8-2011

Statistical Analysis of Fatalities Due to Vehicle Accidents in Las Vegas, NV

Annabelle Marie Mathis
University of Nevada, Las Vegas

STATISTICAL ANALYSIS OF FATALITIES DUE TO VEHICLE ACCIDENTS IN LAS VEGAS, NV

by

Annabelle Marie Mathis

A thesis submitted in partial fulfillment

of the requirements for the

Master of Science in Mathematical Sciences

Department of Mathematical Sciences

College of Sciences

The Graduate College

University of Nevada, Las Vegas

August 2011

THE GRADUATE COLLEGE

We recommend the thesis prepared under our supervision by

Annabelle Marie Mathis

entitled

Statistical Analysis of Fatalities Due to Vehicle Accidents in Las Vegas, NV

be accepted in partial fulfillment of the requirements for the degree of

Master of Science in Mathematical Sciences
Department of Mathematical Sciences

Chih-Hsiang Ho, Committee Chair

Anton Westveld, Committee Member

Kaushik Ghosh, Committee Member

Sarah Catlin, Committee Member

Chad Cross, Graduate College Representative

Ronald Smith, Ph. D., Vice President for Research and Graduate Studies
and Dean of the Graduate College

August 2011

ABSTRACT

STATISTICAL ANALYSIS OF FATALITIES DUE TO VEHICLE ACCIDENTS IN LAS VEGAS, NV

by

Annabelle Marie Mathis

Dr. Chih-Hsiang Ho, Examination Committee Chair
Professor of Mathematical Sciences
University of Nevada, Las Vegas

The goal of this thesis is to investigate factors that affect the odds of having a fatality in a vehicle collision. We will be looking at characteristics of the driver that caused the accident (age, gender, behavior, actions, influences, and seat belt worn), the characteristics of the vehicle the driver drove (type of vehicle, and air bag deployment), the characteristics of the environment in which the accident occurred (weather, road condition, lighting, time of day, the day of the week, and month of the year), the characteristics of the crash (direction of accident and how many vehicles were involved), and the characteristics of the zip code the accident happened (population, median of income per household, and percentage of zipcode that has less than a high school education, a high school education, a two-year degree, a four year degree and a post graduate degree). All of these variables might affect the odds of having a fatality. Modeling will involve the use of multiple logistic regression. We will be addressing the following areas: data management, model fitting, best subset selection, model diagnostics, and model validation.

By identifying the best factors, the selected final model might be a helpful tool in formulating cost-effective safety measures for legislation. Additionally, this model and its findings could potentially be used to develop new social programs that would pinpoint the exact areas that are in need of safety programs, which might save lives in the long term.

ACKNOWLEDGEMENTS

I want to say thanks to my husband; without him I never would have made it through the programming of this thesis. Thanks to my friends for their constant support and overwhelming belief that I could do this, to my committee members, for all their help during each stage of this thesis, to Dr Hualiang Teng, for introducing logistic regression and transportation to me, and to Kim Stalling who gave me the data and helped to explain what the variables meant. Finally, in loving memory of my mother, Rollinda, without whom I would never be the woman that I am today. She is the reason that I work so hard, through difficult times and easy, because no matter what cards were dealt to a person, they can always triumph.

Table of Contents

List of Figures:

Chapter 1

1.1 Introduction

"Every twelve minutes, someone dies in a car crash on U.S. roads," reads the first line on a flier from the National Center for Injury Prevention and Control. A motor vehicle collision occurs when a road vehicle collides with another object, be it a vehicle, pedestrian, animal, road debris, geographical obstacle or architectural obstacle. These collisions can result in injury, property damage, and/or death. In the United States, the definition of road-traffic fatality that is used by the Fatality Analysis Reporting System (FARS), which is run by the NHTSA, is "a fatality in the state of Nevada is defined as any person that dies within 30 days due to a vehicle accident that occurred on a United States public road and the vehicle had an engine." (NHTSA, Fatality Analysis Reporting System, 2010). For 2009, according to FARS, in the United States, there were 30,797 fatal motor vehicle crashes with a total of 33,808 fatalities (FARS, 2010). For Nevada, the corresponding figure in 2009 was 223 fatal crashes with 243 deaths occurring from those crashes. (NHTSA, Fatality Analysis Reporting System, 2010). Several studies have explored the factors that might influence fatal accidents such as these. In one study, it was suggested that there was an association between driver, crash and vehicle characteristics to driver fatalities (Bedard, Guyatt, Stones, & Hirdes, 2002). In another study, driver distractions were reported to have been involved in 16 percent of all fatal crashes in 2008 according to data from the Fatality Analysis Reporting System (Ascone & Lindsey, 2009). What are these characteristics and distractions that these studies refer to? How could these death tolls be brought down? Could prevention policies be used to help reduce motor-vehicle related injuries or fatalities?

In this thesis, we will discuss fatalities due to motor-vehicle collisions. First, we will discuss the variables that other literature talks about. There are several variables to consider when looking at fatalities of any sort. Most of these variables can be broadly placed in the following categories: environment of the deceased individual and characteristics of the deceased individual. However, if that environment is a vehicle collision then we might want to narrow these categories down a bit more precisely. For example, about 37.5% of all nationwide fatalities in vehicle-related incidents in 2006 involved alcohol (AlcoholAlert, 2010). In a study involving all collisions (not exclusively collisions involving fatalities), it was found that 57% of crashes were due solely to driver factors, 27% to combined roadway and driver factors, 6% to combined vehicle and driver factors, 3% solely to roadway factors, 3% to combined roadway, driver, and vehicle factors, 2% solely to vehicle factors and 1% to combined roadway and vehicle factors (Lum & Reagan, 1995). In theory, it seems that there are many variables to look at.

1.1.1 Driver Characteristics

There are several driver characteristics that could affect the odds of a fatality occurring. Several of these are age, gender, distractions and actions. Age is a major contributor to many vehicle collisions. A study, which used multivariate logistic regression, revealed that the odds ratio (OR) of a fatal injury increased with age (Bedard, Guyatt, Stones, & Hirdes, 2002). Gender also seems to be a characteristic that has significance. In the same research study, the majority of fatalities were among male drivers younger than 30 years, which was at 26.6 percent, versus females of the same age range, at a 5.6 percent (Bedard, Guyatt, Stones, & Hirdes, 2002).

Driving is a very complex task, which involves various cognitive, physical, sensory, and psychomotor skills working together. Distractions are defined as any secondary activity that competes for the driver's attention while driving. These distractions have the potential to worsen driving performance and have serious consequences for road safety. According to the NHTSA, it is estimated that 25 percent of police-reported crashes are caused by driver inattention (NHTSA, Fatality Analysis Reporting System, 2010). According to one study, more experienced drivers are often capable of dividing their attention between driving tasks and non-driving tasks without any serious consequences (Young & Regan, 2007). Another study that supports this found that 16 percent of all under-20 year old drivers involved in fatal crashes were reported to have been distracted while driving (Ascone & Lindsey, 2009).

Actions that the driver did before the accident are also important to consider. Actions could involve but are not limited to; changing lanes, following another vehicle improperly, obscured vision, drug usage or fatigue. According to a research article, 65% of reported unsafe car driver acts were because of the improper actions of the driver (Kostyniuk & Zakrajsek, 2002). Another action that other research have discussed was whether the driver wearing a seat belt. In a research paper from 1989, the researchers stated that seat belts should be discounted 12% of the time to reflect actual usage since not wearing a seatbelt has legal consequences in some US jurisdictions (Streff & Wagenaar, 1989).

1.1.2 Vehicle Characteristics

There are many vehicle characteristics for a researcher to examine. The main ones for this research paper are: the type of vehicle and whether it had airbag deployment. The type of vehicle is a classic characteristic to look at, because some vehicles are rated as safer than others. According to one research article, it was reported that a large van had a 9.34 total occupant fatality rate per 100,000 registered vehicles. It would seem by this that a compact car would be more dangerous to be in an accident with than in a van (Subramanian, 2006). Airbag deployment is also an important inquiry topic among researchers. One study revealed that the "airbags did not appear to have a protective effect on drivers younger than forty years old and may have been detrimental to drivers older than 60 years old" (Bedard, Guyatt, Stones, & Hirdes, 2002). This means that airbags, according to this study, are best used for people from forty to sixty years of age.

1.1.3 Environmental Characteristics

There are several types of environmental issues to look at when relating fatalities to vehicle collisions. These could include weather, road condition, lighting, time of day, day of week, and season or quarter of the year of the accident. When describing weather, it could include heavy rain, hail, snowstorms, high winds, blowing sand, fog, and other atmospheric effects. Weather effects often influence the driver in multiple ways; visibility, the ability to control the vehicle, and even the ability to hear. Thus, there is a higher possibility of an accident during these times, which means a higher chance for a fatality to occur.

In a study, it was found that about 34% of serious crashes had contributing factors related to the roadway or its environment (Lum & Reagan, 1995). Road conditions like construction, ice, potholes and wetness can also cause more accidents to occur because it is harder to steer the vehicle if the driver is not familiar with the situation, or is not driving safely.

The lighting and time of day are also things to consider when dealing with fatalities in vehicle collisions. In a study, it was found that the per mile fatal crash rate of 16-year-old male drivers is four times greater at night (9pm to 6am) than during the day (Williams, 1985). If we look at day of the week, it would appear that the weekends (Friday through Sunday) continue to have a higher fatality rate than the weekdays (Monday through Thursday) (Cerrelli, 1996). Also, when looking at months of the year February and March tend to have fewer fatalities than the end of the year; with July 3rd-4th and December 23rd-24th having one of the highest fatality rates (NHTSA, Trend and Pattern Analysis of Highway Crash Fatalities by Month and Day, 2005).

1.1.4 Crash Characteristics

There are two crash characteristics to consider: these include direction of accident and how many vehicles were involved. The directions of an accident are head on, side impact, angle and rear-end impact. In one study they found that 65% of all crashes involved front impacts, which represented the largest source of fatalities. Right-sided impacts were found to be the next most frequent with an occurrence of 17.5% (Bedard, Guyatt, Stones, & Hirdes, 2002). Multiple vehicle collisions are an issue heard across the world. They happen more frequently, but they are not the most deadly. In 2006, single vehicle collisions were 2.8 times as likely to result in a fatality as multiple-vehicle collisions (Hunter, 2006)

1.1.5 Zip Code Characteristics

Some studies have reported that areas with higher population densities might have more accidents due to congestion. In this study, we will look at whether or not the population of a zip code has a significant effect on the odds of a fatality. We will also be looking at the median household income of that zip code to see if that has any significance on the odds of a fatality in a vehicle collision. Lastly, we will look at the percent of the zip code that has one of the following: less than a high school education, a high school diploma, two-year degree, four-year degree, or higher than a four year degree and its effect if any on the odds of a fatality.

1.2 The Data

Three data sets were compiled for use in our analysis. First, the crash information and fatalities for five consecutive years was provided by Kim Stalling, a transportation analyst from the Nevada Department of Transportation. Next, the zip codes were located by using Google Maps, from the information on the intersecting streets in the first set of data. Lastly, the zip code data for median income, zip population, and education levels was found using the website http://realestate.aol.com. Each set is described below.

1.2.1 Crash Information

Let us define a few of the terms used throughout this thesis. A responsible driver is defined as the person who caused the accident to occur. They are the individuals that will normally pay for any property damage and medical bills of all the involved parties. Each accident that occurred had a primary responsible driver; the secondary responsible drivers were not accounted for in this data set. Distractions, as earlier stated, are defined as any secondary activity that competes for the driver’s attention while handling a vehicle. These distractions can range from drinking to falling asleep to illness. A fatality in the state of Nevada is defined as any person that dies within 30 days due to a vehicle accident that occurred on a United States public road and the vehicle had an engine.

The crash information was obtained for the years 2008-2009 from the Nevada Department of Transportation. We will only be looking at the 2009 data to create the model. The 2008 data will be used as the test data set during the validation step of this thesis. The predictor variables that were available for each vehicle collision included gender of the responsible driver, age of the responsible driver, distractions of the responsible driver, seatbelt usage of the responsible driver, vehicle type of the responsible driver, airbag deployment, weather, road condition, lighting, time of day, day of week, season or quarter of the year of the accident, direction of impact, number of vehicles involved, and the intersection of the vehicle collision. The response variable is also inside the crash data; it is whether or not there was a fatality in the accident. More information on the crash data will be provided in Chapter 2.

1.2.2 Zip Code Information

The zip code information was found by using the intersecting streets that was provided by the crash information and entering this information into Google Maps. There were 77 unique zip codes of Las Vegas, Nevada. Some of these zip codes are unincorporated, so they may not have any population and may not be represented in our data. The zip codes were then linked up to the crash information using Microsoft SQL server software. After the zip codes were located, we used AOL Real Estate to locate the population of each zip code, the median household income as reported, and the percent of each type of education in each population.

Each of these categories and variables will be discussed further in the following chapters. In the next chapter, we will discuss treatment of the data, methodology, results, validation and discussion, limitations and conclusion.

Chapter 2:

Developing and Analyzing the Data

In this Chapter we will analyze a multitude of explanatory variables. To do this we will discuss the data in general, the response variable, the explanatory variables, and the univariate logistic regression. We will also review our choice of variables selected for the logistic regression and why each was categorized in such a way.

2.1 Database of Raw Data

Before we get into the data, we will discuss how the data was collected and archived. Firstly, the bulk of the data came from Nevada Department of Transportation. They gain this information from two possible sources; reports made by police officers at the scene of the collision and hospitals who file reports for fatalities that occurred because of a traffic accident. One of the inputs that have to be written into the reports is the two streets that intersect and that are closest to the accident. Because Las Vegas is a growing community and is a combination of several different townships, towns, and cities, these intersections often have several different names and in some cases, perpendicular streets do not actually cross. All of this data was imported into the database software package Microsoft SQL Server. This software is used to create and edit large databases, and to perform queries involving these databases. Since the data that was initially given covered over 200,000 accidents, using SQL was the best method to collect, organize, and parse the data to get what was needed. The 200,000 pieces of data involved all Clark County collisions over a five year period. Our interest consisted of the 2009 data which contained 35,000 pieces of data.

The second part of the data are the relevant zip codes for the accidents imported into MSSQL. The USPS website, http://usps.com, provided the majority of the zip code data for the traffic intersections contained in the accident data, though in some cases the intersections could not be found. In these situations, another geolocating website Google Maps had to be used. Using Microsoft SQL Server, the unique intersections were extracted from the accident data into a separate table. These intersections were then fed into the aforementioned geolocating website to determine their zip code. The resulting lists of databases were then joined with the accident data to create a full list of accidents with the zip code they occurred in. These zip codes were important since part of this thesis involves the zip code data. Unfortunately, not all of the 29,000 were able to be located; only a sample of 18,580 cross streets were found to be useful using two separate programs and many hours of manually entering the streets into Google Maps.

The third part of the data is the zip code information involving population and median of household income. This demographic information was found using a website called http://realestate.aol.com. This website stated that there were 77 different zip codes in Las Vegas, Nevada. The 77 zip codes found did not include zip codes for Henderson or North Las Vegas; however, these 77 zip codes did include unincorporated zip codes that were not in use as of 2009. For the purpose of this thesis, we will only be using the forty-five zip codes defined for the year 2009 in Las Vegas, Nevada.

The traffic data was imported into SQL Server, then, using a custom program to geocode the addresses with the geolocating services mentioned above, we created data relationships between the demographics and traffic accidents using the latitude and longitude that were acquired from the geocoding process. The next step was to look at which variables should be the response and predictor variables.

The majority of the data that was provided from the Nevada Department of Transportation was categorical. Categorical data is a form of discrete data that describes some characteristic or attribute. In most of the data, the variables describe several attributes ranging from which intersections the accidents happened in, to whether there was a fatality or not. The raw data had to be reconstructed into actual categorical groups that made more sense for the type of research that was going to occur in this thesis. Using SQL, a filtered view of the data was created which interpreted the word variables provided from Nevada Department of Transportation into numeric variables which more easily would be used in the next step of the process. In the following sections, each variable will be explained in detail, and whether it was modified, and if so, how and why.

2.2 Variable Definitions

Before we can begin the actual fitting of a model, we need to understand what types of independent variables we have. We have independent variables that are dichotomous, such as gender, whether it occurred at dusk, etc. Dichotomous means that there are two categories, which we used SQL to define as 0 or 1. There are some polychotomous independent variables as well, which means that there were more than two categories being described, such as with the predictor Time of Year. We also have ordinal data; meaning that there is a natural order between the categories, such as with age. Finally, we have continuous independent variables where the observations fall anywhere in a continuum, such as the percent of a zip code with a two year degree. The reason we need to understand each type of explanatory variable we are using is that when it comes to placing them into the software, there are adjustments that need to be made for the model, as well as for a goodness of fit test, which will be discussed further in later chapters. These are important to understand once we get to the interpretation of the results of the fitted final model, which will be discussed in Chapter 5.

2.3 Dummy Variables

The polychotomous predictors having more than two categories will need to be separated into dummy variables to be best represented in the model. We let B be an n x k dummy variable matrix, where $B_{ij} = 1$ if case i falls in class j and zero otherwise. The coding is determined by a contrast matrix C (see Table 2.1) which has the dimensions k x (k-1). The contribution of the factor to the model matrix X is then given by BC (Faraway, 2006). Consider the quarter of the month predictor variable, which is a 4-level factor. The contrast matrix C that describes this coding, where columns represent the dummy variables and the rows represent the levels, is:

[redacted]

This treats level one (1st quarter of month) as the standard level to which all other levels are compared. Each parameter for the dummy variable then represents the difference between the given level and the first level. In our software, the default choice is called treatment coding, which is what is explained above (Faraway, 2006).

2.4 Univariate Statistics

2.4.1 The Simple Logistic Model

Let us begin by discussing the distribution of the response variable. The response variable is binary, taking on the values of 1 and 0 with probabilities of π = P(Y=1) and 1-π = P(Y=0), thus Y ~ Bernoulli(π). If Y_i is the response and X_i is the predictor of the ith case, the logistic regression model I given by

[redacted]

Maximum Likelihood Estimation of β_o and β_1

We know that the response variable follows a Bernoulli distribution. The probability distribution is as follows:

[redacted]

Since the Y_i's are independent, their joint probability function is:

[redacted]

Now, we take the natural log of both sides of the equation; which is called the log-likelihood.

[redacted]

It was stated earlier that $\pi_i = \frac{e^{\beta_o + \beta_1 X_i}}{1+e^{\beta_o + \beta_1 X_i}}$, so it follows that ln $\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_o + \beta_1 X_i$. We can now substitute this information into our log-likelihood function, to get:

[redacted]

where $\ell(\beta_o, \beta_1)$ replaces $g(Y_1, \dots, Y_n)$ to showcase that we are now viewing this function as the likelihood function of the parameters to be estimated, given the sample observations. To find the maximum likelihood estimates, we will need to differentiate with respect to β_o and β_1. These differentiations are given:

[redacted]

where $\pi_i = E[Y_i]$. Now we set these equations to 0 to get:

[redacted]

Which are than solved to get estimates of β_0 and β_1. At this point a computer-intensive numerical search procedure is used to find the actual maximum likelihood estimates b_0 and b_1 of β_0 and β_1 respectively. For more information on this procedure please refer to pages 105-106 of McCulloch, Searle and Neuhaus (2008). After these values are computed, we substitute these values into the response function $\frac{e^{\beta_o + \beta_1 X_i}}{1+e^{\beta_o + \beta_1 X_i}}$ to obtain the fitted response function; which is denoted as:

$$\hat{\pi} = \frac{e^{b_o + b_1 X_i}}{1 + e^{b_o + b_1 X_i}} \, . \tag{2.15}$$

Once the fitted logistic response function has been obtained, we examine the appropriateness of the fitted response function and make predictions and inferences about it. In the following section we will look at several univariate logistic regressions for the continuous explanatory variables that will be used in this thesis.

2.4.2 Deviance Test

If the provided Y is really a binomial and that the n_i are relatively large, the deviance test is approximately a chi-squared distribution with n-s degrees of freedom if the model is correct (Faraway, 2006). Thus we can use the deviance test to examine whether the model is an adequate fit. Deviance is explained by the difference between the observed values (Y) and the expected values ($\hat{Y}$). The greater this difference is the poorer the fit is. The desire is to have a small deviance. As we add more variables to the equation the deviance should get smaller, indicating an improvement in the fit. The deviance test statistic can be found by:

$$\mathcal{X}^2 = \mathrm{D(for\ the\ model\ without\ the\ variable)} - \mathrm{D(for\ the\ model\ with\ the\ variable)}.$$

If the test statistic is smaller than the critical region, we would conclude the null hypothesis; which states that the addition of the variable is a good fit. If the test statistic is larger than the critical region, we would conclude the alternative hypothesis.

2.4.3 Hypothesis Testing

Hypothesis testing produces a decision about any observed difference; either the difference is “statistically significant” or it is “not statistically significant.” In this chapter we will talk about testing a single β_k, using the Wald test. This tests a single regression parameter to see if it is equal to zero, positive or negative. The following is the hypothesis statement we will be using in the next section:

$$H_0 : \beta_k = 0$$
$$H_a : \beta_k \neq 0$$

The test statistic is [redacted] decision rule for this test is: If $|Z^*| > Z(1 - \alpha/2)$, we would reject the null hypothesis at level of significance α (Kutner, Nachtsheim, & Neter, 2004).

2.4.4 The Continuous Explanatory Variables

There are five explanatory variables that could be used to find the final model for this thesis. These variables are: percentage of zip code population that had less than a high school education, only high school education, two-year degree, four year degree, and graduate or doctorate degree. We found the estimates for each of the univariate logistic regressions using R software (R Development Core Team, 2010). Table 2.2, shows both estimates and whether or not β_1 was significant.

[redacted]

As an example, when we look at one of the printouts, we see:

[redacted]

The Wald Stat = $\frac{-0.05966}{0.01644} = -3.629$. The critical region is 1.645, so the absolute value of the test statistic is larger than the critical region; which means that the variable x51 is significant, and does not equal zero. Therefore it should be in the model. If we look at its 95% confidence interval: -0.05966$\pm 1.96 * 0.01644$; it does not include the “no effect” value. Also, the confidence interval is fairly narrow, which means that we have a large sample and a very precise estimate for the true effects.

After looking at the rest of the continuous predictor variables, it was evident that each was significant and made for a better model. Thus each will be added to the model discussed in Chapter 3. In Chapter 3, we will discuss what a multiple logistic regression model is, the full model for the data, model selection and whether the model is a good fit.

2.4.5 The Dichotomous Explanatory Variables

There are forty-two explanatory variables that are referred to as dichotomous. These variables’ results can be found in Appendix A, Section 2 and they are defined in

Appendix D. We performed a univariate logistic regression and the variables that were found to be not significant were: Wet Weather, Cloudy Weather, Road Construction, Road Obstruction, Road Environment, Inappropriate Lane Change, Going the Wrong Way, Passing Other Vehicle, Disregarded Road Signs, Failed to Yield, Hit and Run, Obstructed Visibility, and Median Income. The remaining dichotomous explanatory variables will be used in the main effects model explained in Chapter 3. In Appendix A, Section 2, we have listed the coefficient value, the Wald test, the p-value, the odds-ratio, the 95% confidence interval and the 95% odds ratio confidence interval of each dichotomous explanatory variable.

2.4.6 The Polychotomous Explanatory Variables

There are five explanatory variables that are referred to as polychotomous explanatory variables. These variables' results can be found in Appendix A, Section 3 and they are defined in Appendix D. We performed a univariate logistic model and the variables that were found to be not significant were: Age, Quarter of the Year and Zip Code Population. The two variables that were found to be significant were: The Quarter of the Month and Vehicle Type.

Once we had constructed the models, we then did a deviance test on each. We will look at the printout for Age, perform a deviance test on it and state the results, the remaining polychotomous results can be found in Appendix B. The variable Age's printout is found in Table 2.4.

[redacted]

The deviance for the full model is 1523 and the deviance for the null model is 1529.4. The hypothesis for this test is: the null hypothesis states that the model is a good fit and the alternative hypothesis states that the model is not a good fit. Therefore the deviance test statistic is 6.4, with a p-value of 0.09369079. The critical region is 11.07. The conclusion is that the logistic model is not a good fit with this variable in it.

The significant polychotomous explanatory variables will be used in the main effects model explained in Chapter 3. In Appendix A, Section 3, we have listed the coefficient value, the Wald test, the p-value, the odds-ratio, the 95% confidence interval and the 95% odds ratio confidence interval of each polychotomous explanatory variable.

In Chapter 3, we will discuss the methodology in which we took to construct a final model. All the variables that were found to be significant in this chapter will be used in the main effects model.

Chapter 3

Methodology

In this Chapter we will be looking at the generalized linear model and the use of a specific type of generalized linear model; multiple logistic regression. We will be applying this model to the transportation data and briefly discussing the modeling process from the full model to the final model selection.

3.1 Overview of the Generalized Linear Model

In constructing a generalized linear model, there are three decisions that need to be made. These are:

1. What is the distribution of the data?
2. What function of the mean will be modeled as linear in the predictors?
3. What will be the predictors?

In the case of a generalized linear model, y is assumed to consist of independent measurements from a distribution with density from the exponential family (McCulloch, Searle, & Neuhaus, 2008). The mean of y_i and the linear form of the predictors need to be connected by some function g; we call this function a link function. The decision on which predictors to use and possibly how to transform them, needs to be considered before and during the development of the final model.

Let us talk theory, in the discussion of a generalized linear model; we will use the following matrices to simplify formulas:

[redacted]

Under the General Linear Model (GLM) the mean of y is a linear function of the predictor variables, X, and the to-be estimated model parameters β, so that $\mathrm{E[y]} = \mathrm{X'\beta}$ or in other words: $\mathrm{E[y_i]} = \mu_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{i,k}$. Now, the Generalized Linear Model provides a way to estimate a monotonic function, g, of the mean response as a linear function of the values of the predictor variables, X. This can be written as $g(\mathrm{E[y_i]}) = \mathrm{g}(\mu_i) = \mathrm{g}\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{i,k}\right) = \eta_i$; where $\mathrm{g}(\mu_i)$ is called the link function. The variance of Y may be a function of the mean response μ: $\mathrm{var(Y)} = \lambda \mathrm{V}(\mu)$, where λ represents a constant (McCullagh and Nelder, 1998). In the linear model the $\mu_i = \eta_i$; however in the generalized linear model $\mu_i = \mathrm{g}^{-1}(\eta_i)$.

3.2 Overview of the Logistic Regression

Logistic regression is useful when the outcome is binary, meaning zero or one, with one being a success. A researcher may wish to study the relationship between whether there was a fatality to the age, gender, etc. of the responsible driver. The logistic regression is a specialized case of the generalized linear model. To use the logistic regression, we use the binomial family; which uses the logit link function defined: $\mathrm{g(p)} = \mathrm{logit(p)} = \log\frac{\mathrm{p}}{1-\mathrm{p}}$ and variance function defined by: $\mathrm{var(Y)} = \lambda \mathrm{p}(1-\mathrm{p})$, where p is the probability of success and $\lambda = 1$.

3.3 Fitting a Model

When fitting a model, we utilized the method of maximum likelihood to estimate the parameters of the multiple logistic response function: $\mathrm{E}(\mathrm{Y_i}) = \pi_\mathrm{i} = \frac{\exp(\mathrm{X_i'}\beta)}{1+\exp(\mathrm{X_i'}\beta)}$; where X variables are known constants (Kutner, Nachtsheim, & Neter, 2004). This method yields values for the vector of unknown parameters β, which maximizes the probability of obtaining the observed set of data. The log-likelihood function for simple logistic regression, which was described in Chapter 2, can extend directly into the multiple logistic regression: $\ln \mathrm{L}(\beta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Y_i}(\mathrm{X_i'}\beta) - \sum_{\mathrm{i}=1}^{\mathrm{n}} \ln[1 + \exp(\mathrm{X_i'}\beta)]$. We will be using the R software to find the values of estimates of parameters that maximizes $\ln \mathrm{L}(\beta)$. These maximum likelihood estimates of β will be denoted as b:

$$\underset{px1}{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}. \tag{3.2}$$

The fitted logistic response function and values can be expressed as follows:

$$\hat{\pi} = \frac{\exp(\mathrm{X_i'b})}{1+\exp(\mathrm{X_i'b})} = [1 + \exp(-\mathrm{X_i'b})]^{-1} \tag{3.3}$$

where $\mathrm{X_i'b} = \mathrm{b_0} + \mathrm{b_1X_{i,1}} + \cdots + \mathrm{b_{p-1}X_{i,p-1}}$.

3.3.1 Likelihood Ratio Test

Once we have the full model defined, then we have to select which independent variables should be kept in the model. There are a few ways to decide on this, the methods used in this thesis will be first based on the results of the contingency tables in Chapter 2 and then a step-wise method. The results from the univariate logistic regression explained in Chapter 2, will be used as the second model. This second model will need to be tested to see if the variables that were dropped to create it, should be dropped. This is done by a method called Likelihood Ratio Test. We begin with the full logistic model with response function

$$\pi = [1 + \exp(-X'\beta_F)]^{-1}, \tag{3.4}$$

where

$$\beta_F = \beta_0, \beta_1, \cdots, \beta_{p-1},$$

$$X'\beta_F = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}. \tag{3.5}$$

The reduced logistic model has the response function:

$$\pi = [1 + \exp(-X'\beta_R)]^{-1} \tag{3.6}$$

where

$$\beta_R = \beta_0, \beta_1, \cdots, \beta_{q-1},$$

$$X'\beta_R = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1} X_{q-1}. \tag{3.7}$$

where $q < p$.

Now, we find the maximum likelihood estimates for the both models and evaluate their likelihood functions as explained earlier. Let it be known that p and q are the parameters for the two models.

The hypothesis that will be tested is:

$$H_0: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$$

H_a: not all of the β_k in H_0 equal zero.

As with all hypothesis testing, we need a test statistic for the likelihood ratio test, which is denoted as G^2 (Kutner, Nachtsheim, & Neter, 2004).

$$G^2 = -2\ln\left[\frac{L(R)}{L(F)}\right] = -2[\ln\ L(R) - \ln\ L(F)] \tag{3.8}$$

Large-sample theory states that when n is large, G^2 is distributed approximately as $\mathcal{X}^2_{p-q}$ when H_0 in the hypothesis holds. The degrees of freedom correspond to df_R-df_F = (n-q)-(n-p)=p-q. The appropriate decision rule is:

If $G^2 > \mathcal{X}^2_{1-\alpha,p-q}$, reject H_0 and conclude that the full model is valid.

Once we know if the insignificant variables found in Chapter 2 can be removed, we can begin the step-wise model selection.

3.3.2 Stepwise Model Selection

When the pool of potential X variables contain more than 30 variables, use of a “best” subsets algorithm may not be feasible. Instead, we can use an automatic search procedure which will develop the “best” subset of X variables. This automatic search procedure is known as a stepwise procedure, a useful and effective data analysis tool. Any stepwise procedure for selection or deletion of variables from a model is based on a statistical algorithm that checks for the “importance” of variables, and either includes or excludes them on the basis of a fixed decision rule. Two other criterion are AIC_p, and SBC_p. The stepwise regression procedure is a combination of backward elimination and forward selection.

The backward elimination procedure begins with a model that contains all possible independent variables and identifies the X with the largest p-value. If the maximum p-value is greater than a predetermined limit (such as .05) then it is dropped from the model. The model is refit without this variable and the procedure repeats. This continues until model has the lowest AIC possible. The stepwise procedure eliminates predictors in the pursuit of getting a lower AIC. If after removing a predictor the AIC goes up, this may mean that the variable should not be removed.

The forward selection procedure adds variables to the model one at a time. At each step, each variable that is not already in the model is tested for inclusion in the model. The most significant of these variables is added to the model, so long as it's P-value is below some pre-set level.

The Akaike Information Criterion (AIC) is another method to help with the variable selection process.

$$AIC_p = -2LL + 2p, \tag{3.9}$$

where LL is the maximum log-likelihood and p is the number of parameters in the model, including the constant. The second term is sometimes referred to as the penalty term, and adjusts to the size and complexity of the model. When the number of parameters increase, the first term becomes smaller. This is due to the fact that the more parameters there are, the more the chance of what is observed can happen. So, to adjust for this bias, AIC adds the term 2p to -2LL as a penalty for increasing the number of parameters (Hilbe, 2009). A small value of AIC_p means that the model is a better fit; however, this does not mean that the model is a good fit or a perfect model.

The Schwartz' Bayesian criterion (SBC_p) is a third method to help with the selection process.

$$SBC_p = -2LL+p\log(n), \tag{3.10}$$

where LL is the maximum likelihood of the model, p is the number of parameters in the model, and n is the number of observations in the data set. The second term is the penalty term and does the same as the AIC_p's penalty term. When the number of parameters and data increase the -2LL decreases. The smaller the SBC_p the better the model is for the selection process.

3.3.3 Goodness of Fit Test

Once this new model is selected, we want to be able to do a Goodness of Fit Test. This can be done using several methods, some of these are the Pearson Chi-Square, Deviance and Hosmer-Lemeshow. The first two tests require sufficient replication with the subpopulations for the tests to be valid for a goodness of fit. In other words, these two are only appropriate when there are repeated observations and when the number of replicates at each X is sufficiently large (Kutner, Nachtsheim, & Neter, 2004).

Hosmer and Lemeshow Test

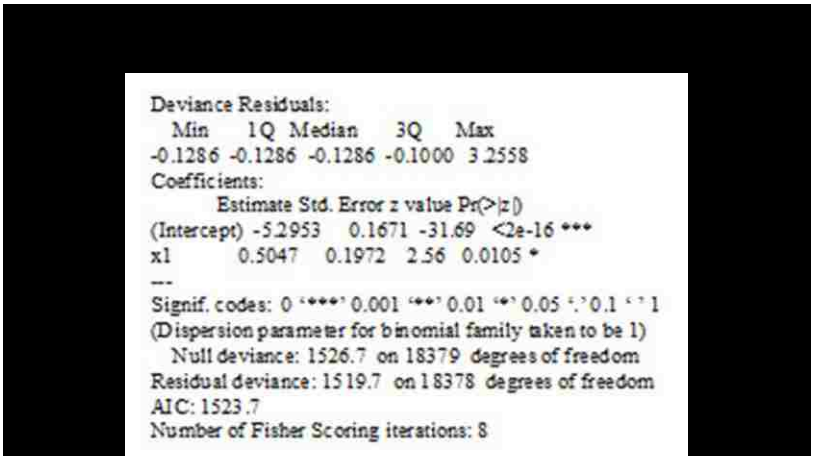
The likelihood ratio statistic mentioned earlier is an appropriate test for comparing two models to each other. If instead of just a full model and a reduced model, we use a saturated larger model where there are as many parameters as there are cases and the fitted values $\hat{\pi}_i = {}^{y_i}/_{n_i}$, then the statistic becomes:

$$D = 2\sum_{i=1}^{n}\left\{y_i \log\frac{y_i}{\hat{y}_i} + \left((n_i - y_i)\log\frac{(n_i - y_i)}{(n_i - \hat{y}_i)}\right)\right\}, \qquad (3.11)$$

where $\hat{y}_i$ are the fitted values from the smaller model (Faraway,2006). Now, since the saturated model fits as well as any model can fit, the deviance D measures how close the smaller model comes to perfection. If Y is binomial and the n_i are relatively large, the deviance is approximately chi-squared with n-s degrees of freedom. Let us look at an example from the data for this thesis: the predictor variable Gender:

[redacted]

Notice, how $n_i = 36$ when the gender is a female and there was a fatality. The thirty-six means that it is a repeated observation. In this situation we are allowed to use the first two tests, since there is repetition or $n_i > 1$. If we setup a model for this situation, we find the following information:



```
Deviance Residuals:
   Min       1Q  Median      3Q      Max
-0.1286  -0.1286  -0.1286  -0.1000  3.2558
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.2953     0.1671  -31.69  <2e-16 ***
x1           0.5047     0.1972    2.56  0.0105 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 1526.7  on 18379  degrees of freedom
Residual deviance: 1519.7  on 18378  degrees of freedom
AIC: 1523.7
Number of Fisher Scoring iterations: 8
```

So, we find that the deviance is 1519.7 with degrees of freedom of 18378 and the null deviance is 1526.7 with degrees of freedom equal to 18379. When we complete the deviance test we get a p-value of 1, which would make us think that the model fits well. When we compare the Null model to that of the full model, we receive a p-value of 0.008, which caused us to conclude that the addition of gender is statistically significant, so we would reject the smaller model (Null).

However, for the predictor, Wrong Way,

[redacted]

notice how in the cell (Wrong Way=1, Fatalities) the $n_i = 1$, this means there is no repetition of the observation of this event occurring in the 2009 data. When $n_i \leq 1$ where $y_i = 0$ or 1, the response is binary, and the deviance reduces to (Faraway, 2006)

$$D = -2\sum_{i=1}^{n}\left\{\hat{\pi}_i \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) + \log(1-\hat{\pi}_i)\right\}. \tag{3.12}$$

For a deviance test to measure fit, it has to be able to compare the $\hat{y}_i$ to the data y_i; however in this equation there is no data, so this deviance statistic no longer follows a chi-squared distribution, and it is no longer a good fit test (Faraway, 2006).

Continuous predictors will cause the data to be too sparse, which means that the covariates are too small, as can be seen with the predictor for Less than a High School education; so we cannot use the first two mentioned methods for a goodness of fit test, instead we would use the Hosmer-Lemeshow test.

The Hosmer-Lemeshow is used when either there are few or no replicated data sets (Hosmer & Lemeshow, 2000). This test is only for binary response models. The following is the Hosmer-Lemeshow statistic:

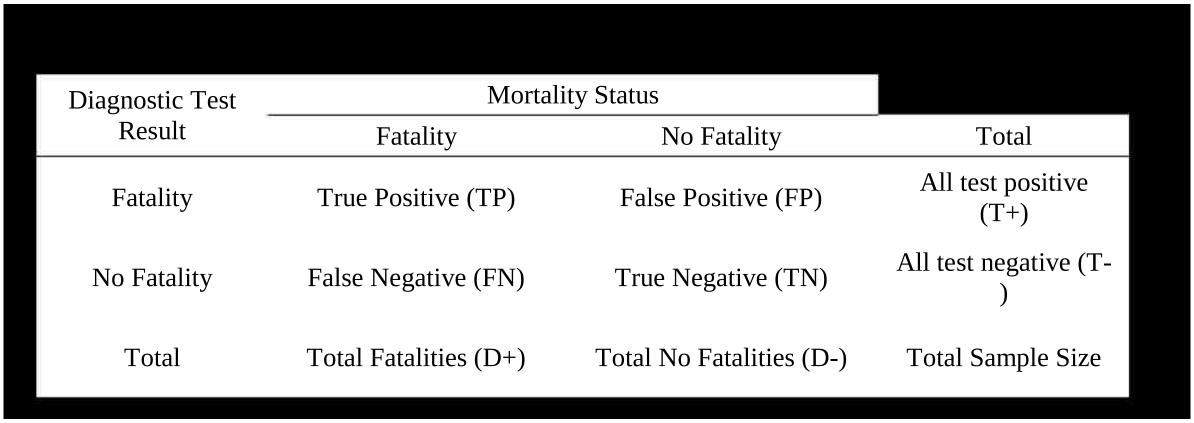
$$H_L = \sum_{g=1}^{G}\frac{\left(O_g - N_g\hat{\pi}_g\right)^2}{N_g\hat{\pi}_g\left(1-\hat{\pi}_g\right)}, \tag{3.13}$$

where N_g = the total frequency of subjects in the gth group, O_g= the total frequency of event outcomes in the gth group, $\hat{\pi}_g$ is the average estimated probability of an event outcome for the gth group and H_L= the approximate chi-square with G-2 degrees of freedom, where G is the number of groups data is split into (Hosmer & Lemeshow, 2000). The smaller values of H_L (and larger p-values) indicates a good fit of the model (Hosmer & Lemeshow, 2000).

This procedure consists of grouping the data into classes with similar fitted values $\hat{\pi}_i$, with approximately the same number of cases in each group (Kutner, Nachtsheim, & Neter, 2004). The number of groups largely depends on the datasets, ten groups are recommended by Hosmer and Lemeshow for large datasets; which is what we have in this thesis.

ROC Curve

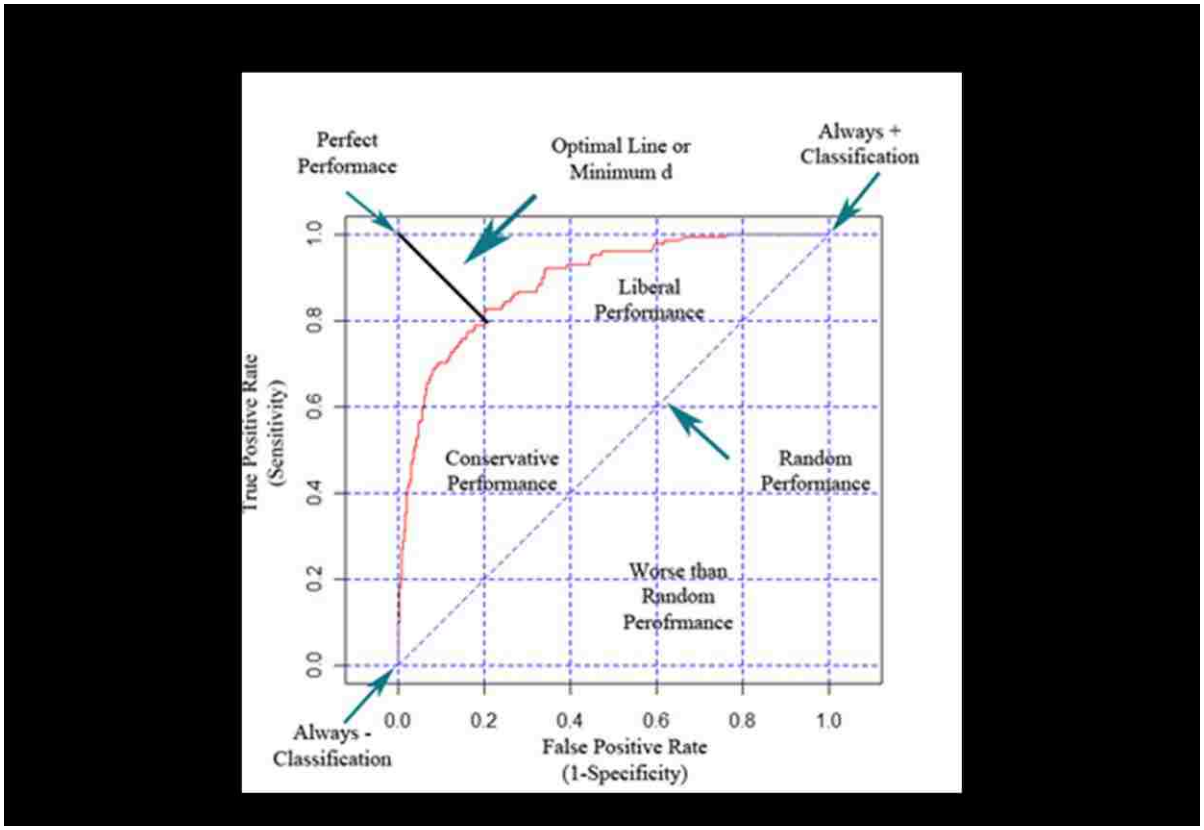
Another way to see if this model is a good fit, is by using the Receiver Operating Characteristic curve (ROC curve). The ROC curve is a plot of the true positive rate against the false positive rate for the different possible cutoffs of a diagnostic test; in other words it depicts the performance and performance trade-off of a classification model.



Diagnostic Test Result	Mortality Status		
	Fatality	No Fatality	Total
Fatality	True Positive (TP)	False Positive (FP)	All test positive (T+)
No Fatality	False Negative (FN)	True Negative (TN)	All test negative (T-)
Total	Total Fatalities (D+)	Total No Fatalities (D-)	Total Sample Size

Table 3.3 shows the results for a diagnostic test. Each cell has an importance to help find several things, some of these include true positive rate, true negative rate, and false positive rate. The true positive rate is also known as sensitivity and it is found by $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$. The true negative rate is known as specificity and is found by $\frac{\mathrm{TN}}{\mathrm{TN+FP}}$. The false positive rate is found by one minus specificity. These are used as the x and y axis of the ROC

curve. They are also used to discover the cutoff points during the prediction phase spoken about in Chapter 5. In order to understand the strength of using this curve we need to understand how it is created and what it may look like. Figure 3.1 represents the ROC curve regions that will be discussed in this section.



Some of the regions of interest are identified in Figure 3.1. The diagonal line from (0,0) to (1,1), known as the random performance line, is the case when there are as many false positive responses as true positive responses. This line has an area under the curve (AUC) of 0.5. If the ROC = 0.5, there is no discrimination, which means it does not showcase a good fit, or the likelihood for some event to occur is based on the flip of a fair coin. If the AUC is between .7 and .9, then this is considered good discrimination and we can confirm a model is well-fit (Hosmer & Lemeshow, 2000). The area under the curve is a useful summary measure of the model's predictive power. Now, to the left bottom of the random performance line we have the conservative performance region; this is where few false positive errors occur. To the right top of the line, we have the liberal performance region; which is where a substantial number of false positive errors occur. The point in the top left corner denotes a perfect classification, which has 100% true positive rate and 0% false positive rate. In Chapter 5 we will discuss how the ROC curve can help with cutoff points for prediction error rates.

3.4 Model Selection for Data

Hosmer and Lemeshow suggest that any variable not selected in the original multivariable model be added back into the model. This is helpful because it will identify variables that, by themselves, are not significantly related to the outcome but make an important contribution in the presence of other variables (Hosmer & Lemeshow, 2000). This model is known as the preliminary main effects model. Finally, we can begin to look for our final model. Let us begin by fitting the full model with all variables added in. The full printout of this model can be found in Appendix G, Section 1.

[redacted]

Appendix D has the full description of each variable in this model. After using the software package R (R Development Core Team, 2010), we were able to find the log-likelihood for the model, which was found to be -559.0268 with 62 degrees of freedom. After looking at the preliminary main effects model, we saw that the variables that were not significant during the univariate logistic regression were also not significant in this model.

Next, we found the model that represented the findings discussed in the univariate logistic regression portion of Chapter 2 was formed. This model is called the main effects model. The full printout of this model can be found in Appendix G, Section 2.

[redacted]

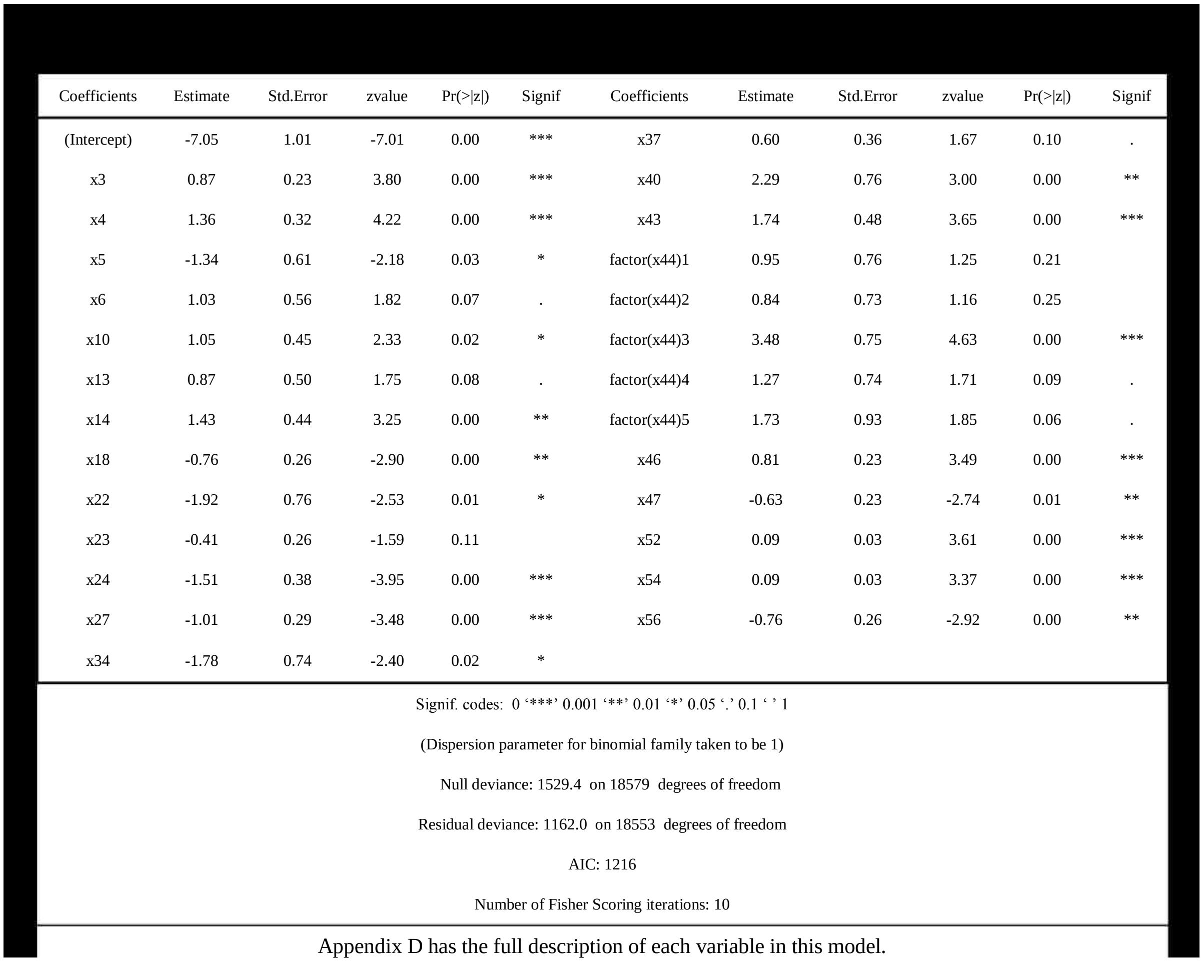
3.4.1 Comparing Log-likelihood Test

This reduced model had a log-likelihood of -575.8864 with degrees of freedom of 40. Now, we can use the likelihood ratio test to see if the variables removed are in fact equal to zero, thus not adding sufficient information to the model. The results are as follows:

$$\mathrm{G}^2 = -2[-575.8864 - (\,-559.0268)] = \; 33.71929$$

For $\alpha = .05$, we have $\mathcal{X}_{22}^{2} = 33.92$. Since $G^2 = 33.72 \leq 33.92$, we conclude that age (x2), wet weather (x7), cloudy weather (x9), road construction (x15) , road obstruction (x16), road environment (x17), quarter of the year (x20), inappropriate lane (x26), driving the wrong way (x30), passing other vehicle (x31), backing up (x32), disregarding road signs (x33), not yielding (x35), hit and run (x39), obstructed visibility (x42), median income (x48), and zip code population (x49) should be dropped from this model. The p-value of this test was 0.052.

Now, we can use the backward elimination process to discover the subset of predictors that would be best in the model. The software R (R Development Core Team, 2010), generated the following model:



Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)	Signif	Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)	Signif
(Intercept)	-7.05	1.01	-7.01	0.00	***	x37	0.60	0.36	1.67	0.10	.
x3	0.87	0.23	3.80	0.00	***	x40	2.29	0.76	3.00	0.00	**
x4	1.36	0.32	4.22	0.00	***	x43	1.74	0.48	3.65	0.00	***
x5	-1.34	0.61	-2.18	0.03	*	factor(x44)1	0.95	0.76	1.25	0.21	
x6	1.03	0.56	1.82	0.07	.	factor(x44)2	0.84	0.73	1.16	0.25	
x10	1.05	0.45	2.33	0.02	*	factor(x44)3	3.48	0.75	4.63	0.00	***
x13	0.87	0.50	1.75	0.08	.	factor(x44)4	1.27	0.74	1.71	0.09	.
x14	1.43	0.44	3.25	0.00	**	factor(x44)5	1.73	0.93	1.85	0.06	.
x18	-0.76	0.26	-2.90	0.00	**	x46	0.81	0.23	3.49	0.00	***
x22	-1.92	0.76	-2.53	0.01	*	x47	-0.63	0.23	-2.74	0.01	**
x23	-0.41	0.26	-1.59	0.11		x52	0.09	0.03	3.61	0.00	***
x24	-1.51	0.38	-3.95	0.00	***	x54	0.09	0.03	3.37	0.00	***
x27	-1.01	0.29	-3.48	0.00	***	x56	-0.76	0.26	-2.92	0.00	**
x34	-1.78	0.74	-2.40	0.02	*						

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1529.4 on 18579 degrees of freedom

Residual deviance: 1162.0 on 18553 degrees of freedom

AIC: 1216

Number of Fisher Scoring iterations: 10

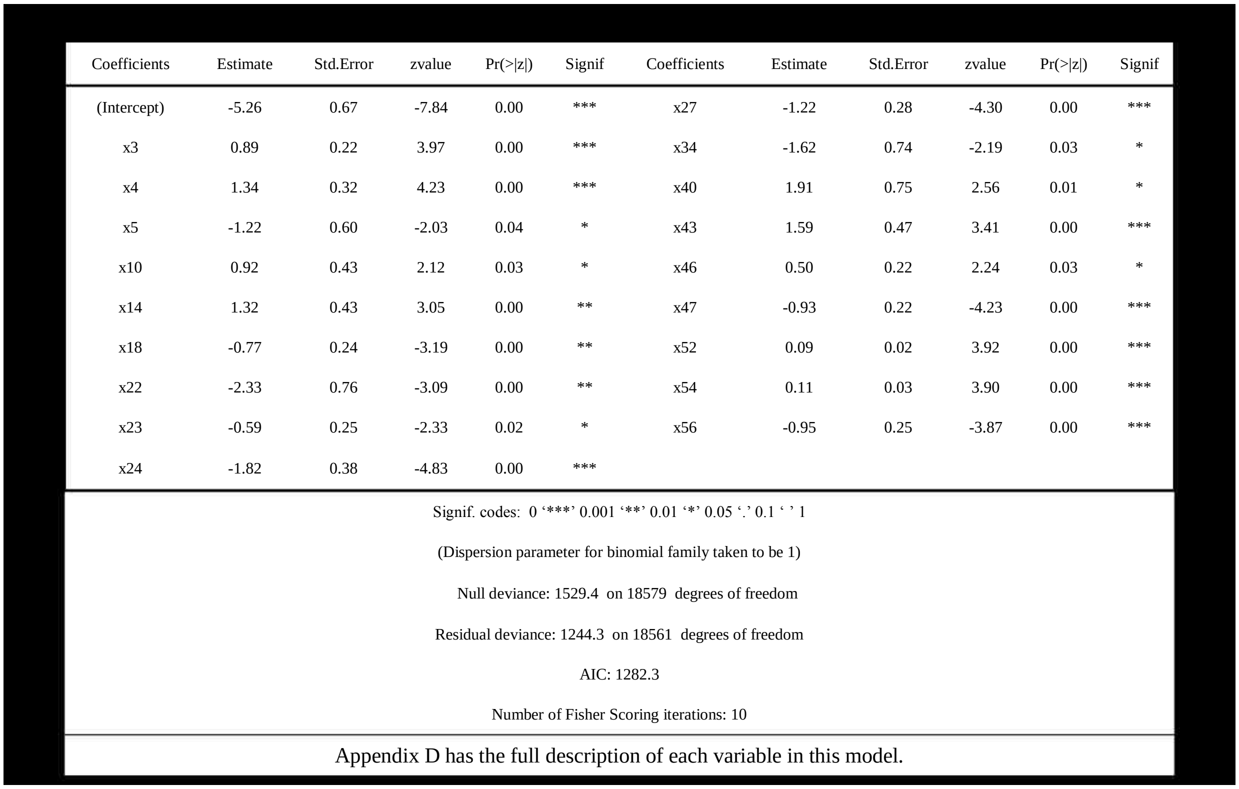
Appendix D has the full description of each variable in this model.

We will test this new model to see if this model is better than the previous one. We will use the log-likelihood test once again to check this. The log-likelihood for this model is -581.0053 with 27 degrees of freedom.

$$G^2 = -2 * [-581.0053 - (-575.8864)] = 10.24$$

For $\alpha = .05$, we have $\mathcal{X}^2_{13,0.95} = 22.36$. Since our test statistic is $10.24 \leq 22.36$, we do not reject the null hypothesis, and conclude that the predictors that were dropped should have been dropped. The p-value for this test is approximately 0.6743813.

This model still shows a few predictors that are not significant, so we decided to continue with the removal of each variable that was not significant, starting with the highest p-value. After this process was finished, we acquired a model with all significant predictor variables; we will call this model AS. The following printout showcases this model.



Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)	Signif	Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)	Signif
(Intercept)	-5.26	0.67	-7.84	0.00	***	x27	-1.22	0.28	-4.30	0.00	***
x3	0.89	0.22	3.97	0.00	***	x34	-1.62	0.74	-2.19	0.03	*
x4	1.34	0.32	4.23	0.00	***	x40	1.91	0.75	2.56	0.01	*
x5	-1.22	0.60	-2.03	0.04	*	x43	1.59	0.47	3.41	0.00	***
x10	0.92	0.43	2.12	0.03	*	x46	0.50	0.22	2.24	0.03	*
x14	1.32	0.43	3.05	0.00	**	x47	-0.93	0.22	-4.23	0.00	***
x18	-0.77	0.24	-3.19	0.00	**	x52	0.09	0.02	3.92	0.00	***
x22	-2.33	0.76	-3.09	0.00	**	x54	0.11	0.03	3.90	0.00	***
x23	-0.59	0.25	-2.33	0.02	*	x56	-0.95	0.25	-3.87	0.00	***
x24	-1.82	0.38	-4.83	0.00	***						

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1529.4 on 18579 degrees of freedom

Residual deviance: 1244.3 on 18561 degrees of freedom

AIC: 1282.3

Number of Fisher Scoring iterations: 10

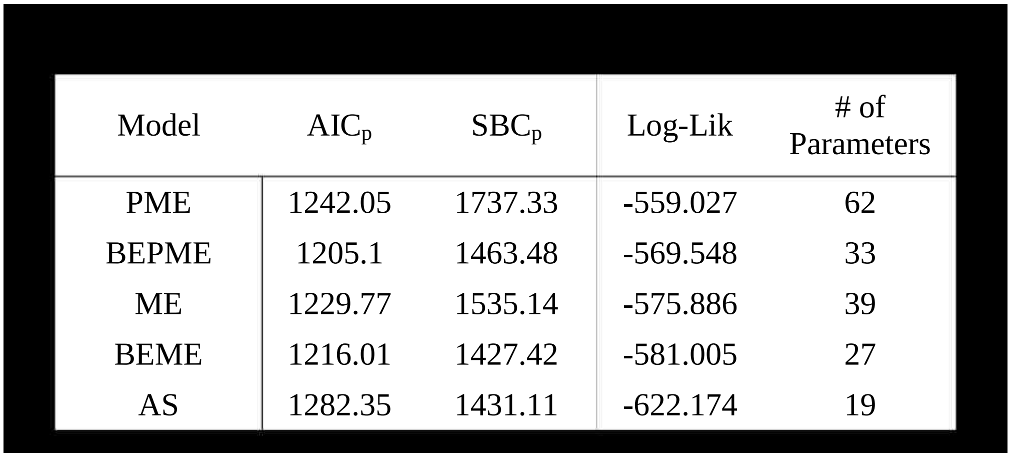
Appendix D has the full description of each variable in this model.

Since this model is based off the BE of Main Effects model, we will check the log-likelihood test to see if the removed predictors should have been indeed removed. The log-likelihood of this model is -622.1736 with 19 degrees of freedom.

$$\mathrm{G}^2 = -2 * [-622.1736 - (-581.0053)] = 82.34$$

For $\alpha = .05$, we have $\mathcal{X}^2_{8,0.95} = 15.51$. Since our test statistic is $82.34 > 15.51$, we conclude that we would reject the null hypothesis. Therefore, at least one of the predictor variables should not have been removed from the model. Even though all the variables are significant, it does not make it a model that should occur. The p-value for this test is 1.654232 * 10^{-14}.

The following table showcases the preliminary main effects model (PME), a backward elimination on the preliminary main effects model (BEPME), the main effects model (ME), the backward elimination of the main effects model (BEME), and a model with all significant predictor variables (AS). Each model had its AIC_p, SBC_p, log-likelihood and number of parameters recorded. This helped to select the final model.



Model	AIC_p	SBC_p	Log-Lik	# of Parameters
PME	1242.05	1737.33	-559.027	62
BEPME	1205.1	1463.48	-569.548	33
ME	1229.77	1535.14	-575.886	39
BEME	1216.01	1427.42	-581.005	27
AS	1282.35	1431.11	-622.174	19

After examining each model type, it would seem that the two backward elimination models have the lowest of both criteria. Although, BEPME may have the lowest AIC it is only a few points away from the BEME model; whereas the SBC of the BEME seems quite a bit smaller than the BEPME. In Chapter 2, we discovered several predictors that were not significant, since the BEPME still incorporates some of these non-significant predictors; we want to use the BEME. Let it be stated that although all the predictors in the fifth model are significant, it's AIC and SBC are quite large. Its AIC is even larger than the PME model's; this is a good indicator that it would not be an ideal model to choose. Thus it does not seem like a good fit for the final model. The printout for the BEPME model can be found in Appendix G, section 3.

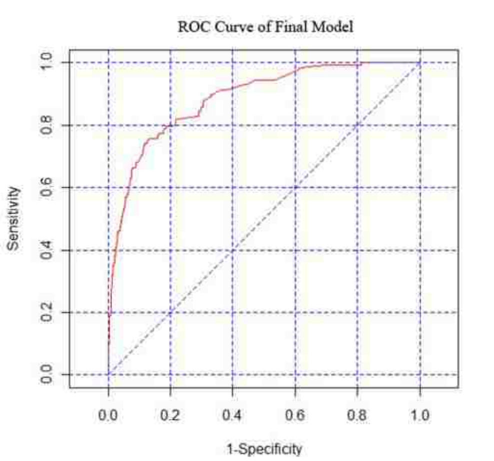
3.4.2 Goodness of Fit Test and ROC Curve

Now, we can test the BEPME model to see if it is a good fit. The test statistic would be calculated as follows:

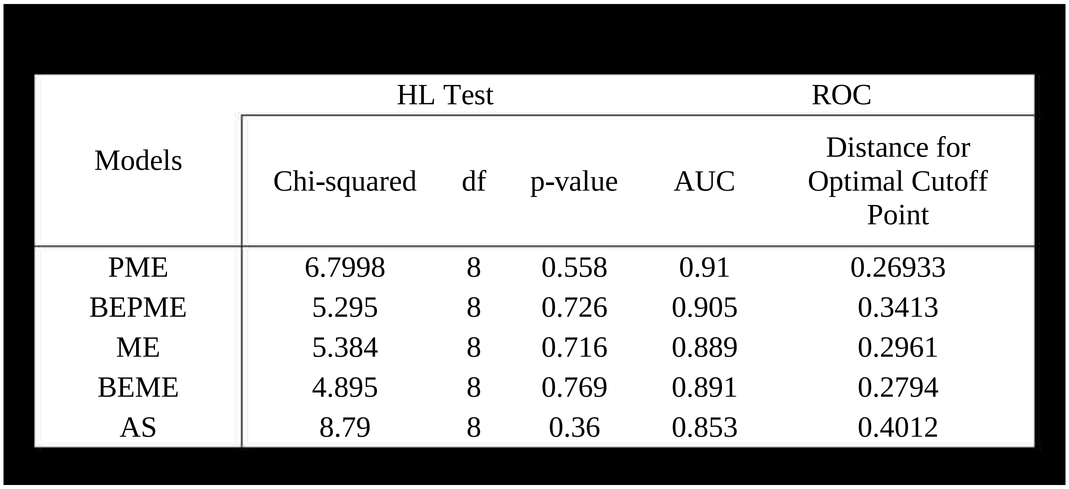
$$\mathcal{X}^2 = \frac{(1888 - 1887.60)^2}{1887.60} + \frac{(0 - 0.398)^2}{0.398} + \cdots + \frac{(1768 - 1775.18)^2}{1775.18} + \frac{(90 - 82.82)^2}{82.82} = 4.895$$

For $\alpha = .05$, we have $\mathcal{X}^2_{8,0.95} = 15.507$. Since $\mathcal{X}^2 = 4.895 \leq 15.507$, we do not reject H_0, and conclude that the logistic response function is appropriate. The p-value was 0.769. Therefore, the Table 3.6 is the final model. To see all the above values for the H-L test please see Appendix F.

If we look at the ROC curve, we get Figure 3.2, which has an AUC = .891. This means that the model is well fit. We will talk more about this figure and how it relates to predictions in Chapter 5, Section 3.



We wanted to see how the other models fared for the goodness of fit test and ROC curve's AUC. The results are found in the Table 3.9.



Models	HL Test			ROC	
	Chi-squared	df	p-value	AUC	Distance for Optimal Cutoff Point
PME	6.7998	8	0.558	0.91	0.26933
BEPME	5.295	8	0.726	0.905	0.3413
ME	5.384	8	0.716	0.889	0.2961
BEME	4.895	8	0.769	0.891	0.2794
AS	8.79	8	0.36	0.853	0.4012

We can see that the AS model had the lowest area under the ROC curve than any of the other models that were fit. Also, it had the highest chi-squared value of all the models. Even though this does not mean that the AS model is not a good model, it does show some possible issues with the model and why we chose the BEPME model over it.

3.5 Results

In Chapter 2 and Chapter 3, a classical regression approach to fit the generalized linear model, logistic regression, was used. The subset selection procedures discovered the best predictors of fatalities due to vehicle collisions given that a collision occurred. We are primarily interested in knowing if these best predictors will give accurate results for the odds of having a fatality in Las Vegas due to a vehicle collision.

3.6 Final Model Selection

Using a logistic regression model, we found that drinking (x3), drugs (x4), inattention (x5), ailment (x6), dark (x10), dawn (x13), dusk (x14), roadway (x18), sideswipe (x22), angle (x23), rear (x24), improper turning (x27), followed too closely (x34), ran off the road (x37), driverless (x40), over evaluation (x43), factor of vehicle type (x44), factor of airbag deployment (x46), seatbelt usage (x47), high school diploma only (x52), four-year degree (x54), and total vehicles involved in accident (x56) were regressed on the whether there was a fatality due to a vehicle collision. The results are shown above in Table 3.6.

In Chapter 4, we will be discussing if there are any influential cases amongst the data and how this affects the model. We will also interpret the final results and talk about what each predictor means to the model.

Chapter 4

Model Diagnostics

In Chapter 3, we discussed a final model, Table 3.6. In this chapter, we will be using a number of different diagnostic procedures to check the adequacy of the model in Table 3.6. As with the standard linear model, it is important to check the adequacy of the assumptions that support the GLM.

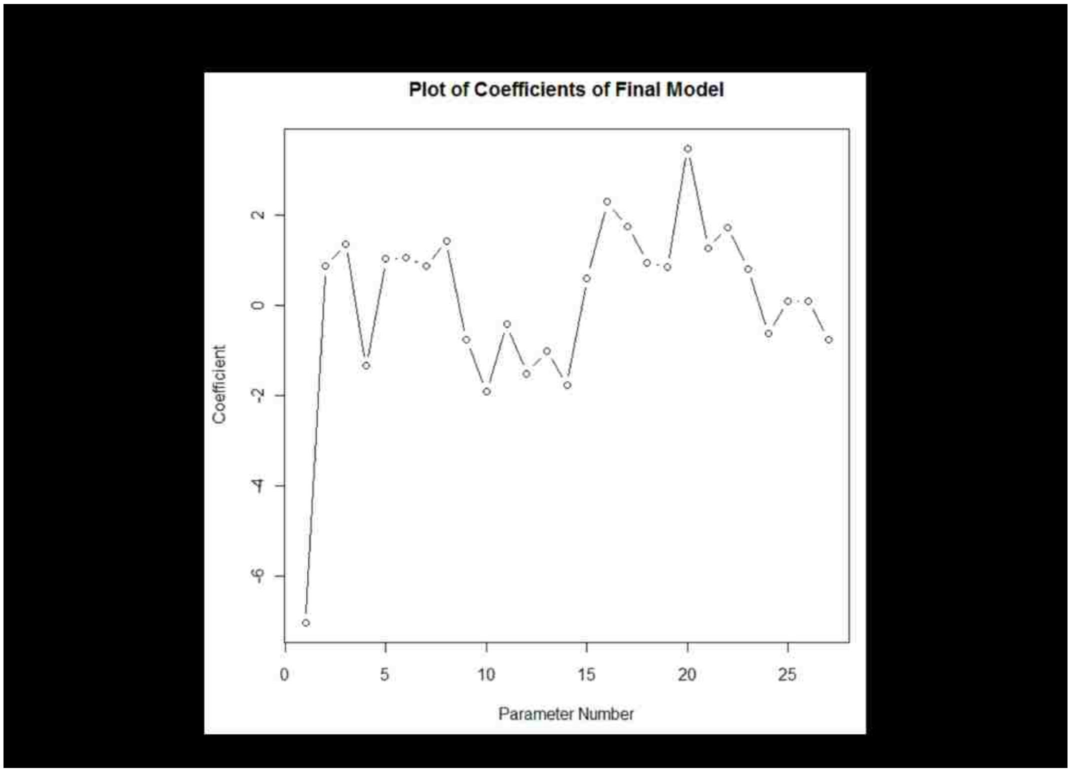
4.1 Leverage

First, we will look at leverage. Leverage points are observations that are discrepant or distant from the other values of the x variable. They may or may not also be outliers. Leverage is the potential for an observation to affect the fit of the model. Leverages h_i are given by the diagonal of the hat matrix, H, given in (4.1) and represent the potential of the point to influence the fit. The h_i is a function of only the X values, so h_i measures the role of the X values in determining how important Y_i is in affecting the fitted value $\widehat{Y}_i$ (Kutner, Nachtsheim, & Neter, 2004). The GLM model, which is what we are using in this thesis, uses weights, W, to fit the model. Leverage is based on the function of X and the response through the weights W (Faraway, 2006). We form a matrix W = diag(w) and the hat matrix is:

$$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2} \qquad (4.1)$$

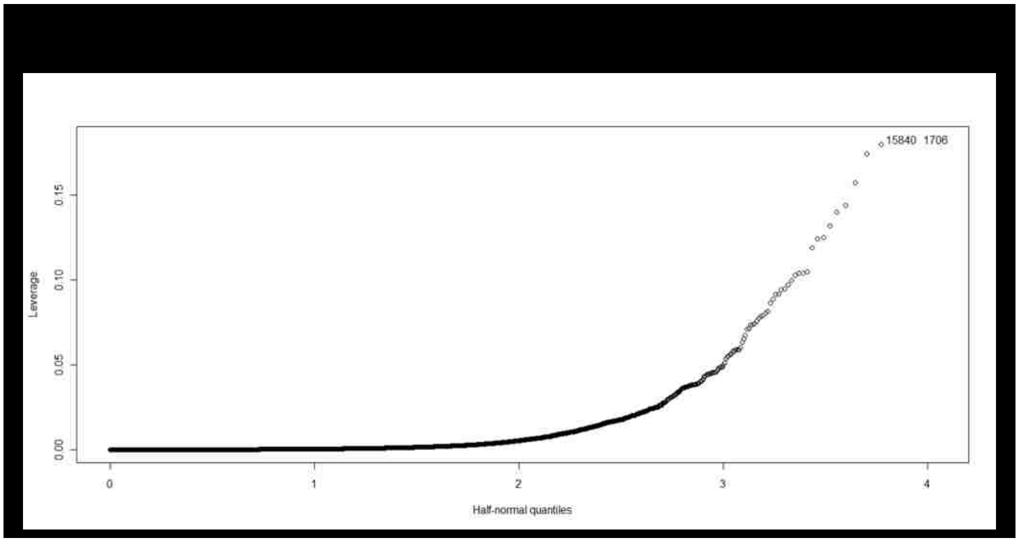
Now, we extract the diagonals of H to get the actual leverages of h_i. A large value of this h_i indicates that the fit may be sensitive to the response in some case i. As a rule of thumb, if h_i is greater than two or three times that of p/n, the observation may be of concern.

Before we look for any influential observations, we will graph the coefficients of the final model found in Chapter 3, to compare it to any models without these possible influential observations.

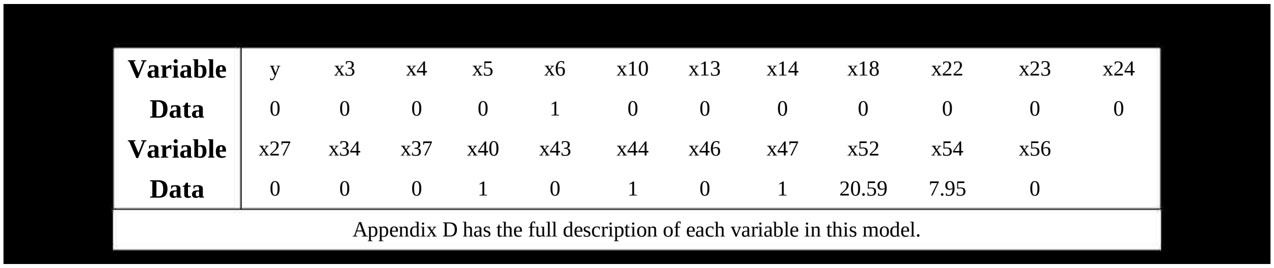


To be able to see if BEME model has any leverage or influential cases, we will look at the half-normal plot of these estimated effects. Along the y-axis of this plot, we have the ordered absolute value of either the leverage or influence. Along the x-axis, we have the theoretical order statistic medians from a half-normal distribution. The outputs are a rank of a list of factors and interactions from the most important to the least important. Thus the half-norm probability plot is a graphical tool that uses these ordered estimates to help assess which factors are important and which are not.

So, first we decided to graph the leverages in a half-normal plot to see whether there seemed to be any outlying cases.



As we can see case 1706 seems to be the farthest out so it may have some leverage. Therefore, we looked at this particular case; the following values were given for each of the variables.



Variable	y	x3	x4	x5	x6	x10	x13	x14	x18	x22	x23	x24
Data	0	0	0	0	1	0	0	0	0	0	0	0
Variable	x27	x34	x37	x40	x43	x44	x46	x47	x52	x54	x56	
Data	0	0	0	1	0	1	0	1	20.59	7.95	0	

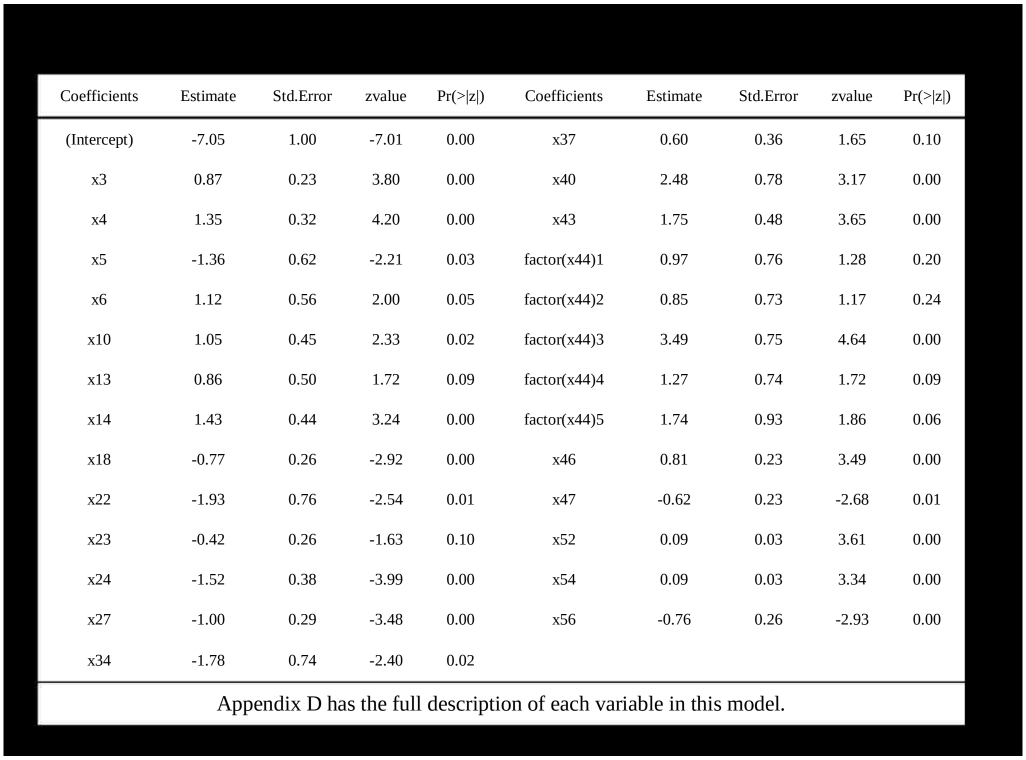
Appendix D has the full description of each variable in this model.

Looking at this printout, we see that none of the variables look out of place; however, it is harder to see something with binary data. Therefore we decided to check the leverage value against 2*p/n; where p is the number of parameters in the model and n is the number of observations.

$$2\left(\frac{27}{18580}\right) = 0.002906$$

When we look at the leverage for this observation, we found it to be 0.1829278; which is larger than the rule of thumb. This means that the observation may be of some concern.

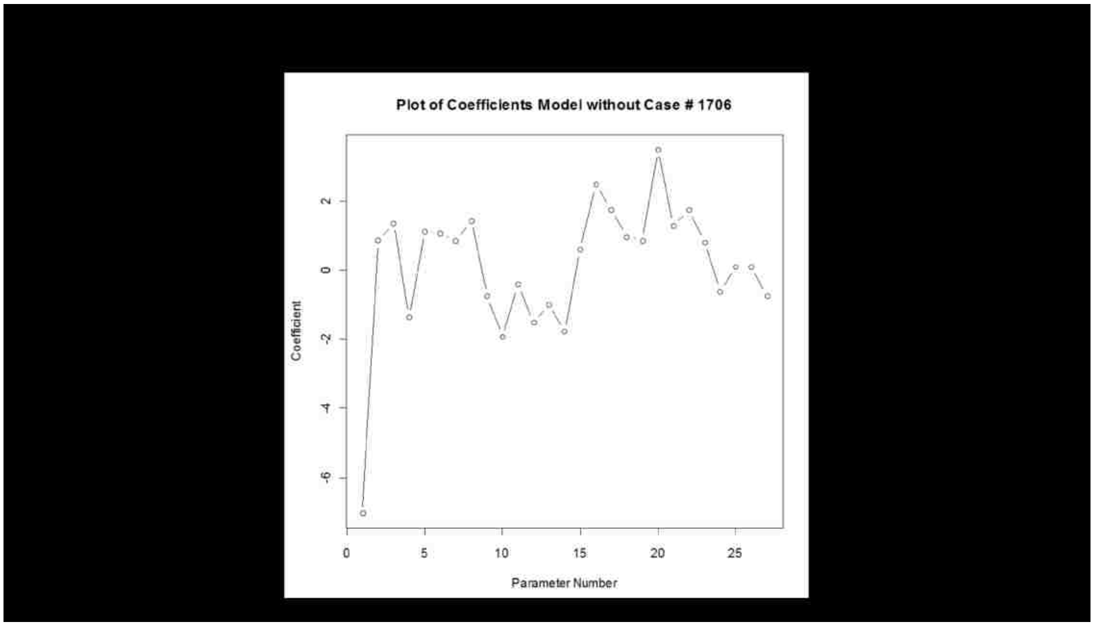
Therefore, it is good practice to remove this case from the model and refit it. What we are looking for is a possible significant change in the coefficients of the model.



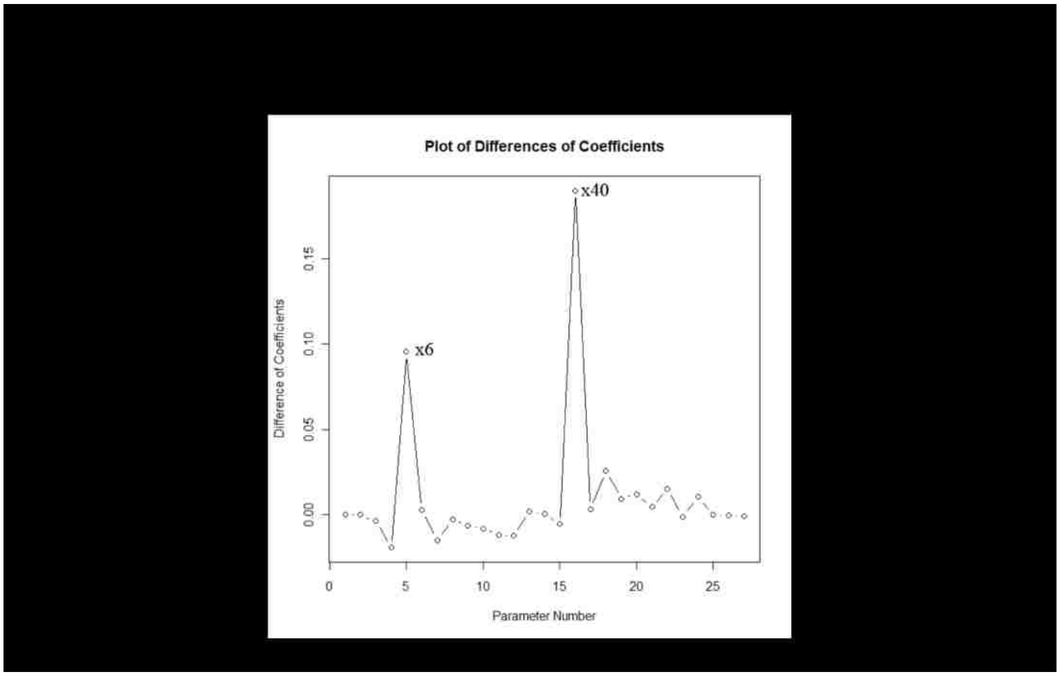
Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)	Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)
(Intercept)	-7.05	1.00	-7.01	0.00	x37	0.60	0.36	1.65	0.10
x3	0.87	0.23	3.80	0.00	x40	2.48	0.78	3.17	0.00
x4	1.35	0.32	4.20	0.00	x43	1.75	0.48	3.65	0.00
x5	-1.36	0.62	-2.21	0.03	factor(x44)1	0.97	0.76	1.28	0.20
x6	1.12	0.56	2.00	0.05	factor(x44)2	0.85	0.73	1.17	0.24
x10	1.05	0.45	2.33	0.02	factor(x44)3	3.49	0.75	4.64	0.00
x13	0.86	0.50	1.72	0.09	factor(x44)4	1.27	0.74	1.72	0.09
x14	1.43	0.44	3.24	0.00	factor(x44)5	1.74	0.93	1.86	0.06
x18	-0.77	0.26	-2.92	0.00	x46	0.81	0.23	3.49	0.00
x22	-1.93	0.76	-2.54	0.01	x47	-0.62	0.23	-2.68	0.01
x23	-0.42	0.26	-1.63	0.10	x52	0.09	0.03	3.61	0.00
x24	-1.52	0.38	-3.99	0.00	x54	0.09	0.03	3.34	0.00
x27	-1.00	0.29	-3.48	0.00	x56	-0.76	0.26	-2.93	0.00
x34	-1.78	0.74	-2.40	0.02					

Appendix D has the full description of each variable in this model.

There does not seem to be a sign change or a significant change in any of the coefficients, so it is not believed that this observation has much leverage to make a substantial difference on the fit of the model; however, we could look at a graph of this model’s coefficients and then compare it to the original in Figure 4.1.



When looking at the coefficients for the model without this case, it does not appear that this is much difference; however, when we look at a graph that plots the difference between the new coefficients and the original, we see that parameters x6 and x40 do have a huge leap in them.



This may show that we should have some concern of this case number, that it might have some leverage. So we need to examine why these two variables are affected so much. The predictor x6 has a 1 in it for a fatality due to ailment. This is one of only four fatalities for this predictor; which could be causing the problem. A very similar story can be said about x40, which has a 1 for fatalities due to driverless vehicle. It is only one of three fatalities for this predictor. We should not remove this case number, because it could be misleading if we did remove it. Due to its potential in the models loss of the number of fatalities, it is scientifically best to keep it.

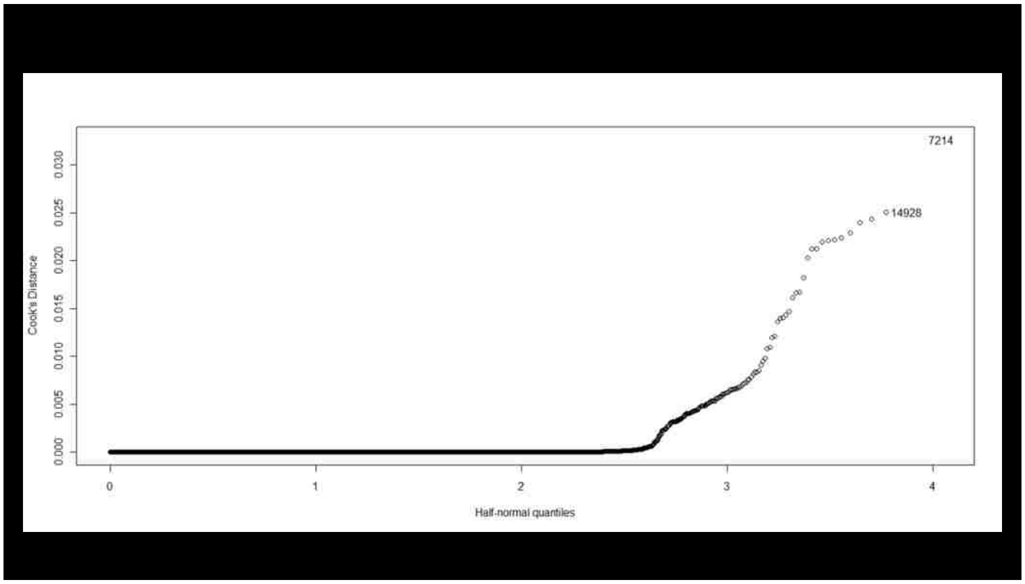
4.2 Influence

Leverage only measures the potential to affect the fit of the model, whereas measures of influence more directly assess the effect of each case on the fit (Faraway, 2006). An influential observation is one which has a relatively large effect on inferences based on the model and removing them would markedly change the statistical analysis. These may or may not be outliers. Outliers are observations that pull the linear regression line in one direction or another. These are typically relatively extreme values of the y variable. One way to find influential cases is by finding the Cook's distance. This measures the standardized change in the linear predictor when the ith case is deleted. The formula for this is:

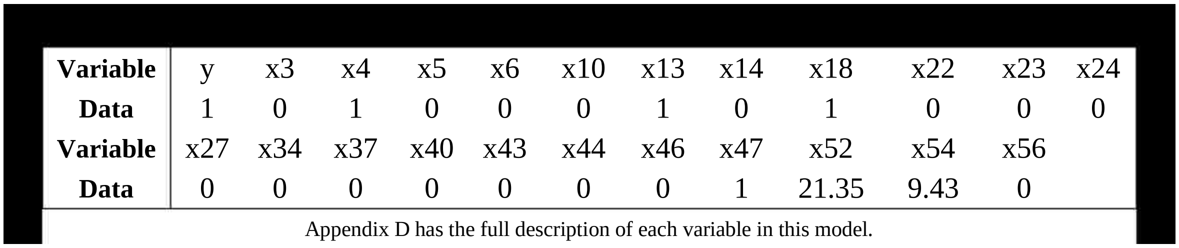
$$D_i = \frac{r_{P_i}^2 h_{ii}}{p(1-h_{ii})^2}, \tag{4.2}$$

where $r_{P_i}^2$is the square of the Pearson's residual, h_{ii} is the leverage factor, and p is the number of parameters. Now, in order for this value to be classified to have a high influence is if D_i is greater than 1 (Hosmer & Lemeshow, 2000).

In order to see if there are any influential observations, we graphed a half-normal quartile graph based on Cook's distance. The following graph shows case number 7214 as the most influential.



We decided to look at this particular data observation and this is what we found:



Variable	y	x3	x4	x5	x6	x10	x13	x14	x18	x22	x23	x24
Data	1	0	1	0	0	0	1	0	1	0	0	0
Variable	x27	x34	x37	x40	x43	x44	x46	x47	x52	x54	x56	
Data	0	0	0	0	0	0	0	1	21.35	9.43	0	

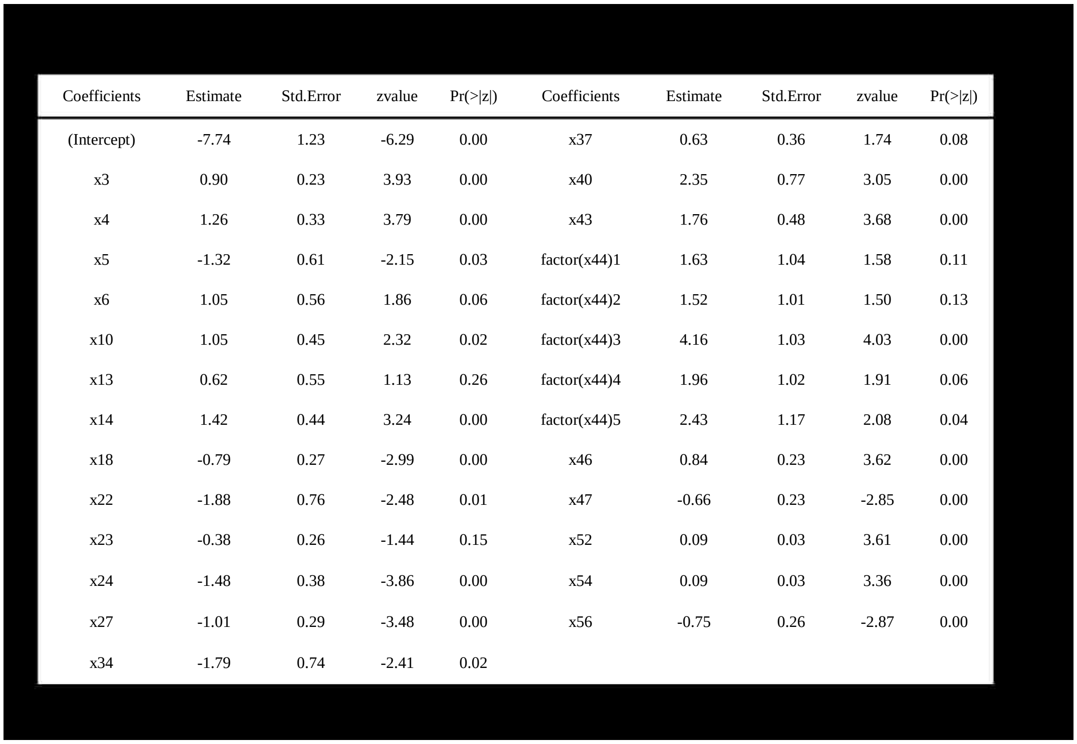
Appendix D has the full description of each variable in this model.

By looking at this observation, the only one that seems off is x13, Dawn, and this could be due to this predictor having 1 fatality out of five fatalities for Dawn. However, we should keep this in mind as we look at the graphs and printouts of the model without this observation.

Now, we will find the Cook's distance for this case:

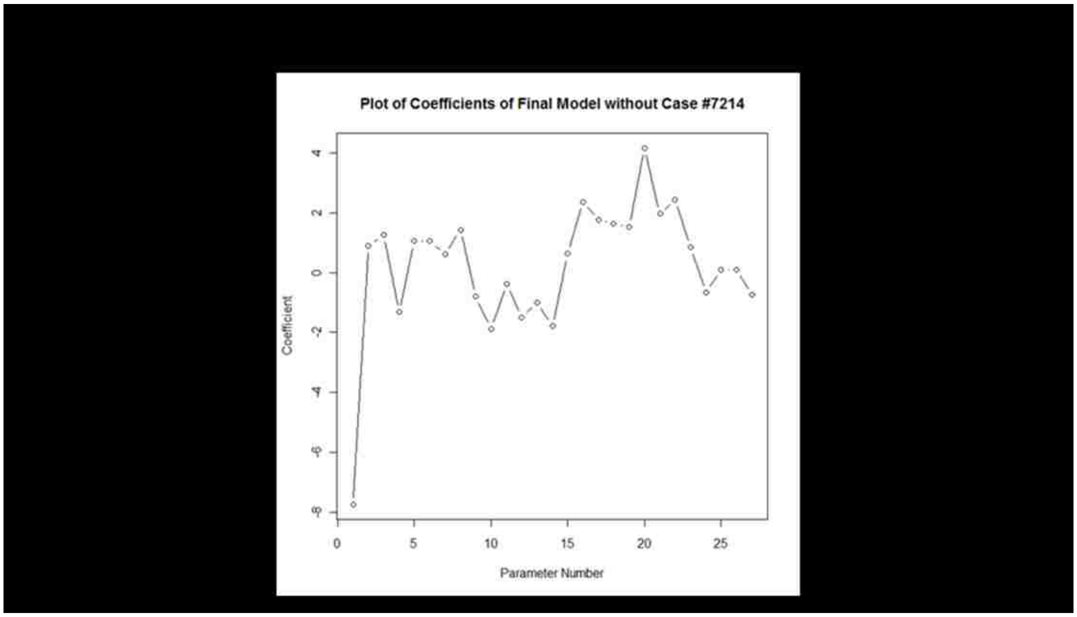
$$D_{7214} = \frac{(5.448)^2 * 0.028}{27 * (1 - 0.028)^2} = 0.0326.$$

D_{7214} is not greater than one, so it may not be influential; however, to further investigate we removed this case from the model and refit the model to see if the coefficients were changed dramatically.

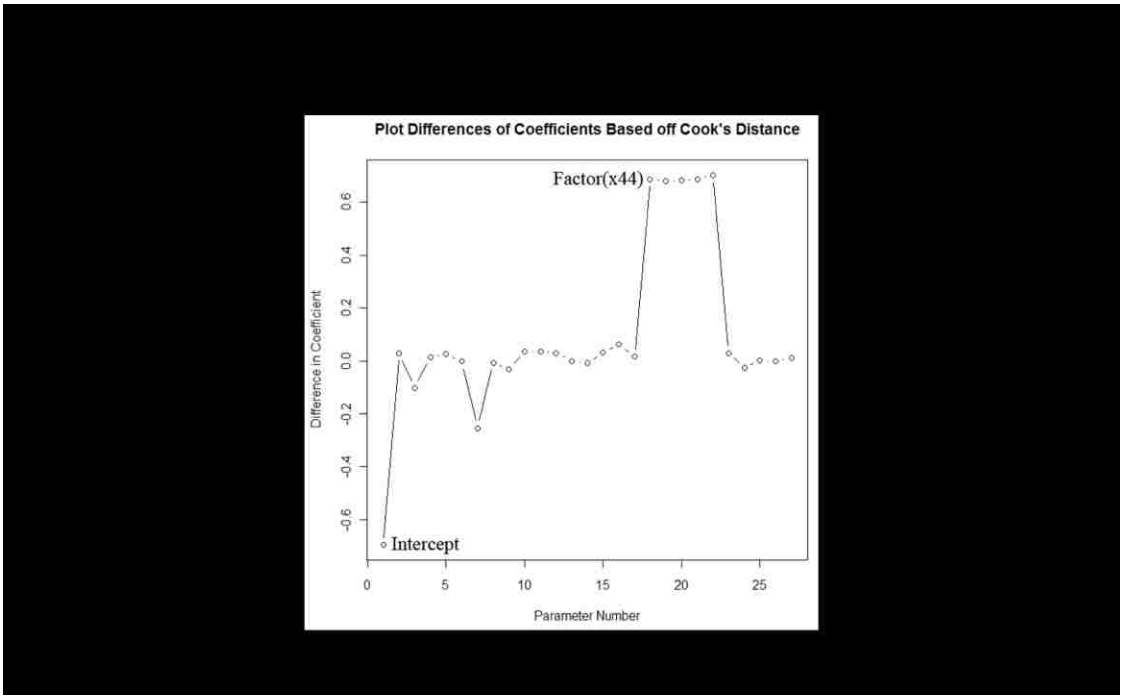


Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)	Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)
(Intercept)	-7.74	1.23	-6.29	0.00	x37	0.63	0.36	1.74	0.08
x3	0.90	0.23	3.93	0.00	x40	2.35	0.77	3.05	0.00
x4	1.26	0.33	3.79	0.00	x43	1.76	0.48	3.68	0.00
x5	-1.32	0.61	-2.15	0.03	factor(x44)1	1.63	1.04	1.58	0.11
x6	1.05	0.56	1.86	0.06	factor(x44)2	1.52	1.01	1.50	0.13
x10	1.05	0.45	2.32	0.02	factor(x44)3	4.16	1.03	4.03	0.00
x13	0.62	0.55	1.13	0.26	factor(x44)4	1.96	1.02	1.91	0.06
x14	1.42	0.44	3.24	0.00	factor(x44)5	2.43	1.17	2.08	0.04
x18	-0.79	0.27	-2.99	0.00	x46	0.84	0.23	3.62	0.00
x22	-1.88	0.76	-2.48	0.01	x47	-0.66	0.23	-2.85	0.00
x23	-0.38	0.26	-1.44	0.15	x52	0.09	0.03	3.61	0.00
x24	-1.48	0.38	-3.86	0.00	x54	0.09	0.03	3.36	0.00
x27	-1.01	0.29	-3.48	0.00	x56	-0.75	0.26	-2.87	0.00
x34	-1.79	0.74	-2.41	0.02					

It does not appear as though any of the coefficients changed a large amount once this case was removed. So to see an image of the new model's coefficients we can graph them. If we look at x13, it was not changed dramatically.



If we look at the graph of this new model compared to the original and we see many variables that have changed, the predictor with the most significant change is factor(x44), this is the type of vehicle predictor. It seems that the removal affected the coefficient positively.



This may mean it is influential; however, after checking the Cook's distance, it is not over one, therefore it does not seem to be statistically influential.

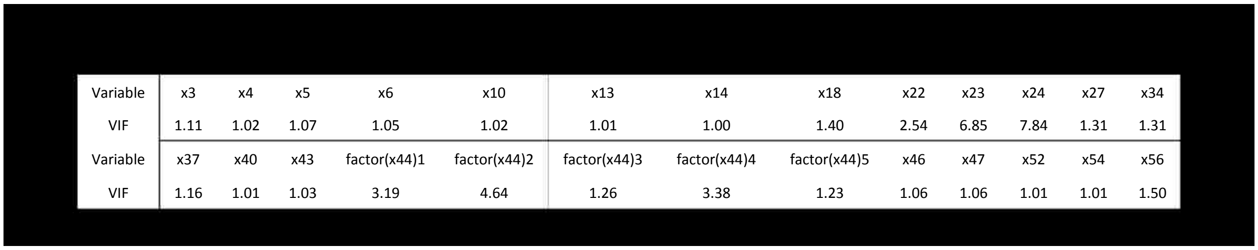
4.3 Multicollinearity

In multiple regression analysis, we want to check the nature and significance of the relations between predictor variables and the response variables. Multicollinearity helps us to understand is predictor variables are related among themselves. If two predictor variables are highly correlated, they both convey essentially the same information. If both of these predictors are included into the model, neither may contribute significantly to the model and corresponding coefficients would be hard to estimate. This could cause some serious issues in determining the validity of whether some variables are influential or not.

If we add or delete a predictor variable that is highly correlated with another predictor variable, it could change the regression coefficients. This could cause a problem when interpreting the results or the model. Having multicollinearity could cause the estimated standard deviation of the regression coefficients to be larger, which could lead to the estimated regression coefficients to not be significant in the model. These are just some of the issues if we have pairwise multicollinearity in the model.

To help assess the multicollinearity, we could create a correlation matrix or a variance inflation factor (VIF). A correlation matrix showcases all the correlations between each variable and also correlations between each predictor variable and the response variable. A correlation matrix may reveal large pairwise collinearities. A VIF measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. When used in the diagnostic methods, if a VIF is larger than 10, then it is typically an indication that multicollinearity may be influencing the least squares estimates (Kutner, Nachtsheim, & Neter, 2004).

With the BE of CSAB model, the correlation matrix revealed a moderate pairwise correlation between x24 and x23 with a correlation of -0.79, the whole matrix can be found in Appendix H . When the VIF was calculated for each of the predictors, there was no VIF greater than 10, thus not showcasing a multicollinearity issue in this model. The list of VIFs are shown below:



Variable	x3	x4	x5	x6	x10	x13	x14	x18	x22	x23	x24	x27	x34
VIF	1.11	1.02	1.07	1.05	1.02	1.01	1.00	1.40	2.54	6.85	7.84	1.31	1.31
Variable	x37	x40	x43	factor(x44)1	factor(x44)2	factor(x44)3	factor(x44)4	factor(x44)5	x46	x47	x52	x54	x56
VIF	1.16	1.01	1.03	3.19	4.64	1.26	3.38	1.23	1.06	1.06	1.01	1.01	1.50

In conclusion, the diagnostic portion of this thesis shows some leverage, but after checking the data for possible data-entry errors, examining the physical case number and excluding the point from the model, it was not found to significantly alter the model. However, to exclude the outlier would be dangerous to the model; scientifically it gives important information about the successes for the logistic regression. We also discovered that there was one pairwise correlation; however after looking at the VIF it was not seen as extreme or existing for multicollinearity. Therefore, this model is ready for the interpretation and prediction steps that will occur in Chapter 5.

Chapter 5

Results

5.1 Inference

Statistical inference is the process of drawing conclusions from data that are subject to random variation, for example, observational errors or sampling errors (Upton & Cook, 2008). One way inferences occur in statistics is through confidence intervals (CI). Confidence intervals usually are used to indicate the reliability of an estimate.

In our data, it might interest us to look at the confidence intervals around our parameters. These were calculated using R software and are shown in table 5.2, but could have been done using the following formula for each coefficient:

$$\mathrm{b_i} \pm 1.96\mathrm{se}\{\mathrm{b_i}\}, \mathrm{i} = 0,1, \dots, \mathrm{p} \qquad (5.1)$$

Let's calculate the confidence interval around the predictor variable: Ran off Road. This is shown below:

$$0.6018330 \pm 1.96 * 0.1194533 = (\,-0.13, 1.29)$$

This confidence interval means that we are 95% confident that the true coefficient is in that interval. To be able to interpret this result, we would want to learn more about odds ratio.

5.2 Interpretation

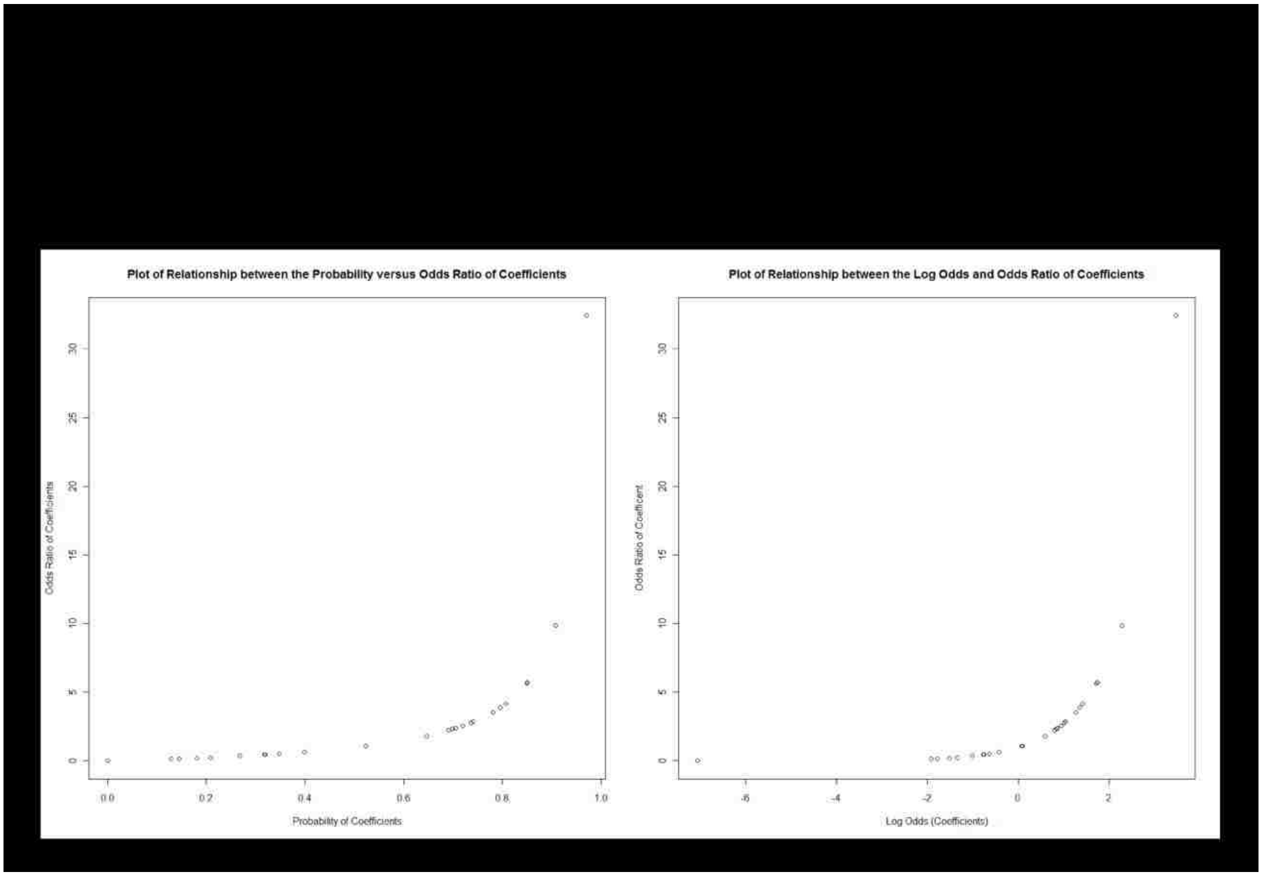
The interpretation of the results that we gained in both the parameters and the confidence intervals are left to this section of Chapter 5. Although we could discuss the log odds of some event occurring, it is easier to discuss the odds of an event. In logistic regression, the odds of an event, and the log odds of an event occurring are given below:

$$\mathrm{odds} = \mathrm{o} = \frac{\pi}{1-\pi} \qquad (5.2)$$

and

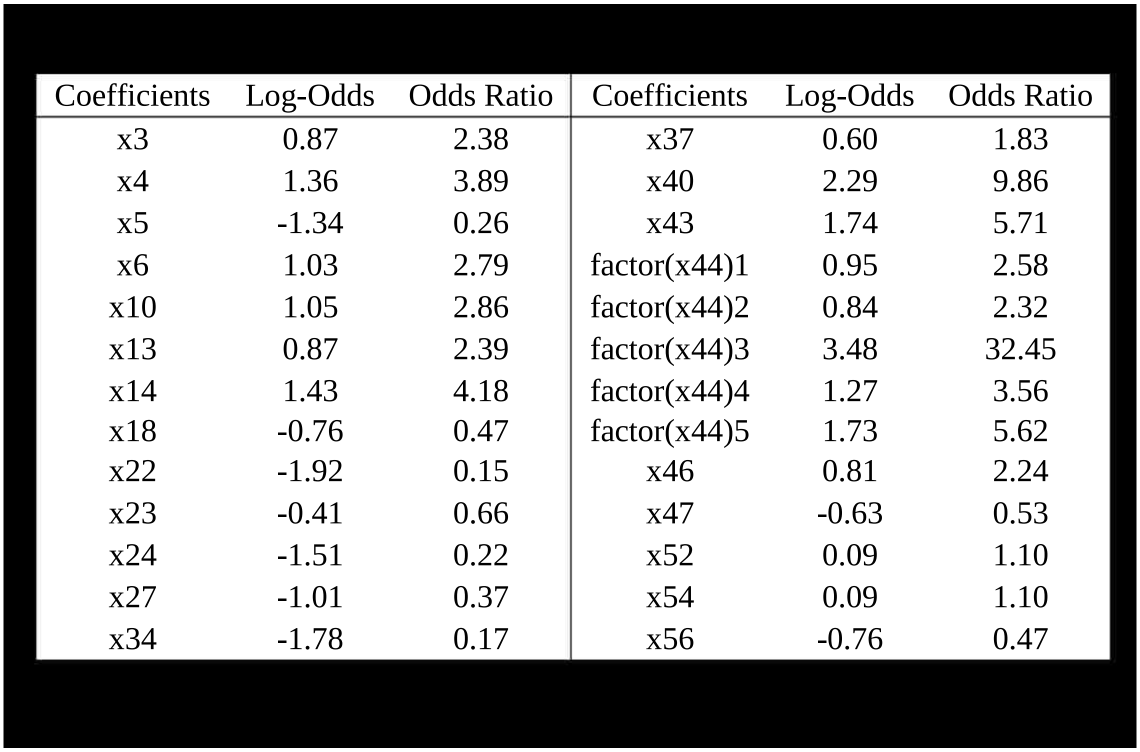
$$\log - \text{odds} = \pi' = \log_e(o) = X'\beta \tag{5.3}$$

where π is the probability of the event, given by $\pi = \frac{e^{X'\beta}}{1+e^{X'\beta}}$. The log-odds is estimated using x'b. The transformation between probability and odds ratio is a monotonic transformation, meaning the odds ratio increase as the probability increases. Probability ranges from 0 to 1; whereas the odds ratio ranges from 0 and positive infinity. The transformation from the odds ratio to log odds is the log transformation. This transformation is again a monotonic transformation. That is, the greater the odds ratio, the greater the log of odds. In Figure 5.1, we can see these relationships involving our data.



Let us look at this relationship in our own data. When we look at the predictor variable Run off Road, $b_{14} = 0.601833$. The interpretation of this is that per 1 unit increase in x_{37}

with all other variables held fixed, the log-odds of success increases by 0.61. Now, we transform this into the odds ratio, which is found by $\widehat{\mathrm{OR}} = \exp(0.601833) = 1.83$. Thus the odds of having a fatality are increased by 83%, if someone is run off the road. Now, if we follow the formula 5.3, we find that the estimated probability is 0.646 or 64.6%. Therefore, the chance of having a fatality will increase by 64.6% if someone is run off the road. In Table 5.1, the log-odds, odds ratio, and probability are listed for each parameter.

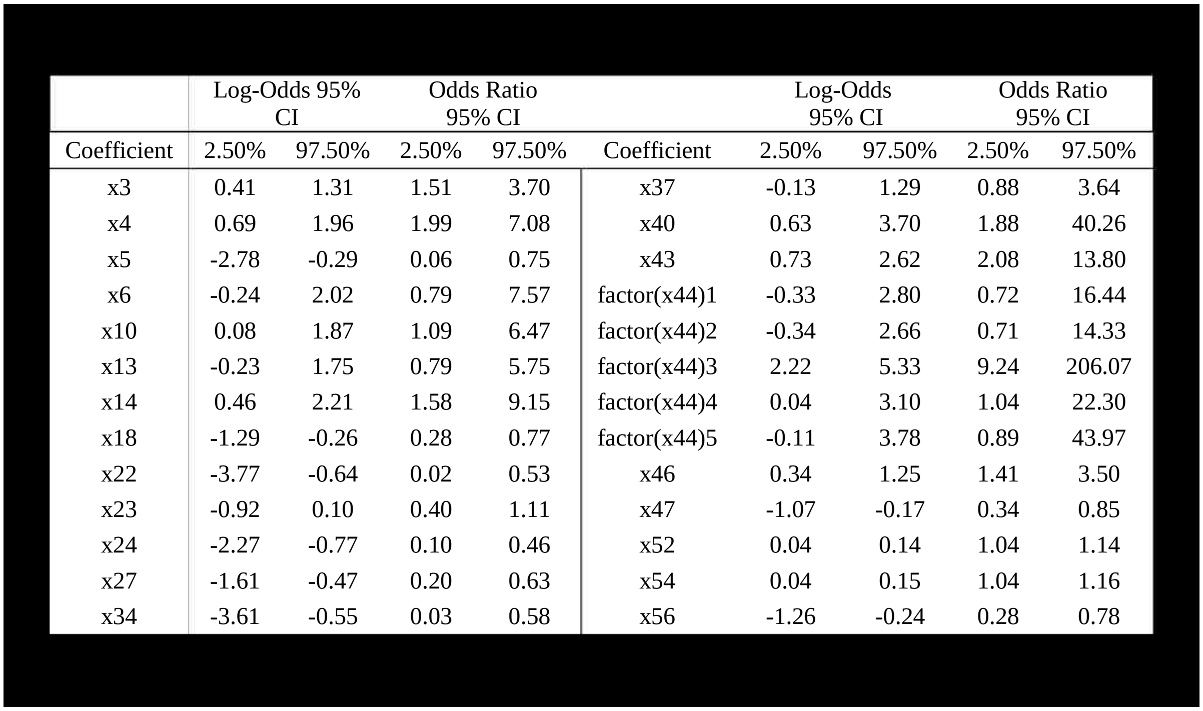


Coefficients	Log-Odds	Odds Ratio	Coefficients	Log-Odds	Odds Ratio
x3	0.87	2.38	x37	0.60	1.83
x4	1.36	3.89	x40	2.29	9.86
x5	-1.34	0.26	x43	1.74	5.71
x6	1.03	2.79	factor(x44)1	0.95	2.58
x10	1.05	2.86	factor(x44)2	0.84	2.32
x13	0.87	2.39	factor(x44)3	3.48	32.45
x14	1.43	4.18	factor(x44)4	1.27	3.56
x18	-0.76	0.47	factor(x44)5	1.73	5.62
x22	-1.92	0.15	x46	0.81	2.24
x23	-0.41	0.66	x47	-0.63	0.53
x24	-1.51	0.22	x52	0.09	1.10
x27	-1.01	0.37	x54	0.09	1.10
x34	-1.78	0.17	x56	-0.76	0.47

In Section 5.1, we discussed confidence intervals of a logistic regression; however, the interpretation is a bit more difficult to discuss without first talking about the odds ratio in relation to log-odds. We stated that the odds ratio was calculated by first finding the estimate of the parameter and then taking the exponential of it. This is, also, true for the odds ratio of a confidence interval centered on that parameter. We would want to take the exponential of the confidence interval that we just found. We will use the example from the beginning of Section 5.2 that dealt with running off the road. First we found the parameter, then the CI around this parameter, now we will take the exponential of this. This will look like the following:

$$\exp(0.6018330 \pm 1.96 * 0.1194533) = (0.876, 3.64)$$

The corresponding 95% confidence limits for the odds ratio are 0.88 and 3.64. In Table 5.2, all the log-odds and odds ratio confidence intervals listed.



	Log-Odds 95% CI		Odds Ratio 95% CI			Log-Odds 95% CI		Odds Ratio 95% CI	
Coefficient	2.50%	97.50%	2.50%	97.50%	Coefficient	2.50%	97.50%	2.50%	97.50%
x3	0.41	1.31	1.51	3.70	x37	-0.13	1.29	0.88	3.64
x4	0.69	1.96	1.99	7.08	x40	0.63	3.70	1.88	40.26
x5	-2.78	-0.29	0.06	0.75	x43	0.73	2.62	2.08	13.80
x6	-0.24	2.02	0.79	7.57	factor(x44)1	-0.33	2.80	0.72	16.44
x10	0.08	1.87	1.09	6.47	factor(x44)2	-0.34	2.66	0.71	14.33
x13	-0.23	1.75	0.79	5.75	factor(x44)3	2.22	5.33	9.24	206.07
x14	0.46	2.21	1.58	9.15	factor(x44)4	0.04	3.10	1.04	22.30
x18	-1.29	-0.26	0.28	0.77	factor(x44)5	-0.11	3.78	0.89	43.97
x22	-3.77	-0.64	0.02	0.53	x46	0.34	1.25	1.41	3.50
x23	-0.92	0.10	0.40	1.11	x47	-1.07	-0.17	0.34	0.85
x24	-2.27	-0.77	0.10	0.46	x52	0.04	0.14	1.04	1.14
x27	-1.61	-0.47	0.20	0.63	x54	0.04	0.15	1.04	1.16
x34	-3.61	-0.55	0.03	0.58	x56	-1.26	-0.24	0.28	0.78

5.3 Prediction of a New Observation

Multiple logistic regression is frequently used for making predictions for new observations. In the model for this thesis, we wanted to be able to predict whether a fatality will occur given the responsible driver’s distractions, vehicle type, seatbelt usage, airbag deployment, time of day, type of weather, type of lighting, roadway structure, angle of accident, and reason for accident. Forecasting a binary outcome for given levels

X_h of the X variables is simple in the sense that the outcome 1 will be predicted if the estimated value of $\widehat{\pi}_h$ is large or 0 if this value is small. The cutoff that has the lowest proportion of incorrect predictions is the one to use.

Let us begin by stating that there were 18580 accidents in our sample of 2009, of those 128 had at least one fatality within 30 days after the collision. If we take 128/18580, we get approximately .007. This proportion can be used as the starting point in search for the best cutoff in the prediction rule. So, the following is our first rule:

$$\text{Predict 1 if } \widehat{\pi}_h \geq .007\text{; predict 0 if } \widehat{\pi}_h < .007.$$

If we use this cutoff, then we noticed that for $\widehat{\pi}_1 = 0.014$ (case 1) the prediction rule would predict that the person would have a fatality, but the observation shows no fatality occurred. So, it is predicted incorrectly. However, for $\widehat{\pi}_4 = 0.002$ the prediction would be correct, because it would say there is no fatality, and the observation had none. See Appendix E for more of the $\widehat{\pi}_i$ and their observed values of y. Table 5.3a provides a summary of the number of correct and incorrect classifications based on the stated prediction rule. Of the 18452 collisions that were not fatal, 3200 would be incorrectly predicted to have had a fatality, or an error of 17.3%. Of the 128 collisions with a fatality, 29 would be incorrectly predicted to not have a fatality, or 22.7%. Altogether, 29+3200=3229 of the 18580 predictions would be incorrect, so the prediction error rate for this rule is approximately 17.38%.

These analyses were made for other cutoff points. Let us say we know nothing about our data, normally we would choose a prediction rule like the following:

$$\text{Predict 1 if } \widehat{\pi}_h \geq .5\text{; predict 0 if } \widehat{\pi}_h < .5.$$

We performed this and the information can be found in Table 5.3b. The prediction error for this rule is (123+4)/18580 = .007 or 0.7%. Some worry is that the true positive may be too small and therefore the original rule might want to be followed.

[redacted]

Back in Chapter 3, during the goodness of fit test, the ROC curve was introduced. One way that this curve can be used is to find the optimal cutoff. Recall that the line from the center of the random performance line diagonally to the top left is balance between the conservative and liberal classification model. Therefore, it may make sense to find a cutoff close to this line, since it best represents the balance between too many false positives and too few false positives. So, it was decided to look at the two rules in Table 5.3 to see which might be the better rule for prediction and which is closest to the optimal cutoff points. The calculations for the two prediction rules can be found in Table 5.4a and b.

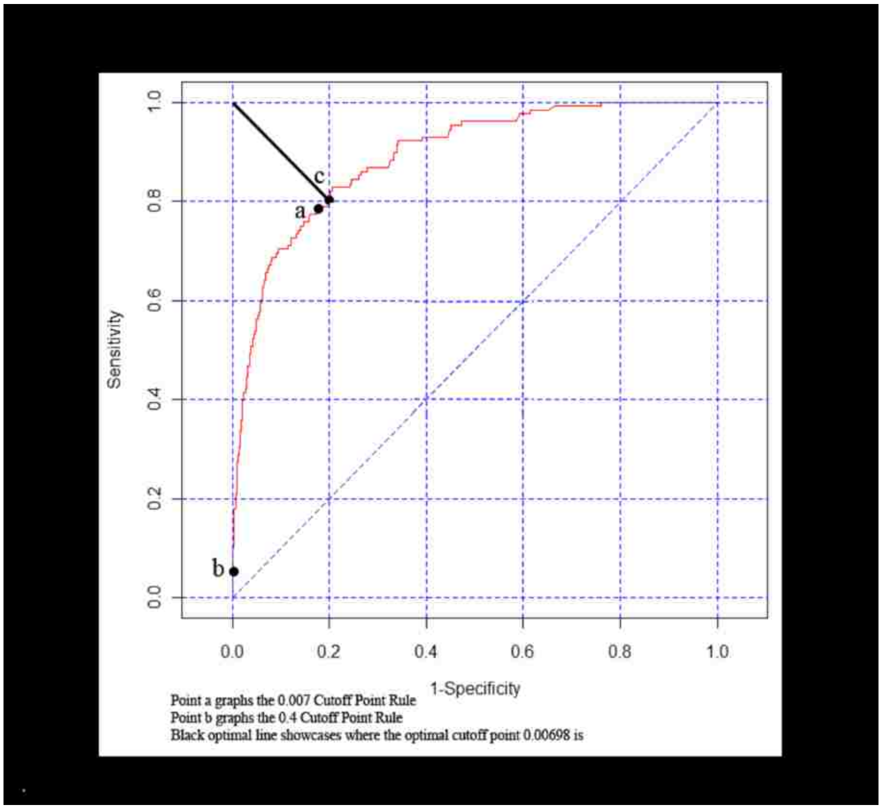
[redacted]

When we graphed these two points onto the ROC curve, we noticed that Rule 0.007, seems to be closer to the optimal line, which makes it a better rule than Rule 0.5, but not the optimal one.

In Chapter 3, we discussed the ROC curve. Figure 3.1, illustrates a line called the minimum d, sometimes this line is called the optimal line. The optimal point is located by finding the smallest distance from the point (0,1) and some cutoff point. This distance has the following formula:

$$\mathrm{d}^2 = [(1 - \mathrm{S_N})^2 + (1 - \mathrm{S_P})^2] \tag{5.5}$$

where S_N is the sensitivity and S_P is the specificity for some cutoff point. The optimal cutoff point is the point that maximizes the correct classification (Kumar & Indrayan, 2011). We located the optimal point at coordinates (0.174, 0.781), which is at cutoff point 0.00698. This rule would have a prediction error of 17.42%. There are two error rates, one is called a false negative and the other is false positive. The false negative rate is 0.22 and the false positive rate is 0.17. The risks for the two groups are pretty balanced, which is the desirable outcome when looking at an optimal cutoff point. As we moved further from this optimal point in either direction, we noticed that these error rate values became significantly unbalanced. These three rules are placed onto Graph 5.2 for comparison.



Point a graphs the 0.007 Cutoff Point Rule
Point b graphs the 0.4 Cutoff Point Rule
Black optimal line showcases where the optimal cutoff point 0.00698 is

5.4 Validation of Prediction Error Rate

The reliability of the prediction error rate observed in the model-building set of data is examined by applying the prediction rules to a validation set. If the new prediction rate is about the same as that for the model-building data set, then the latter gives a reliable indication of the predictive ability of the fitted logistic regression model and the chosen prediction rule (Kutner, Nachtsheim, & Neter, 2004).

Our validation set will be a sample of the data provided for the year 2008 collisions in Las Vegas. In Table 5.4, we see the summary output for both rules that were discussed in the prediction portion of this thesis.

[redacted]

None of the rules have prediction error rates that are considerably higher than the error rates based on the model-building data set. Therefore, they may be reliable. In our final chapter, we will discuss some of the possible changes that could be made at a later time to the model.

Chapter 6

Discussion, Limitations, and Conclusions

6.1 Discussion

The principal objective of this study was to examine the transportation data to advance understanding and appreciation of the causes of fatalities related to a vehicle collision. This study looked at several different independent predictors to see which ones would best be able to predict the odds of having a fatality in a vehicle in Las Vegas, Nevada.

We hypothesized that the following predictors would either increase or decrease the odds of having a fatality in a vehicle collision: Responsible driver age and gender, what distracted the responsible driver, the reason for the accident, the weather and type of lighting for the day, vehicle information, angle of impact, zip code characteristics, and date of the collision.

For this hypothesis, we used a logistic regression analysis to examine the best predictors of odds of fatalities occurring. The predictor variables can be broken up into two categories. These are known as risk factors, which increase the chances of a fatality, and protector factors, which decrease or prevent the chances of a fatality. We found that the drinking, drugs, ailment, dark, dawn, dusk, ran off road, driverless vehicle, over evaluating, driving a car versus a van, driving a truck versus a van, driving a motorcycle compared to a van, driving a government vehicle vs. a van, and driving an unusual vehicle vs. a van, airbag not deploying, HS diploma education and a four-year degree seem to increase the odds of having a fatality in a vehicle collision. The predictors that are classified risk factors are: drugs, ailment, darkness, dusk, driverless vehicle, over evaluating, driving a car verses a van, driving a government vehicle verses a van and driving an unusual vehicle verses a van. These variables all increase the chances of having a fatality significantly more than the other variables ($b_i > 1$). The rest of the variables are referred to as protector variables, since their coefficient is less than one.

We found that the wearing of a seatbelt, the more vehicles involved, following too closely, making an illegal turn, being rear ended, getting hit at an angle, getting sideswiped, roadway environment, and inattention to driving all decrease the log-odds of a fatality. These are all referred to as protector factors since they decrease the chances of a fatality occurring. After we found the predictors that would pass the Goodness of Fit test and did diagnostics on this model, we used it to create a prediction of a new observation and then validated this with data from 2008. It seems that the final model that was created is a moderately good model for prediction purposes.

Some of the possible reasons why the 2008 data may not have a better fit for the prediction rules is because 1) in 2009, there were vehicles released with more updated safety measures, 2) in 2009 there could have been new policies put into action that could increase the safety of driving, 3) some of the zip codes between 2008 and 2009 were changed, and 4) some intersections may have become safer due to lights being placed in, medians developed, 4-way stop signs, etc. These are all possible reasons why the predictions are a bit different.

6.2 Limitations

Some of the data was unusable due to not being able to locate proper zip codes to help with the analysis of the zip code characteristics, in particular the cross streets. An item that would have been helpful is the longitude and latitude of the collision. This would have pinpointed exactly where the collision occurred and there would have been no guess work needed. However, this could only be done if police officers had a GPS in their vehicles that would easily input the coordinates into their paperwork. This possibly would require a multi-million dollar budget just across Nevada, let alone the rest of the country; however, if the government was really interested in where these fatalities are absolutely occurring and implementing strategies to decrease the collisions, the possible lives saved could be worth the expense.

Part of this same limitation is the fact that there may be an issue with too few successes (i.e., fatalities) in the data. One way to check this is by comparing the means of the unusable information (since it is missing the zip codes) to the information that has zip codes. If there appears to be substantial differences in the means for the two groups, there could be an issue. In Appendix I, we have the results of the two situations; it does not appear as though there are any extreme situations that would indicate an issue. A second way to check for the issue of too few successes is by comparing the cutoff point with the actual predicted rate of having a fatality in comparison to accidents. The optimal cutoff point was 0.00698 and the rate of successes was 128/18580 = .00689; thus not showing an issue since the cutoff point is very close to the true value of number of successes.

A second limitation to using data from police reports is the police themselves. Many are very busy throughout the day and do not get to finish their paperwork until their shift is over; sometimes not even until the end of the week. Because of this, some details may become fuzzier or the officer may have completely forgotten where the accident occurred or other details such as if the person was a female or male that was responsible. The other limitation with the data we began with could be the direct result of the health care provider, as they may not see the actual injury that a person has as a fatal injury and thus send them home. Days later, the person may die from the complications related to the initial injury and it may not be reported as a fatality due to the original vehicle collision.

A third limitation of this study is that it is only using a straightforward method of logistic regression. It might be possible that the data was spatially dependent, since zip code information is provided. This means that if the accident had occurred in one zip code, it could have had an effect on a neighboring zip code. Again, having the exact coordinates of a vehicle collision would have been extremely helpful in allowing us to be able to estimate the distance between several accidents within a certain time to see if there was any dependency. For example, an accident could have occurred in one zip code, but caused a pile up in a second or third zip code. All of these accidents would all have had an estimated time of accident that was relatively the same. However, because we cannot see the exact geographic location of the collisions, there is no telling whether these multiple collisions could be dependent.

Fourthly, since there are repeated measures in a zip code, there could be a need for a random effects model. Even though during Chapter 2, the majority of the zip code was removed from the model, the population and median income were not. However, they did have repeated values for zip codes. To fix this issue they were placed in ordinal categories. By the final model, these two were removed due to not being significant. They might have been significant, if we used a different model such as a random effects model; which is a type of hierarchical linear model.

6.3 Conclusion

In conclusion, although this particular model is moderately useful at this stage, it could be just the beginning of a model that could really detect where the odds of a fatality increase and where certain types of educational programming could occur. It may be beneficial to eventually look at the two different modeling types spoken about in the limitations section. The hope of this thesis is to become part of a larger study that may be able to be used to help further policies to create safer roads for all people involved.

Appendix A: Description Statistics

Section 1: Continuous Explanatory Variable

	Coefficient	Wald Test	P-Value	OR of Coefficient	Lower 95% CI	Upper 95% CI	Lower OR 95% CI	Upper OR 95% CI
x51	-0.06	-3.63	0.00	0.94	-0.09	-0.03	0.91	0.97
x52	0.06	2.32	0.02	1.07	0.01	0.12	1.01	1.12
x53	0.33	4.27	0.00	1.39	0.18	0.48	1.20	1.62
x54	0.09	3.97	0.00	1.10	0.05	0.14	1.05	1.15
x55	0.07	2.11	0.04	1.07	0.00	0.14	1.00	1.15

Section 2: Dichotomous Explanatory Variable

	Coefficient	Wald Test	P-Value	OR of Coefficient	Lower 95% CI	Upper 95% CI	Lower OR 95% CI	Upper OR 95% CI
x1	0.51	2.56	0.01	1.66	0.12	0.89	1.13	2.44
x3	1.51	7.69	0.00	4.54	1.13	1.90	3.09	6.68
x4	1.88	6.52	0.00	6.56	1.32	2.45	3.73	11.55
x5	-1.37	-2.34	0.02	0.26	-2.51	-0.22	0.08	0.80
x6	1.06	2.08	0.04	2.90	0.06	2.07	1.06	7.92
x7	-1.13	-1.12	0.26	0.32	-3.09	0.84	0.05	2.33
x9	0.11	0.40	0.69	1.11	-0.42	0.63	0.66	1.88
x10	1.84	4.64	0.00	6.30	1.06	2.62	2.90	13.69
x11	-0.97	-5.48	0.00	0.38	-1.32	-0.63	0.27	0.54
x12	0.60	3.27	0.00	1.82	0.24	0.96	1.27	2.61
x13	1.24	2.69	0.01	3.46	0.34	2.15	1.40	8.56
x14	1.02	2.41	0.02	2.77	0.19	1.85	1.21	6.33
x15	-0.73	-1.02	0.31	0.48	-2.13	0.67	0.12	1.95
x16	0.06	0.10	0.92	1.06	-1.09	1.21	0.34	3.35
x17	-0.42	-0.71	0.48	0.66	-1.57	0.73	0.21	2.07
x18	1.06	5.71	0.00	2.88	0.69	1.42	2.00	4.14
x19	-0.48	-2.67	0.01	0.62	-0.83	-0.13	0.44	0.88

Appendix A: Section 2 continued

	Coefficient	Wald Test	P-Value	OR of Coefficient	Lower 95% CI	Upper 95% CI	Lower OR 95% CI	Upper OR 95% CI
x22	-1.46	-2.05	0.04	0.23	-2.86	-0.07	0.06	0.93
x23	0.37	2.06	0.04	1.45	0.02	0.72	1.02	2.05
x24	-1.53	-5.86	0.00	0.22	-2.04	-1.02	0.13	0.36
x25	1.27	2.48	0.01	3.58	0.27	2.28	1.31	9.79
x26	-0.91	-1.78	0.08	0.40	-1.90	0.09	0.15	1.10
x27	-0.88	-3.27	0.00	0.42	-1.40	-0.35	0.25	0.70
x29	2.78	2.63	0.01	16.14	0.71	4.85	2.03	128.30
x30	1.31	1.29	0.20	3.72	-0.68	3.31	0.51	27.27
x31	0.25	0.24	0.81	1.28	-1.73	2.22	0.18	9.22
x33	-0.03	-0.09	0.93	0.97	-0.75	0.68	0.47	1.98
x34	-2.49	-3.49	0.00	0.08	-3.88	-1.09	0.02	0.34
x35	-0.04	-0.17	0.86	0.96	-0.47	0.39	0.63	1.48
x37	2.17	7.48	0.00	8.76	1.60	2.74	4.96	15.48
x38	-0.05	-0.14	0.89	0.95	-0.77	0.67	0.46	1.95
x40	3.53	5.45	0.00	34.04	2.26	4.80	9.58	120.93
x41	1.30	2.20	0.03	3.67	0.14	2.46	1.15	11.68
x42	0.77	1.51	0.13	2.16	-0.23	1.77	0.79	5.90
x43	2.51	6.23	0.00	12.35	1.72	3.30	5.60	27.24
x46	0.94	4.60	0.00	2.55	0.54	1.33	1.71	3.80
x47	-1.51	-7.62	0.00	0.22	-1.90	-1.12	0.15	0.33
x48	0.32	1.77	0.08	1.37	-0.03	0.67	0.97	1.95
x56	-2.06	-11.45	0.00	0.13	-2.41	-1.71	0.09	0.18

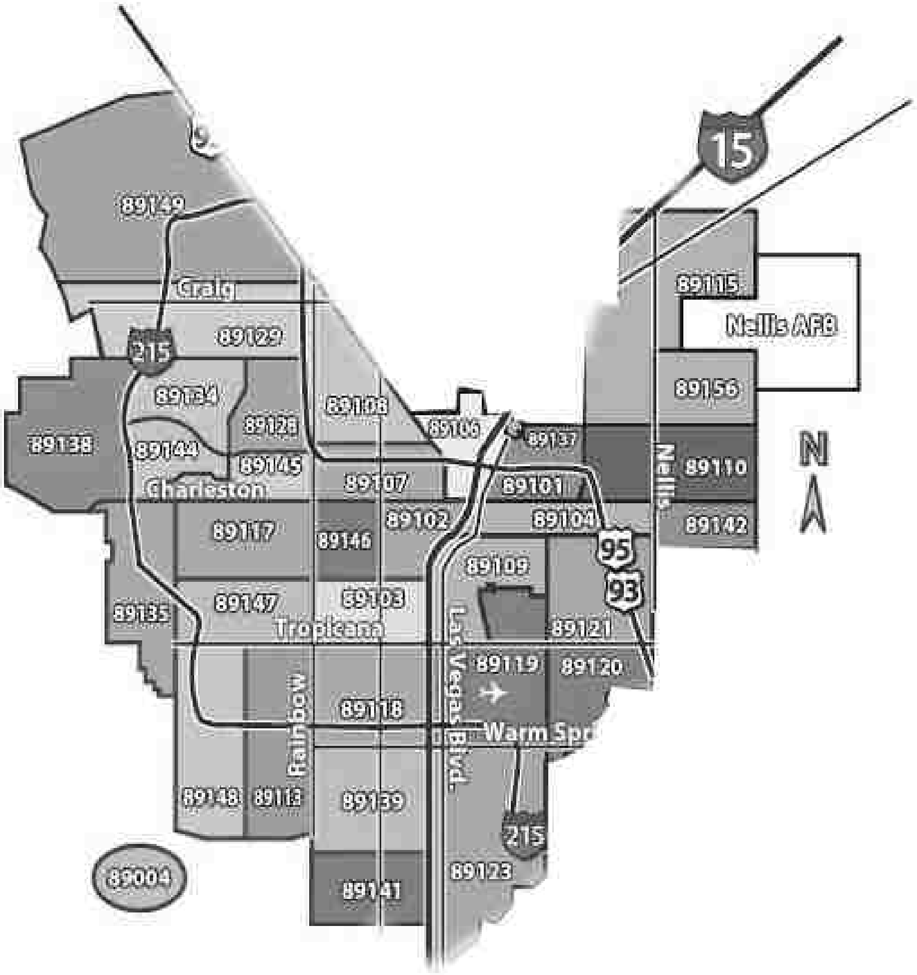
Section 3: Polychotomous Explanatory Variable

	Coefficient	Wald Test	P-Value	OR of Coefficient	Lower 95% CI	Upper 95% CI	Lower OR 95% CI	Upper OR 95% CI
factor(x2)1	-0.34	-1.65	0.10	0.71	-0.73	0.06	0.48	1.06
factor(x2)2	-0.36	-1.36	0.17	0.70	-0.88	0.16	0.41	1.17
factor(x2)3	0.50	1.30	0.19	1.65	-0.25	1.26	0.78	3.53
factor(x20)1	-0.19	-0.71	0.48	0.83	-0.71	0.33	0.49	1.40
factor(x20)2	-0.14	-0.61	0.54	0.87	-0.60	0.31	0.55	1.37
factor(x20)3	-0.17	-0.72	0.47	0.84	-0.63	0.29	0.53	1.34
factor(x21)1	-0.09	-0.41	0.68	0.91	-0.53	0.34	0.59	1.41
factor(x21)2	-0.52	-2.00	0.05	0.60	-1.02	-0.01	0.36	0.99
factor(x21)3	-0.17	-0.64	0.52	0.84	-0.69	0.35	0.50	1.42
factor(x44)1	1.11	1.49	0.14	3.04	-0.35	2.58	0.70	13.18
factor(x44)2	1.00	1.39	0.16	2.73	-0.41	2.42	0.66	11.20
factor(x44)3	4.10	5.59	0.00	60.61	2.66	5.54	14.36	255.82
factor(x44)4	1.43	1.95	0.05	4.19	-0.01	2.87	0.99	17.65
factor(x44)5	1.7982	1.95	0.05	6.04	-0.004	3.83	0.996	46.02
factor(x49)1	-0.15	-0.52	0.61	0.86	-0.71	0.41	0.49	1.51
factor(x49)2	-0.02	-0.07	0.94	0.98	-0.63	0.58	0.53	1.79
factor(x49) 3	0.24	0.82	0.41	1.27	-0.33	0.81	0.72	2.25

Appendix B: Results from Deviance Test

Coefficient	Residual Deviance	Null Deviance	Chi-Squared Test	P-value
factor(x2)	1523	1529.4	6.4	0.09
factor(x20)	1528.6	1529.4	0.8	0.85
factor(x21)	1524.9	1529.4	4.5	0.21
factor(x44)	1415.6	1529.4	113.8	0
factor(x49)	1415.6	1529.4	113.8	0

Appendix C: Zip Code Map of just Las Vegas



outlying areas = 89124

Appendix D: Description of Variables

Description of the Variables Obtained from the Transportation Data, 18580 Observations		
Variable	Description	Codes/Values
x1	Gender of Responsible Driver	1 = Male, 0 = Female
x2	Age of Responsible Driver	0 = 0-25, 1 = 26-50, 2 = 51-75, 3 = 76-100
x3	Drinking or Drunk while Driving	0 = No Alcohol, 1 = Alcohol is involved
x4	Under Influence of a Drug	0 = No Drugs, 1 = Drugs Involved
x5	Inattention (due to radio, cell phone, etc.)	0 = Not due to inattention, 1 = Due to inattention
x6	Ailment (due to illness)	0 = Feeling Fine, 1 = Feeling Ill, Faint, or Fatigued
x7	Wet Weather	0 = not wet weather, 1 = wet weather
x9	Cloudy Weather	0 = Clear sky, 1 = Cloudy, Smog, or Fog
x10	Dark	1 = Complete Darkness, 0 = Other
x11	Daylight	1 = Daylight, 0 = Other
x12	Non-Daylight	1 = Light Source Not Sun, 0 = Other
x13	Dawn	1 = Dawn, 0 = Other
x14	Dusk	1 = Dusk, 0 = Other
x15	Road Construction	1 = Road Construction, 0 = Other
x16	Road Obstruction	1 = Road Obstructions, 0 = Other
x17	Road Environment	1 = Snow, water, Ice on Road, 0 = Other
x18	Roadway	1 = Road Construction & Debris, 0 = Other

Description of the Variables Obtained from the Transportation Data, 18580 Observations

Variable	Description	Codes/Values
x19	Time of Day	0 = AM, 1 = PM
x20	The Quarter of the Year	0 = 1st Quarter of Year, 1 = 2nd Quarter of Year, 2 = 3rd Quarter of Year, 3 = 4th Quarter of Year
x21	The Quarter of the Month	0 = 1st Quarter of Month, 1 = 2nd Quarter of Month, 2 = 3rd Quarter of Month, 3 = 4th Quarter of Month
x22	Sideswipe	1 = Sideswipe, 0 = Other
x23	Angle	1 = Hit at an Angle, 0 = Other
x24	Rear	1 = Rear Hit, 0 = Other
x25	Head on Collision	1 = Head on Collision, 0 = Other
x26	Inappropriate Lane Change	1 = Lane Change, 0 = Other
x27	Illegal Turn	1 = Improper Turn, 0 = Other
x29	Racing or Speeding	1 = Racing/Speeding, 0 = Other
x30	Going the Wrong Way	1 = Drove Wrong Way, 0 = Other
x31	Passing Other Vehicle	1 = Passing Vehicle, 0 = Other
x32	Backing up	1 = Backing Up, 0 = Other
x33	Disregarded Road Signs	1 = Disregarded Road Signs, 0 = Other
x34	Followed too Closely	1 = Followed too Close, 0 = Other
x35	Failed to Yield	1 = Failed to Yield, 0 = Other
x37	Ran off the Road	1 = Ran off Road, 0 = Other
x38	Hit and Run	1 = Hit and Run, 0 = Other
x40	Driverless Vehicle	1 = Driverless Vehicle, 0 = Other

Description of the Variables Obtained from the Transportation Data, 18580 Observations

Variable	Description	Codes/Values
x41	Mechanical Failure	1 = Mechanical Failure, 0 = Other
x42	Obstructed Visibility	1 = Obstructed Visibility, 0 = Other
x43	Over Evaluated	1 = Over Evaluated, 0 = Other
x44	Vehicle Type	0 = Van, 1 = Truck, 2 = Car, 3 = Motorcycles, 4 = Government Vehicle, 5 = Unusual Vehicle
x46	Airbag Deployment	0 = Airbag Deployed, 1 = Airbag Did Not Deploy
x47	Seatbelt Usage	1 = Seatbelt Used Correctly, 0 = Other
x48	Median Income	0 = Median Income less than $51,000, 1 = Median Income Greater Than or Equal to $51,000
x49	Zip code Population	1 = Less than 19,951, 2 = Between 19,951 and 38,301, 3 = 38,302 and 56,652, 4 = more than 56,652
x51	Less than a HS Diploma	0-100 Percent
x52	Hs Diploma Only	0-100 Percent
x53	Two-Year Degree	0-100 Percent
x54	Four- Year Degree	0-100 Percent
x55	Graduate School	0-100 Percent
x56	Total number of Vehicles involved in Collision	0 = Less than Three Vehicles, 1 = More than or Equal to Three Vehicles

Appendix E: Response Variable, Predictor Variable & Rule Passing

			Pass the Prediction Rule		
Case i	y_i	$\hat{\pi}_i$	Rule 0.007	Rule 0.4	Rule 0.0064
1	0	0.014240	N	Y	N
2	0	0.002143	Y	Y	Y
3	0	0.002143	Y	Y	Y
4	0	0.002143	Y	Y	Y
5	0	0.002143	Y	Y	Y
⋮	⋮	⋮	⋮	⋮	⋮
18575	0	0.004604	Y	Y	Y
18576	0	0.005765	Y	Y	Y
18577	0	0.026400	N	Y	N
18578	0	0.002259	Y	Y	Y
18579	0	0.005921	Y	Y	Y
18580	0	0.010761	N	Y	N

Appendix F: Hosmer-Lemeshow Goodness of Fit Test for Logistic Regression Function

		Number of Non-Fatal Collisions		Number of Fatal Collisions	
Group i	Interval	Observed	Expected	Observed	Expected
1	[9.45e-06,0.000339]	1888	1887.602	0	0.3984134
2	(0.000339,0.000648]	1833	1832.138	0	0.8615453
3	(0.000648,0.0011]	1880	1879.357	1	1.6434339
4	(0.0011,0.00164]	1831	1830.438	2	2.5618473
5	(0.00164,0.0022]	1865	1863.405	2	3.59502
6	(0.0022,0.00276]	1873	1872.4	4	4.5998515
7	(0.00276,0.00412]	1821	1822.732	8	6.2682024
8	(0.00412,0.00644]	1853	1853.279	10	9.7205015
9	(0.00644,0.0116]	1840	1835.47	11	15.5299185
10	(0.0116,0.681]	1768	1775.179	90	82.8212662

Appendix G: Models

[redacted]

```
Number of Fisher Scoring iterations: 15
```

[redacted]

Section 3: BEPE Model

Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)	Signif	Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)	Signif
(Intercept)	-7.05985	1.01236	-6.974	3.09E-12	***	x27	-1.16166	0.29213	-3.977	6.99E-05	***
factor(x2)1	-0.42524	0.21949	-1.937	0.052695	.	x34	-1.72697	0.742	-2.327	0.019941	*
factor(x2)2	-0.19078	0.28494	-0.67	0.503155		x37	0.604	0.36232	1.667	0.095502	.
factor(x2)3	1.14173	0.41504	2.751	0.005943	**	x40	2.22307	0.7597	2.926	0.003431	**
x3	0.91288	0.23447	3.893	9.88E-05	***	x42	0.98249	0.5463	1.798	0.072106	.
x4	1.41549	0.3246	4.361	1.30E-05	***	x43	1.91313	0.47814	4.001	6.30E-05	***
x5	-1.37695	0.6153	-2.238	0.025231	*	factor(x44)1	1.02798	0.75819	1.356	0.175149	
x6	1.0915	0.56642	1.927	0.053978	.	factor(x44)2	0.79294	0.72941	1.087	0.276991	
x10	1.01858	0.46014	2.214	0.026854	*	factor(x44)3	3.56536	0.75376	4.73	2.24E-06	***
x13	0.96971	0.49596	1.955	0.05056	.	factor(x44)4	1.31096	0.74314	1.764	0.07772	.
x14	1.45818	0.44052	3.31	0.000932	***	factor(x44)5	1.68857	0.93821	1.8	0.071895	.
x17	-1.03455	0.63867	-1.62	0.105263		x46	0.73317	0.23059	3.18	0.001475	**
x18	-0.8207	0.26599	-3.085	0.002033	**	x47	-0.64272	0.23256	-2.764	0.005716	**
x22	-1.41483	0.72822	-1.943	0.052034	.	x52	0.09386	0.02549	3.682	0.000231	***
x24	-1.20403	0.31101	-3.871	0.000108	***	x54	0.09739	0.02781	3.502	0.000462	***
x25	0.88209	0.54875	1.607	0.107957		x56	-0.96479	0.23471	-4.111	3.95E-05	***
x26	-1.0078	0.53171	-1.895	0.058039	.						

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 1529.4 on 18579 degrees of freedom
Residual deviance: 1139.1 on 18547 degrees of freedom
AIC: 1205.1
Number of Fisher Scoring iterations: 10

Section 4: All Significant Model

Coefficients	Estimate	Std.Error	zvalue	Pr(>|z|)	Significance
(Intercept)	-5.25856	0.67037	-7.844	4.35E-15	***
x3	0.89277	0.22481	3.971	7.15E-05	***
x4	1.33962	0.31652	4.232	2.31E-05	***
x5	-1.22409	0.60336	-2.029	0.042478	*
x10	0.92194	0.43406	2.124	0.033669	*
x14	1.32248	0.43351	3.051	0.002283	**
x18	-0.77403	0.24241	-3.193	0.001408	**
x22	-2.33038	0.75538	-3.085	0.002035	**
x23	-0.59354	0.25491	-2.328	0.019891	*
x24	-1.81672	0.37594	-4.832	1.35E-06	***
x27	-1.21895	0.28333	-4.302	1.69E-05	***
x34	-1.61862	0.74076	-2.185	0.028882	*
x40	1.91411	0.74777	2.56	0.010475	*
x43	1.59249	0.46759	3.406	0.00066	***
x46	0.49767	0.22233	2.238	0.025192	*
x47	-0.93301	0.22065	-4.229	2.35E-05	***
x52	0.09148	0.02334	3.919	8.90E-05	***
x54	0.10544	0.02706	3.897	9.75E-05	***
x56	-0.95176	0.24583	-3.872	0.000108	***

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1529.4  on 18579  degrees of freedom
Residual deviance: 1244.3  on 18561  degrees of freedom
AIC: 1282.3
Number of Fisher Scoring iterations: 10
```

Appendix H: Correlation Matrix

	x3	x4	x5	x6	x10	x13	x14	x18	x22	x23	x24	x27	x34
x3	1	0.1	-0.04	0	0.05	0.07	-0.01	0.22	-0.01	-0.03	-0.01	-0.06	-0.03
x4	0.1	1	-0.02	0.02	0	0.01	0.01	0.09	0	-0.01	0	-0.04	-0.02
x5	-0.04	-0.02	1	0.15	0.01	0.01	0.03	-0.01	-0.03	-0.14	0.16	-0.08	0
x6	0	0.02	0.15	1	-0.01	0.03	0.01	0.12	0	-0.02	-0.01	-0.04	-0.02
x10	0.05	0	0.01	-0.01	1	-0.01	-0.01	0.06	0	0	-0.03	-0.03	-0.03
x13	0.07	0.01	0.01	0.03	-0.01	1	-0.01	0.05	-0.01	0.01	-0.02	-0.02	-0.01
x14	-0.01	0.01	0.03	0.01	-0.01	-0.01	1	0	0	0.01	-0.01	0	-0.02
x18	0.22	0.09	-0.01	0.12	0.06	0.05	0	1	0.05	0.18	-0.32	-0.1	-0.18
x22	-0.01	0	-0.03	0	0	-0.01	0	0.05	1	-0.25	-0.22	-0.05	-0.11
x23	-0.03	-0.01	-0.14	-0.02	0	0.01	0.01	0.18	-0.25	1	-0.79	0.41	-0.37
x24	-0.01	0	0.16	-0.01	-0.03	-0.02	-0.01	-0.32	-0.22	-0.79	1	-0.38	0.47
x27	-0.06	-0.04	-0.08	-0.04	-0.03	-0.02	0	-0.1	-0.05	0.41	-0.38	1	-0.21
x34	-0.03	-0.02	0	-0.02	-0.03	-0.01	-0.02	-0.18	-0.11	-0.37	0.47	-0.21	1
x37	0.13	0.07	0	0.07	0.04	0.01	0	0.27	-0.01	0.01	-0.09	-0.03	-0.05
x40	0.01	0.02	0	0.01	0	0.03	0	-0.01	-0.01	0.02	-0.02	-0.02	-0.01
x43	0.04	0.01	0.03	0.01	0.07	0	0	0.12	-0.01	0	-0.04	0	-0.03
factor(x44)1	0.03	0.01	-0.01	0.01	-0.02	0.01	0	0	0	-0.02	0.03	0	0.01
factor(x44)2	0.01	-0.01	0.01	-0.01	0.03	-0.01	-0.01	0.01	-0.02	0.03	-0.02	0.01	-0.01
factor(x44)3	0.04	0.01	-0.02	-0.01	0	-0.01	-0.01	0.02	0	0	-0.04	-0.04	-0.01
factor(x44)4	-0.02	0	0	0	0	0	0.02	-0.01	0.01	-0.01	0.01	0	0.01
factor(x44)5	-0.01	0	0.01	0	0	0.01	0	-0.02	0	0	0	-0.01	0.03
x46	0.1	0.01	-0.01	0.05	0.03	0.03	-0.01	0.17	-0.07	0.1	-0.11	0.05	-0.07
x47	-0.17	-0.07	0.03	0	-0.03	-0.03	0	-0.13	-0.02	0	0.04	0.03	0.02
x52	0.01	0.01	0	-0.01	-0.01	0	0	-0.03	0.01	-0.07	0.07	-0.03	0.05
x54	0	0.01	0.02	0.02	0.03	-0.01	0.01	0.03	0	-0.01	0	0	-0.01
x56	-0.14	-0.06	0.03	-0.07	-0.07	-0.03	0.01	-0.3	0.04	-0.03	0.24	0.01	0.13

	x37	x40	x43	factor (x44)1	factor (x44)2	factor (x44)3	factor (x44)4	factor (x44)5	x46	x47	x52	x54	x56
x3	0.13	0.01	0.04	0.03	0.01	0.04	-0.02	-0.01	0.1	-0.17	0.01	0	-0.14
x4	0.07	0.02	0.01	0.01	-0.01	0.01	0	0	0.01	-0.07	0.01	0.01	-0.06
x5	0	0	0.03	-0.01	0.01	-0.02	0	0.01	-0.01	0.03	0	0.02	0.03
x6	0.07	0.01	0.01	0.01	-0.01	-0.01	0	0	0.05	0	-0.01	0.02	-0.07
x10	0.04	0	0.07	-0.02	0.03	0	0	0	0.03	-0.03	-0.01	0.03	-0.07
x13	0.01	0.03	0	0.01	-0.01	-0.01	0	0.01	0.03	-0.03	0	-0.01	-0.03
x14	0	0	0	0	-0.01	-0.01	0.02	0	-0.01	0	0	0.01	0.01
x18	0.27	-0.01	0.12	0	0.01	0.02	-0.01	-0.02	0.17	-0.13	-0.03	0.03	-0.3
x22	-0.01	-0.01	-0.01	0	-0.02	0	0.01	0	-0.07	-0.02	0.01	0	0.04
x23	0.01	0.02	0	-0.02	0.03	0	-0.01	0	0.1	0	-0.07	-0.01	-0.03
x24	-0.09	-0.02	-0.04	0.03	-0.02	-0.04	0.01	0	-0.11	0.04	0.07	0	0.24
x27	-0.03	-0.02	0	0	0.01	-0.04	0	-0.01	0.05	0.03	-0.03	0	0.01
x34	-0.05	-0.01	-0.03	0.01	-0.01	-0.01	0.01	0.03	-0.07	0.02	0.05	-0.01	0.13
x37	1	0.01	0.09	0.02	-0.01	0.03	-0.01	-0.01	0.07	-0.1	-0.02	0.01	-0.3
x40	0.01	1	0	0	-0.01	0	0	0	-0.01	-0.05	0	0	-0.03
x43	0.09	0	1	0	0.01	0	0	0	0.01	-0.02	-0.01	0.01	-0.08
factor(x44)1	0.02	0	0	1	-0.5	-0.05	-0.21	-0.05	-0.04	-0.01	0	-0.03	0
factor(x44)2	-0.01	-0.01	0.01	-0.5	1	-0.14	-0.54	-0.14	0.07	0.02	0.01	0.01	0.01
factor(x44)3	0.03	0	0	-0.05	-0.14	1	-0.06	-0.02	-0.04	-0.09	0	0.01	-0.07
factor(x44)4	-0.01	0	0	-0.21	-0.54	-0.06	1	-0.06	-0.03	0	-0.01	0.03	0
factor(x44)5	-0.01	0	0	-0.05	-0.14	-0.02	-0.06	1	-0.02	-0.01	-0.01	-0.02	-0.01
x46	0.07	-0.01	0.01	-0.04	0.07	-0.04	-0.03	-0.02	1	-0.03	-0.03	0.06	-0.1
x47	-0.1	-0.05	-0.02	-0.01	0.02	-0.09	0	-0.01	-0.03	1	-0.01	0.01	0.1
x52	-0.02	0	-0.01	0	0.01	0	-0.01	-0.01	-0.03	-0.01	1	0.02	0.03
x54	0.01	0	0.01	-0.03	0.01	0.01	0.03	-0.02	0.06	0.01	0.02	1	-0.02
x56	-0.3	-0.03	-0.08	0	0.01	-0.07	0	-0.01	-0.1	0.1	0.03	-0.02	1

Appendix I: Missing Data Analysis

Variables	y	x1	x2	x3	x4	x5	x6	x7	x9	x10	x11
No Missing Data	0.01	0.61	0.97	0.08	0.02	0.09	0.01	0.02	0.11	0.01	0.71
Missing Data	0.00	0.62	0.95	0.08	0.02	0.10	0.01	0.02	0.12	0.02	0.70
Difference	0.01	-0.02	0.03	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01	0.01
Variables	x12	x13	x14	x15	x16	x17	x18	x19	x20	x21	x22
No Missing Data	0.25	0.01	0.02	0.03	0.02	0.04	0.16	0.68	1.40	1.34	0.06
Missing Data	0.25	0.01	0.02	0.05	0.03	0.04	0.18	0.67	1.40	1.34	0.07
Difference	0.00	0.00	0.00	-0.02	-0.01	-0.01	-0.02	0.01	0.00	0.00	-0.01
Variables	x23	x24	x25	x26	x27	x29	x30	x31	x32	x33	x34
No Missing Data	0.47	0.41	0.01	0.07	0.25	0.00	0.00	0.01	0.02	0.06	0.16
Missing Data	0.37	0.46	0.01	0.07	0.21	0.00	0.00	0.01	0.02	0.05	0.13
Difference	0.10	-0.05	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.01	0.03
Variables	x35	x37	x38	x40	x41	x42	x43	x44	x46	x47	x48
No Missing Data	0.21	0.01	0.07	0.00	0.01	0.01	0.01	2.14	0.12	0.92	0.48
Missing Data	0.14	0.02	0.05	0.00	0.01	0.01	0.01	2.15	0.13	0.93	Missing
Difference	0.07	-0.01	0.01	0.00	-0.01	0.00	-0.01	0.00	-0.01	-0.01	N/A
Variables	x49	x50	x51	x52	x53	x54	x55	x56			
No Missing Data	2.57	12.88	14.16	20.63	4.21	8.57	4.91	0.91			
Missing Data	Missing	Missing	Missing	Missing	Missing	Missing	Missing	0.87			
Difference	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.03			

Bibliography

AlcoholAlert. (2010). *200g Drunk Driving Statistics*. Retrieved from AlcoholAlert! Intervention at the Point of Consumption: http://www.alcoholalert.com/drunk-driving-statistics-2006.html

Ascone, D., & Lindsey, T. (2009). *An Examination of Driver Distraction as Recorded in NHTSA Databases.* Washington DC: NHTSA.

Bedard, M., Guyatt, G. H., Stones, M. J., & Hirdes, J. P. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis & Prevention*, 717-727.

Cerrelli, E. C. (1996). *Trends in Daily Traffic Fatalities, 1975-1995.* Washington DC: NHTSA.

City-Data. (2010). Retrieved from Onboard Informatics: http://www.city-data.com/city/Las-Vegas-Nevada.html

Crombie, I. K. (2009, April). What are confidence intervals and p-values? *What is...? series*.

Daniel, W. W. (2009). *Biostatistics: A Foundation for Analysis in the Health Sciences.* Hoboken, NJ: John Wiley & Sons, INC.

Everitt, B. S. (2002). *The Cambridge Dictionary of Statistics.* New York, New York: Cambridge University Press.

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models.* Boca Raton, Florida: Taylor & Francis Group, LLC.

FARS. (2010). *Highway Safety Research & communications*. Retrieved from Insurance Institute for Highway Safety: http://www.iihs.org/research/fatality_facts_2009/statebystate.html

Hawkes, J. S., & Marsh, W. H. (2005). *Discovering Statistics: Second Edition.* Charleston: Hawkes Learning Systems and Quant systems, Inc.

Hilbe, J. M. (2009). *Logistic Regression Models.* Boca Raton: Chapman & Hall/CRC Press.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression: Second Edition.* Danvers, Massachusetts: John Wiley & Sons, Inc.

Hunter, M. (2006). *Idaho Transportation Department*. Retrieved from Idaho Government: http://itd.idaho.gov/ohs/2006data/06Units.pdf

Kostyniuk, L., & Zakrajsek, J. (2002). *Identifying Unsafe Driver Actions*. Retrieved from AAA Foundation for Traffic Safety: www.aaafoundation.org

Kumar, R., & Indrayan, A. (2011, April). Reciever Operating Characteristic (ROC) Curve for Medical Researchers. *Indian Pediatrics, 48*, 277-287.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied Linear Regression Models: Fourth Edition.* New York: McGraw-Hill Irwin.

Lum, H., & Reagan, J. A. (1995). Interactive Highway Design Model: Accident Predictive Module. *Public Roads Magazine*, 14-17.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models: Second Edition.* Hoboken, New Jersey: John Wiley & Sons.

NHTSA. (2005). *Trend and Pattern Analysis of Highway Crash Fatalities by Month and Day*. Washington DC: National Center for Statistics and Analysis.

NHTSA. (2010). *Fatality Analysis Reporting System*. Retrieved from National Highway Traffic Safety Administration: http://www.nhtsa.gov/

R Development Core Team. (2010). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, 2.11.1*. Vienna, Austria.

Streff, F. M., & Wagenaar, A. C. (1989). Are there reakky shortcuts? Estimating seat belt use with self-reporting measures. *Accident Analysis and Prevention*, 509-516.

Subramanian, R. (2006). *Passenger Vehicle Occupant Fatality Rates by Type and Size of Vehicle.* Washington DC: NHTSA.

Upton, G., & Cook, I. (2008). *Oxford Dictionary of Statistics.* New York: Oxfor University Press.

Williams, A. F. (1985). Nighttime driving and fatal crash involvement of teenagers. *Accident Analysis and Prevention, 17*, 1-5.

Young, K., & Regan, M. (2007). Driver Distraction: A Review of the Literature. *Australasian College of Road Safety*, 379-405.

VITA

Graduate College

University of Nevada, Las Vega

Annabelle Mathis

Degree:

Bachelor of Science in Mathematics and Secondary Education, 2001

Concordia University of WI, Mequon, WI

Thesis Title: Statistical Analysis of Fatalities Due To Vehicle Accidents in Las Vegas, NV

Thesis Examination Committee:

Chairperson: Chih-Hsiang Ho, Ph.D.

Committee Member: Kaushik Ghosh Ph.D.

Committee Member: Sandra Catlin Ph.D.

Committee Member, Anton Westveld, Ph.D.

Graduate Faculty Representative, Chad Cross, Ph.D., PStat(R), NCC, MAC, SAP, CCH, LCADC, MFT