypothesized in section 3.2.1, that the determination of sole fault is less convoluted, that the relationship is clearer, than that for partial fault.

Further interpretation of this model is unnecessary as these variables maintain their relationship in the stronger, bivariate probit model that follows this discussion of the binary probit models.

5.2.4 Application

While these binary probit models are not the strongest models that will be presented for these explanatory variables, it does have its own set of applications. For starters, it may be used to see how well the variables independently affect the assignment of fault to a single party. The binary probit may also be used as a comparison by which to see how great of role the variation of the entity assigning fault has on the precision of that assignment.

5.3 Bivariate Probit

The sole-fault estimated bivariate probit model's estimates are shown in Table 5-3. Figure 5-6 shows the result of simulation for the dataset on which fault was acknowledged. For comparison, Figure 5-7 shows the result of simulation on the same dataset for the partial-fault estimated bivariate probit model.

TABLE 5-3. Bivariate Probit Model

Variable	Motorist Solely at Fault			Cyclist Solely at Fault		
	Coeff.	**St. Err.**	**t-score**	**Coeff.**	**St. Err.**	**t-score**
Constant						
Constant	-0.6371	0.23	-2.8	0.8209	0.24	3.36
Actions						
Speed was a Factor	0.3946	0.12	3.22	—	—	—
Motorist Overtook Cyclist	1.3226	0.11	12.37	—	—	—
Motorist Turned	1.1364	0.10	11.51	—	—	—
Motorist Intoxicated	0.5185	0.13	3.99	—	—	—
Cyclist Turned	—	—	—	1.3474	0.12	11.27
Helmet Used	—	—	—	-0.3644	0.11	-3.44
Conditions						
Clear Weather	0.0455	0.07	0.63	—	—	—
Dawn or Dusk	0.1902	0.12	1.56	-0.0585	0.13	-0.46
Dark, Street Lighting	-0.0601	0.11	-0.54	0.3090	0.11	2.73
Dark, no Street Lighting	-0.3765	0.12	-3.25	0.0912	0.11	0.82
Weekend	0.0430	0.07	0.62	-0.0027	0.07	-0.04
Position						
Accident on Driveway	-1.5864	0.30	-5.22	1.8862	0.23	8.27
Cyclist in Bike Lane	—	—	—	-0.3062	0.11	-2.79
Cyclist on Bike Path	—	—	—	0.1403	0.11	1.26
Cyclist Facing Traffic	—	—	—	0.6603	0.07	9.24
Posted Speed Limit (mph)	-0.0084	0.00	-2.45	—	—	—
Square Root Speed Limit	—	—	—	0.0051	0.04	0.14
Signage						
Traffic Sign (Stop, Yield, etc.)	0.1244	0.07	1.68	0.5033	0.08	6.37
Traffic Signal	-0.2100	0.11	-1.99	0.6590	0.10	6.28
Other Traffic Control (RR Crossing, etc.)	-0.2951	0.80	-0.37	—	—	—
Land Use						
Commercial Land Use	0.2642	0.07	3.98	-0.1433	0.07	-2.07
Institutional Land Use	0.2260	0.30	0.75	—	—	—
Agricultural Land Use	0.0791	0.09	0.9	—	—	—
Industrial Land Use	—	—	—	0.2631	0.42	0.63
Cyclist Demographics						
Cyclist Age	—	—	—	-0.0478	0.01	-5.76
Cyclist Age Squared	—	—	—	0.0005	0.00	4.05
Cyclist Gender (0 Female, 1 Male)	—	—	—	0.0503	0.06	0.85
Cyclist Black	—	—	—	0.2773	0.06	4.54

(*continued*)

TABLE 5-3. Bivariate Probit Model (cont'd)

Variable	Motorist Solely at Fault			Cyclist Solely at Fault		
	Coeff.	St. Err.	t-score	Coeff.	St. Err.	t-score
Motorist Demographics						
Motorist Age	-0.0080	0.01	-1.09	—	—	—
Motorist Age Squared	0.0001	0.00	1.18	—	—	—
Vehicle Type: Motorcycle	-0.5304	0.65	-0.82	—	—	—
Sports Utility Vehicle	-0.0468	0.20	-0.24	—	—	—
Motorist Black	-0.1147	0.05	-2.12	—	—	—
Motorist Native-American	-0.2457	0.12	-1.98	—	—	—
Motorist Hispanic	0.2952	0.10	2.98	—	—	—
Motorist of Other Non-White/Asian Ethnicity	0.1142	0.25	0.46	—	—	—
Number of Observations		2959				
η		-1.000				
For Motorist Equation:						
Log-Likelihood at Constants		-1641				
Log-Likelihood at Convergence		-1050				
For Cyclist Equation:						
Log-Likelihood at Constants		-1640				
Log-Likelihood at Convergence		-987				
For Full Model:						
Log-Likelihood at Comparison		-2037				
Log-Likelihood at Convergence		-1408				

*unless otherwise specified, variables coded: 1 = yes; 0 = no

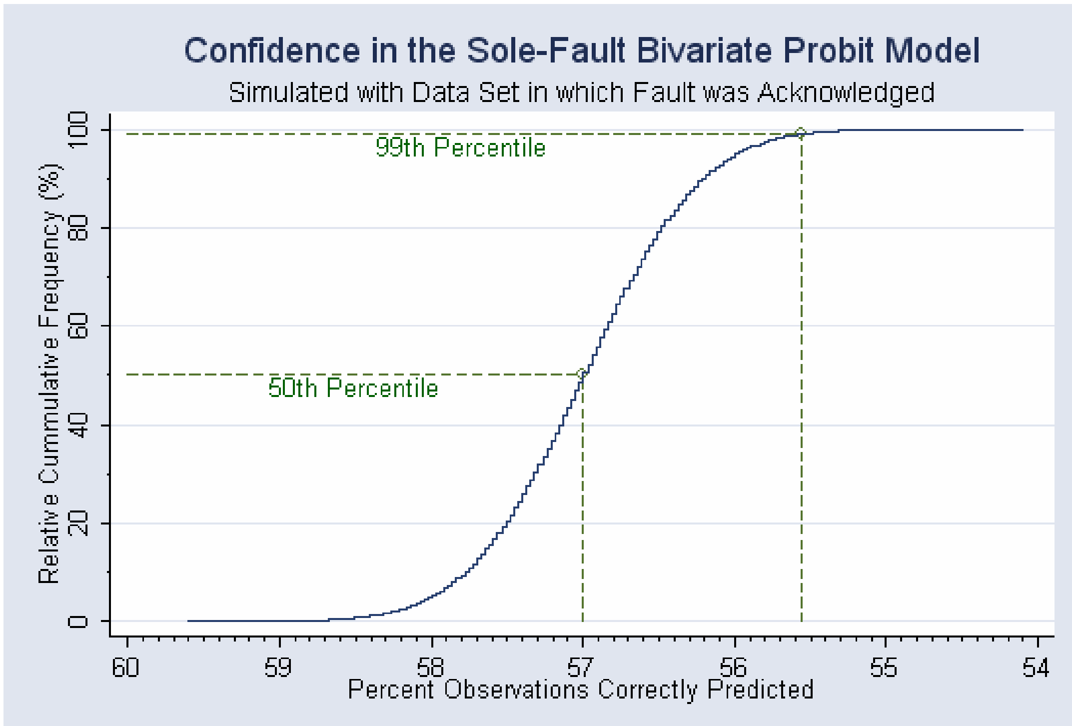


FIGURE 5-6. Bivariate Probit Sole-Fault on Data where Fault Acknowledged

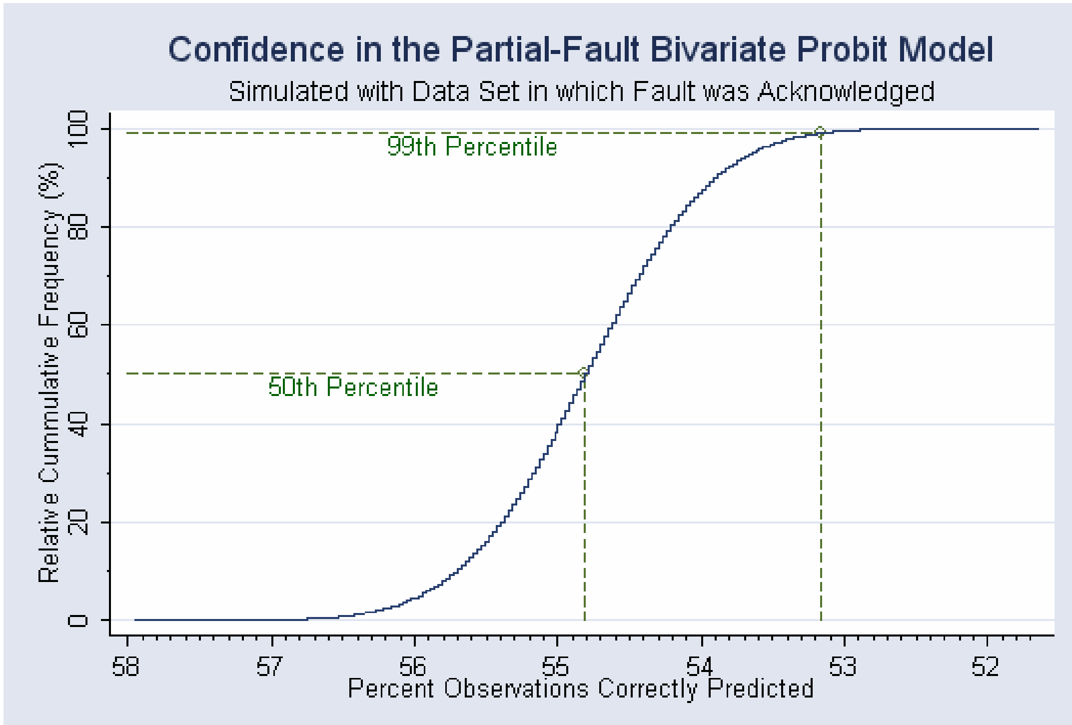


FIGURE 5-7. Bivariate Probit Partial-Fault on Data where Fault Acknowledged

5.3.1 Accuracy

The sole-fault estimated bivariate probit model was, on average, 57.0% accurate; additionally one could expect, with 99% confidence, the model to be at least 55.6% accurate. The partial-fault estimated bivariate probit models was, on average, 54.8% accurate; additionally one could expect, with 99% confidence, the model to be at least 53.2% accurate. Clearly the sole-fault estimated model was better.

Whether it was worthwhile estimating the two binary probit models with a single bivariate probit model is shown by whether η, the covariance in the errors of each binary probit model, is significantly different from zero. For both the partial-fault estimated and sole-fault estimated models, η was found to be significantly different from zero at the 99.99% confidence level. Thus, it is shown that the bivariate probit model is statistically better than the two binary probit models because the two binary probit decisions are not independent – that is they share something such as, for example, a common decision maker.

It is worth noting the stability of the bivariate probit model. How the explanatory variables affect the outcome remained similar during model construction. Also, the core model is strong; for example there was less than a percentage point drop in accuracy given the omission of such seemingly important variables as the presence of traffic signs, traffic signals and other controls.

5.3.2 Assumptions

Before interpreting the model, a quick recap of the assumptions on which it is based. Like the binary probit models of the previous section, these are based on the assumption of an officer's perspective by independently assigning fault to each party, the combination of which encompasses all fault assignment outcomes. The difference is that the bivariate probit also takes into account that the determination of fault it made by a single entity and, through simultaneous estimation, it estimates the observation-specific variations. The accuracy of the model seems to suggest that it performed as expected.

Another assumption, derived from the results of the binary probit models that the sole-fault estimation is better than the partial-fault estimation also holds true by the comparison of the simulations of the bivariate probit models estimated by each.

5.3.3 Interpretation

There are a few stages to interpreting the bivariate probit model. The first is to see whether it describes the relationship it is intended to well; something the mean accuracy of 57.0% suggests it does. The second is to separate the estimation sets of the two parties' fault as done in Table 4-3, above. The third is to acknowledge the coefficient's sign for each of the variables by estimation set; here negative suggests the party is not at fault and positive suggests the party is at fault. The fourth step is to provide some sort of justification based on personal experience, etc. for the variable to affect in the way that it is purported to.

As with the inferences made upon any model, those here reflect a single plausible interpretation of many plausible interpretations. Before implementing such interpretations, they should be further considered outside of speculation by experts or they should be regarded as hypotheses to be tested. With that caveat, possible interpretations are here provided for the motorist's equation and the cyclist's equation.

Consider the factors contributing to the motorist's fault. The net affect of the motorist's age terms suggest a driver less and less likely to be found in violation up to the age of forty whence he or she becomes more and more likely to be found in violation; this suggests that young and elderly drivers are particularly prone to committing violations in cyclist-motorist collisions. A confirmed state of motorist intoxication implicates the driver, probably because this is itself a violation. When speed is a factor, the motorist is implicated; this may be because speeding is dangerous and correlates to issuance of a violation. Likewise, such hazardous movements as the motorist overtaking the cyclist and turning strongly implicate the motorist; this is probably because these are themselves acts warranting issuance of a violation. If the accident occurred on a driveway, the motorist is less likely to be found at fault; this may be related to acknowledgment of the motorist's limited field of vision. Clear weather suggests the driver is more likely to be found in violation; this may be because the motorist behaves less cautiously in good weather. The lighting conditions of dark suggest the motorist is not in violation whereas that of twilight implicates the motorist; this is possibly because the motorist was not expected to see a cyclist in the dark but was expected to see the cyclist given adequate light.

Commercial land use, institutional land use, and agricultural land use all, to varying degrees, implicate the motorist as compared to residential land use; this may be a result of the mentality of a driver when within a zone – a function of trip purpose. Weekend also implicates the motorist; a result potentially stemming from the same reasons land use implicates the motorist. The vehicle types of sports utility vehicle and motorcycle both suggest their drivers are not to be found in violation; this may be a result of the trip purpose as reflected by the choice of vehicles. Before continuing, it is worth restating that ethnicity/race is merely an indicator variable to represent various potential correlations external to this study. So, as compared to white and Asian motorists, black and Native-American motorists are less likely to be found in violation; one possible, however extremely speculative, interpretation of this phenomena is that the ethnicity of a driver of a vehicle reflects upon a set of values that encourage a more vigilant command of the vehicle. Likewise, in comparison to white and Asian motorists, Hispanic and other ethnicities are more likely to be found in violation; the reasons for this may also reflect on a set of values governing control of the vehicle.

Greater posted speed limits absolve the motorist; this may be because the higher speed limits reflect better driving facilities providing the motorist more control over the vehicle. The presence of traffic signs implicates the motorist; it is possible that this is merely a reflection of the presence of an intersection or other change in traffic flows for which the driver may not properly observe a cyclist. The presence of traffic signals and other traffic controls suggest that the motorist is not in violation; in contrast with the potential cause of the phenomena observed with traffic signs, signals demand compliance by the motorists independent of what they may fail to see.

Consider the factors contributing to the cyclist's fault. The net affect of the cyclist's age terms, as with the motorist, suggest that with age a cyclist is less and less likely to be found in violation up to about 49 whence he or she becomes more and more likely to be found in violation; and, as with the motorist, this suggests that young and elderly cyclists are particularly prone to committing violations in cyclist-motorist collisions. The use of the helmet suggests the cyclist to not be at fault; while lack of helmet use does not likely warrant issuance of a violation, it may indicate a more responsible cyclist. Cyclist turning implicates the cyclist; this suggests as with the ordered probit, the action is executed in a dangerous matter. The lighting conditions of dark suggest the cyclist is in violation whereas that of twilight suggests otherwise; this is possibly because the cyclist may not provide adequate warning illumination at night which does not itself explain why a lit street still suggests fault be with the cyclist but may be explained by a legal requirement to maintain adequate warning lighting between sunset and the next sunrise.

A cyclist positioned in a bike lane is less likely to be found at fault; perhaps because the bike lane is a legal force-field similar to a crosswalk for a pedestrian. A cyclist positioned on a bike path is more likely to be found at fault; perhaps this is related to an ill-founded sense that the cyclist always has the right-of-way. An accident on a driveway implicates the cyclist; potentially because cyclists are not expected on this facility, particularly given the motorist's generally reduced field of vision on such facility. Commercial land use suggests the cyclist is not at fault; this may reflect on the hazards of facilities within the land use. Industrial land use suggests the cyclist is at fault; this may potentially reflect on mentality of the cyclist given the trip's purpose.

The square root of the posted speed limit implicates the cyclist for greater speeds; while not a strong factor, it may indicate, for example, that cyclists underestimate the time they have to make maneuvers. A male cyclist is more likely to be found in violation than his female counterpart; this may point to the aggressive nature of males in their cycling. While ethnicity/race in itself does not implicate the cyclist, being black may, via correlations external to this study[5], reflect upon some characteristics that do implicate the cyclist. Facing traffic implicates the cyclist; this may be because facing traffic, while the law for pedestrians on the roadway, is generally illegal for cyclists. Weekend suggests that the cyclist is not at fault; this may be a result of a potential shift in trip purpose. Both traffic signs and traffic signals implicate the cyclist to a degree; perhaps this stems from the bike's limited stopping and strenuous restarting and, hence, cyclist's general disregard for traffic controls.

The interpretations are similar to those provided for the ordered probit. For the variables that appear in both models, all affect the outcome in the same way except cyclist positioned on a bike path and presence of a traffic sign. These discrepancies may not indicate a true inconsistency between the models: for example, the presence of a traffic sign positively affects both equations in the bivariate probit the relative magnitude there between demonstrates the results of the ordered probit. Similarity in interpretations is good because it shows the factors identified share a distinct relationship. This goes to the strength of interpretations made upon the variables. However, it should again be noted that such interpretations are merely a subset of the many viable interpretations that could be made for each variable. The greater value of these brief interpretations is a place to start the process of finding which factors implicate which party, how, and why. The reader is encouraged to further ponder the physical meaning of the variables.

[5] For example, ethnicity/race may, for historical reasons external to this study, be correlated with wealth.

5.3.4 Application

The bivariate probit model has many applications. As with the application of the ordered probit model discussed in section 5.1.4, the variables may be interpreted in terms of their respective coefficients in order to determine which factors implicate the driver or cyclist, how they do so, and why the do so. From this, policy decisions may be made to, say, supply lights and reflective vests to increase visibility of the cyclist at night, if this were the reason that darkness implicates the cyclist.

5.4 Comparison of Models

As described in Chapter 1, an advantage of decision oriented modeling is that the decision itself is modeled. Thus a comparison of the strongest models representing each perspective may be compared to describe how much one perspective accounts for the variance in the accuracy between that and another perspective. The statistic by which to make this comparison is the ω^2, described in section 3.3.2.

The best model representing the civilian's perspective is the ordered probit model with a mean of 62.3, a sample standard deviation of 0.599 and a sample of size two hundred thousand. The best model representing the officer's perspective is the sole-fault estimated bivariate probit model with a mean of 57.0, a sample standard deviation of 0.612, and a sample size of two hundred thousand. The two-sample t-statistic for independent samples as computed in equation 3-15 is $t = 2141.59$. Thus, by equation 3-14, $\omega^2 = .918$.

The form of the model thus reflects 92% of the variation in the accuracy of the models in simulation. This is not trivial per the terms of the methodology which set the threshold for significance at 50%. Clearly, one model accounts for greater variation than the other. While no hypothesis was made as to which perspective was to do better, knowing that there is such significance for the non-directional case allows one to state that, by inspection, the ordered probit representing the civilian's perspective is stronger than the bivariate probit representing the officer's perspective.

The finding that the civilian's perspective better explains the determination of the reporting police officer could be interpreted by some people to mean that the police officer is more likely to miss the violations of the less offensive party – an accusation that suggests bias in the testimony of witnesses, or worse, bias in the collection and interpretation of evidence and other forms of corruption. Certainly, the severity associated with an improper interpretation of this phenomenon is great. It is therefore important to brainstorm and ponder the viability of many possible other factors contributing to the prevalence of the civilian's perspective. For example, take into consideration the possibility of the presence of important factors not included in the model that qualify the magnitude of violations. Furthermore, consider whether the models of the decisions themselves *best* reflect the decision making process rather than simply *better* reflecting the process.

Some food for thought, consider why determinations of fault in the legal sense are reserved for the insurance companies and the courts. This process allows parties involved a chance, if required, to sort out the soundness of testimony by witnesses, to confront the logic of the law enforcement personnel involved, etc. Perhaps, in this manner, a model that always finds fault may be productive for insurance companies to assign legal fault to the parties involved and allow the courts to clarify; although a model based on the actual determination of fault by such companies and courts would be more powerful.

Chapter 6

Conclusions

To determine the relationship of factors contributing to the fault of parties in cyclist-motorist collisions, there are two perspectives on which to base models: the civilian's perspective, and the officer's perspective. The civilian's perspective recognizes the role of the severity of the infraction in the determination of the party at fault. The relative severity of the parties' infractions may be represented by an ordered fault line and fitted by an ordered probit model. The officer's perspective recognizes the limited and reserved role of an "officer" making the determination of fault independently of each party. The independent nature of this determination allows each party's fault to be determined individually by a binary probit model and combined in simulation. To acknowledge that the "officer" is a single entity, the variations thereby may be taken into account through simultaneously estimating the two outcomes by a bivariate probit model.

Table 6-1 summarizes the performance of the various models included in this study. The ordered probit and sole-fault estimated bivariate probit capture the relationships well, with median 64% and 57% accuracies respectively. The model's strong capturing of the relationship allows for interpretations to be made upon their coefficients to provide a start for future in-depth studies for such purposes as setting policy.

TABLE 6-1. Summary of Models' Performance

Model	Percent Accuracy with at least 99% Confidence
Ordered Probit	
On Estimation Dataset	62.5%
On Simulation Dataset	60.9%
Binary Probit	
Partial-Fault Estimation	53.2%
Sole-Fault Estimation	54.8%
Bivariate Probit	
Partial-Fault Estimation	53.2%
Sole-Fault Estimation	55.6%

As stated in the data description, "fault" is not present in the legal sense of the term but rather in the sense that a party committed a violation. Thus the inferences made are upon factors contributing to the propensity of a party being in violation. These inferences are made with respect to interests in the public sector such as public policy or education and the private sector such as market research, but before making those inferences, the best model on which to base them should be selected.

Again, the best models here are the ordered probit and the sole-fault estimated bivariate probit. Since the coefficients in both models are similar, either model may be used for making inferences upon those coefficients. In consideration of only the coefficients, it is simpler to interpret them with respect to the motorist and cyclist separately as presented in the bivariate probit. An example of a policy recommendation that may come from this is the interpretation of the positive coefficient (implicating the party) on the factor of driveway in the cyclist equation to mean that cyclists are not granted adequate legal protections on these facilities. Another example of a public policy is educating the public of the increased likelihood of a violation to be committed in turning by the cyclist as indicated by the positive coefficient in the cyclist equation suggesting that cyclists should have a rearview mirror affixed to their helmets. There is also interest in making inferences based on the relevant decision process.

Comparisons of the models provides insights into which perspective better describes the decision making process. Based on the ω^2 statistic, the difference in models between the ordered probit representing the civilian's perspective and the sole-fault bivariate probit representing the officer's perspective accounts for 92% of the variation there between. Since the ordered probit is the stronger of the two, further studies may be undertaken with regard to the presence of the civilian's perspective in police officer's reporting of violations, and what, if anything, should be done about it. Additionally, the model's high accuracy in prediction may quite possibly be extrapolated for usage by insurance companies for adjusting rates – particularly because the rate adjustment process need not be transparent to the insured party; however, without basing fault present in the legal sense, it is unlikely that an insurance company may try to utilize this in their assignment of fault in claims.

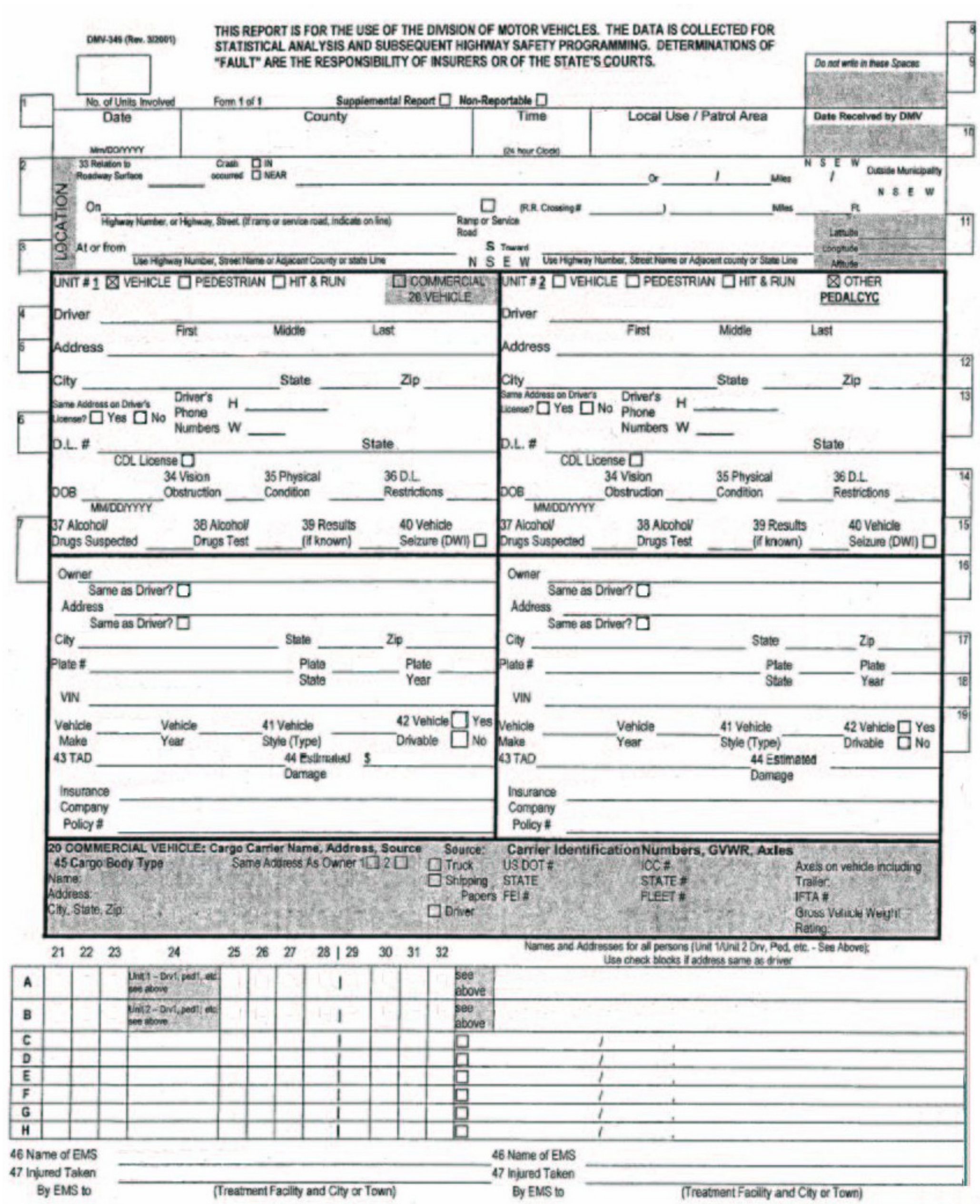
There is more to want: a dataset containing determinations in the legal sense would yield a more powerful model; knowledge that the observations are representative outside of the boundaries of North Carolina; and others. Yet, the methods presented of utilizing decision oriented models based on hypothesized perspectives, simulating and comparing models as independent variables themselves are transferable to these other applications. Indeed these methods are the greatest contribution of this study to the field.

The future should look to modeling fault in its legal sense. With this more powerful definition of fault, conclusions may provide insight into matters of greater interest:

courtroom selection, actuarial sciences, lawmaking, stronger interpretations of factors, etc. Additionally, improvements may be made by further studying more complex decision making processes and working with their associated models to appropriately select the best distribution or combinations of distributions. Implementation may begin by compiling data from courthouses. Meanwhile hypotheses may be made regarding the plausible decision making processes and models selected to represent them. Following creation of the strongest models for each perspective, simulation may occur and accuracies compared; thus allowing insight into the decision making process. Furthermore, the factors present within each model may be interpreted for such uses as policy formulation.

Appendix A

Police Report Form



DMV-349 (Rev. 3/2001)

THIS REPORT IS FOR THE USE OF THE DIVISION OF MOTOR VEHICLES. THE DATA IS COLLECTED FOR STATISTICAL ANALYSIS AND SUBSEQUENT HIGHWAY SAFETY PROGRAMMING. DETERMINATIONS OF "FAULT" ARE THE RESPONSIBILITY OF INSURERS OR OF THE STATE'S COURTS.

Do not write in these Spaces

No. of Units Involved | Form 1 of 1 | Supplemental Report ☐ | Non-Reportable ☐

Date (MM/DD/YYYY) | County | Time (24 hour Clock) | Local Use / Patrol Area | Date Received by DMV

LOCATION

33 Relation to Roadway Surface | Crash occurred ☐ IN ☐ NEAR | Or / Miles | N S E W | Outside Municipality

On | Highway Number, or Highway, Street. (if ramp or service road, indicate on line) | ☐ Ramp or Service Road | (R.R. Crossing #) | Miles | Ft. | Latitude | Longitude | Altitude

At or from | Use Highway Number, Street Name or Adjacent County or state Line | S Toward N S E W | Use Highway Number, Street Name or Adjacent county or State Line

UNIT # 1 ☒ VEHICLE ☐ PEDESTRIAN ☐ HIT & RUN ☐ COMMERCIAL 20 VEHICLE

UNIT # 2 ☐ VEHICLE ☐ PEDESTRIAN ☐ HIT & RUN ☒ OTHER PEDALCYC

Driver (First, Middle, Last)

Address

City | State | Zip

Same Address on Driver's License? ☐ Yes ☐ No | Driver's Phone Numbers H | W

D.L. # | State

CDL License ☐

DOB (MM/DD/YYYY) | 34 Vision Obstruction | 35 Physical Condition | 36 D.L. Restrictions

37 Alcohol/ Drugs Suspected | 38 Alcohol/ Drugs Test | 39 Results (if known) | 40 Vehicle Seizure (DWI) ☐

Owner | Same as Driver? ☐

Address | Same as Driver? ☐

City | State | Zip

Plate # | Plate State | Plate Year

VIN

Vehicle Make | Vehicle Year | 41 Vehicle Style (Type) | 42 Vehicle Drivable ☐ Yes ☐ No

43 TAD | 44 Estimated Damage $

Insurance Company Policy #

20 COMMERCIAL VEHICLE: Cargo Carrier Name, Address, Source | Source: ☐ Truck ☐ Shipping Papers ☐ Driver | Carrier Identification Numbers, GVWR, Axles

45 Cargo Body Type | Same Address As Owner 1 ☐ 2 ☐ | US DOT # | ICC # | Axels on vehicle including Trailer:

Name: | STATE | STATE # | IFTA #

Address: | FEI # | FLEET # | Gross Vehicle Weight Rating:

City, State, Zip:

21 22 23 24 25 26 27 28 | 29 30 31 32 | Names and Addresses for all persons (Unit 1/Unit 2 Drv, Ped, etc. - See Above); Use check blocks if address same as driver

A | Unit 1 - Drv1, ped1, etc. see above | see above

B | Unit 2 - Drv1, ped1, etc. see above | see above

C

D

E

F

G

H

46 Name of EMS 47 Injured Taken By EMS to | (Treatment Facility and City or Town)

46 Name of EMS 47 Injured Taken By EMS to | (Treatment Facility and City or Town)

Date	County	Time	Local Use / Patrol area	Page # 2

48 Points of Initial Contact (Write in Codes) Unit # 1 __ Unit # 2 __

CRASH SEQUENCE (Unit Level)	Unit # 1	Unit # 2
49 Vehicle Maneuver / Action		
50 Non - Motorist Action		
51 Non - Motorist Location Prior to Impact		
52 Crash Sequence - First Event for This Unit		
53 Crash Sequence - Second Event *		
54 Crash Sequence - Third Event *		
55 Crash Sequence - Fourth Event *		
56 Most Harmful Event for This Unit		
57 Distance / Direction to Object Struck		
58 Vehicle Under ride / Over ride		
59 Vehicle Defects		

VEHICLE INFO.	Veh # 1	Veh # 2
60 Authorized Speed Limit		
61 Estimate of Original Traveling Speed		
62 Estimate of Speed at Impact		
63 Tire Impressions Before Impact (ft.)		
64 Distance Traveled After Impact (ft.)		
65 Emergency Vehicle Use		
66 Post Crash Fire (If "Yes" check block)	☐	☐
67 School Bus - Contact Vehicle *	☐	☐
68 School Bus - Noncontact Vehicle *	☐	☐

ROADWAY INFO.	
69 Road Feature	
70 Road Character	
71 Road Classification	
72 Road Surface Type	
73 Road Configuration	
74 Access Control	
75 Number of Lanes	
76 Traffic Control Type	
77 Traffic Control Oper	

WORKZONE RELATED	
78 Work zone Area	
79 Work Activity	
80 Work Area Marked	
81 Crash Location	

TRAILER INFO.	Unit 1	Unit 2
82 Trailer Type		
1st Trailer No. Axles		
Width (inches)		
Length (feet)		
2nd Trailer No. Axles		
Width (inches)		
Length (feet)		

83 Unit # Over width Trailer and Over width Mobile Home | Over-width Permit #

COMMERCIAL VEHICLE: Hazardous Materials Involvement

HazMat Placard ☐ Yes ☐ No From Placard Indicate:

Hazardous Cargo Released (does not include fuel from fuel tank) ☐ Yes ☐ No 4-digit placard number or name from diamond or box 1-digit number from Bottom of Diamond

Carrying HazMat ☐ Yes ☐ No

84 DIAGRAM

Unit # 1 was: ☐ Traveling ☐ Parked facing N S E W SE On ________ Unit # 2 was: ☐ Traveling ☐ Parked facing N S E W On ________

85 NARRATIVE (Include pertinent and unusual aspects, which are not listed elsewhere on the form)

ADDITIONAL PROPERTY DAMAGE

86 Type /Owner / Address & Phone / State Property? ☐ Estimated Damage $____

2nd Type /Owner / Address & Phone / ☐ $____

WITNESSES

Name ________ Address ________ Phone No. ________

Name ________ Address ________ Phone No. ________

TRAFFIC VIOLATION (S)

Name ________ Charge(s) ____

Name ________ Charge(s) ____

Officer Title & Name | Officer Number | Department | Date of Report

Appendix B

Stata® Modeling Commands

Data Preparation

`DataPrep11-09-2005.do`

```
clear
set memory 10m
set matsize 100

insheet using "C:\DATA\Bike2005-09-29\bikeacc.csv", comma

gen fault = 1 * faulttwo + 0*faultnon + 2*faultdr + 3*faultbik
gen ofault = -1*faultdr + 0*faulttwo + 1*faultbik

gen weekend = 0
replace weekend = 1 if(day == 1 | day == 7)

gen bikage2 = bikage^2
gen lnbikage = ln(bikage)
gen drage2 = drage^2
gen lndrage = ln(drage)

//Sharing the blame
gen partbik = 0
replace partbik = 1 if(faulttwo == 1 | faultbik == 1)
gen partdr = 0
replace partdr = 1 if(faulttwo == 1 | faultdr == 1)

gen spdlimit05 = spdlimit^0.5
```

Estimation of the Ordered Probit

oprobit 11-15-2005 with export.do

```
quietly do "C:\thesis_calcs\DataPrep11-09-2005.do"

oprobit ofault helmetus vehspeed bikage lnbikage drdrink drage lndrage
    dawndusk darklit darknlit speed bturn movert mturn rain comm facetrf
    bikblack drway bikelane bikepath sidecros trfsign trfsigl trfothr weekend
    if(faultnon == 0) //11-13-2005
predict r1 r2 r3
sum r1 r2 r3

set more off
biprobit (faultdr = drage drage2 drdrink speed movert mturn drway clear
    dawndusk darklit darknlit comm inst farm weekend suv motcycle drblack
    drnativ drhisp drother spdlimit) (faultbik = bikage bikage2 helmetus bturn
    dawndusk darknlit darklit bikelane bikepath drway comm indu spdlimit05
    bikmale bikblack facetrf weekend) if(faultbik == 1 | faultdr == 1)
predict temp, p11
set more on

keep if (r1 < .)
sum crsh_id fault r1 r2 r3
outsheet crsh_id fault r1 r2 r3 using "C:\oprobit-FullDataset-sm.raw", comma
    replace

keep if(fault<. & temp <.)
sum crsh_id fault r1 r2 r3
outsheet crsh_id fault r1 r2 r3 using "C:\oprobit-FaultDataset-sm.raw", comma
    replace

quietly do "C:\thesis_calcs\DataPrep11-09-2005.do"
oprobit ofault helmetus vehspeed bikage lnbikage drdrink drage lndrage
    dawndusk darklit darknlit speed bturn movert mturn rain comm facetrf
    bikblack drway bikelane bikepath sidecros trfsign trfsigl trfothr weekend
    if(faultnon == 0) //11-13-2005
predict r1 r2 r3
sum r1 r2 r3

keep if(e(sample))
sum crsh_id fault r1 r2 r3
outsheet crsh_id fault r1 r2 r3 using "C:\oprobit-eSampleDataset-sm.raw",
    comma replace
```

Estimation of the Partial-Fault Binary Probit

probit 11-15-2005.do

```
quietly do "C:\thesis_calcs\DataPrep11-09-2005.do"

quietly probit partbik bikage bikage2 helmetus bturn dawndusk darknlit darklit
    bikelane bikepath drway comm indu spdlimit05 bikmale bikblack facetrf
    weekend trfsign trfsigl if(faultnon == 0)
predict part1 // part biker
gen samplebik = 0
replace samplebik = 1 if(e(sample))

probit partdr drage drage2 drdrink speed movert mturn drway clear dawndusk
    darklit darknlit comm inst farm weekend suv motcycle drblack drnativ
    drhisp drother spdlimit trfsign trfsigl trfothr if(faultnon == 0 & part1 < .)
predict pbdr // part driver
gen sampledr = 0
replace sampledr = 1 if(e(sample))


probit partbik bikage bikage2 helmetus bturn dawndusk darknlit darklit
    bikelane bikepath drway comm indu spdlimit05 bikmale bikblack facetrf
    weekend trfsign trfsigl if(faultnon == 0 & pbdr < .)
predict pbbik // part biker

sum crsh_id fault pbdr pbbik

keep if(pbdr < . & pbbik < .)
sum crsh_id fault pbdr pbbik
outsheet crsh_id fault pbdr pbbik using "C:\probit-FullDataSet-sm.raw",
    comma replace

keep if(fault < .)
sum crsh_id fault pbdr pbbik
outsheet crsh_id fault pbdr pbbik using "C:\probit-FaultDataSet-sm.raw",
    comma replace


keep if(samplebik == 1 & sampledr == 1)
sum crsh_id fault pbdr pbbik
outsheet crsh_id fault pbdr pbbik using "C:\probit-eSampleDataSet-sm.raw",
    comma replace
```

Estimation of the Sole-Fault Binary Probit

probit sm 11-15-2005.do

```
quietly do "C:\thesis_calcs\DataPrep11-09-2005.do"

quietly probit faultbik bikage bikage2 helmetus bturn dawndusk darknlit
    darklit bikelane bikepath drway comm indu spdlimit05 bikmale bikblack
    facetrf weekend trfsign trfsigl if(faultbik == 1 | faultdr == 1)
predict part1
gen samplebik = 0
replace samplebik = 1 if(e(sample))

probit faultdr drage drage2 drdrink speed movert mturn drway clear dawndusk
    darklit darknlit comm inst farm weekend suv motcycle drblack drnativ
    drhisp drother spdlimit trfsign trfsigl trfothr if((faultbik == 1 | faultdr == 1)
    & part1 < .)
predict pbdr
gen sampledr = 0
replace sampledr = 1 if(e(sample))

probit faultbik bikage bikage2 helmetus bturn dawndusk darknlit darklit
    bikelane bikepath drway comm indu spdlimit05 bikmale bikblack facetrf
    weekend trfsign trfsigl if((faultbik == 1 | faultdr == 1) & pbdr < .)
predict pbbik

sum crsh_id fault pbdr pbbik

keep if(pbdr < . & pbbik < .)
sum crsh_id fault pbdr pbbik
outsheet crsh_id fault pbdr pbbik using "C:\probit_sm-FullDataSet-sm.raw",
    comma replace

keep if(fault < .)
sum crsh_id fault pbdr pbbik
outsheet crsh_id fault pbdr pbbik using "C:\probit_sm-FaultDataSet-sm.raw",
    comma replace


keep if(samplebik == 1 & sampledr == 1)
sum crsh_id fault pbdr pbbik
outsheet crsh_id fault pbdr pbbik using "C:\probit_sm-eSampleDataSet-
    sm.raw", comma replace
```

Estimation of the Partial-Fault Bivariate Probit

`biprobit 11-13-2005.do`

```
quietly do "C:\thesis_calcs\DataPrep11-09-2005.do"

probit partdr drage drage2 drdrink speed movert mturn drway clear dawndusk
    darklit darknlit comm inst farm weekend suv motcycle drblack drnativ
    drhisp drother spdlimit trfsign trfsigl trfothr if(faultnon == 0)
probit partbik bikage bikage2 helmetus bturn dawndusk darknlit darklit
    bikelane bikepath drway comm indu spdlimit05 bikmale bikblack facetrf
    weekend trfsign trfsigl if(faultnon == 0)

set more off

biprobit (partdr  = drage drage2 drdrink speed movert mturn drway clear
    dawndusk darklit darknlit comm inst farm weekend suv motcycle drblack
    drnativ drhisp drother spdlimit trfsign trfsigl trfothr) (partbik  = bikage
    bikage2 helmetus bturn dawndusk darknlit darklit bikelane bikepath drway
    comm indu spdlimit05 bikmale bikblack facetrf weekend trfsign trfsigl)
    if(faultnon == 0)

set more on

predict pb11, p11
predict pb10, p10
predict pb01, p01
predict pb00, p00
sum crsh_id fault pb11 pb10 pb01 pb00

keep if (pb11 < .)
sum crsh_id fault pb11 pb10 pb01 pb00
outsheet crsh_id fault pb11 pb10 pb01 pb00 using "C:\biprobit-FullDataset-
    sm.raw", comma replace

keep if(fault<.)
sum crsh_id fault pb11 pb10 pb01 pb00
outsheet crsh_id fault pb11 pb10 pb01 pb00 using "C:\biprobit-FaultDataset-
    sm.raw", comma replace

keep if(e(sample))
sum crsh_id fault pb11 pb10 pb01 pb00
outsheet crsh_id fault pb11 pb10 pb01 pb00 using "C:\biprobit-
    eSampleDataset-sm.raw", comma replace
```

Estimation of the Sole-Fault Bivariate Probit

biprobit sm 11-13-2005.do

```
quietly do "C:\thesis_calcs\DataPrep11-09-2005.do"

probit faultdr drage drage2 drdrink speed movert mturn drway clear dawndusk
    darklit darknlit comm inst farm weekend suv motcycle drblack drnativ
    drhisp drother spdlimit trfsign trfsigl trfothr if(faultbik == 1 | faultdr == 1)
probit faultbik bikage bikage2 helmetus bturn dawndusk darknlit darklit
    bikelane bikepath drway comm indu spdlimit05 bikmale bikblack facetrf
    weekend trfsign trfsigl if(faultbik == 1 | faultdr == 1)

set more off

biprobit (faultdr = drage drage2 drdrink speed movert mturn drway clear
    dawndusk darklit darknlit comm inst farm weekend suv motcycle drblack
    drnativ drhisp drother spdlimit trfsign trfsigl trfothr) (faultbik = bikage
    bikage2 helmetus bturn dawndusk darknlit darklit bikelane bikepath drway
    comm indu spdlimit05 bikmale bikblack facetrf weekend trfsign trfsigl)
    if(faultbik == 1 | faultdr == 1)

set more on

predict pb11, p11
predict pb10, p10
predict pb01, p01
predict pb00, p00
sum crsh_id fault pb11 pb10 pb01 pb00

keep if (pb11 < .)
sum crsh_id fault pb11 pb10 pb01 pb00
outsheet crsh_id fault pb11 pb10 pb01 pb00 using "C:\biprobit_sm-
    FullDataset-sm.raw", comma replace

keep if(fault<.)
sum crsh_id fault pb11 pb10 pb01 pb00
outsheet crsh_id fault pb11 pb10 pb01 pb00 using "C:\biprobit_sm-
    FaultDataset-sm.raw", comma replace

keep if(e(sample))
sum crsh_id fault pb11 pb10 pb01 pb00
outsheet crsh_id fault pb11 pb10 pb01 pb00 using "C:\biprobit_sm-
    eSampleDataset-sm.raw", comma replace
```

Appendix C

Perl Simulation Scripts

Simulate the Ordered Probit Model

```
#Input:  Oprobit prediction file set in .raw format below
#        "ID_No,Fault,r1,r2,r3"
#Output: Percent correct predictions
#        "Iteration,Pct_Correct"

my @oprobit_files = ("oprobit-FaultDataset-sm.raw","oprobit-
eSampleDataset-sm.raw");
my $itterations  = 200000;

foreach my $file (@oprobit_files)
{
    if(open(IN,"<$file"))
    {
        if(open(OUT, ">pred_$file"))
        {
            print OUT "ITT_No,Pct_Correct\n";
            my @table = ();
            while(<IN>)
            {
                chomp;
                push @table, [ split /,/ ];
            }

            for(my $i=1; $i<$itterations+1; $i++)
            {
                my $correct=0;
                for(my $row=1;$row<$#table+1;$row++)
                {
                    my $rnd1 = rand();
                    if($rnd1<$table[$row][2])
                    {
                        if($table[$row][1] == 2)
```

Simulate the Ordered Probit Model (cont'd)

```
                    {
                        $correct++;
                    }
                }
                elsif($rnd1<$table[$row][2]+$table[$row][3])
                {
                    if($table[$row][1] == 1)
                    {
                        $correct++;
                    }
                }
                else #Only 3 Options
                {
                    if($table[$row][1] == 3)
                    {
                        $correct++;
                    }
                }
            }
            my $pct_correct = $correct*100/$#table;
            print OUT "$i,$pct_correct\n";
        }
        close OUT;
    }
    close IN;
  }
}
```

Simulate Ordered Probit Model for Case Study: Neither at Fault Treated as Both at Fault

```
#Input:  Oprobit prediction file set in .raw format below
#        "ID_No,Fault,r1,r2,r3"
#Output: Percent correct predictions
#        "Iteration,Pct_Correct"

my @oprobit_files = ("oprobit-eSampleDataset-sm.raw");
my $itterations  = 200000;
```

Simulate Ordered Probit Model for Case (cont'd)

```
foreach my $file (@oprobit_files)
{
    if(open(IN,"<$file"))
    {
        if(open(OUT, ">faultnon_eq_faulttwo_pred_$file"))
        {
            print OUT "ITT_No,Pct_Correct\n";
            my @table = ();
            while(<IN>)
            {
                chomp;
                push @table, [ split /,/ ];
            }
            for(my $i=1; $i<$itterations+1; $i++)
            {
                my $correct=0;
                for(my $row=1;$row<$#table+1;$row++)
                {
                    my $rnd1 = rand();
                    if($rnd1<$table[$row][2])
                    {
                        if($table[$row][1] == 2)
                        {
                            $correct++;
                        }
                    }
                    elsif($rnd1<$table[$row][2]+$table[$row][3])
                    {
                        if($table[$row][1] == 1)
                        {
                            $correct++;
                        }
                    }
                    else #Only 3 Options
                    {
                        if($table[$row][1] == 3 || $table[$row][1] == 0)
                        #both at fault or neither at fault "correct"
                        {
                            $correct++;
                        }
                    }
```

Simulate Ordered Probit Model for Case (cont'd)

```
                }
                my $pct_correct = $correct*100/$#table;
                print OUT "$i,$pct_correct\n";
            }
            close OUT;
        }
        close IN;
    }
}
```

Simulate the Binary Probit Models

```
#Input:  Probit prediction file set in .raw format below
#          "ID_No,Fault,Driver,Bike"
#Output: Percent correct predictions
#          "Iteration,Pct_Correct"

my @probit_files = ("probit_sm-FaultDataset-sm.raw","probit_sm-
eSampleDataset-sm.raw", "probit-FaultDataset-sm.raw","probit-
eSampleDataset-sm.raw");
my $itterations  = 200000;

foreach my $file (@probit_files)
{
    if(open(IN,"<$file"))
    {
        if(open(OUT, ">pred_$file"))
        {
            print OUT "ITT_No,Pct_Correct\n";
            my @table = ();
            while(<IN>)
            {
                chomp;
                push @table, [ split /,/ ];
            }

            for(my $i=1; $i<$itterations+1; $i++)
            {
                my $correct=0;
                for(my $row=1;$row<$#table+1;$row++)
```

Simulate the Binary Probit Models (cont'd)

```
                {
                    my $rnd1 = rand();
                    my $rnd2 = rand();
                    if($rnd1<$table[$row][2] && $rnd2<$table[$row][3])
                    {
                        if($table[$row][1] == 1)
                        {
                            $correct++;
                        }
                    }
                    elsif($rnd1<$table[$row][2])
                    {
                        if($table[$row][1] == 2)
                        {
                            $correct++;
                        }
                    }
                    elsif($rnd2<$table[$row][3])
                    {
                        if($table[$row][1] == 3)
                        {
                            $correct++;
                        }
                    }
                    else
                    {
                        if($table[$row][1] == 0)
                        {
                            $correct++;
                        }
                    }
                }
                my $pct_correct = $correct*100/$#table;
                print OUT "$i,$pct_correct\n";
            }
            close OUT;
        }
        close IN;
    }
}
```

Simulate the Bivariate Probit Models

```
#Input:  Biprobit prediction file set in .raw format below
#        "ID_No,Fault,pb11,pb10,pb01,pb00"
#Output: Percent correct predictions
#        "Iteration,Pct_Correct"

my @biprobit_files = ("biprobit_sm-FaultDataset-sm.raw","biprobit_sm-
eSampleDataset-sm.raw","biprobit-FaultDataset-sm.raw","biprobit-
eSampleDataset-sm.raw");
my $itterations  = 200000;

foreach my $file (@biprobit_files)
{
    if(open(IN,"<$file"))
    {
        if(open(OUT, ">pred_$file"))
        {
            print OUT "ITT_No,Pct_Correct\n";
            my @table = ();
            while(<IN>)
            {
                chomp;
                push @table, [ split /,/ ];
            }

            for(my $i=1; $i<$itterations+1; $i++)
            {
                my $correct=0;
                for(my $row=1;$row<$#table+1;$row++)
                {
                    my $rnd1 = rand();
                    if($rnd1<$table[$row][2])
                    {
                        if($table[$row][1] == 1)
                        {
                            $correct++;
                        }
                    }
                    elsif($rnd1<$table[$row][2]+$table[$row][3])
                    {
                        if($table[$row][1] == 2)
                        {
```

Simulate the Bivariate Probit Models (cont'd)

```
                    $correct++;
                }
            }

    elsif($rnd1<$table[$row][2]+$table[$row][3]+$table[$row][4])
            {
                if($table[$row][1] == 3)
                {
                    $correct++;
                }
            }
            else
            {
                if($table[$row][1] == 0)
                {
                    $correct++;
                }
            }
        }
        my $pct_correct = $correct*100/$#table;
        print OUT "$i,$pct_correct\n";
    }
    close OUT;
  }
  close IN;
 }
}
```

Create Densities

```
#Input:  Prediction file set in .raw format below
#        "ITT_No,Pct_Correct"
#Output: Percent correct predictions
#        "Cumdensity,Pct_Correct"

my @files = ("pred_biprobit_sm-FaultDataset-sm.raw","pred_biprobit_sm-
eSampleDataset-sm.raw", " pred_biprobit-FaultDataset-
sm.raw","pred_biprobit-eSampleDataset-sm.raw", "pred_probit_sm-
FaultDataset-sm.raw","pred_probit_sm-eSampleDataset-sm.raw", "
pred_probit-FaultDataset-sm.raw","pred_probit-eSampleDataset-sm.raw ",
```

Create Densities (cont'd)

```
"pred_oprobit_sm-FaultDataset-sm.raw","pred_oprobit_sm-eSampleDataset-
sm.raw");
foreach my $file (@files)
{
    if(open(IN,"<$file"))
    {
        if(open(OUT, ">cum_$file"))
        {
            my $head1;
            my $head2;
            my @itt = ();
            my @pct = ();
            my @a = ();
            while(<IN>)
            {
                chomp();
                @a = split /,/;
                push @itt, $a[0];
                push @pct, $a[1];
            }
            $head1 = shift @itt;
            $head2 = shift @pct;
            while(($#itt+2)%100)
            {
                pop @itt;
                pop @pct;
            }
            @pct = sort @pct;
            foreach my $i (@itt)
            {
                $i = 100-($i/($#itt+2))*100;
            }
            print OUT "cumdensity,pct_correct\n";
            for(my $j=0;$j<$#itt+1;$j++)
            {
                print OUT "$itt[$j],$pct[$j]\n";
            }
            close OUT;
        }
    }
}
```

References

Al-Balbissi, A. H. (2003). Role of gender in road accidents. *Traffic Injury Prevention*, *4*(1), 64-73.

Ball, Karlene, & Owsley, C., & Stalvey, B. & Roenker, D. L., & Sloane, M. E., & Graves, M. (1998). Driving avoidance and functional impairment in older drivers. *Accident Analysis and Prevention*, *30*(3), 313-322.

Calisir, F., & Lehto, M. R. (1996). Drivers' risk assessments and their impact on seat belt use. *Computers and Industrial Engineering*, *31*(1-2), 495-498.

Chen, Wenjum, & Cooper, P. & Pinili, Mario. (1995). Driver accident risk in relation to the penalty point system in British Columbia. *Journal of Safety Research*, *26*(1), 9-18.

Clarke, David D., & Ward, P. J., & Jones, J. (1998). Overtaking road –accidents: differences in manoeuvre as function of driver age. *Accident Analysis and Prevention*, *30*(4), 455-467.

Cooper, P. J., & Pinili, M., & Chen, W. (1995). An examination of the crash involvement rates of novice drivers aged 16 to 55. *Accident Analysis and Prevention*, *27*(1), 89-104.

Dissanayake, S. & Lu, J. J. (2002). Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashes. *Accident Analysis and Prevention*, *34*(5), 609-618.

Elliot, Michael R., & Waller, P. F., & Raghunathan, T. E., & Shope, J. T., & Little, R. J. A. (2000). *Journal of Safety Research*, *31*(4), 229-242.

Ferrari, R., & Russell, A. S. (2001). Why blame is a factor in recovery from whiplash injury. *Medical Hypotheses*, *56*(3), 372-375.

van der Flier, H., & Schoonman, W. (1988). Railway signals passed at danger: situational and personal factors underlying stop signal abuse. *Applied Ergonomics, 19*(2), 135-141.

Garretson, Marilee, &Peck, R. C. (1982). Factors associated with fatal accident involvement among California drivers. *Journal of Safety Research*, *13*(4), 141-156.

Gruder, Charles D., & Romer, D., & Korth, B. (1978). Dependency and fault as determinants of helping. *Journal of Experimental Social Psychology*, *14*(2), 227-235.

Hanowski, R. J., & Wierwille, W. W., & Dingus, T. A. (2003). An on-road study to investigate fatigue in local/short haul trucking. *Accident Analysis and Prevention*, *35*(2), 153-160.

Hing, J. Y. C., & Stamatiadis, N., & Aultman-Hall, L. (2003). Evaluating the impact of passengers on the safety of older drivers. *Journal of Safety Research*, *34*(a), 343-351.

Kim, K, & Li, L. (1996). Modeling fault among bicyclists and drivers involved in collisions in Hawaii, 1986-1991. *Transportation Research Record*, 1538, 75-80.

Kim, K., & Richardson, J., & Nitz, L. (1998). Drivers at fault: influences of age, sex, and vehicle type. *Journal of Safety Research*, *29*(3), 171-179.

Kim, K. E. (2000). Differences between male and female involvement in motor vehicle collisions in Hawaii, 1986-1993. *Women's Travel Issues: Proceedings from the Second National Conference October 1996* (chap. 27). Retrieved November 18, 2005, from http://www.fhwa.dot.gov/ohim/womens/chap27.pdf

Legree, Peter J., & Heffner, T. S., & Psotka, J., & Martin, D.E., & Medsker, G. J. (2003). Traffic crash involvement: experiential driving knowledge and stressful contextual antecedents. *Journal of Applied Psychology*, *88*(1), 15-26.

McCartt, Anne T., & Northrup, V. S., & Retting, R. A. (2004). Types and characteristics of ramp-related motor vehicle crashes on urban interstate roadways in Northern Virginia. *Jornal of Safety Research*, *35*(1), 107-114.

McGwin, Gerald Jr., & Brown, David B. (1999). Characteristics of traffic crashes among young, middle-aged, and older drivers. *Accident Analysis and Prevention*, *31*(3), 181-198.

National Safety Council. (2004). *Injury Facts*® (2004 ed.). Itasca, IL: The Council.

Peltzer, K., & Walter, R. (2004). Psychosocial correlates of the impact of road traffic accidents among South African drivers and passengers. *Accident Analysis and Prevention*, *36*(3), 367-374.

Pruesser, David F., & Ferguson, S. A., & Williams, A. F. (1998). The effect of teenage passengers on the fatal crash risk of teenage drivers. *Accident Analysis and Prevention*, *30*(2), 217-222.

Quddus, M. A., & Nolan R. B., & Chin, H. C. (2002). An analysis of motorcycle injury and vehicle damage severity using ordered probit models. *Journal of Safety Research*, *33*(4), 445-462.

Raedt, R. D., & Ponjaert-Kristoffersen, I. (2001). Predicting at-fault car accidents of older drivers. *Accident Analysis and Prevention*, *33*(6), 809-819.

Retting, R. A., & Weinstein, H. B., & Solomon, M. G. (2003). Analysis of motor-vehicle crashes at stop signs in four U.S. cities. *Journal of Safety Research*, *34*(5), 485-489.

Runyon, Richard P., & Coleman, K. A., & Pittenger, D. J. (2000). *Fundamentals of Behavioral Statistics* (9th ed.). New York: McGraw-Hill.

Sagberg, Fridulv. (1999). Road accidents caused by drivers falling asleep. *Accident Analysis and Prevention*, *31*(6), 639-649.

Sjögren, Harmeet, & Eriksson, A., & Öström, M. (1996). Role of disease in initiating the crashes of fatally injured. *Accident Analysis and Prevention*, *28*(3), 307-314.

Solnick, Sara J., & Hemenway, D. (1995). The hit-and-run in fatal pedestrian accidents: victims, circumstances and drivers. *Accident Analysis and Prevention*, *27*(5), 643-649.

Stamatiadis, Nikiforos, & Deacon, J. A. (1995). Trends in highway safety: effects of an aging population on accident propensity. *Accident Analysis and Prevention*, *27*(4), 443-459.

StataCorp. (2003). *Stata Statistical Software: Release 8.0*. College Station, TX: Stata Corporation

Stewart, Alan E. (2005). Attributions of responsibility for motor vehicle crashes. *Accident Analysis and Prevention*, *37*(4), 681-688.

Train, Kenneth E. (2003). *Discrete Choice Methods with Simulation*. New York: Cambridge University Press.

Underwood, Geoffrey, & Chapman, P., & Crundall, D. (1999). Anger while driving. *Transportation Research Part F: Psychology and Behaviour*, *2*(1), 55-68.

Vernon, D. D., & Diller, E. M., & Cook, L. J., & Reading, J. C., & Suruda, A. J., & Dean, J. M. (2002) Evaluating the crash and citation rates of Utah drivers licensed with medical conditions, 1992 1996. *Accident Analysis and Prevetion*, *34*(2), 237-246.

Waller, Patricia F., & Elliot, M. R. & Shope, J. T., & Raghunathan, T. E., & Little, R. J. A. (2001). Changes in young adult offense and crash patterns over time. *Accident Analysis and Prevention*, *33*(1), 117-128.

Yannis, G., & Golias, J., & Papadimitriou, E. (2005). Driver age and vehicle engine size effects on fault and severity in young motorcyclists accidents. *Accident Analysis and Prevention*, *37*(2), 327-333.

Vita

Craig Wademan Schiller

Born in St. Louis, Craig Schiller is a 2002 graduate of Marquette High School in Clarkson Valley, Missouri. In the Fall, he entered Washington University in St. Louis where he pursues a Bachelor of Science in Civil Engineering under the advisement of Dr. Phillip L. Gould. He concurrently enrolled in the Master of Science in Civil Engineering (Transportation) under the advisement of Dr. Gudmundur Freyr Ulfarsson.

Craig Schiller's academic honors include being named to Dean's List from Spring '03 to Spring '05, membership into the National Society of Collegiate Scholars, Sigma Xi and Chi Epsilon. He is a member of the American Planning Association.

Short Title: Fault in Cyclist-Motorist Accidents Schiller, M.S. 2005