1-1-2008

Use of single vehicle collisions to model fatigue-related crashes on rural two-lane highways

Xu Lin
Ryerson University

USE OF SINGLE VEHICLE COLLISIONS TO MODEL FATIGUE-RELATED CRASHES ON RURAL TWO-LANE HIGHWAYS

By

XU LIN

Bachelor of Science, Chuang Chun, China, July, 2000

Master of Science, Kuala Lumpur, Malaysia, April 2003

A thesis presented to
Ryerson University
in partial fulfillment of the
requirements for the degree of
Master of Applied Science
in the Program of
Civil Engineering

Toronto, Ontario, Canada, 2008

Author's Declaration

I hereby declare that I am the sole author of this thesis.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Signature

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signature

Borrower

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Name	Signature of Borrower	Address	Date

ABSTRACT

Xu Lin

Thesis Title: Use of Single Vehicle Collisions to Model Fatigue-Related Crashes on Rural Two-Lane Highways

Master of Applied Science, Civil Engineering, Ryerson University, 2007

Fatigue-related crashes are believed to be more common on rural highways than on urban roads and on two-lane roads than on other rural road types. Thus an understanding of how design factors affect fatigue-related crashes on rural two-lane roads is vital. The problem is that fatigue is rarely reported as a cause of crashes, since it is rarely suspected by the police as a possible cause and since potential liability may motivate the drivers not to reveal the real causes of the crash. Thus, getting a handle on these crashes through modeling is a formidable challenge. Fortunately, there is research to suggest that single-vehicle run-off-road crashes, particularly those during periods of low circadian rhythm, can be used as a reasonable surrogate in modeling fatigue-related crashes. The paper is based on research to examine how fatigue-related crashes rural on two-lane roads, as represented by single vehicle crashes, are affected by various engineering design factors. This study's goal is to explore the effects of fatigue on driving on rural two-lane roads in North America, and to consider how we can work towards mitigating the effects of fatigue on traffic safety. For this investigation, generalized linear and logistic regression modelling were used on US Highway Safety Information System (HSIS) data

for Ohio. Models were developed separately and combined for periods of high and low circadian rhythm and for single-vehicle run-off-road and other crashes. The results show, for example, that after controlling for traffic volumes, increases in speed limit, average curvature and average gradient and decreases in surface width and average shoulder width were found to be associated with increased fatigue related crashes. Important differences were found in the effects of factors for periods of low and high circadian rhythm.

ACKNOWLEDGMENTS

I am deeply indebted to my supervisor Prof. Dr. Bhagwant N. Persaud whose help, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis. My sincere thanks also go to Dr. Khaled Sennah and Dr. Ali Mekky who were my committee members in this research.

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I want to thank the Highway Safety Information System (HSIS) office for giving me permission to use their database to extract data for this research. For its financial support, I wish to thank the Operating Grant received from Canadian Institute for Health research, Team Planning & Development Grants – Toward Enhanced Quality of Life through Injury Prevention, Acute Response and Rehabilitation, administered through the Institute of Population and Public Health.

My final thanks go to my dearest parents – LIN, HAI BO and ZHOU, FENG who are encouraging me all the time.

Table of Contents

List of Tables

List of Figures

Notation and Abbreviations

AAADT –variable representing the average value of AADT in five years per segment

AADT –Average Annual Daily Traffic

AASHTO –American Association of State Highway and Traffic Officials

ADT –Average Daily Traffic

AIC -Akaike's information criterion

AGE_GROUP–variable representing the Driver Age Group for each traffic crash

AVG_OUTSH –variable representing the average outside shoulder width per segment

CURV_HI –variable representing the average curvature for each segment

DOF –Degrees of Freedom

DRV_SEX–variable representing the Driver Sex for each traffic crash

FR –Fatigue-Related

GLM –Generalized Linear Model

GRAD_HI –variable representing the average gradient for each segment

LIGHT –variable representing the Light Condition for each traffic crash

NFR –Non-Fatigue-Related

NSP-FR –Non-Sleepy-Period Fatigue-Related

RD_CHAR–variable representing the Road Characteristic for each traffic crash

RD_WIDTH–variable representing the Road Width for each traffic crash

SAS –Statistical Analysis Software

SPDLIMT –variable representing the posted speed limit per segment

SPD_LIMIT–variable representing the Speed Limit for each traffic crash

SPFR – Sleepy-Period Fatigue-Related

SP-FR – Sleepy-Period Fatigue-Related

SPNFR – Sleepy-Period Non-Fatigue-Related

SURF_WID –variable representing the roadway surface width per segment

1 INTRODUCTION

This thesis is concerned with the problem of fatigued driving, and how transportation safety engineers can use an understanding of the road and traffic factors that contribute to fatigue related crashes to mitigate the effects of fatigue on our roads. If we can improve our understanding of fatigued driving, we may be able to develop countermeasures that can lead to significant improvements in road safety.

1.1 Background: Driver Fatigue

Fatigue, which refers to a phenomenon of both physiology and psychology, is a general term commonly used to describe the experience of being "sleepy", "tired" or "exhausted" (RTA, 2001). Driver fatigue can severely impair judgment and reaction, and can affect any person operating a motor-vehicle, leading to potentially dangerous situations.

A number of symptoms can suggest driver fatigue. Typical symptoms consist of yawning, poor concentration, tired or sore eyes, restlessness, drowsiness, slow reactions, boredom, feeling irritable, making fewer and larger steering corrections, missing road signs, having difficulty in staying in the lane, and micro-sleeps. Driver fatigue is not only a function of time spent on driving, but also corresponds to many other factors such as hours since last slept (hours of wakefulness) and time of day or night.

Fatigue-related crashes at certain times of the day may coincide with dips in the body's circadian rhythms (http://roadsafetydirectory.com). Circadian rhythms are physiological cycles that follow a daily pattern, and program us to sleep at night and to be awake during the day. During nighttime hours, and to a lesser extent during afternoon "siesta" hours, most types of human performance, including the ability to drive, may be significantly impaired,

According to research by the Mayo Foundation for Medical Education and Research (1995), the human body has more than 100 circadian rhythms. During each 24-hour cycle, these rhythms influence numerous aspects of the human body's function, including body

temperature, hormone levels, heart rate, blood pressure, and even pain threshold, as shown in Figure 1.1. It is natural that humans sleep when tired and awake when rested, but the underlying pattern of fatigue and alertness follow a circadian rhythm.

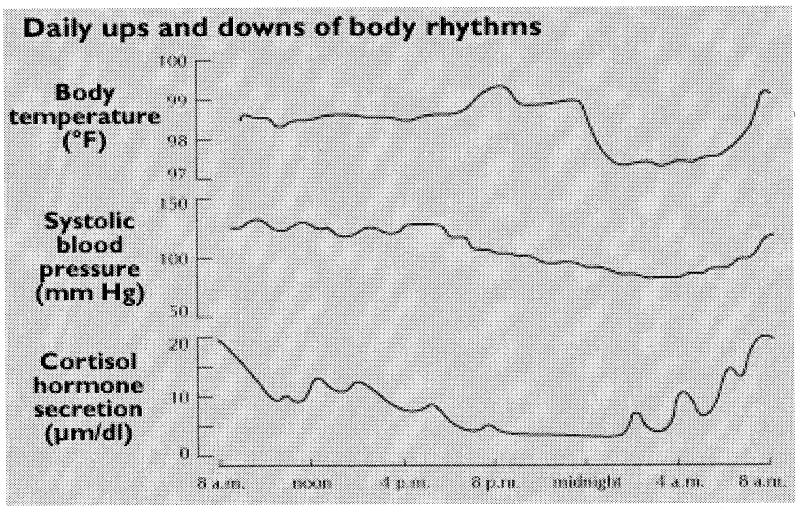

Figure 1.1 Body Circadian Rhythms (Mayo Foundation for Medical Education and Research, 1995)

Circadian rhythms are very powerful: circadian rhythms cannot be reversed even for those working nightshifts for many years (NATA, 2006). If natural sleep cycles are disrupted (e.g., staying awake during the night, not getting enough sleep, or poor quality of sleep), we easily become fatigued.

The problem of fatigue, and the problem of driving while fatigued, is obvious not only because fatigue makes us feel tired, but also because fatigue affects the body in many ways (RTA, 2001), including slower reaction time, loss in concentration, and fatal 'micro-sleeps.' Fatigue makes it more difficult to recognize how tired we are, and therefore more difficult to recognize when fatigue is setting in while driving.

1.2 Problem Statement

Fatigue is a more common cause of traffic crashes than alcohol or prescription drugs (Flemons, 1999). The estimation of the proportion of crashes attributable to driver fatigue

varies from 5 percent to 50 percent. The difficulty in measuring fatigue contributes to this variation, but most experts estimate that 20 to 30 percent of fatal road crashes could result from driver fatigue (HORSCOCTA, 2000). The identification of fatigue-related traffic crashes is based on police reports, but it is acknowledged that these reports underestimate the incidence of fatigue as police officers do not judge fatigue to be a contributory factor unless there is strong evidence that the driver was asleep at the time of the crash (Haworth, 1998a).

Driving requires giving continuous attention to the driving environment, conducting complex dynamic tasks, and searching for and detecting potential hazards. Fatigue is both a physiological and a psychological state which can undermine a driver's performance should fatigue set in while a driver is operating a motor vehicle. Perception of hazards may be affected by driver reaction (Haworth, 2006), which is impaired by driver fatigue. With repeated exposure to stressful driving environment over the course of a journey, fatigue symptoms may develop (Lal and Craig, 2001). While driver fatigue has been conceptualized as a consequence of circadian disruption and sleep deprivation which can lead to reduced alertness and impaired performance, driver stress has not been linked with circadian disruption or sleep deprivation (Taylor and Dorn, 2006). Fatigue-related crashes are complex and driver fatigue has been widely identified as a major cause of serious crashes through reduced driving performance efficiency (Akerstedt, 2000).

The risk of being involved in road crashes is increased by fatigue or falling asleep at the wheel (Tzamalouka et al., 2005). Although fatigue is a well-known risk factor for traffic crashes, many drivers drive while fatigued (RTA, 2001). An understanding of how road design factors might affect fatigue-related crashes is becoming a key issue to traffic engineers in the prevention and mitigation of such traffic crashes.

1.3 Significance of this Study

In recent years, there has been much interest in the role of driver impairment in the causation of road crashes. Recognition of the role of alcohol in driver impairment and efforts taken to reduce the incidence of alcohol impairment have played an important part

in reducing the number of road injuries and fatalities in the past decades (Haworth, 1998a). An understanding of the relationship between driver fatigue and crashes has the potential to lead to further improvements in road safety.

As rural roadways tend to have high crash rates, adequate crash models for these roadways are especially desirable (Vogt and Bared, 1998). Fatigue-related crashes are more common on rural highways than on urban roads (Haworth, 1998b). One of the reasons for this is that average trip lengths are likely to be longer on rural roads, and inattention and drowsiness are brought on by the constant speeds and monotony. In Canada, rural roads account for approximately 40 percent of all motor vehicle travel, but 60 percent of all fatal crashes (NHTSA, 2003a). Approximately 90 percent of all rural fatal crashes occur on rural two-lane roads (NHTSA, 2003b).

Rural two-lane roads in North America generally lack physical measures such as wide medians or barriers to separate opposing traffic flows (Persaud et al., 2004). Fatigue related (FR) crashes on rural two-lane roads tend to be severe, possibly because of the higher speed involved, a lack of escape options such as wide paved shoulders on many roads, and driver inability to take appropriate and timely avoiding action, such as safe braking, due to the driver's fatigued state.

Research in Canada on the problem of fatigued driving is limited, but there is considerable research in other countries where the problem of sleepiness or falling asleep at the wheel has long been recognized and studied in detail (Maycock, 1997; Mccartt et al., 1996). There is, however, very little research on how engineering design factors contribute to FR traffic crashes on rural two-lane roads in North America. An understanding of the role played by highway design factors in FR crashes will help jurisdictions to improve road safety.

Given that rural two-lane roads in North America tend to have high crash rates, and that there is a need to understand the relationship between FR traffic crashes and highway design factors, it seems appropriate to study the factors that affect the incidence of FR

traffic crashes on rural two-lane roads. This study considers a variety of engineering, environmental, and driver factors, but concentrates on engineering factors.

The findings of the study can be applied to the design of engineering countermeasures that can mitigate FR traffic crashes. Such a study can develop FR crash models that will give highway engineers a clear understanding of the relationship between engineering factors and FR crashes on rural two-lane roads.

1.4 Study Objectives

This study's goal is to explore the effects of fatigue on driving on rural two-lane roads in North America, and to consider how we can work towards mitigating the effects of fatigue on traffic safety. The study's objectives are:

1. Identify road design, traffic, environmental, and demographic factors that contribute to fatigue-related (FR) and non fatigue-related (NFR) crashes on rural two-lane roads.
2. Identify road design and traffic factors that contribute to Sleepy Period FR (SPFR), and Sleepy Period NFR (SPNFR) crashes. The sleepy periods are identified from research into circadian rhythms.
3. Develop safety performance functions (SPFs) to predict FR crashes on rural two-lane roads, and apply the SPFs to the task of mitigating the effects of fatigue on traffic safety at sites where the potential for improving fatigue crashes related to fatigue is greatest. This objective involves:
    - using the SPFs to conduct a network screening exercise designed to identify sites where the rate of FR crashes suggests a high potential for FR safety improvement;
    - identifying and ranking the sites by their potential for FR safety improvement; and
    - identifying the effects of road design and traffic factors on sites found to have a high potential for FR safety improvement.

Two modelling approaches are used. The first approach is generalized linear modelling (GLM). GLM techniques were used to relate the fatigue crash rate to important road and

traffic variables. The second approach is logistic regression modeling. Logistic regression models were used to analyze FR crash occurrence, and to identify important road, traffic, environmental, and driver variables.

1.5 Study Scope and Limitations

Unlike alcohol, fatigue is difficult to measure. There is no objective test for determining the level of fatigue or sleepiness of drivers involved in crashes (Connor et al., 2000), and there are no definitive criteria for establishing the level of fatigue that leads to crashes.

Crash outcomes may complicate the identification of fatigue-related crashes. In fatal crashes, there may be no surviving witnesses to give an account of the crash, or the surviving driver may be influenced and inhibited by possible legal consequences of the crash. The crash itself is likely to alter the driver's level of arousal, and may eliminate any evidence of impairment due to fatigue. The contribution of fatigue as a cause for road crashes cannot be estimated because crashes rarely leave any sign to indicate that the driver had been impaired by fatigue at the wheel.

Crash data collection methods also complicate the identification of FR crashes. Canada has no official crash-reporting document on which the driver, soon after the crash, could report his or her point of view about the cause of their crash.

The absence of an official document means that fatigue is rarely reported as a cause of road crashes. As there is no official data on the prevalence of fatigued driving, no "material" for scientific analysis is available. It is, therefore, necessary to consider alternative approaches to the collection of data for this study.

The Horne and Reyner (1995) study has already been mentioned. Horne and Reyner (1995) found that a crash in which a single vehicle drove off the road was frequently related to fatigue and sleeping at the wheel. A recent FHWA study (2007) found that FR crashes tend to be single-vehicle crashes in which a car or truck leaves the roadway and then turns over or hits a fixed object. The findings of these two studies suggest that it is

reasonable to use single-vehicle run-off-road crashes as a substitute for FR crashes in the proposed research.

This study focuses only on single vehicle run-off-road traffic crashes on rural two-lane highways. Other types of traffic crash which may result from driver fatigue are not included. The effect of this omission is unknown.

1.6 Outline of the Thesis

This thesis consists of six chapters:

CHAPTER 1 presents the background, problem statement, significance of the study, study objectives, and study scope and limitations.

CHAPTER 2 presents the literature review relevant to this study. Various factors influencing traffic crashes are reviewed. The Chapter also provides a review of the Generalized Linear Models (GLM) Regression and the Logistic Regression approaches used in this study to model FR traffic crashes.

CHAPTER 3 presents the data collection, data editing, and data characteristics used for the GLM analysis

CHAPTER 4 presents the GLM methodology, analysis, and results for FR, NFR, SPFR, and SPNFR traffic crashes.

CHAPTER 5 presents the logistic regression methodology, analysis, and results.

CHAPTER 6 presents a discussion and conclusions, with recommendations for further study.

2 LITERATURE REVIEW

2.1 Introduction

In order to better understand the issues related to this research into FR crashes, a literature review was conducted. Numerous resources were consulted including the Transportation Research Record, Accident Analysis and Prevention, and sources identified in the Transportation Research Information System.

The main purpose of the literature review was to gain insight on how FR crashes relate to various roadway design factors, and how to best analyze this type of crash. Motor vehicle crashes are thought to be caused by a combination of factors such as driver characteristics (attention, mood, eyesight, reaction times, driving skills, etc.), roadway characteristics (sight distance, pavement surface, roadway alignment, signing and striping, traffic control, roadside environment, etc.), environmental factors (weather conditions, visibility, wind, etc.), and traffic characteristic (traffic volume, traffic combination, etc.).

The literature review concentrated on the factors that influence FR crashes, as these were the major focus of the study. Special emphasis was also given to statistical methods in traffic crash analysis.

This Chapter presents the following sections:

- Roadway features and traffic crashes (Section **Error! Reference source not found.**);
- Roadway features and rural two-lane traffic crashes (Section **Error! Reference source not found.**);
- Time of day (Section **Error! Reference source not found.**);
- Single vehicle run-off-road crashes (Section **Error! Reference source not found.**);
- Age and sex effects related to FR traffic crashes (Section **Error! Reference source not found.**);;
- Criteria for Identification of FR Crashes in the United Kingdom and United States (Section 2.7);
- GLM models (Section **Error! Reference source not found.**);

- Traffic crashes and GLMs (Section **Error! Reference source not found.**);
- Logistic regression analysis (Section **Error! Reference source not found.**);
- Traffic crashes and logistic regression models (Section 2.11); and
- Summary of literature review findings (Section 2.12).

2.2 Roadway Features and Traffic Crashes

Roadway conditions are an important factor in determining the cause of a traffic crash. For example, a roadway's condition (quality of pavements, presence of shoulders, and intersections, traffic control devices, etc) can be a factor in a crash (Garber and Hoel, 2001). Inadequate traffic control and complex intersections with excessive signage can lead to confusion, and may also be factors in a crash. It is vital that the roadway design allows enough perception-reaction time. Sight distances must be adequate for the design speed, and road alignments should be appropriate to a driver's decision sight distance and provide enough time. Superelevation on highways, especially on-ramps, should be designed carefully with correct radii and appropriate transition zones for vehicles to negotiate curves safely. Another factor is frictional force between the pavement and tires. Frictional force is important in maintaining vehicle stability.

Numerous studies have investigated the relationships between traffic crashes and the geometric design of roadways. These studies have indicated that improvements to highway geometric design can significantly reduce the number of vehicular crashes.

For example, Miaou et al. (1992) established empirical relationships between truck crashes and key highway geometric design components by using a Poisson regression approach. The researchers found that annual average daily traffic (AADT) per lane, horizontal curvature, and vertical grade were significantly correlated with truck crash rates.

Shankar et al. (1995) estimated overall crash models to evaluate the effects of roadway geometric variables and environmental factors on crash frequencies. The study concluded that separate regression models for specific types of crashes have the potential to provide

greater explanatory power than can single overall crash frequency models. Because of the suspected heterogeneity in underlying causal mechanisms associated with different crash types, the authors found that it is reasonable to expect that the probabilities of crash occurrence by crash type are associated with roadway, traffic, and environmental factors in different ways.

Curves have been shown to contribute to crashes, whether they be horizontal curves or vertical curves. Various elements of curves may affect the likelihood of a crash. Zegeer et al. (1991) conducted a study called Cost Effective Geometric Improvements for Safety Upgrading of Horizontal Curves. The study was conducted in Washington State, and identified the crash types that are more strongly associated with curves than with adjacent straight-aways. The researchers found that curves were associated with a higher percentage of fatal crashes, head-on crashes, opposite sideswipe crashes, fixed-object and rollover crashes, nighttime crashes, and crashes involving drinking drivers. Vertical curves have also been associated with higher crash rates, but to a less extent than horizontal curves. An important design element regarding vertical curve safety is the need to provide drivers with adequate stopping sight distance.

A study by Choueiri et al. (1994) showed a negative relationship between radius of curve and accident rate, meaning that the smaller the radius, the more accidents occurred. When there is space available, large radii should be used on horizontal curves. Once radii became greater than 400 m to 500 m, the marginal increase in safety per increase in radius is, however, very low.

Joshua and Garber (1990) studied the relationship between highway geometric factors and truck crashes in Virginia using both linear and Poisson regression models. The paper presented mathematical relationships, obtained through multiple linear and Poisson regression analyses, and related the number of truck involved crashes per year on a section of highway to the highway's traffic and geometric variables. The models showed that the slope change rate (absolute curve of slope changes in the vertical direction divided by the highway segment), the average daily traffic, and the difference in speed

between trucks and non-trucks influence the number of truck-involved crashes on a given stretch of highway.

Hadi et al. (1995) used a negative binomial regression analysis to estimate the effects of cross-sectional design elements. They found that increasing lane width, shoulder width, centre shoulder width, and median width were significant in reducing crashes.

Caliendoa et al. (2007) presented a study of crash-prediction models for multilane roads. Crash-prediction models for a four-lane median-divided Italian motorway were developed using crash data observed during a 5-year monitoring period (1999 to 2003). The Poisson, negative binomial, and negative multinomial regression models, were applied separately to tangents and curves, and were used to model the frequency of crash occurrence. Model parameters were estimated by the Maximum Likelihood Method, and the Generalized Likelihood Ratio Test was applied to detect the significant variables that were to be included in the model equation. Goodness of fit was measured by means of both the explained fraction of total variation and the explained fraction of systematic variation. The candidate set of explanatory variables was: length (L), curvature (1/R), annual average daily traffic, sight distance (SD), side friction coefficient (SFC), longitudinal slope (LS), and the presence of a junction (J). Separate prediction models for total crashes and for fatal and injury crashes were developed. For curves, the significant variables were L, 1/R, and AADT. For tangents, the significant variables were L, AADT, and junctions.

Ivan and O'Mara (1997) used crash data from the Connecticut Department of Transportation and Poisson regression for the prediction of traffic crashes. The model showed that a highway's posted speed limit and AADT were critical crash prediction variables. The authors preferred the Poisson regression model to the linear regression model.

2.3 Roadway Features and Rural Two-Lane Traffic Crashes

Although urban areas experience the highest rates of motor vehicle crashes (Insurance

Research Council, 2001), fatal crashes are more likely to occur in rural areas. Approximately 90 percent of all rural fatal crashes occur on two-lane highways (NHTSA, 2003b). Agent et al. (2001) reported that fatal crash rates for rural two-lane highways in Kentucky were approximately twice the overall fatal crash rate for all state maintained roads.

Many factors contribute to the higher fatal crash rates in rural areas including higher traffic speeds, which can increase crash severity (Joksch, 1993), lower rates of seat belt use (NHTSA, 1995), and longer response times for emergency medical assistance (NHTSA, 2002). Roadway design is another important factor.

Qin et al. (2003) studied the selection of exposure measures for crash rate prediction for two-lane highway segments. The paper investigated crash and physical characteristics data for highway segments in Michigan using data from the Highway Safety Information System (HSIS). They found that the relationship between crashes and daily volume (AADT) was non-linear, varied by crash type, and was significantly different from the relationship between crashes and segment length for all crash types.

Persaud et al. (2000) presented one of the earliest studies for separate analyses of curves and tangents on rural two-lane roads. The dependent variable was crash frequency, and the independent variables included traffic flow and road geometry. Regression models were calibrated using GLM. A dummy variable for "flat" or "undulating" terrain was also used. For curves, crash frequency was found to increase with AADT, section length, and curvature (1/R). For tangents, the number of crashes per year increased with AADT and section length. The results also showed a higher crash frequency on undulating terrain than that on flat terrain.

2.4 Time of Day

Human beings have a neurobiological cycle based on a circadian rhythm (EPDFS, 1997). Research by Hartley et al. (2000) has shown that there are two periods during the 24 hour circadian cycle where the level of sleepiness is high. The first period is during the night

and early morning, and the second period is in the afternoon.

During these periods of sleepiness, functions such as alertness, performance, and subjective mood are degraded (Rosekind, 1999). Crash risk reaches its highest level in the early hours of the morning, with a secondary peak in the early afternoon. The secondary peak corresponds to the "post-lunch dip" (Folkard, 1997; Hartley et al., 2000).

Pack et al. (1995) conducted research into the effects of circadian rhythms on traffic crashes. They used North Carolina traffic crash data, and found that FR crashes corresponded to circadian variation in sleepiness, with a major crash peak during the night and a secondary crashes peak in the mid afternoon.

The loss of alertness and degraded performance of 80 male truck drivers from the United States and Canada were examined by Wylie et al. (1996). The study used continuous video to monitor a driver's face for eyelid droop and other fatigue induced facial expressions. The results showed that drowsiness peaked between late evening and dawn. The study also found that "time of day" was the most consistent factor influencing driver fatigue.

In an attempt to develop an algorithm to identify crashes attributable to driver fatigue, Chipman and Jin (2007) analyzed single-vehicle crashes using Ontario data (1999 to 2004). Crashes occurring at times of low circadian rhythm (2 a.m. to 5 a.m. and 2 a.m. to 4 p.m.) were compared with crashes occurring at times when light conditions were similar, but circadian rhythm was higher (9 p.m. to 11 p.m. and 10 a.m.- noon). Logistic regression was used to predict which single-vehicle crashes would occur at times of low circadian rhythm (when fatigue is more likely). The initial results indicated many circumstance associated with crash occurrence at these times of low circadian rhythm including the age and sex of the driver and reported driver condition as well as weather.

2.5 Single Vehicle Run-off-Road Traffic Crashes

Studies of the effects of design factors on crash frequency have provided many insights

(Hildebrand et al., 2007; Leggett, 1988), but studies of the factors that influence single vehicle run-off-roadway crashes have been less successful. Although US national statistics indicate that about one-third of fatal traffic crashes are associated with vehicles running off the road (Vogt and Bared, 1998), little attention has been given to the relationships between single-vehicle run-off-road crash frequency and road features. The statistics for run-off-roadway crashes indicate a continued need for research to develop cost-effective ways to reduce single-vehicle run-off-roadway crash frequency.

Hildebrand et al. (2007) set out to develop a better understanding of the relationship between the frequency/severity of single-vehicle run-off-road collisions and certain geometric/operational characteristics of the corresponding highway sections. The study found a strong relationship between collision rates and the width of the clear zone provided (CZP).

In a Tasmanian study, Leggett (1988) found that the contribution of fatigue increased with crash severity, mainly because of the often fatal nature of single-vehicle run-off-road crashes.

2.6 Age and Sex Effect Related to FR Traffic Crashes

Knipling and Wang (1994) examined United States crash statistics between the years of 1989 and 1993, and found that the age and sex of drivers were strongly related to involvement in FR crashes. In 1990, 77 per cent of fatigued drivers were male and 62 per cent of fatigued drivers were under 30 years of age. When comparing vehicle kilometres travelled, male drivers were nearly twice more likely to be involved in fatigue-related crashes than female drivers. Drivers under 30 years of age were four times more likely to be involved in fatigue crashes than drivers over 30 years of age.

2.7 Criteria of Identification of FR Crashes in the United Kingdom and United States

In the United Kingdom, Horne and Reyner (1995) identified FR crashes by using the following criteria:

- Vehicle ran off the road and/or collided with another vehicle or object;
- Absence of skid marks or braking;
- Driver only saw the point of run-off or the object hit just prior to the crash;
- Witnesses reported lane drifting prior to the crash; and
- Other possible causes of the crash, e.g., mechanical defect, speeding, excess alcohol, bad weather, etc., were excluded.

Similarly, in the United States, the Expert Panel on Driver Fatigue and Sleepiness (1997) selected the following criteria to characterize a FR crash:

- Occurred late at night, early morning, or mid-afternoon;
- Resulted in higher than expected severity;
- Involved a single-vehicle leaving the roadway;
- Occurred on a high speed road;
- Driver did not attempt to avoid the crash; and
- Driver was the sole occupant in the vehicle.

2.8 GLM Model

Regression methods have become an integral component of any data analysis concerned with the relationship between a response variable and one or more explanatory variables. A generalized linear model (McCullagh and Nelder, 1989) is described in terms of the following sequence of assumptions:

- There is a response variable, y, of interest and stimulus variables x_1, x_2, *etc.*, whose values influence the distribution of the response variable.
- The stimulus variables influence the distribution of y through a single linear function only. This linear function is called the linear predictor, and is usually written as:

$$\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots\ldots + \beta_p x_p \qquad (2.1)$$

Hence x_i has no influence on the distribution of y if and only if $\beta_i = 0$

- The distribution of y has the following form:

$$f_Y(y;\mu,\varphi)=\exp\left[\frac{A}{\varphi}\{y\lambda(\mu)-\gamma\lambda(\mu)+\tau(y,\varphi)\}\right] \qquad (2.2)$$

Where φ is a scale parameter (possibly known) and is constant for all observations, *'A'* represents a prior weight, which is assumed to be known, but which may possibly vary within the observations, and μ is the mean of y. It is assumed that the distribution of y is determined by its mean and possibly a scale parameter as well.

- The mean, μ, is a smooth invertible function of the linear predictor

$$\mu = \mathrm{m}(\eta) \qquad \eta = \mathrm{m}^{-1}(\mu) = \iota(\mu) \qquad (2.3)$$

And this inverse function $\iota(\mu)$ is called the link function.

These assumptions are loose enough to encompass a wide class of useful models in statistical practice, but tight enough to allow the development of a unified methodology of estimation and inference, at least approximately.

GLMs represent a unifying framework which includes classical regression models with normally distributed dependent variables, categorical regression models like logistic regression or Poisson regression, and various other nonstandard regression type models. A main feature of GLMs is the presence of a linear predictor which is built from explanatory variables. This linear predictor is linked to the mean response by the so-called link function, which may have various forms.

Many linear regression ideas carry over to the wider class of GLM models: likelihood based on inference techniques, especially estimation of parameters; goodness of fit tests; tests for the significance of covariates; and diagnostic tools. Extensions to quasi likelihood models are described where only first and second moments of the response variable are specified. Tools for inference for the multi-categorical case are derived and applied.

GLMs are generally most useful when used in an exploratory way. They are a powerful tool for experimenting when fitting various models to a great variety of types of data. SAS can be is used for the statistical analysis. SAS is a primary and powerful software design tool for analyzing data when the main variable and response variable of interest are one-dimensional univariates (SAS, 1998a). SAS can be used to fit regression, logistic regression, and the analysis of variance or covariance models.

The SAS GENMOD Procedure for GLM is largely based on the assumption that there is a random sample of independent observations among which the variations are to be modeled. The observations, however, may not be statistically independent. In this case, it is still possible to use SAS for the analysis. The observations cannot be assumed to be statistically independent when the likelihood of the observations can be expressed as a product of two functions, the first observation being a function of the data alone, and the second observation having the form of likelihood of the independent observations. A very important situation that involves non-independent observations occurs when the data consist of frequencies in a multi-way contingency table. For such data, in the form of count, log-linear models are often the most appropriate. SAS can be used for this analysis, and makes the fitting of such models relatively straightforward.

The SAS specification requires the following:

- the term included in the linear predictor GLM;
- the link function connecting the linear predictor to the theoretical value (log function); and
- the distribution of the random component---negative binomial distribution.

Poisson and Negative Binomial Models

Miaou and Lump (1993) suggested using Poisson regression as an initial step in the modelling effort, with the negative binomial model then being applied where appropriate. For the Poisson regression model, the probability of section i having y_i crashes per year (where y_i is a non-negative integer) has the following form (Washington et al., 2002):

Mass Function: $$P(Y = y \mid X_1, X_2, X_3) = \frac{e^{-\mu(X)}[\mu(X)]^y}{y!} \qquad y = 0,1,2,... \tag{2.4}$$

Link Function: $g(\mu) = \log(\mu)$ (2.5)

Systematic Component:

$$g(\mu) = \log(\mu) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n$$
$$\Rightarrow \quad \mu = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n) \tag{2.6}$$
$$n = 0,1,2,...$$

The mean parameter: $E\left[\frac{y_i}{x_i}\right] = \mu = \exp(\beta x_i)$, $Variance = \mu$ (2.7)

where y_i= a random variable representing number of crashes or crash rate,

x_i= parameter which is related to the occurrence of crash (vector of explanatory variable)

β = the coefficient of the corresponding factor (vector of estimable parameter).

n = the number of observations

The Pearson residuals are obtained by computing:

$$e_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(Y_i)}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} = \frac{\text{observed-fitted}}{\sqrt{\text{fitted}}} \qquad \chi^2 = \sum e_i^2 \tag{2.8}$$

Under the hypothesis that the model is adequate, χ^2 is approximately chi-square with n-p degrees of freedom, where p = the number of model parameters.

Negative binomial regression has the following form (Cameroon and Trivedi, 1998):

Mass Function:

$$P(Y = y \mid X_1, X_2, X_3, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y)}\left(\frac{k}{k+\mu}\right)^k\left(\frac{\mu}{k+\mu}\right)^y \qquad y = 0,1,2,... \tag{2.9}$$

$$E(Y) = \mu \tag{2.10}$$

$$Var(y_i) = \mu + (\frac{\mu^2}{k}) \tag{2.11}$$

Link Function: $g(\mu) = \log(\mu)$ (2.12)

Systematic Component:

$$g(\mu) = \log(\mu) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$
$$\Rightarrow \quad \mu = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3) = \exp(x'\beta) \qquad (2.13)$$
$$\left(x' = [1 \quad X_1 \quad X_2 \quad X_3]\right)$$

k is the overdispersion parameter.

Goodness of fit for this model is based on Pearson Residuals, and is obtained with:

$$e_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(Y_i)}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + \hat{\mu}_i^2 / \hat{k}}} \qquad \chi^2 = \sum e_i^2 \qquad (2.14)$$

Under the hypothesis that the model is adequate, χ^2 is approximately chi-square with m-n degrees of freedom.

The appropriateness of the negative binomial model compared with the Poisson model is determined by the statistical significance of the estimated coefficient k. If k is not significantly different from zero, the negative binomial model simply reduces to Poisson. If k is significantly different from zero, the negative binomial is the correct choice and Poisson becomes inappropriate (Poch and Mannering, 1996).

Apart from the parameter k, the between using a Poisson or negative binomial distribution is also based on the dispersion parameter, σ_d by Poisson error structure (Sawalha, 2003).

$$\sigma_d = \frac{Pearson\chi^2}{n-p} \qquad (2.15)$$

$$Pearson\chi^2 = \sum_{i=1}^{n} \frac{[y_i - E(y_i)]}{Var(y_i)} \qquad (2.16)$$

where n is the number of observations, p is the number of model parameters, and y_i is the observed number of crashes on section i, $E(Y_i)$ is the predicted crash frequency for section i, $Var(y_i)$ is the variance of crash frequency for section i. If σ_d turns out to be

significantly greater than 1.0, then the data has greater dispersion than is explained by Poisson distribution, and a negative binomial regression model is fitted to the data (Sawalha, 2003).

2.9 Traffic Crashes and GLM Analysis

Three approaches to relating traffic crashes to geometric and traffic related explanatory variables have been attempted by researches: multiple linear regression (MLR), Poisson regression, and negative binomial regression. Crash data are typically random, discrete, nonnegative, and sporadic. These characteristics mean that MLR models have a number of limitations when applied to crash data (Joshua and Garber, 1990; Zegeer et al., 1990; Miaou and Lump, 1993).

The Poisson model has been applied to crash analyses (Joshua and Garber, 1990; Miaou and Lump, 1993). In the Poisson model, the K constraint of the Poisson model specifies that the mean must be equal to the variance. If this constraint is not valid, the statistical model is incorrect and leads to the incorrect estimation of the likelihood of crash occurrences. A number of studies have reported crash data that were significantly overdispersed which indicated that the variance was greater than the mean (Miaou, 1994; Shankar et al., 1995; Vogt and Bared, 1998).

Over the years, numerous models have been developed to relate traffic crashes to various measures of traffic flow, site characteristics, and geometry. While early models used classical least squares regression modeling, and were based on the assumption of a normal error structure, it is now accepted that the use of GLM, with a negative binomial error structure, is more appropriate (Mountain et al., 1998). The negative binomial model overcomes the problem of overdispersion, and handles the discrete and nonnegative events typical of crash data by relaxing the constraint that the mean must be equal to the variance (Kulmala, 1995). The use of the negative binomial approach has become a popular tool for the analysis of road crash data.

The popularity of GLMs is primarily due to their statistical advantages. One of the first

accident-prediction models for multilane roads using GLMs was developed by Persaud and Dzbik (1993). Persaud and Dzbik examined the relationships between crash data and traffic flow. Traffic flow was expressed both as average daily traffic (ADT) and as hourly volume (VH). The results showed that the crash rate increases with increasing traffic flow, whether measured as ADT or as VH. The accident risk on four-lane freeways was found to be lower than the accident risk on freeways with more than four lanes, reflecting the fact that, for the same traffic volume, freeways with more than four lanes had conditions of freer-flow and greater freedom for drivers to maneuver than did the four-lane freeways. The findings suggested that VH was a more appropriate measure of traffic flow than ADT for explaining crashes since VH takes into account the degree of congested or free-flow traffic conditions at the time of crashes, but as accurate measures of VH are difficult to obtain, ADT is often used in crash-prediction models.

Radin et al. (1996) used a GLM to analyze conspicuity-related motorcycle crashes in Seremban and Shah Alam, Malaysia. The purpose of the study was to understand the possible effects of introducing regulations and a supporting campaign for the use of running headlights. The GLM was developed to describe the relationship between the frequency of conspicuity-related motorcycle crashes and a range of explanatory variables. The model developed revealed that the running headlight intervention reduced the conspicuity-related motorcycle accidents by about 29%. It is concluded that the intervention has been successful in improving conspicuity-related motorcycle accidents in Malaysia.

Abdel-Aty et al. (2000) used negative binomial regression to predict crash frequency as a function of AADT, degree of horizontal curvature, section length, lane, shoulder and median widths, and urban/rural designation. The results showed that crash frequency increases with AADT, degree of horizontal curvature, and section length, and that crash frequency decreases with lane, shoulder, and median width.

Shankar et al. (1995) used negative binomial regressions to model the effects of roadway geometrics and environmental factors on rural crash frequency on sections of highway in

Washington State. Shankar et al. modeled overall crash frequency and the frequency of specific types of crash. They confirmed statistically that separate regression models for each specific type of crash had greater explanatory power than did the model for all crash types. .

Poch and Mannering (1996) used negative binomial regression to predict crash frequency on sections of principal arterials in Washington State. The results showed that the negative binomial regression is a powerful predictive tool and one that should be more widely applied in crash frequency studies.

Miaou (1994) studied the relationship between highway geometric variables and crashes using negative binomial regression. He evaluated the performance of the Poisson regression, zero-inflated Poisson regression, and negative binomial regression. Maximum likelihood was used to estimate the coefficients of the models. Miaou suggested that the Poisson regression model should be used initially to establish the relationship between highway geometric variables and crashes, and that if overdispersion is found to be moderate or high, the negative binomial and zero inflated Poisson regression models can be explored.

Khattak and Council (2002) used the negative binomial model to investigate crash frequencies by analyzing the effect of work zone duration on crashes. Khattak and Council created a unique dataset for California freeway work zones. The dataset included crash data (crash frequency and injury severity), road inventory data (ADT and urban/rural character), and work zone related data (duration, length, and location). The researchers used the negative binomial model for their statistical analyses of crash rates and crash frequencies in the pre-work zone and during-work zone periods. The results showed that crash frequencies increased with increasing work zone duration, length, and average daily traffic. The important finding was that after controlling for various factors, longer work zone duration significantly increased both injury and non-injury crash frequencies.

2.10 Logistic Regression

Logistic regression analysis has special relevance to traffic crash modelling. This is because a traffic crash is discrete or qualitative in nature, i.e. a crash either occurs or does not occur. A crash is either an event or a non-event. A traffic crash therefore usually takes the form of a dichotomous (binary) indicator or dummy variable.

Binomial (or binary) logistic regression is used when the dependent variable is dichotomous and when the independent variables are of any type (Hosmer and Lemeshow, 2000). The regression coefficients show an increase or decrease in the predicted probability of having a characteristic or of experiencing an event due to a one-unit change in the independent variables. If the independent variables are categorical, or a mix of continuous and categorical, logistic regression is preferred.

Logistic regression is an extension of simple bivariate regression where, given a linear relationship between two variables, we can calculate the value of the dependent variable given the value of the independent variable. In logistic regression, there can be one or more independent variables and a single binary dependent variable. The variables are used to establish the probability that the dependent variable belongs to either the event (1) or to the non-event (0) group for a particular crash.

Figure 2.1 shows an idealized logistic regression (Maschner, 1996) using two independent variables (x and y) and two mutually exclusive entities: site (1) and non-site (0). The presence of sites is related to high values for both independent variables x and y. The probability of finding a non-site increases with the values for independent variables x and y. With these two pieces of information, the logistic regression equation can calculate the probability of getting either a site (1) or a non-site (0) given the values of the two independent variables (x and y). The s-shaped curve typically found with logistic regression shows how the probability of finding a site increases as the values for x and y increase. The range of probabilities falls between 0 and 1.

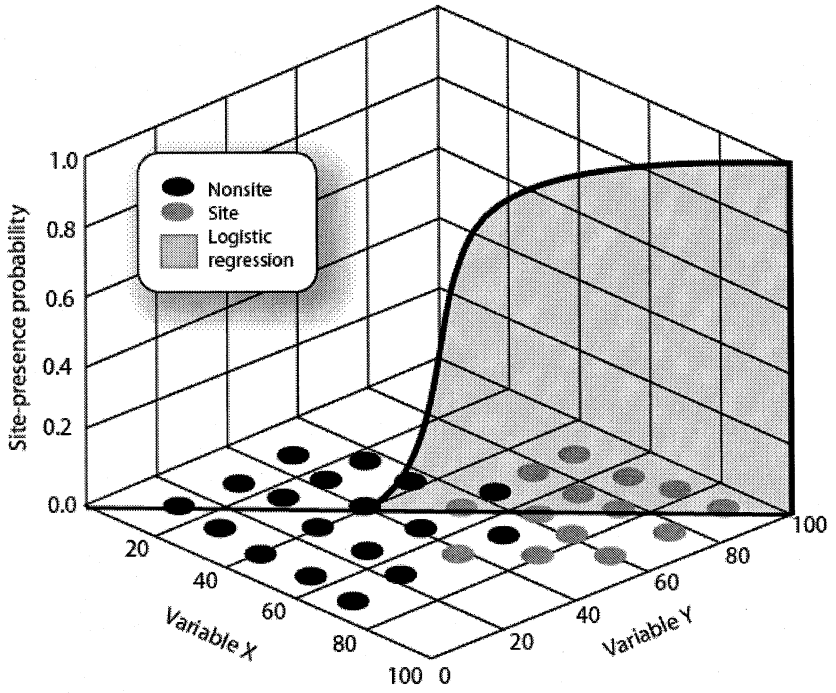

Figure 2.1 Idealized Logistic Regression (Maschner, 1996)

Assumptions of Logistic Regression

Logistic regression does not assume a linear relationship between the dependent variables and the independent variables. Because the logit link function on the left-hand side of the logistic regression equation is non-linear, logistic regression can handle nonlinear effects even when exponential and polynomial terms are not explicitly added as additional independents. Logistic regression makes the following four assumptions:

- The dependent variable need not be normally distributed.
- The dependent variable need not be homoscedastic for each level of the independents, that is, there is no homogeneity of variance assumption.
- The error terms need not be normally distributed.
- The independents need not be unbounded.

The Logistic Regression Model

In logistic regression, the dependent variable is usually dichotomous, that is, the dependent variable can take the value 1 with a probability of success θ, or the value 0 with a probability of failure 1-θ (Balakrishnan, 1991). This type of variable is called a

Boolean (or binary) variable. The following sections draw extensively from

http://userwww.sfsu.edu/~efc/classes/biol710/logistic/logisticreg.htm

The advantage of using the logistic regression model rather than linear or any other type of model is that the independent or predictor variables in logistic regression can take any form, that is, logistic regression makes no assumption about the distribution of the independent variables. The variables do not have to be normally distributed, linearly related, or of equal variance within each group. The relationship between the predictor and response variables is not a linear function in logistic regression. The logistic regression function is used instead, and is the logit transformation of θ:

$$\theta = \frac{e^{(\alpha+\sum \beta X_i)}}{1+e^{(\alpha+\sum \beta X_i)}} \tag{2.17}$$

where α = the constant of the equation and, β = the coefficient of the predictor variables.

An alternative form of the logistic regression equation is:

$$\log it\left[\theta(x)\right] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \sum \beta X_i \tag{2.18}$$

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, the model created includes all the predictor variables that are useful in predicting the response variable. Several different options are available during model creation. Variables can be included or removed from the regression model by the testing the significance. This is known as stepwise regression.

Stepwise regression is used in the exploratory phase of research, but it is not recommended for theory testing (Menard, 1995). Theory testing is the testing of a-priori theories or hypotheses about the relationships between variables. Backward stepwise regression appears to be the preferred method of exploratory analyses. In backward

stepwise regression, the analysis begins with a full or saturated model, and variables are eliminated from the model in an iterative process. The fit of the model is tested after the elimination of each variable to ensure that the model still adequately fits the data. When no more variables can be eliminated from the model, the analysis has been completed.

Model Testing

The process by which coefficients are tested for significance for inclusion or elimination from the model involves several different techniques (Green, 2003). The techniques used in this thesis are discussed below.

Likelihood-Ratio Test

The probability of the observed results given the parameter estimates is known as the likelihood. Since the likelihood is a small number (less than 1), it is customary to use -2 times the log of the likelihood (*-2LL*). *-2LL* is a measure of how well the estimated model fits the likelihood. A good model is one that results in a high likelihood of the observed results. This translates to a small number for *-2LL*. (If a model fits perfectly, the likelihood is 1, and -2 times the log likelihood is 0.)

The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_1) over the maximized value of the likelihood function for the simpler model (L_0). The likelihood-ratio test statistic is a log transformation:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2\left[\log(L_0) - \log(L_1)\right] = -2(L_0 - L_1) \qquad (2.19)$$

This log transformation of the likelihood functions yields a chi-squared statistic. This is the recommended test statistic to use when building a model through backward stepwise elimination.

Hypothesis Testing

Hypothesis testing in logistic regression involves reasoning by contradiction. The first assumption, or the null hypothesis, is that the predictor coefficient is zero in the

population. Hypothesis testing is used to decide whether there is sufficient evidence in the sample data to reject the null hypothesis and, therefore, to accept the alternative hypothesis that the predictor variable coefficient differs from zero.

A large test statistic means that the coefficient probability differs from zero. The criterion for this test is the probability (p-value) associated with the test statistic. Typically, if the observed statistic occurs in 5% or fewer of random samples from a population in which the coefficient is zero, the null hypothesis is rejected. The cutoff probability for hypothesis testing is usually 0.05 and is known as alpha.

Confidence Intervals

Confidence intervals can be used for hypothesis testing as well as for regression coefficients. The interpretation of the 95% confidence interval is that if a 95% confidence interval were computed for an infinite number of random samples from a population, 95% of the intervals would contain the population value on average. In addition, when the 95% confidence interval includes 0, the coefficient is not significant at the 5% alpha level. A 95% confidence interval can be computed for the odds ratio by raising e to the power of the upper and lower confidence bounds for the regression coefficients. If the 95% confidence interval includes the number 1, the odds ratio is non-significant at the 5% level of significance, meaning that the change from one category to another does not reliably increase the odds of membership in the target group.

Interpretation of Logistic Regression Coefficients

The interpretation of the effect of the independent variables on the response variable has intuitive appeal. The logistic regression coefficients show the change in the predicted logged odds of experiencing an event or having a characteristic for a one-unit change in the independent variables (Wong, 2003). Odds in logistic regression express the likelihood of an occurrence relative to the likelihood of a nonoccurrence. To illustrate this, assume that each independent variable has a probability of experiencing an event, defined as P_i. Given this probability, the logit transformation involves two steps. First, take the ratio of P_i and 1-P_i, or odds of experiencing the event. Second, take the natural logarithm

of the odds. The logged odds or logit is:

$$L_i = \ln\left[\frac{P_i}{1 - P_i}\right] \tag{2.33}$$

For example, if P_i equals 0.2, its odds equal 0.25 or 0.2/0.8, and its logit equals −1.386, the natural log of the odds. Manipulating the above formula for odds will give further insight into one of the inherent problems faced with nonlinear functions in linear regression and how logistic regression performs in dealing with these functions. The manipulated logged odds or logit is:

$$\frac{P_i}{1 - P_i} = O_i \tag{2.20}$$

and implies that:

$$P_i = \frac{O_i}{1 + O_i} \tag{2.21}$$

The formula shows that the probability can never be equal or exceed one. Conversely, the probability can never fall below zero. The first property of logit is that it has no upper or lower boundary. Logits vary from negative infinity to positive infinity. The second property is that the logit transformation is symmetric around the midpoint probability of 0.5. The third property is that the same change in the probabilities translates into different changes in the logits. As P_i comes closer to 0 and 1, the same change in the probability translates into a greater change in the logged odds. This means that the general principle is that small differences in probabilities result in increasingly larger differences in the logit when the probabilities are near the bounds of 0 and 1.

Fitting a Binary Logistic Regression Model

Binary logistic regression models model the relationship between a binary response variable and one or more explanatory variables. The logistic regression model uses the explanatory variables to predict the probability that the response variable takes on a given

value. In the case of binary logistic regression models, the response variable takes one of the two binary values (0 or 1). For a binary response variable y, the logistic regression model has the following form (Dissanayake, 2003):

$$Logit\ (P_i) = \log(\frac{P_i}{1-P_i}) = \alpha + \sum \beta X_i \qquad (2.22)$$

where $P_i = Prob(y_i=y_1/X_i)$ is the response probability to be modeled, and y_1 is the first ordered level of y, α is the Intercept parameter, β is the Vector of slope parameters, and X_i is the vector of explanatory variables.

This logistic regression equation models the logit transformation of the i^{th} individual's event probability, P_i, as a linear function of the explanatory variables in the vector, i.

2.11 Traffic Crashes and the Logistic Regression Models

Using logistic regression modelling to model crash occurrence as a discrete outcome involves estimating the probability (conditional probability in nature) that a vehicular crash has a certain type by determining the likelihood of outcomes given that a crash (i.e., FR crash in this thesis) has occurred. Despite the many limitations of statistical models, the models are helpful for identifying factors associated with motor vehicle crashes.

Kim and Yamashita (2000) developed a logistic regression model to explain the likelihood of alcohol impairment among motorcycle riders who had a crash that was reported by police. A likelihood ratio was used to assess the model fit by testing the null hypothesis that the covariates had no effect on the response variable. The likelihood ratio was calculated by subtracting the log-likelihood values of the full model from the loglikelihood values of a model with only the intercept term. The results indicated that impairment was more likely to be associated with middle-aged riders, and unlicensed riders who did not wear a helmet. The results also showed that impaired related crashes are more likely to occur at night, on weekends, and in rural areas.

Krull et al. (2000) developed logistic regression models to investigate the driver, roadway,

and crash characteristics that influence the likelihood of fatal or incapacitating injuries given that a single-vehicle run-off-road crash has occurred. Two models were developed: one for all single-vehicle run-off-road crashes; and one for single-vehicle run-off-road crashes in which the vehicle rolled over. The study found that the use of safety belts lead to a reduced likelihood of fatal or incapacitating injury.

Multivariate logistic regression models were developed by Donelson et al. (1999) to explore the effect of various factors on the likelihood of a fatality in single-vehicle rollover crashes involving light duty trucks. The study objectives were to quantify the effect of fatality risk factors, adjust fatality-based rates for that influence, and assess how the adjusted rates measured differences for various groupings of vehicles. The models were calibrated using crashes of all severities. Counts and rates were then adjusted for the effect of the higher risk of the conditions that resulted in a fatality. This was done by multiplying the observed number of fatalities by the ratio of the sum of probabilities of a fatality occurring at a base condition (i.e., each variable was set to its safest level) to the sum of probabilities of a fatality.

McGinni (1999) used logistic regression to identify factors that differentiate run-off-road crashes from non-run-off-road crashes. McGinni created 56 separate models to analyze the data by combinations of seven age groups, both gender groups, and four highway classification groups. The author found that the effect of the variables on drivers was associated with differences in driver age and gender. Variables that showed an increased likelihood of a run-off-road crash included a non-intersection location, presence of a horizontal curve, rural highway, alcohol involvement, slippery pavement condition, no street lighting, and a high speed limit.

A study by Lin (1993) developed a time-dependent logistic regression model of the crash risk of truck drivers. The model included multiday (i.e., hours on and off-duty) and continuous driving time as factors. A cluster analysis was used to define 10 time-based driving patterns. The results showed that driving time had the most influence on crash risk. Drivers with less than 9 hours of "off-time" had a higher risk than drivers with a

longer rest period. Drivers with at least 10 years of driving experience had a smaller crash risk. Drivers with infrequent driving patterns and a tendency towards night driving showed a higher crash risk.

2.12 Summary of Literature Review Findings

The literature review suggests that appropriate tools for modelling FR crashes would be GLM modelling using the negative binomial distribution, and logistic regression modelling.

Many studies have investigated the relationship between possible explanatory factors and traffic crashes. The findings are not necessarily consistent. Past studies discuss the safety effects of numerous variables, many of which could be investigated in this thesis's study of FR crashes. AADT, section length, speed limit, shoulder width, pavement width, roadway curvature, age, sex, and time of day were found to have a relationship with traffic crashes.

3 DATA PREPARATION FOR GLM ANALYSIS

3.1 Introduction

The following sections describe the characteristics of the data used in this study, and the methodology used to capture crash and geometric data from the Highway Safety Information System (HSIS) for the development of the study's GLM crash model for FR crashes. The HSIS is a multi-state database that contains crash, roadway inventory, and traffic volume data for a select group of States. The HSIS is operated by the University of North Carolina Highway Safety Research Center (HSRC), LENDIS Corporation, and the Federal Highway Administration (FHWA).

The data for the FR crash model development were extracted from the Ohio Traffic Crashes Database obtained from the HSIS Database. The Ohio data in the HSIS are derived from databases of police-reported crashes in Ohio State. Ohio safety data were used because Ohio's safety information is more comprehensive and complete than that of many other states.

Further information about the data is provided in Section 3.2.

3.2 Data Preparation for GLM Analysis

Data Description

In order to develop the GLM model, detailed information on crash data, traffic flow and road design was required. Single-vehicle run-off-road crashes on rural two-lane roadway segments were selected from the Ohio data for the five years from 2000 to 2004. The site of the crashes was related to the specific road link by the road ID number, and crash, roadway, curve and grade data were collected for the segments where a crash occurred.

The Ohio data system provided to HSIS includes the following four basic files:

- Accident Data File (accident, vehicle and occupant);
- Roadway Inventory File;
- State Supplemental Inventory; and

- "Points" File.

Accident data file stores the basic crash information on a case-by-case basis. The data for each crash are found in three separate subfiles: crash, vehicles, and occupants.

Roadway inventory file contains the general roadway characteristics. The information includes all functional classes of roads within the state system: freeways, arterials, and collectors, both rural and urban. The file contains information on approximately 1,500 miles of Interstates, 4,000 miles of U.S. Routes, and 14,000 miles of State Routes. The file contains general cross section information related to travel way widths, number of lanes, median width, and other variables. Traffic information in the form of ADT is included for each section in the file.

State supplemental inventory data file contains curve, grade, and other geometric data.

The fourth file is the points file which contains information on intersections, railroad grade-crossings, and underpasses.

Each accident file record is referenced to a point on the roadway. Roadlog file is an additional file which contains information on a homogeneous section of the roadway (i.e. a stretch of road which is consistent in terms of certain characteristics). Each new section is defined by a new beginning reference point. Each record in the roadlog file contains the current characteristics of the road system including surface type and width, shoulder and median information, lane information, etc. Information on curves and grades is captured in separate files for curves and grades. The Curve File has data on all horizontal curves, and the Grade File has information on grades greater than 3 percent.

Segmentation of Roadway Sections

To analyze the FR crash rate, it was important to define a roadway section. Three main approaches were considered: absolute homogeneous sections, fixed-length non-homogeneous sections and limited homogeneous sections.

Absolute homogeneous sections are defined by changes to any geometric or roadway variable (e.g. a new section would be identified when the shoulder width changed from 1.8m to 2.0m). Section-defining information for the absolute homogeneous method includes changes to district number, state route number, roadway type, numbers of lanes, roadway width, shoulder width, presence of curb or retaining wall, divided or undivided highway, speed limit, AADT, truck percentage, peak hour factors, and vertical and horizontal curve characteristics (Miaou et al., 1991).

The disadvantage of the absolute homogeneous sections is that roadways with numerous horizontal curves and grades tend to produce sections that are less than 0.1 km in length. These short road sections can result in undesirable impacts on the estimation of their regression models (Zegeer et al., 1991). The use of fixed-length non-homogeneous sections, however, ignores the effect of the change of geometric or roadway variables and leads to non-homogeneity in geometric design variables (Miaou et al., 1993).

On consideration of the advantages and disadvantages of the two approaches, this study decided to use unequal-length limited homogeneous sections. The sections were homogeneous for only ten variables: Route Number, AADT, Speed Limit, Access Control, Functional Class, Median Width, Median Type, Number of Lanes, Surface Width, and Roadway Type. As the length of the sections varied, the crash frequencies and associated geometric data refer to roadway sections of unequal length.

To overcome non-homogeneity in the geometric design variables, crash frequencies, and roadway geometrics for both roadway directions were collected. Alignment indices of average horizontal curvature and average vertical gradient were also collected. To overcome inequity between two outside shoulder widths, the variable 'average outside shoulder width' was applied.

Data Editing

In order to develop the dataset for GLM analysis, the data from the Ohio HSIS data files

were sorted and merged by their SAS variables (CASENO, VEHNO, CNTYRTE, and MILEPOST). Descriptions of SAS variables are provided in Appendix B.

As noted above, the crash data in the accident data file are subdivided into three subfiles (accident, vehicle, and occupant). The accident and vehicle subfiles were linked together using the crash report number (i.e., CASENO). When linking the occupant subfile, the additional linking variable related to vehicle number (i.e., VEHNO) was matched so that the occupants were associated with the vehicle in which they were traveling. To link vehicles with accidents, both subfiles were first sorted by CASENO. The separate subfiles were linked by specifying a SQL JOIN operation of the SAS Procedure with the constraining condition that the case number and vehicle number from each table are equal. SQL processing does not require the data to be presorted, and the output will not be in any particular sort order unless ORDER BY is specified.

The accident subfile was linked to the roadlog file using the CNTYRTE and MILEPOST variables in the crash record, and the CNTY_RTE, BEGMP and ENDMP variables in the roadlog file. Similarly, the accident subfiles were linked to curve, grade, and angle points using similar variables found in each respective file. To link the accident file and the points file, CNTYRTE and MILEPOST variables from the accident file were matched with CNTY_RTE and MILEPOST variables of the points file. To extract data on the intersecting state-system route in the points file, the roadlog file was linked to the XMILEPST and XCNTYRTE variables.

To prepare the accident subfile for linking with the roadlog file using a SAS data step process, both the accident and the roadway file were sorted into location order by CNTYRTE and MILEPOST on the accident file and by CNTY_RTE and BEGMP on the roadlog file. Similar sorts were done with other files to be merged.

In order to develop the dataset suitable for GLM analysis, homogeneous road segments were created from the roadlog file using changes in given variables. The road segments were then identified by using a SAS RETAIN Statement to compare the current record to

the previous record. If the two records were different, a flag was set to indicate the start of a new segment. Next, the road segments were sorted in descending order to save the data contained in the first record of a homogeneous segment, especially the beginning milepost number. Finally, the homogeneous segments were assembled and the ending milepost of the last record was assigned to the first record. The average AADT for the five years from 2000 to 2004 was calculated for each road section and transformed to log form for the GLM analysis. Average horizontal curvature and average vertical gradient were also calculated.

Single-vehicle run-off-road traffic crashes were selected from the accident data files for 2000 to 2004. The rate of the selected crashes within each road section was calculated as a response variable in GLM. Eventually, the roadway segment file was merged with the new variables LOGAADT (log form of AADT), average horizontal curvature, average vertical gradient, and crash frequency. The new SAS dataset was suitable for the GLM analysis.

Figure 3.1 shows the procedure used for developing the dataset. The SAS codes for the dataset development for the GLM analysis are listed in Appendix C.

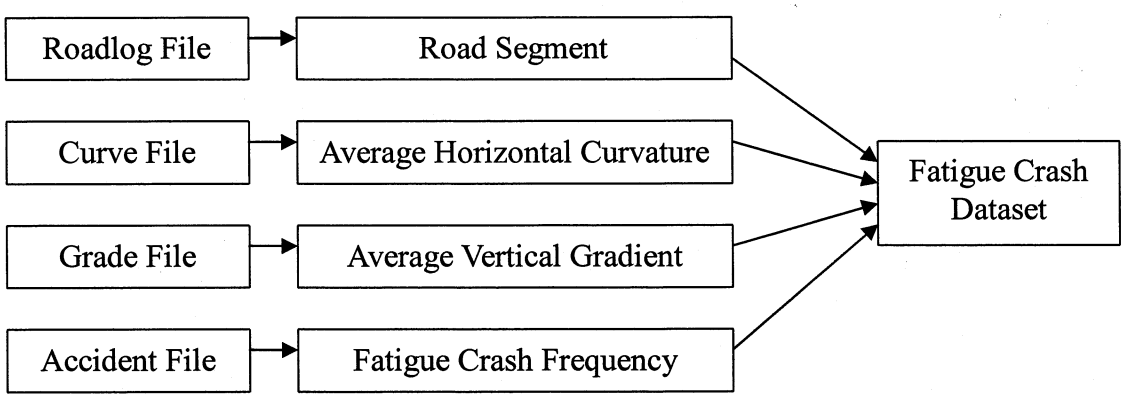

Figure 3.1 Fatigue Crash Dataset Developing Procedure

3.3 Variables for GLM Analysis

The following sections discuss the variables used in the GLM analysis. The response

variable is the FR crash rate. The explanatory variables are: traffic volume, posted speed limit, horizontal and vertical alignment indices, pavement surface width, and average width of outside shoulders.

Response Variable

The review of crash models reported in Chapter 2 showed that crash rate and crash frequency are commonly used as dependent (or response) variables. In this study, the crash rate is defined as the number of crashes per mile per year. The use of the crash rate as the response variable means that the section length is not treated as an independent variable. If section length has to be considered as an independent variable, crash frequency (and not crash rate) should then be considered a response variable.

Traffic Volume

Several studies (Caliendoa et al., 2007; Martin, 2002; Miaou et al., 1992) have attempted to quantify the relationship between traffic volume and the traffic crash rates, and the relationship between traffic volume and the severity of crashes. The effects of traffic volume vary and may be influenced by speed, density, and flow. Research indicates that traffic volume is positively correlated with incidences of traffic crashes: as the number of vehicles on a highway increases, the potential for conflicts within a traffic stream also increases.

Posted Speed Limit

The posted speed limit is the speed set by state, city, or county to inform motorists of the highest safe speed under favorable road, traffic, and weather conditions. Experience and research are used to develop procedures for setting speed limits. These procedures differ from country to country. The posted speed limit is usually based on the results of an engineering and traffic investigation. The speed limit is usually set at the 85th percentile of operating speeds.

The posted speed limit is a function of several roadway parameters, sight distance, and roadway condition. In a crash prediction model, the speed limit is, therefore, a numerical

variable rather than a categorical variable.

Table 3.1 shows the number of road segments in each posted speed limit group. The posted speed limit ranges from 20 mph to 65 mph. Most of the road segments (61.5%) have a posted speed limit of 55mph. The second largest group (20.9%) has a posted speed limit of 55mph.

The last column of Table 3.1 shows the mileage of the road segments for each speed limit. The 55 mph road segments have a total mileage of 10,732 miles (86.87% of the total mileage).

Table 3.1 Posted Speed Limits on Road Segments selected for Study

Speed Limit (MPH)	Number of Observations (Road Segments)	% of Observations Percentage (Road Segments)	Total Mileage by Speed Limit (Miles)
20	8	0.1	1.2
25	520	4.2	94.1
30	8	0.1	1.31
35	2,586	20.9	760.3
40	289	2.3	138.6
45	981	7.9	732.1
50	347	2.8	188.3
55	7,607	61.5	10,731.7
60	2	<0.1	0.4
65	6	0.1	9.4
Total	**12,354**	**100**	**12,657.3**

Horizontal and Vertical Alignment Indices

Alignment indices are quantitative measures of the general characteristics of a roadway segment's alignment (Fitzpatrick et al., 2000.). The measures incorporated into the alignment indices include average radius per roadway section, average vertical curve, and curvature change rate per kilometer.

Curvature on roads is not expressed in terms of radius, as it is on model layouts. The curvature on roads is the angle between two lines drawn from the centre of the circle of which the curve is a part to two points on the circumference 100 feet apart. Curvature can be expressed in terms of the number of degrees traversed by 100 feet of road. The radius (distance from centre point to edge) of a curve is obtained with the following conversion equation: radius in feet equals 5,729 divided by the degrees of curvature.

Polus et al. (1998) suggested several alignment indices to quantify the general character of the alignment of a roadway. Indices based on horizontal alignment characteristics included the average curvature in degrees per kilometer, the average radius of curvature, the ratio of the maximum to the minimum radius, and two horizontal alignment indices using the radii of curves. They also suggested one index based on vertical alignment: the average gradient along a section of roadway.

Average curvature was selected for this study because it expresses what motorists typically encounter on curved sections of the road. Large average curvature indicates sharp curves.

Average gradient was also selected for this study. Average gradient represents the absolute change in vertical direction along a roadway. As the average gradient increases, there is either a large change in elevation between the vertical points of intersection, or the vertical alignment is hilly.

Pavement Surface Width

Pavement surface width measures the total paved road width of all of a road segment's through lanes in both directions. The measurement is in feet. Pavement surface width does not include left or right turn lanes, parking lanes, or acceleration or deceleration lanes.

This study used total pavement surface width because individual lane width is not available in the HSIS database. It is, however, noted that wide lanes imply that vehicles

moving in adjacent lanes can be widely separated. The advantage of wide lanes is that they may provide a buffer which may absorb small random deviations of vehicles from their intended path. The disadvantage of wide lanes is that drivers may adapt to the wide lanes and tend to drive faster than in narrow lanes. The effects on safety require empirical statistical evidence. Pavement surface width was included in the modelling to shed light on the lane width issue.

Table 3.2 shows the number of road segments in each pavement surface width group (14 feet to 44 feet). Pavement surface width ranges from 14 feet to 44 feet. The most common pavement surface width is 24 feet (33% of the road segments), followed by 20 feet (30%). The third most common pavement surface width is 18 feet (11%). The last column of Table 3.2 shows the mileage of the road segments for each pavement surface width. 4,297 miles have a pavement surface width of 20 feet, 3,744 of 24, and 2,287 of 18.

Table 3.2 Pavement Surface Width Frequencies on Road Segments selected for Study

Pavement Surface Width (Feet)	Number of Observations (Road Segments)	% of Observations (Road Segments)	Total Mileage by Pavement Surface Width
14	1	0.01	0.35
15	1	0.01	0.40
16	55	0.44	53.48
17	19	0.15	21.81
18	1326	10.69	2,286.73
19	562	4.53	839.70
20	3,665	29.54	4,297.4
21	238	1.92	222.59
22	1,128	9.09	1035.3
23	64	0.52	65.90
24	4,111	33.14	3,744.27
25	33	0.27	15.05
26	69	0.56	20.68
27	51	0.41	9.19
28	118	0.95	25.23
29	13	0.10	1.91
30	199	1.60	45.42
31	30	0.24	5.60
32	144	1.16	28.39
33	16	0.13	3.00
34	69	0.56	18.45
35	26	0.21	4.95

Pavement Surface Width (Feet)	Number of Observations (Road Segments)	% of Observations (Road Segments)	Total Mileage by Pavement Surface Width
36	233	1.88	49.49
37	15	0.12	4.90
38	51	0.41	9.40
39	18	0.15	3.48
40	79	0.64	15.23
41	18	0.15	2.45
42	26	0.21	4.96
43	5	0.04	0.99
44	23	0.19	3.94
Total	**12,383**	**100.00**	**12,657.3**

Average Width of Outside Shoulders

The average width of outside shoulders variable provides outside shoulder information. Shoulder width is the portion of the roadway between the outer edge of the through pavement and the end of the shoulder, or the portion of the roadway between the outer edge of the through pavement and the intersection of the slope lines of the outer edge of the roadway with the ditch.

Outside shoulders provide for stopped vehicles, emergency use, and lateral support of the roadbed. The average width of the outside shoulders was used in this analysis because of the consideration of possible unequal widths of both outside shoulders.

Table 3.3 shows the number of road segments in each the average width of outside shoulders ranges from 0 feet to 15 feet. The three most common average outside shoulders widths are 4 feet (20% of the road segments), followed by 3 feet (17%), and 2 feet (16%). The next three most common average outside shoulders widths are 0 feet (11% of the road segments), 8 feet (9%), and 6 feet (9%).

The last column of Table 3.3 shows the mileage of the road segments for each average width of outside shoulders. 2,613 miles have an average outside shoulders width of 4 feet, 2,601 of 2, and 2,566 of 3.

Table 3.3 Average Outside Shoulder Width Frequencies on Road Segments selected for Study

Average Outside Shoulder Width (Feet)	Number of Observations (Road Segments)	% of Observations (Road Segments)	Total Mileage by Average Outside Shoulder Width
0	1,310	10.56	558.70
0.5	8	0.06	15.16
1	490	3.95	546.51
1.5	44	0.35	42.23
2	1,951	15.73	2,601.43
2.5	59	0.48	72.60
3	2,041	16.45	2,565.74
3.5	22	0.18	50.89
4	2,414	19.46	2,613.07
4.5	30	0.24	13.94
5	651	5.25	763.34
5.5	29	0.23	41.08
6	1,116	9.00	1,118.70
6.5	4	0.03	0.67
7	92	0.74	78.01
7.5	11	0.09	7.72
8	1,131	9.12	941.25
8.5	5	0.04	2.31
9	73	0.59	56.57
9.5	15	0.12	6.68
10	869	7.00	718.24
11	5	0.04	3.46
11.5	4	0.03	0.20
12	18	0.15	17.36
13	3	0.02	0.53
14	1	0.01	0.08
15	10	0.08	4.17
Total	**12,406**	**100.00**	**12,657.3**

4 GLM ANALYSIS

Chapter 4 describes the GLM analysis:

- Section 4.1 outlines SAS analysis for generalized linear models;
- Section 4.2 describes the generalized linear regression model;
- Section 4.3 discusses the goodness of fit testing (Deviance, Scaled Deviance and Pearson Chi-Square);
- Section 4.4 discusses the appropriateness of variables used in the model, and the selection of the study's six explanatory variables;
- Section 4.5 introduces the modelling procedure used in the generalized linear modelling, and the four GLM models estimated in this study;
- Section 4.6 presents the FR and NFR results obtained from the generalized linear modelling.
- Section 4.7 presents the sleepy-period FR traffic crashes and sleepy-period NFR results obtained from the generalized linear modelling.
- Section 4.8 summarizes the generalized linear modelling results.
- Section 4.9 applies the FR GLM model to network screening. Safety performance functions (SPFs) and empirical Bayes (EB) principles are introduced and illustrated, and then applied to the Ohio data.

4.1 SAS Analysis for GLM

Figure 4.1 presents a flow chart showing the application of GLM regression analysis using SAS software.

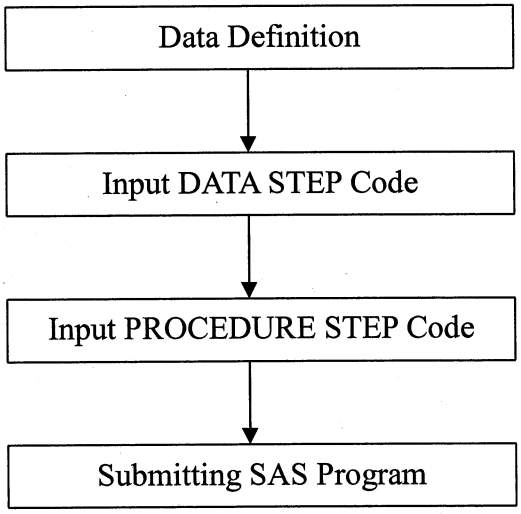

Figure 4.1 Flow Chart of the Application of GLM Analysis by using SAS

Analysis using SAS assumes that there is a random sample of independent observations, and that the modelling will address the variation in the observations. For accident count data and modelling, the generalized linear model with negative binomial error distribution is often most appropriate, and SAS makes the fitting of such models relatively straightforward. The log function is the link function that enables the calibration of a linear predictor. A negative binomial distribution is used with a constant over dispersion parameter that is calibrated using a quasi-likelihood approach. SAS Codes for GLM analysis are listed in Appendix D.

4.2 Description of GLM

The generalized linear regression model used in this study relates the number of observed FR crashes to traffic flow and road design. Generalized linear modelling techniques were used to fit the model, and the distributions of crash counts were assumed to follow a negative binomial distribution (as mentioned above). The regression analyses were performed using the GENMOD procedure in SAS. Several studies (Khattak and Council, 2002; Radin et al., 1996; Poch and Mannering, 1996; Miaou, 1994; Persaud and Dzbik, 1993) have proved that the negative binomial distribution might be more appropriate for traffic crash analysis because it allows for non-constant variance, or overdispersion in the

data.

Previous studies have investigated a number of different ways in which to relate crash frequencies to traffic flows (Caliendoa et al., 2007; Khattak and Council, 2002; Persaud et al., 2000; Shankar et al., 1995). For road links, the general opinion is that crash frequencies can be described by a flow variable raised to a power. The flow variable often consists simply of the motor vehicle traffic flow along the link (AADT). In this study, crash rate is used for GLM analysis.

The model structure is:

$$E(\mu) = AADT^{\alpha} e^{(\Sigma \beta_i X_i)} \qquad (4.1)$$

Where $E(\mu)$ is the expected number of FR crashes per year per mile,
AADT is the annual average daily traffic,
The X variables describe the road variables,
α, β_i are estimated parameters.

Type 3 analysis was selected for the analysis of FR crashes. The GENMOD procedure generates a Type 3 analysis analogous to Type III sums of squares in the GLM procedure. A Type 3 analysis does not depend on the order in which the terms for the model are specified. A GENMOD Type 3 analysis consists of specifying a model and computing likelihood ratio statistics for Type III contrasts for each term in the model. The contrasts are defined in the same way as they are in the GLM procedure. The GENMOD procedure optionally computes Wald statistics for Type III contrasts. This is computationally less expensive than likelihood ratio statistics, but it is thought to be less accurate because the specified significance level of hypothesis tests based on the Wald statistic may not be as close to the actual significance level as it is for likelihood ratio tests.

A Type 3 analysis generalizes the use of Type III estimable functions in linear models. Briefly, a Type III estimable function for an effect is a linear function of the model parameters that involves the parameters of the effect and any interactions with that effect.

A test of the hypothesis that the Type III contrast for a main effect is equal to 0 is intended to test the significance of the main effect in the presence of interactions (SAS, 1998a).

4.3 Goodness of Fit Tests

As available in SAS manual, the main goodness of fit measures for GLM analysis are Deviance, Scaled Deviance, and Pearson Chi-Square (SAS, 1998a).

The Deviance has an approximate chi-square distribution with n-p degrees of freedom, where n is the number of observations, p is the number of predictor variables (including the intercept), and the expected value of a chi-square random variable is equal to the degrees of freedom (DF). Then, if the model fits the data well, the ratio of the Deviance to DF, Value/DF, should be about one. Large ratio values may indicate model misspecification or an over-dispersed response variable; ratios less than one may also indicate model misspecification or an under-dispersed response variable. A consequence of such dispersion issues is that standard errors are incorrectly estimated, implying an invalid chi-square test statistic, superscript p. Importantly, however, assuming the model is correctly specified, the regression estimates remain unbiased in the presence of over or under-dispersion. A "fix" is to adjust the standard error of the estimates. The standard error correction corresponds to the approach for the scaled criterion. A naive explanation is that when the scale option is specified (*scale* = *dscale*), the Scaled Deviance is forced to equal one. By forcing Value/DF to one (dividing Value/DF by itself), the model becomes "optimally" dispersed; however, what actually happens is that the standard errors are adjusted ad hoc. The standard errors are adjusted by a specific factor, namely the square root of Value/DF. Below are summaries of the various measurements used to assess the model fit.

- Deviance - This is the deviance for the model. The deviance is defined as two times the difference of the log-likelihood for the maximum achievable model (i.e., each subject's response serves as a unique estimate of the negative binomial parameter), and the log likelihood under the fitted model. The difference in the Deviance and degrees of freedom of two nested models can be used in likelihood ratio chi-square

tests. McCullagh and Nelder (1989) caution against the use of the deviance alone to assess model fit.

- Scaled Deviance - The scaled deviance is equal to the deviance since the *scale=dscale* option is not specified on the model statement.
- Pearson Chi-Square - This is the Pearson chi-square statistic. The Pearson chi-square is defined as the squared difference between the observed and predicted values divided by the variance of the predicted value summed over all observations in the model.
- Scaled Pearson χ^2 - This is the scaled Pearson chi-square statistic. The scaled Pearson χ^2 is equal to the Pearson chi-square since the *scale=pscale* option is not specified on the model statement.
- Log Likelihood - This is the log likelihood of the model. Instead of using the Deviance, two times the difference between the log likelihood for nested models are taken to perform a chi-square test.
- DF and Value - These are the degrees of freedom DF and the respective Value for the Criterion measures. The DF equals *n-p*, where *n* is the Number of Observation Used and *p* is the number of parameters estimated.
- Value/DF - This is the ratio of Value to DF.
- Algorithm Converged - This is a note indicating that the algorithm for parameter estimation has converged, implying that a solution has been found.

4.4 Appropriateness of Variables in the Model

Different approaches are used to assess the significance of the variables in the model. This section discusses the two main ways in which the validity of including each variable in the model is judged. Section 4.4.1 discusses the *p*-values, and Section 4.4.2 discusses the sign of the coefficients. Section 4.4.3 then explains the procedure used to select the six explanatory variables used in this study.

p-Values

The *p*-value is the primary decision criterion for variables which are significant in the crash model. The *p*-value is a statistical test associated with the null hypothesis. The null

hypothesis used in this analysis is that a variable has zero coefficients, that is, the variable has little effect on traffic crashes compared with the effect of other variables. In general, the *p*-value is the probability that the sample could have been drawn from the population(s) being tested (or that a more improbable sample could be drawn) given the assumption that the null hypothesis is true.

In this analysis, a *p*-value tolerance of up to 0.05 seems acceptable and follows earlier research (Radin et al., 1996). A *p*-value less than 0.05 signifies that the null hypothesis is false, and the variable has an effect to crash frequency. *p*-values larger than 0.05 imply that the null hypothesis cannot be rejected, which means that the variable has no effect on the crash frequency. All *p*-values for this analysis are calculated using Chi-Square and the Z-statistic.

In addition to testing individual variables, an aggregate *p*-value is used to test the whole model. The selection of one model rather than another model in this analysis is also based on this aggregate *p*-value.

Sign of the Coefficients

The sign of the variable coefficient can sometimes cause the coefficient to be omitted in the model. This happens when the direction of the effect of the variable is known in advance, and the magnitude is the one that is tested. For example, if a variable for section length has a negative coefficient, meaning that the longer the roadway section the lower the crash frequency, then the variable can be dropped since this inference makes no sense. Unexpected coefficients can sometimes result from unreliable data for that variable or from errors that arise from other sources.

Selection of Explanatory Variables

The choice of explanatory variables should primarily be based on the theory used, the question to be answered, and on professional knowledge rather than on ambitions regarding multiple correlations and curve-fitting (Harnen et al., 2006; OECD, 1997). Six explanatory variables were selected for building the model in this study. Table 4.1 lists

the six variables, and shows the SAS variable name and the unit of measurement.

Table 4.1 Explanatory Variables included in GLM Analysis

Description of Variable	**SAS Variable Name**	**Unit of Measurement**
Average AADT in five years	AADT	Vehicles per day
Speed limit	Spdlimt	Miles per hour
Average curvature of road section	Curv_hi	Degrees per hundred feet
Average absolute gradient of road section	Grad_hi	Percent
Surface width	Surf_wid	Feet
Average width of outside unpaved shoulder	Avg_outsh	Feet

The selection of variables to fit a model is a very important procedure. Variable selection refers to removing variables which are not significant and adding variables which are significant. There are different methods for selecting independent variables when running a model: sequential forward selection, backward variable elimination, and stepwise selection.

Sequential forward selection starts the model with a constant term. All the variables that have not yet been selected are considered for selection, and their p-values, are recorded. The variable that produces the best p-value is included in the set. Then, a new step is started, and the remaining variables are considered. This is repeated until a pre-specified number of variables have been included.

The sequential forward selection procedure first estimates parameters for variables forced into the model. These variables are the intercepts and the first explanatory variables in the model. Next, the adjusted chi-square statistics for each variable in the model are computed, and the largest of these statistics is assessed for significance. If the statistic is significant at the specified *p*-value, the corresponding variable is added to the model.

Once a variable is entered into the model, it is never removed from the model. The process is repeated until none of the remaining variables meet the specified level for entry, or until the final model is reached. The drawback of forward variable selection is that once a variable has been selected, it cannot be excluded.

In backward variable elimination, variables are removed, one at a time, from the list of explanatory variables. The model starts with all the explanatory variables included, and then eliminates those which do not seem to improve on the explanation provided by the other variables. At any step, the variable to be removed is determined by the *p*-values. The removed variable is the one with the correspondingly largest *p*-value. Each time a variable is removed, the Deviance and *p*-values are checked. The procedure stops when the elimination of another term cannot further improve these goodness of fit measures.

Stepwise selection is modified forward and backward selection. The variables are removed and added randomly. Variables are entered into and removed from the model in such a way that each forward selection step can be followed by one or more backward elimination steps. The stepwise selection process terminates when no more variables can be added to the model, or when the variable just entered into the model is the only variable removed in the subsequent backward elimination.

Variable selection sometimes results in the elimination of very useful variables which should be in the model. At such a point, some variables can be transformed and used in the model. Common transformations include taking the inverse, square, square root, natural log, or base logarithm of the variable. Other transformations include converting numerical variables into categorical variables by combining and grouping some values or by confounding two variables as products or as quotients (e.g. LOG_AADT = LOG(AADT). The variables of average curvature, average gradient, and average outside shoulder width have been thus transformed in this analysis.

4.5 Modelling Procedure

Backward selection was used in this study's modelling procedures because all six

selected variables were expected to be included in the model. At first, all six variables were included in the model for the regression analysis. Insignificant variables which proved insignificant were excluded one by one, starting with the least significant variables. Insignificant variables were identified on the basis of the likelihood ratio statistics and standard errors of the estimated parameter values. The SAS codes created for the logistic regression modelling in this research are listed in Appendix A.

Separate models were estimated for FR crashes and NFR crashes, and also for sleepy-period fatigue-related (SPFR) crashes and sleepy-period non fatigue-related (SPNFR) crashes. Two sleepy periods were defined in this study: 2 a.m. to 5 a.m., and 2 p.m. to 4 p.m. These two periods are consistent with research into circadian rhythms and the definitions of sleepy periods provided by Chipman and Jin (2007) in their study of drowsy drivers (as mentioned in Chapter 2). The purpose of the GLM modelling SPFR and SPNFR crashes was to examine the effect of road design and traffic factors on fatigue crashes during sleepy periods.

As mentioned earlier, as GLM techniques were used to fit the models, the variation in FR, NFR, SPFR, and SPNFR crashes was assumed to follow a negative binomial distribution. Four separate models are calibrated in order to understand the effect of each variable on FR, NFR, SPFR, and SPNFR crashes. Table 4.2 lists the four models. The models are discussed in detail in Sections 4.6 and 4.7.

Table 4.2 The Four Calibrated GLM Models

Model	Period	Crash Type
1	All Day	FR
2	All Day	NFR
3	Sleepy Period	SPFR
4	Sleepy Period	SPNFR

The analysis procedure began by considering the exact form that the dependent variable would take. Traffic crash rate was investigated and found to be the best variable to be used as the dependent variable. Once the dependent variable was determined, a prediction model was developed.

The calibrated model in SAS has the following structure:

$$Crashes/Mile/Year=(AAADT)^{\alpha} * exp[\beta_0+\beta_1(SPDLIMT)+\beta_2(CURV\_HI)+\beta_3(GRAD\_HI) +\beta_4(SURF\_WID)+\beta_5(AVG\_OUTSH)] \quad (4.2)$$

where α, β_n are estimated parameters.

4.6 GLM Estimation Results: FR and NFR Traffic Crashes

This section presents the results of the FR and NFR traffic crash models. The independent variables were included in the models if they were significant at the 95 percent confidence level. All six independent variables were statistically significant. The coefficients estimated provide a good indication of the strength of each of the independent variable in contributing to FR and NFR crashes.

Table 4.3 shows the estimated results for the FR and NFR models. All the variables have the expected sign (a positive sign indicates an increase in the crash rate, and a negative sign indicates a decrease).

Table 4.3 Negative Binomial Models for FR and NFR Traffic Crashes

	FR		NFR	
Parameter/variable	**Estimate**	**Pr > ChiSq**	**Estimate**	**Pr > ChiSq**
β_0 (Constant)	-4.2481	<.0001	-5.4211	<.0001
α (AAADT)	0.4165	<.0001	0.5985	<.0001
β_1 (SPDLIMT)	0.0166	<.0001	0.0028	.0225
β_2 (CURV_HI)	0.0143	.0136	0.0125	<.0001
B_3 (GRAD_HI)	0.0434	<.0001	0.0106	.0003
B_4 (SURF_WID)	-0.0251	<.0001	0.0422	<.0001
β_5 (AVG_OUTSH)	-0.0453	<.0001	0.0158	<.0001
Dispersion parameter	0.6816		0.7465	

Table 4.4 shows the goodness of fit tests. The goodness of fit results show that the models fit the data well. A *p*-value is not computed for the deviance shown in Table 4.4, but a deviance that is approximately equal to its degrees of freedom is a possible indication of a good model fit. The value of the deviance divided by its degrees of freedom is about 1 in FR and NFR models, which shows that the models fit the data well,

Table 4.4 Model Goodness of Fit for FR and NFR Traffic Crashes

	FR			NFR		
Criteria For Assessing Goodness Of Fit						
Criterion	**DF**	**Value**	**Value/DF**	**DF**	**Value**	**Value/DF**
Deviance	1.20E+04	11422.185	0.9251	1.20E+04	13752.354	1.1138
Scaled Deviance	1.20E+04	11422.185	0.9251	1.20E+04	13752.354	1.1138
Pearson Chi-Square	1.20E+04	15299.171	1.2391	1.20E+04	23030.8165	1.8653
Scaled Pearson $\chi 2$	1.20E+04	15299.171	1.2391	1.20E+04	23030.8165	1.8653
Log Likelihood	36861.5005			206961.8427		

The overdispersion parameters in Table 4.3 were significant (indicating that the mean is larger than the variance), and confirm the appropriateness of the Negative binomial relative to the Poisson formulation for predicting FR and NFR crashes.

The six sections that follow discuss the effects of each explanatory variable.

Effect of Annual Average Daily Traffic

An increase in AADT on the roadway section had a positive impact on the likelihood of crashes (i.e. was associated with an increase in the likelihood of crashes). AADT was expected to be statistically significant because numerous studies (Caliendoa et al., 2007; Qin et al., 2003; Persaud et al., 2000) have found AADT to be a very strong crash predicting variable.

Both the FR and the NFR models indicated that an increase in AADT had a positive effect on crash involvement on rural two-lane highways, but the effect was higher for NFR traffic crashes (0.5985) than for FR traffic crashes (0.4165).

Effect of Posted Speed Limit

The posted speed limit is a significant factor. As a higher speed limit implies higher vehicle speeds on the road, it was likely that both FR and NFR traffic crashes occurred on roads with high speed limits.

Table 4.3 indicates that the effect of the speed limit was more positive for FR traffic crashes (0.0166) than for NFR traffic crashes (0.0028).

Effect of Average Horizontal Curvature

Because many roadway sections have straight portions or contain more than one curve, the average horizontal curvature was used as the Alignment Index to represent the average horizontal curvature of each whole roadway section. In this study, average horizontal curvature is expressed as the degree of the curve per hundred feet.

An increase in the degree of curvature was associated with an increased crash rate. There was little difference in the contribution of average horizontal curvature to FR crashes (0.0143) and NFR crashes (0.0125).

Effect of Average Vertical Gradient

Average vertical gradient is expressed as the average absolute value of the vertical gradients of a given roadway section. As the vertical alignment of a roadway section may include more than one gradient, the average vertical gradient was used as the Alignment Index to represent the average vertical gradient of each whole roadway section.

An increase in the average gradient was associated with an increased crash rate. As mentioned in Section 3.3.4, as the average gradient represents the absolute change in vertical direction, a large value indicates either that there is a large change in elevation between the vertical points of intersection, or that the vertical alignment is hilly. The association between an increased average gradient and an increased crash rate conforms to expectations.

Table 4.3 shows that the effect of the average gradient on FR crashes (0.0434) was more pronounced than the effect on NFR crashes (0.0106).

Effect of Pavement Surface Width

Pavement surface width was used in this study as a substitute for lane width because the Ohio HSIS database does not include a lane width variable. The coefficient of pavement surface width in this study is negative (-0.0251) in the FR model with p-value=0.0001. This means that wider surface widths have fewer FR crashes compared to narrow surface widths. The result may suggest that wider pavement surface widths provide an errant vehicle with more chances to return to the roadway. Roads with a wide pavement surface width also provide a buffer between vehicle lanes, and this buffer may provide additional room and time for a driver to avoid running off the road. The results for NFR crashes (0.0422) were different: an increase in pavement surface width increased the NFR crash rate.

Effect of Shoulder Width

Average outside shoulder width had a negative coefficient (-0.0453) in the FR model, but a positive coefficient (0.0158) in the NFR model. The negative coefficient in the FR model indicates that as the width of the shoulder increased, the crash rate decreased. This result suggests that drivers could use the wider shoulder widths to avoid running off the road and hitting roadside obstacles.

The positive coefficient in the NFR model appears surprising, but does not imply that NFR crashes are less likely to occur on roads with narrow shoulders than on roads with wider shoulders.

4.7 GLM Results: Modelling SPFR and SPNFR Traffic Crashes

This section presents the results of the sleepy-period FR (SPFR) and sleepy-period NFR (SPNFR) crash models. The independent variables were included in the models if they were significant at the 95 percent confidence level. The coefficients estimated provide a good indication of the strength of each of the independent variable in contributing to

SPFR and SPNFR crashes.

Table 4.5 shows the estimated parameters for the SPFR and SPNFR models. The overdispersion parameters in both models are significant (indicating that the mean is larger than the variance), and confirm the appropriateness of the negative binomial as opposed to the Poisson formulation for predicting SPFR and SPNFR crash rates.

Table 4.5 Negative Binomial Models for SPFR and SPNFR Traffic Crashes

	SPFR		SPNFR	
Parameter/variable	**Estimate**	**Pr > ChiSq**	**Estimate**	**Pr > ChiSq**
β_0 (Constant)	-5.8259	<.0001	-4.6900	<.0001
α (AAADT)	0.4196	<.0001	0.5506	<.0001
β_1 (SPDLIMT)	0.0165	<.0001	0.0061	<.0001
β_2 (CURV_HI)	NS		0.0236	<.0001
B_3 (GRAD_HI)	0.0526	<.0001	0.0166	<.0001
B_4 (SURF_WID)	-0.0334	<.0001	0.0323	<.0001
β_5 (AVG_OUTSH)	-0.0557	<.0001	0.0030	<.0001
Dispersion parameter	0.6599		0.6994	

NS--- Not Significant at 95 percent confidence level

Table 4.5 shows the following results for the effects of the explanatory variables:

- AADT has a positive effect on SPFR and SPNFR crashes. The effect of AADT is greater for SPNFR crashes (0.5985) than SPFR crashes (0.4196);
- The posted speed limit has a small positive effect on both SPFR and SPNFR crashes. The effect for SPFR crashes (0.0165) is greater than the effect for SPNFR crashes (0.0028).
- Horizontal curvature is not significant for SPFR traffic crashes, but has a small positive effect on SPNFR crashes (0.0125).
- Average vertical gradient has a positive effect on SPFR and SPNFR crashes. The effect for SPNFR crashes (0.0526) is greater than the effect for SPFR crashes (0.0106).
- Pavement surface width has a negative effect on SPFR crashes (-0.0334), and a positive effect on SPNFR crashes (0.0422).

- Average outside shoulder width also has a negative effect on SPFR crashes (-0.0557), and a small positive effect on SPNFR crashes (0.0158).

Table 4.6 shows the goodness of fit tests for the SPFR and SPNFR crashes. A *p*-value is not computed for the deviance shown in Table 4.6, but a deviance that is approximately equal to its degrees of freedom is a possible indication of a good model fit. The value of the deviance divided by its degrees of freedom is about 1 in both models, which shows that the models fit the data well.

Table 4.6 Model Goodness of Fit for SPFR and SPNFR Traffic Crashes

	SPFR			**SPNFR**		
Criteria For Assessing Goodness Of Fit						
Criterion	**DF**	**Value**	**Value/DF**	**DF**	**Value**	**Value/DF**
Deviance	1.20E+04	7447.229	0.6031	1.20E+04	13997.39	1.1336
Scaled Deviance	1.20E+04	7447.229	0.6031	1.20E+04	13997.39	1.1336
Pearson Chi-Square	1.20E+04	13765.48	1.1148	1.20E+04	21613.17	1.7503
Scaled Pearson Chi-Square	1.20E+04	13765.48	1.1148	1.20E+04	21613.17	1.7503
Log Likelihood	5113.5503			300153.6093		

4.8 Summary of the GLM Results

Six variables were included in the FR, NFR and SPNFR models, and five in the SPFR model. The six variables included in the FR, NFR and SPNFR models were AADT, posted speed limit, average horizontal curvature, average vertical gradient, pavement surface width, average outside shoulder width. Average horizontal curvature was not included in the SPFR model as it was not statistically significant.

The models showed that an increase in AADT, a higher posted speed limit, an increase in average horizontal curvature, and an increase in average vertical gradient all increase the crash rate in all the models (with the exception of average horizontal curvature in the case of SPFR crashes as this variable was not included in the SPFR model). An increase in pavement surface width or average outside shoulder width increases the crash rate in the NFR and SPNFR models, and decreases the crash rate in the FR and SPFR models.

The detailed results (Tables 4.3 and 4.5) showed that an increase in AADT was by far the most critical factor. As AADT increased, the crash rate for FR, NFR, SPFR, and SPNFR all increased substantially (and especially for NFR and SPNFR crashes). In general, average outside shoulder width, average vertical gradient, and pavement surface width had more effect on crash rates than did the speed limit or average horizontal curvature.

AADT had more effect on the non fatigue crashes (NFR and SPNFR) than on the fatigue crashes (FR and SPFR), but most of the other variables had more effect on the fatigue crashes than on the non fatigue crashes. Most of the variables had more effect on crashes that occurred during the sleepy periods (both fatigue and non fatigue crashes) than on crashes that occurred during other periods of the day.

4.9 Application of the GLMs to Network Screening

The third objective of this research was to "develop safety performance functions (SPFs) to predict FR crashes on rural two-lane roads, and apply the SPFs to the task of mitigating the effects of fatigue on traffic safety at sites where the potential for improving fatigue crashes related to fatigue is greatest" (Section 1.4). GLMs can be used to predict FR crashes by estimating the relationship between safety (number of crashes) and exposure (AADT). GLMs that estimate the relationship between safety and exposure are known as SPFs. SPFs can be used by state and local highway agencies as highway safety management tools, and especially as Network Screening tools (www.safetyanalyst.org).

The purpose of Network Screening is to review the entire roadway network, or portions of the roadway network, and to identify and prioritize those sites that have potential for safety improvement (PSI). The identification of a site by network screening means that an opportunity to improve safety may exist at the site, but does not necessarily indicate that there is a correctable safety problem at the site (www.safetyanalyst.org).

SPFs can be used to screen the network for high FR crash frequency, and to identify sites with potential for FR safety improvement. Network screening methodology uses

empirical Bayes (EB) principles to estimate the PSI of a site (Persaud et al., 1999). In general terms, the EB network screening approach combines observed FR crash frequencies with FR crash frequencies predicted by the SPFs to estimate the expected FR crash frequency for a site. The EB-adjusted expected crash frequency is used for ranking and comparing the PSI among road segments of equal length. Road segments with higher than expected (or excess) FR crash frequencies are considered to have potential for safety improvement (i.e., they are PSI sites). These sites are ranked higher than sites where expected FR crash frequencies are low. The network screening procedure then applies a cut off to create a list of sites that will be considered for appropriate countermeasures.

EB methods are a class of methods that use empirical data to evaluate/approximate the conditional probability distributions that arise from Bayes' theorem (George, 1985). The methods allow one to estimate quantities (probabilities, averages, etc.) about an individual member of a population by combining information from empirical measurements of the individual with information from empirical measurements of the entire population. The use of EB methods for the estimation of road safety increases the precision of estimation, and corrects for regression to mean bias (Hauer, 2001).

The following sections provides a simple numerical example of the calculation of the PSI for FR crashes for a site (road segment), and applies network screening to the Ohio data using this study's FR models.

Simple Numerical Example of PSI Estimation

This section provides of simple numerical example of a PSI estimation for a road segment. In this example, there is one year of FR crash counts. The road segment is part of a two-lane rural road, is 2.0 miles long, has an AADT of 5,000, a speed limit of 55 mile/hr, an average curvature of 2, an average gradient of 3, a surface width of 24 feet, average outside shoulder width of 8 feet, and 2 FR crashes recorded in the last year. The estimation of the PSI of this road segment follows the three steps presented below.

Step 1: Calculate the Mean and Variance of the Site (Road Segment)

The SPF prediction for the frequency of FR crashes on similar roads is $5000^{0.4165} \times e^{(-4.2481+0.0166\times55+0.0143\times2+0.0434\times3-0.0251\times24-0.0453\times8)} = 0.55$ FR crashes/mile-year with an overdispersion parameter =1.47 ($1/k$). Therefore, segments that are 2 miles long are expected to have 1.1 (2×0.55=1.1) crashes in one year.

Step 2: Calculate the Weight for the Site

A 'weight' is needed for combining the 2 crashes recorded on this road with the 1.1 crashes predicted for an average road of this kind. In general, the 'weight' is given by:

$$Weight = \frac{1}{1 + (\mu \times Y) / \varphi} \tag{4.3}$$

where, μ is crashes/(km-year), and 'Y' is the number of years of crash count data. Here μ=0.55 FR crashes/(mile-year), Y=1 and the estimate of φ=1.47. Therefore, weight = 1/[1+(0.55×1)/1.47] =0.72.

Step 3: Calculate the Expected Crash Frequency of the Site

Use the equation below to estimate expected crash frequency for the specific road segment at hand.

Estimate of the Expected Accidents for an entity =
Weight × Crashes expected on similar entities + (1-Weight)×Count of crashes on this entity

where 0=<Weight=<1 (4.4)

The "Estimate" is 0.72×1.1+0.28×2=1.36 FR crashes in one year. Note that 1.36 is between the average for similar sites (1.1) and the crash count for this site (2). The EB estimator pulls the crash count towards the mean and thereby accounts for regression to mean bias.

The "Estimate" (1.36 crashes/year) is the PSI value and can be used to rank the 2 mile

long segments in terms of their potential for FR crashes and, therefore, their potential for FR safety improvement.

Application of the Study's FR Crash Prediction Models to Ohio Network Screening

It is common practice for agencies with jurisdiction over extensive road infrastructure to identify and rectify hazardous locations. A two-stage process is usually used to identify and rectify the hazardous locations.

The first stage is the network screening. The network screening reviews the past crash history of the roadway network of interest, and screens and prioritizes a limited number of high risk locations which have promise as sites for potential safety improvements and, therefore, merit further investigation.

The second stage conducts detailed engineering studies on the selected sites. Only a limited number of the PSI studies can usually be investigated by a highway agency in any one year because the investigation process is expensive. The studies assess possible countermeasures for the individual sites, and recommend the most cost-effective remedial actions.

This section applies the FR SPF developed in this study to the network screening of the Ohio data. The SPFs were calculated using Equation 4.2 and the coefficients in Table 4.3.The objective of this application was to perform a network screening of FR crashes, to rank the sites by their PSI, and to identify the 50 road segments with the greatest potential for FR safety improvement.

The PSI was estimated as the EB expected crash frequency, and then normalized by dividing by the segment length. The EB expected crash frequency was estimated using the following equation:

Estimate of expected crashes per mile for an entity (EB) = [α×Crashes expected on similar entities + (1 – α) ×Observed crash frequency on this entity]/ Segment length of the entity

Where:
Crashes expected on similar entities = The SPF prediction for the entity
α (0 =<α=<1) is calculated as a function of k, the dispersion parameter of the regression model used as the SPF,

$$\alpha = 1/(1+k\times(SPF\ Prediction^2)/SPF\ Prediction) \quad (4.5)$$

The roadway segments were ranked by their PSI (EB expected crash frequency per mile). The 50 sites with the greatest potential for FR safety improvement were selected. Figure 4.2 is a flow chart showing the network screening process followed to rank the Ohio FR crashes.

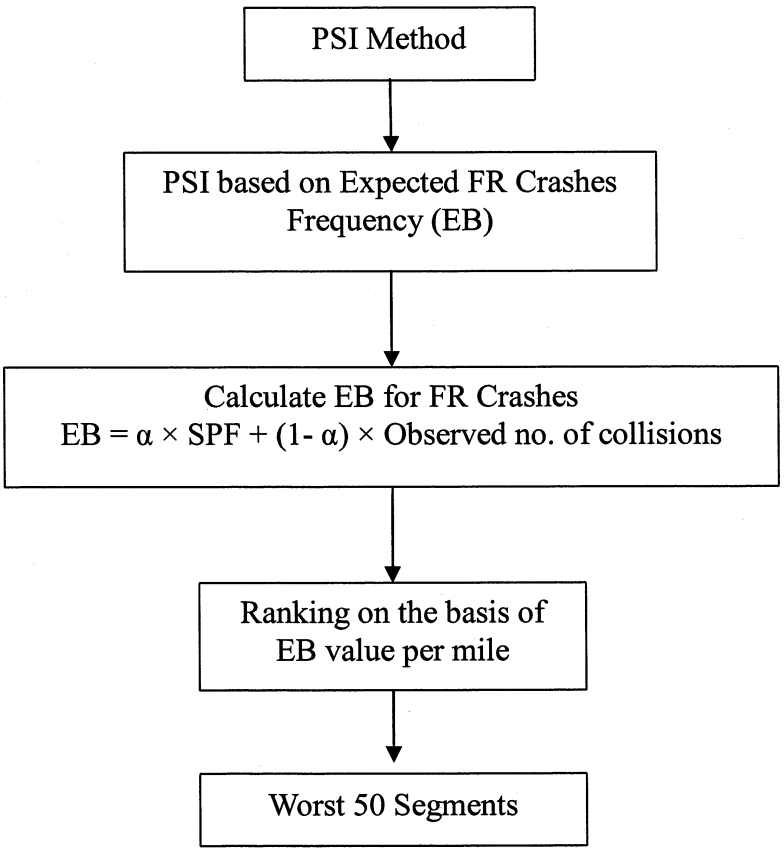

Figure 4.2 Flow Chart of Network Screening of FR crashes

The results of the network screening are shown in Table 4.7. The sites with the greatest

PSI were compared with sites which were similar (defined as having AADT of +/- 10 percent), but which had the lowest PSI. The objective of this exercise was to investigate the difference between 50 comparable sites with the highest and lowest PSI values. Table 4.8 lists the comparable 50 sites with the lowest PSI values.

Table 4.9 shows the weighted averages of with the exception of average horizontal curvature in the case of SPFR crashes the modelling variables in Tables 4.7 and 4.8. As might be expected, the 50 sites with the highest PSI values have higher speed limits, greater average curvature, higher average gradients, narrower pavement surface width, and narrower average outside shoulder width than the 50 comparable sites with the lowest PSI values. The differences appeared most marked for average curvature, average gradient, and average outside shoulder width.

Table 4.7 Network Screening and Ranking of FR Crash Sites (Road Segments): 50 sites with highest PSI values

Rank	Site No	EB value per Mile	Observed FR Crashes	SPF Calibrated Crashes Per Mile	Average AADT	Speed Limit (MPH)	Average curvature	Average gradient	Pavement Surface Width (Feet)	Average Outside shoulder Width (Feet)	Roadway Segment Length (Miles)
1	1233	16.5363	7	0.66682	8,234	55	0.2500	0	24	5	0.16
2	10246	15.3767	6	0.72092	2,640	55	1.3333	9	20	4	0.16
3	2292	13.3779	3	0.82301	8,193	55	0	0	19	3	0.12
4	6872	12.8345	3	0.67793	8,278	55	1.2500	0	24	5	0.11
5	3537	12.5399	14	0.60471	3,481	55	0.1429	5	24	4	0.36
6	9068	11.8118	35	0.93283	8,642	55	0.4650	5	22	4	1.20
7	773	11.5482	9	0.41539	5,539	35	0.3333	0	24	4.5	0.20
8	5462	11.4516	10	1.19743	3,828	55	1.2769	15	19	2.5	0.45
9	3536	11.3718	9	0.81998	3,498	55	1.5312	7	18	3	0.33
10	9323	11.3208	33	0.81816	5,046	45	3	7	22	1	1.09
11	9984	11.1678	3	0.55650	1,530	55	0	5	19	1	0.11
12	9792	11.1359	3	1.19416	15,453	55	0.4167	6.2857	24	4	0.18
13	4745	11.1005	3	0.87182	11,011	55	0	0	18	5	0.15
14	2370	11.0550	2	0.68677	10,684	45	0	0	24	3	0.10
15	1634	10.7127	2	0.86451	17,271	55	0	0	24	6	0.12
16	5763	10.5076	12	0.76698	4,594	55	1.6694	5	20	4	0.44
17	146	10.2950	2	0.71164	9,169	55	1.6667	0	24	5	0.11
18	9229	10.2607	5	0.73307	4,793	55	0.4286	6.0476	20	6	0.21
19	1218	9.9934	2	1.11744	27,973	65	0	0	32	4	0.15
20	3208	9.8962	5	0.85155	5,504	55	4.2500	9	20	8	0.24
21	210	9.7509	39	0.93939	7,307	55	0.3157	5.8906	24	2	1.62
22	6145	9.7485	6	0.62958	5,752	55	0	0	22	4	0.23
23	3468	9.6610	5	0.71822	6,816	60	0.7083	4	32	3	0.22
24	4083	9.6485	3	0.70313	2,549	55	0.3116	8.0541	20	3	0.15

Rank	Site No	EB value per Mile	Observed FR Crashes	SPF Calibrated Crashes Per Mile	Average AADT	Speed Limit (MPH)	Average curvature	Average gradient	Pavement Surface Width (Feet)	Average Outside shoulder Width (Feet)	Roadway Segment Length (Miles)
25	1750	9.6200	7	0.48893	5,377	50	2.0400	0	24	3	0.22
26	1311	9.5545	2	0.92157	22,422	45	0	0	30	0	0.14
27	516	9.4983	8	0.39162	2,056	50	0.8182	0	18	2	0.21
28	9844	9.4688	1	0.91379	4,270	55	0	6	18	1	0.10
29	2356	9.4623	8	0.6899	6,181	55	0.5586	4	34	0	0.32
30	786	9.4412	6	0.49657	7,049	55	0	0	24	10	0.20
31	2785	9.4405	76	1.11313	8,149	55	0.0769	6.5192	20	2	3.54
32	4139	9.3338	3	1.14705	5,088	55	0.2125	14	24	2	0.21
33	598	9.2960	5	0.67853	4,780	55	0.7857	0	20	2	0.22
34	7769	9.1340	24	0.88627	6,741	55	0	3	18	3	1.05
35	4154	9.1127	4	0.40585	3,537	45	0	0	20	3	0.13
36	4045	9.0855	2	0.67297	3,554	55	0	5	20	4	0.12
37	7781	9.0330	38	0.91655	5,068	55	0	8.8256	22	3	1.68
38	1083	9.0221	1	0.84583	16,019	55	7	0	24	8	0.10
39	8828	9.0202	6	0.43704	1,562	55	2.2500	0	18	3	0.19
40	5458	9.0131	16	0.74042	6,390	55	1.6000	0	20	3	0.65
41	114	8.9452	3	0.90098	12,631	55	2.5	0	24	3	0.19
42	4699	8.9303	3	0.51517	1,530	45	27	0	24	0	0.13
43	6986	8.9200	5	0.52103	3,652	55	0	0	22	4	0.19
44	7790	8.8313	15	0.56738	3,233	55	0	0	22	1	0.52
45	1020	8.818	2	0.72468	10,038	35	0	6	30	0	0.13
46	6268	8.7447	12	1.25497	10,701	55	0	8.7805	20	4	0.71
47	7771	8.7232	10	1.15844	7,110	55	2	8	18	3	0.58
48	730	8.6717	10	0.48788	8,247	50	0	0	24	10	0.33
49	3063	8.6596	8	0.63433	1,590	55	1.075	11	20	4	0.33
50	2231	8.6214	2	0.62397	4,015	55	0	0	20	2	0.12

Table 4.8 Network Screening and Ranking of Comparable FR Crash Sites (Road Segments): 50 sites with lowest PSI values

Rank	Original Rank	Site No	EB value per Mile	Observed FR Crashes	SPF Calibrated Crashes per Mile	Average AADT	Speed Limit (MPH)	Average curvature	Average gradient	Pavement Surface Width (Feet)	Average Outside shoulder Width (Feet)	Roadway Segment Length (Miles)
1	9092	181	0.44740	1	0.73469	11,083	55	0.85710	0	20	8	1.84
2	9104	1578	0.44538	1	1.01108	14,302	55	0.13670	5.37356	24	6	2.26
3	9109	1039	0.44481	2	0.84097	10,501	55	0.11780	6.18750	24	8	2.84
4	9641	1038	0.30003	0	0.54248	9,472	50	0.34122	6.02990	36	10	1.32
5	9661	1791	0.29423	0	0.68052	7,897	40	0.07725	5.97100	24	8	1.58
6	9661	1791	0.29423	0	0.68052	7,897	50	0.07725	5.97100	24	8	1.58
7	9755	2078	0.26618	2	0.70804	7,138	55	0.07404	3.85450	24	6	4.24
8	9801	7871	0.24944	3	0.57778	6,403	55	0.17122	4.35710	24	10	5.06
9	9801	7871	0.24944	3	0.57778	6,403	55	0.17122	0	24	10	5.06
10	9801	7871	0.24944	3	0.57778	6,403	55	0.17122	0	24	10	5.06
11	9801	7871	0.24944	3	0.57778	6,403	35	0.17122	4.35710	24	10	5.06
12	9814	1636	0.24579	0	0.92623	17,647	55	0.55360	5.37500	24	10	2.31
13	9883	2633	0.22135	0	1.01475	8,978	55	0.21680	0	24	4	2.71
14	9884	6887	0.22080	0	1.20766	10,313	55	0.30680	5.48280	19	2	3
15	10009	5338	0.1762	0	0.68869	5,979	55	0.09400	4.90980	24	6	2.66
16	10025	1098	0.17112	1	0.74370	5,459	55	0.02750	5.48760	24	4	4.85
17	10025	1098	0.17112	1	0.74370	5,459	55	0.02750	5.48760	24	4	4.85
18	10034	6857	0.16825	0	0.91524	5,856	55	0.13740	5.16000	20	2	3.35
19	10034	6857	0.16825	0	0.91524	5,856	55	0.13740	5.16000	20	2	3.35
20	10034	6857	0.16825	0	0.91524	5,856	55	0.13740	5.16000	20	2	3.35
21	10048	5407	0.16144	0	0.63661	8,618	45	0.26880	0	20	5	2.75
22	10048	5407	0.16144	0	0.63661	8,618	45	0.26880	0	20	5	2.75
23	10048	5407	0.16144	0	0.63661	8,618	45	0.26880	0	20	5	2.75
24	10048	5407	0.16144	0	0.63661	11,083	45	0.26880	0	20	5	2.75

Rank	Original Rank	Site No	EB value per Mile	Observed FR Crashes	SPF Calibrated Crashes per Mile	Average AADT	Speed Limit (MPH)	Average curvature	Average gradient	Pavement Surface Width (Feet)	Average Outside shoulder Width (Feet)	Roadway Segment Length (Miles)
25	10070	1061	0.15092	0	1.02144	12,118	55	0.34259	4	24	3	3.99
26	10136	4346	0.12642	0	0.85093	4,681	55	0.01964	6.6003	18	4	4.26
27	10136	4346	0.12642	0	0.85093	4,681	55	0.01964	6.6003	18	4	4.26
28	10136	4346	0.12642	0	0.85093	4,681	55	0.01964	0	18	4	4.26
29	10136	4346	0.12642	0	0.85093	4,681	55	0.01964	0	18	4	4.26
30	10136	4346	0.12642	0	0.85093	4,681	55	0.01964	6.6003	18	4	4.26
31	10136	4346	0.12642	0	0.85093	4,681	55	0.01964	6.6003	18	4	4.26
32	10136	4346	0.12642	0	0.85093	4,681	55	0.01964	0	18	4	4.26
33	10137	3621	0.12585	0	0.77511	3,974	55	0.33551	5.9167	18	4	4.03
34	10149	1873	0.11892	0	0.73295	3,557	55	0.11591	6.1034	24	1	4.11
35	10154	7115	0.11783	0	0.59903	3,250	55	0.13976	0	20	1	3.61
36	10154	7115	0.11783	0	0.59903	3,250	55	0.13976	0	20	1	3.61
37	10154	7115	0.11783	0	0.59903	3,250	55	0.13976	0	20	1	3.61
38	10154	7115	0.11783	0	0.59903	3,250	55	0.13976	0	20	1	3.61
39	10180	3990	0.1055	0	0.52020	2,111	55	0.02536	5.7951	23	4	3.64
40	10188	9642	0.10017	1	0.42778	1,342	45	2.02000	0	18	2	5.56
41	10195	2601	0.09618	0	0.55740	4,228	55	0.41701	5.3444	24	8	4.2
42	10195	2601	0.09618	0	0.55740	4,228	55	0.41701	0	24	8	4.2
43	10202	1678	0.093429	0	0.87581	10,018	55	0.05045	3.4223	24	4	5.87
44	10202	1678	0.093429	0	0.87581	10,018	55	0.05045	3.4223	24	4	5.87
45	10211	3531	0.084751	0	0.55963	3,351	55	0.39646	5.5865	24	6	4.78
46	10220	4812	0.08077	0	0.44400	2,509	55	0.08480	0	24	3	4.22
47	10220	4812	0.08077	0	0.44400	2,509	55	0.08480	0	24	3	4.22
48	10232	9270	0.068644	2	0.39809	1,434	45	0.09689	0	20	8	10.78
49	10232	9270	0.068644	2	0.39809	1,434	55	0.09689	0	20	8	10.78
50	10232	9270	0.068644	2	0.39809	1,434	45	0.09689	0	20	8	10.78

Table 4.9 Weighted Averages of the Modelling Variables in Tables 4.8 and 4.9

	Average AADT	Average Speed Limit (MPH)	Average curvature	Average gradient	Pavement Surface Width (Feet)	Average Outside Shoulder Width (Feet)
Sites with Highest PSI Values	6,056	53.2	1.345	3.758	22.34	3.48
Comparable Sites with Lowest PSI Values	6,196	52.7	0.208	2.411	22.8	5.12

5 LOGISTIC REGRESSION ANALYSIS

5.1 Introduction

In addition to the GLM analysis discussed in Chapter 4, this study used logistic regression models to analyze FR crash occurrence, and identify potential explanatory variables for FR crashes. The logistic regression analysis included potential environment and driver variables as well as the road variables included in the GLM analysis. Three fatigue crash models were developed: FR crashes (FR), sleepy-period fatigue-related crashes (SP-FR), and non-sleepy-period fatigue-related crashes (NSP-FR).

This Chapter describes the data collection, data editing, and data analysis used in the logistic regression modelling, and presents the results of the logistic regression analysis. Figure 5.1 is a flowchart showing the logistic regression model study process.

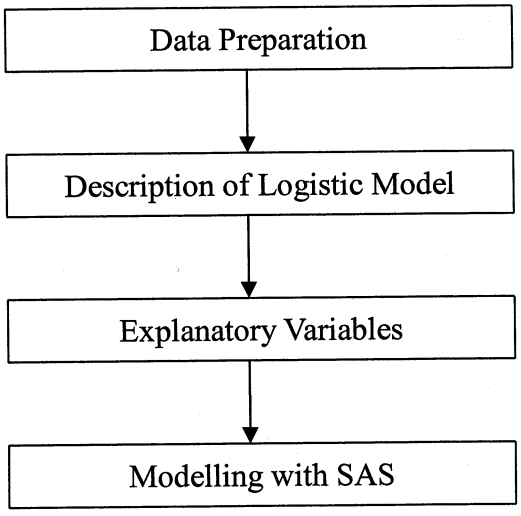

Figure 5.1 Flow Chart showing the Study Process of the Logistic Regression Model

The discussion of the logistic regression analysis is presented as follows:

Section 5.2 describes the data preparation for the logistic regression analysis;

Section 5.3 describes the logistic regression model;

Section 5.4 discusses the goodness of fit testing;

Section 5.5 discusses the explanatory variables;

Section 5.6 discusses modelling with SAS;

Section 5.7 presents the estimation results; and

Section **Error! Reference source not found.** outlines limitations of the logistic regression analysis.

5.2 Data Preparation for Logistic Regression Analysis

The data used in the logistic regression analysis were extracted from the Ohio Traffic Accident Database obtained from the HSIS Database: 36,341 crashes on rural two-lane roads were extracted for the logistic regression analysis. Each crash was regarded as an observation in the logistic regression analysis. The single vehicle run-off-road crashes were assumed to be FR crashes. To develop the logistic model, detailed data on crash data, traffic flow and road design were also extracted.

The logistic regression analysis used two mutually exclusive entities, (1) and (0), for the FR and NFR crash respectively. A FR crash was assigned a value of 1, and a NFR crash a value of 0.

The process of creating a dataset suitable for the logistic regression analysis is discussed below. The SAS Codes are listed in Appendix E.

The accident and vehicle subfiles were linked together using the accident report number (i.e., CASENO). The accident subfile was linked to the roadlog file using the CNTYRTE and MILEPOST variables in the accident record, and the CNTY_RTE, BEGMP, and ENDMP variables in the roadlog file. Similarly, the accident subfiles were linked to Curve, Grade, and Angle Points using similar variables found in each respective file.

To prepare the accident subfiles for linking with the roadlog file using a SAS data step process, both of them were sorted into location order by CNTYRTE and MILEPOST in the crash subfiles, and by CNTY_RTE and BEGMP in the roadlog file. Similar sorts were completed for other files that had to be merged.

5.3 Description of Logistic Regression Model

The purpose of the model development was to identify highway, environment, and driver characteristics related to FR crashes. As the outcome under investigation (i.e. FR/NFR) is of a discrete nature, logistic regression was identified as a suitable approach for identifying important explanatory factors. This is because logistic regression modelling uses a set of independent (explanatory) variables to predict the probability of the occurrence of a discrete dependent variable.

As Dissanayake (2003) points out, "This logistic regression equation models the logit transformation of the individual event probability, p_i, as a linear function of the explanatory variables in the vector, X_i. A more general class of models shares the feature that a function $g=g(\mu)$ of the response variable is assumed to be linearly related to the explanatory variables. The function g is known as the 'link function'. " Other common link functions include the Normit function (used in probit analysis), and the complementary log-log function. Since the logit function has advantages, such as being relatively easy to interpret, it is used in developing event/non-event models such as for FR/non-FR crashes. Ben-Akiva and Lerman (1993) provide additional theoretical information on link functions.

5.4 Goodness of Fit Tests

The SAS programming language allows the user to print tables and statistics to help analyze and evaluate the estimated logistic regression models developed using the LOGISTIC Procedure (SAS, 1998b). The results of testing the null hypothesis provide two criteria (AIC and SC) that are useful for comparing models. Two other criteria ($-2logL$ and Score) test the null hypothesis that all regression coefficients are zero. Except for the score statistic, all of the criteria are based on the likelihood for fitting a model with intercepts only, or fitting a model with intercepts and explanatory variables. The details of the four criteria are as follows:

- AIC is the Akaike information criterion which is a goodness of fit measure that could be used to compare one model to another. Lower values indicate a more desirable

model.

- SC is the Schwarz criterion which is also a goodness of fit measure that can be used to compare one model to another. Lower values indicate a more desirable model.
- *–2logL* is the –2log likelihood statistic which has a chi-square distribution under the null hypothesis that all regression coefficients of the model are zero which provides a *p*-value for the chi-square statistic. A significant *p*-value (for example a *p*-value less than 0.05) provides evidence that at least one of the regression coefficients for an explanatory variable is non-zero.
- Score is a score statistic which also outputs the chi-square value, degrees of freedom, and a *p*-value for this statistic.

5.5 Explanatory Variables

Table 5.1 lists the seven potential explanatory variables considered in the model development. Light Condition, Road Characteristic (alignment), Driver Age Group, and Driver Sex were treated as categorical variables. Speed Limit, Road Width, and AADT were treated as continuous variables.

Table 5.1 Details of Variables selected as Potential Explanatory Variables for FR Crashes

Variable Characteristic		Frequency	Percent
LIGHT	Light Condition		
	DAYLIGHT	19,332	53.20
	DARK-NO-LIGHTS	14,468	39.81
	DAWN	957	2.63
	DARK-LIGHTED	886	2.44
	DUSK	698	1.92
RD_CHAR	Road Characteristic		
	STRAIGHT-LEVEL	16,655	45.83
	CURVE-GRADE	7,155	19.69
	STRAIGHT-GRADE	6,379	17.55
	CURVE-LEVEL	6,152	16.93
AGE_GROUP	Driver Age Group		
	16-25	14,395	41.07
	26-45	13,065	37.28
	46+	7,881	21.65
DRV_SEX	Driver Sex		
	Male	23,809	65.52
	Female	12,532	34.48
SPD_LIMIT	Speed Limit (mph)		
	55	32,515	88.87
	45	2,094	5.76
	35	710	1.95
	40	484	1.33
	50	333	0.92
	25	120	0.33
	65	41	0.11
	60	23	0.06
	20	8	0.02
	15	7	0.02
	30	6	0.01
RD_WIDTH	Road Width (Feet)	Range from 24 to 48	
AADT	Annual Average Daily Traffic (Veh)	Range from 250 to 11,750	

The information presented in Table 5.1 provides only simple descriptions of the variables,

making no allowance for inter-correlations among the variables and ignoring possibly important confounding factors, but the following points may be noted:

- Light Condition: Most crashes occurred during DAYLIGHT (53%) and DARK-NO-LIGHTS (40%).
- Road Characteristic: STRAIGHT-LEVEL accounted for 46 percent of the crashes.
- Driver Sex: 65 percent of the drivers were male.
- Age: 41 percent of the drivers were young (16 to 25), and 37 percent were 26 to 45 years old.
- Speed Limit: The great majority of the crashes occurred on roads with a posted speed limit of 55 mph.
- Road Width: The road width ranged from 24 to 48 feet.
- AADT: The AADT ranged from 250 to 11,750.

5.6 Modelling with SAS

SAS was applied to the logistic regression modelling using the PROC LOGISTIC module. This module can be used to develop multivariate logistic regression models, and allows for a variety of input parameters to be used for model building and significance testing. The SAS programs created for the logistic regression modelling used in this study are listed in Appendix F.

Since the models created in this study use both continuous and categorical independent variables, it was necessary to specify in SAS which variables were categorical. This was done using the CLASS statement. Any variable not defined as categorical was assumed to be a continuous variable. The CLASS statement was used to define a reference category within each categorical independent variable. The reference category is used in logistic regression modelling as a way of redefining categorical variables as a series of dichotomous variables. For any independent categorical value with n categories, the use of the CLASS statement in PROC LOGISTIC will convert the independent variable to a series of $n-1$ dichotomous variables.

The MODEL statement in PROC LOGISTIC was used to specify the modelling parameters,

and to conduct any additional significance tests. A stepwise modelling procedure was used for the data used in the logistic regression modelling. Specified significance levels were required for a variable to be entered into the model. After the MODEL statement, the dependent variable to be tested was listed, followed by an equal sign, and followed by the full list of independent variables to be tested in the model, separated by spaces. The MODEL statement supports a variety of additional options which instruct SAS to perform additional tests of significance on the data. A combination of the options SCALE=NONE and AGGREGATE prompt SAS to display the Pearson goodness of fit test and the Deviance goodness of fit test. Although these two tests are appropriate for assessing logistic regression models in SAS, they typically perform better when all of the variables are continuous rather than categorical.

In the stepwise selection modelling procedure, the final regression model was created by starting with a flat intercept model with no variables, and successively adding the independent variables to the model one at a time in order of significance. By specifying *p*-value cutoff levels within SAS, the point at which no more variables will be entered into the model can be determined using a significance level. A *p*-value cutoff level can also be specified, as required, for a *p*-value that is already in the model to remain in the model. In other words, a variable can be removed from the model at a later step if the variable is longer found not to be significant when placed with other variables. For the stepwise selection model, it was necessary to use the SELECTION = STEPWISE option. For these logistic models, a *p*-value of 0.05 was used as the cutoff for a variable to be input into the model, and as the cutoff for a variable to remain in the model. The small *p*-value for these models was selected to make the models as simple as possible.

5.7 Estimation Results

Three sets of results are presented: FR crashes, SP-FR crashes, and NSP-FR crashes. The modeling examined the crashes and the effects of differences in the seven independent variables selected for the modelling. Independent variables that were significant at the 95 percent confidence level were included in the models.

Tables 5.2, 5.5, and 5.8 show the results for FR, SP-FR, and NSP-FR crashes respectively. The results presented in these Tables are the odds ratios, the estimate, and the Pr > ChiSq, The odds ratio measures the odds of the outcome increasing if the value of the independent variable increases. The ratio provides a good indication of the strength of the effect of each of independent variable on FR crashes. The coefficients of the independent variables estimated by the models are directly related to the probability of having a FR crash: a positive coefficient indicates that a variable increases the probability of having a more serious crash outcome, and vice versa. Model fit statistics and null hypothesis tests are also provided for the model for each crash type.

FR Crashes

Table 5.2 shows the results of the logistic regression for FR crashes. Road Width was not significant.

Table 5.2 Logistic Regression Model Results for FR Crashes

Variables				Odds Ratio	Estimate	Pr > ChiSq
INTERCEPT					2.206400	<.0001
LIGHT	DAYLIGHT	vs.	DARK-LIGHTED	1.116	-0.142700	<.0001
	DAWN	vs.	DARK-LIGHTED	1.622	0.231200	<.0001
	DUSK	vs.	DARK-LIGHTED	1.631	0.236500	<.0001
	DARK-NO-LIGHTS	vs.	DARK-LIGHTED	1.197	-0.072500	<.0001
RD_CHAR	STRAIGHT-LEVEL	vs.	CURVE-GRADE	4.518	0.820100	<.0001
	STRAIGHT-GRADE	vs.	CURVE-GRADE	3.378	0.529300	<.0001
	CURVE-LEVEL	vs.	CURVE-GRADE	1.027	-0.661400	<.0001
AGE_GROUP	16-25	vs.	46+	0.433	-0.445700	<.0001
	26-45	vs.	46+	0.715	0.055100	<.0001
DRV_SEX	Female	vs.	Male	1.094	0.044800	<.0001
SPD_LIMIT	5 mile/hr increase			0.863	-0.029400	<.0001
RD_WIDTH	Not significant					
AADT	1,000 veh increase			1.049	0.000048	<.0001

The odds of a FR crash being related to fatigue are highest for:

- road sections with straight and level alignment (OR=4.518) compared with curved and graded sections; and

- road sections with straight and level alignment (OR=3.378) compared with curved and graded sections.

Table 5.2 also shows that the odds of a FR crash being related to fatigue are higher for:

- dusk conditions (OR=1.631) compared with dark, but lighted conditions;
- dawn conditions (OR=1.622) compared with dark, but lighted conditions;
- dark with no light conditions (OR=1.197) compared with dark, but lighted conditions;
- daylight conditions (OR=1.116) compared with dark, but lighted conditions;
- female drivers (OR=1.094) compared with male drivers;
- road sections with curved and level alignment (OR=1.027) compared with curved and graded sections; and
- older drivers (46+) compared with younger drivers (26-45 and 16-25) (OR=0.715 and 0.433 respectively).

Table 5.2 shows that, for the continuous variables, the odds of a crash being related to fatigue change as follows:

- as the speed limit increases, each increase of 5 miles per hour decreases the odds of a FR crash occurring by 13.7% (OR=0.863); and
- as AADT increases, each increase of 1,000 vehicles increases the odds of a FR crash occurring by 4.9% (OR=1.049).

These results do not suggest that we should increase speed limits and welcome increases in AADT as a way to reduce the odds of a FR crash occurring. When speeds are higher and when there are more vehicles on the road, the additional demands required of the driver may increase driver concentration and overcome fatigue. On the other hand, fatigue undermines the driver's ability to react quickly and this is clearly a problem when speeds and the demands made on maneuvering the vehicle are high. It is not possible to draw firm conclusions as each driver's fatigue levels are not known.

Table 5.3 shows the model fit statistics for the FR model. The Table shows the AIC, SC, and -2 Log L statistics. Table 5.4 shows the null hypothesis tests for the FR model. The results of

three tests are presented: the Likelihood Ratio, Score, and Wald tests.

Table 5.3 Logistic Model Fit Statistics for FR Crashes

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	159864.57	144641.28
SC	159874.45	144769.69
-2 Log L	159862.57	144615.28

Table 5.4 Global Null Hypothesis Tests of Logistic Regression Model for FR Crashes

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	15247.2884	12	<.0001
Score	15994.9079	12	<.0001
Wald	13872.3391	12	<.0001

The significant *p*-values (<0.05) for –2logL (Table 5.3) and Score (Table 5.4) indicate that at least one of the regression coefficients is non-zero.

The residual chi-square values were examined to evaluate the effectiveness of the explanatory variables that were not entered into the models. Their *p*-values less than 0.05 indicate that at least one of the excluded variables' parameter coefficients is non-zero. As all the residual chi-square values examined were smaller than 0.05, the effects of the variables that were entered into the models appear to be minimal.

The Score and Wald tests have *p*-values less than 0.05, indicating that the model fit is adequate. The *p*-value of the likelihood ratio chi-square test in Table 5.4 is less than 0.001 ($\chi 2$ = 15247.2884, DF=12) which means that the global null hypothesis for the whole model is rejected. Statistically, this result indicates that the model's predictor variables affect the occurrence of FR crashes.

SP-FR Crashes

Table 5.5 shows the results of the logistic regression for SP-FR crashes. Road Width was not significant.

Table 5.5 Logistic Regression Model Results for SP-FR Crashes

Variables				Odds Ratio	Estimate	Pr > ChiSq
INTERCEPT					1.951800	<.0001
LIGHT	DAYLIGHT	vs.	DARK-LIGHTED	3.230	0.209300	<.0001
	DAWN	vs.	DARK-LIGHTED	4.184	0.468100	
	DUSK	vs.	DARK-LIGHTED	5.285	0.701700	
	DARK-NO-LIGHTS	vs.	DARK-LIGHTED	1.729	-0.415900	
RD_CHAR1	STRAIGHT-LEVEL	vs.	CURVE-GRADE	4.017	0.753700	<.0001
	STRAIGHT-GRADE	vs.	CURVE-GRADE	3.201	0.526400	
	CURVE-LEVEL	vs.	CURVE-GRADE	0.994	-0.643300	
AGE_GROUP	16-25	vs.	46+	0.505	-0.367100	<.0001
	26-45	vs.	46+	0.767	0.050800	
DRV_SEX	Female	vs.	Male	1.069	0.033300	0.0446
SPD_LIMIT	5 mile/hr increase			0.859	-0.030200	<.0001
RD_WIDTH	Not significant					
AADT	1000 veh increase			1.059	0.000057	<.0001

Table 5.5 shows that the odds of a SP-FR crash being related to fatigue are higher for:

- dusk conditions (OR=5.285) compared with dark, but lighted conditions;
- dawn conditions (OR=4.184) compared with dark, but lighted conditions;
- road sections with straight with a grade (OR=4.017) compared with curved and graded sections;
- daylight conditions (OR=3.230) compared with dark, but lighted conditions;
- road sections with straight alignment and a grade (OR=3.201) compared with curved and graded sections;
- dark with no light conditions (OR=1.729) compared with dark, but lighted conditions;
- female drivers (OR=1.069) compared with male drivers.
- older drivers (46+) compared with younger drivers (26-45 and 16-25) (OR=0.767 and 0.505 respectively).

Table 5.5 shows that, for the continuous variables, the odds of a crash being related to fatigue

change as follows:

- as the speed limit increases, each increase of 5 mile per hour decreases the odds of a FR crash occurring by 14.1% (OR=0.859); and
- as AADT increases, each increase of 1,000 vehicles increases the odds of a FR crash occurring by 5.9% (OR=1.059).

These findings require careful interpretation as discussed in Section 0.

Table 5.6 shows the model fit statistics for the SP-FR model. The Table shows the AIC, SC, and -2 Log L statistics. Table 5.7 shows the null hypothesis tests for the SP-FR model. The results of three tests are presented: the Likelihood Ratio, Score, and Wald tests.

Table 5.6 Logistic Regression Model Fit Statistics for SP-FR Crashes

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	28058.104	25081.088
SC	28066.171	25177.897
-2 Log L	28056.104	25057.088

Table 5.7 Global Null Hypothesis Tests of Logistic Regression Model for SP-FR Crashes

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	2999.0158	12	<.0001
Score	2990.4107	12	<.0001
Wald	2571.7912	12	<.0001

The significant *p*-values (<0.05) for –2logL (Table 5.6) and Score (Table 5.7) indicate that at least one of the regression coefficients is non-zero. The results indicate that the model fit is adequate. The *p*-value of the likelihood ratio chi-square test in Table 5.7 is less than 0.001 ($\chi 2 = 2999.0158$, DF=12) which means that the global null hypothesis for the whole model is rejected. Statistically, this result indicates that the model's predictor variables affect the occurrence of SP-FR crashes.

NSP-FR Crashes

Table 5.8 shows the results of the logistic regression for NSP-FR crashes. Road Width was not significant.

Table 5.8 Logistic Regression Model Results for NSP-FR Crashes

Variables				Odds Ratio	Estimate	Pr > ChiSq
INTERCEPT					2.243100	<.0001
LIGHT	DAYLIGHT	vs.	DARK-LIGHTED	0.937	-0.209600	<.0001
	DAWN	vs.	DARK-LIGHTED	1.387	0.182500	
	DUSK	vs.	DARK-LIGHTED	1.394	0.187700	
	DARK-NO-LIGHTS	vs.	DARK-LIGHTED	1.137	-0.016300	
RD_CHAR1	STRAIGHT-LEVEL	vs.	CURVE-GRADE	4.619	0.820100	<.0001
	STRAIGHT-GRADE	vs.	CURVE-GRADE	3.396	0.524700	
	CURVE-LEVEL	vs.	CURVE-GRADE	1.039	-0.659300	
DRV_SEX	Female	vs.	Male	1.078	0.037500	<.0001
AGE_GROUP	16-25	vs.	46+	0.426	-0.458500	<.0001
	26-45	vs.	46+	0.718	0.063700	
SPD_LIMIT	5 mile/hr increase			0.866	-0.029200	<.0001
RD_WIDTH	Not significant					
AADT	1,000 veh increase			1.045	0.000046	<.0001

The odds of a NSP-FR crash being related to fatigue are highest for:

- road sections with straight and level alignment (OR=4.619) compared with curved and graded sections; and
- road sections with straight alignment with a grade (OR=3.396) compared with curved and graded sections.

Table 5.8 also shows that the odds of a NSP-FR crash being related to fatigue are higher for:

- dusk conditions (OR=1.394) compared with dark, but lighted conditions;

- dawn conditions (OR=1.387) compared with dark, but lighted conditions;
- dark with no light conditions (OR=1.137) compared with dark, but lighted conditions;
- female drivers (OR=1.078) compared with male drivers;
- road sections with curved and level alignment (OR=1.039) compared with curved and graded sections;
- daylight conditions (OR=0.937) compared with dark, but lighted conditions;
- older drivers (46+) compared with younger drivers (26-45 and 16-25) (OR=0.718 and 0.426 respectively).

Table 5.8 shows that, for the continuous variables, the odds of a crash being related to fatigue change as follows:

- as the speed limit increases, each increase of 5 mile per hour decreases the odds of a FR crash occurring by 13.4% (OR=0.866); and
- as AADT increases, each increase of 1,000 vehicles increases the odds of a FR crash occurring by 4.5% (OR=1.045).

These findings require careful interpretation as discussed in Section 0.

Table 5.9 shows the model fit statistics for the NSP-FR model. The Table shows the AIC, SC, and -2 Log L statistics. Table 5.10 shows the null hypothesis tests for the logistic model. The results of three tests are presented: the Likelihood Ratio, Score, and Wald tests.

Table 5.9 Logistic Regression Model Fit Statistics for NSP-FR Crashes

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	131578.05	118831
SC	131587.75	118957.09
-2 Log L	131576.05	118805

Table 5.10 Global Null Hypothesis Tests of Logistic Regression Model for NSP-FR Crashes

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	12771.0482	12	<.0001
Score	13501.9585	12	<.0001
Wald	11660.2157	12	<.0001

The results indicate that the NSP-FR model fit is adequate. The AIC and SC values of the NSP-FR model indicates that the NSP-FR model has a better goodness of fit than the FR or SPFR models.

5.8 Summary of the Logistic Regression Results

The four statistically significant categorical variables were analyzed in terms of ten comparisons. For FR and NSP-FR crashes, the odds of a crash being related to fatigue were highest for the road characteristics variables followed by the environment (light levels) variables.

The logistic regression analysis used AIC, SC, and *-2LogL* tests of model fit, and Likelihood Ratio, Score and Wald tests of the global null hypothesis. The model fit was found to be adequate in all three models. The NSP-FR model had the best goodness of fit of the three models. As the global null hypothesis for each model was rejected, the predictor variables of each model affected the occurrence of FR, SP-FR, and NSP-FR crashes.

The results for FR and NSP-FR crashes were very similar. The top five odds were for the same variable comparison for both types of crash. The odds were highest for road sections with straight and level alignment (compared with curved and graded sections), and second highest for road sections with straight alignment and a grade (compared with curved and graded sections). The third and fourth highest odds were for dusk conditions (compared with dark, but lighted conditions), and for dawn conditions (compared with dark, but lighted conditions) respectively. The fifth highest odds for both types of crash were for dark with no light conditions (compared with dark, but lighted conditions).

For SP-FR crashes, the odds of a crash being related to fatigue were generally higher than for FR and NSP-FR crashes, and the order was different. The highest odds were for dusk conditions (compared with dark, but lighted conditions), followed by dawn conditions (compared with dark, but lighted conditions), straight and level alignment (compared with curved and graded sections), daylight conditions (compared with dark, but lighted conditions), and road sections with straight alignment and a grade (compared with curved and graded sections).

The demographic variables tended to be near the bottom of the list for all three crash types. Female drivers appeared to be more involved than male drivers in fatigued driving crashes, and older drivers appeared to be more involved than younger drivers, but the odds were generally less than for the road characteristics and environment variables.

The two continuous variables, speed limit and AADT, were both statistically significant. The results were similar for all three crash types. Each increase of 5 mile per hour in the speed limit decreased the odds of a fatigued driving crash occurring by about 13% to 14%. Each increase of 1,000 vehicles in AADT increased the odds of a fatigued driving crash occurring by 4.5% (NSP-FR) to 5.9% (SP-FR).

These results all clearly need careful analysis and interpretation. For example, we cannot interpret the findings to suggest that higher speed limits and more vehicles on a two-lane rural road will help to improve the road's safety in terms of fatigued driving crashes.

6 LIMITATIONS OF THE STUDY

This study used GLM models and logistic regression analysis as two approaches to analyze rural two lane road crashes in which fatigue was suspected to play a role. Both model building techniques aimed to find the best fitting and most parsimonious, yet logically reasonable, model to describe the relationship between an outcome (dependent variable or response variable) and a set of independent (predictor or explanatory) variables. The results are encouraging, but it is important to note a number of limitations in the study and the modeling.

The first six limitations refer to general issues. Limitations 7 to 9 refer to GLM, and limitations 10 to 12 refer to logistic regression analysis.

General limitations

1. The analysis could be applied only to the variables selected for the study. The study did not account for numerous additional factors that might influence the occurrence of FR crashes. Additional factors might include individual human characteristics, individual circadian rhythms, weather conditions, and specific crash related issues such as sight distance. Some such additional factors may be important.

2. The definition of FR crashes in this study was restricted to single-vehicle run-off-road crashes. In reality, driver fatigue may also lead to some multi-vehicle crashes, to on-road crashes, and to rear-end crashes, head-on crashes, side-swipe crashes, etc., but all these crash types were ignored in the analyses. The road design and other factors related to single-vehicle run-off-road crashes may be different from the factors related to other crashes that result from fatigue. The effect of restricting the definition of FR crashes to single vehicle run-off-road crashes is not known, but it appears likely that the FR crash definition used in this paper may underestimate the incidence of FR crashes.

3. The models developed for this work were developed with traffic, crash, and road data for rural two lane roads. The models are, therefore, appropriate for rural two lane roads, but

other models would have to be developed for other types of road.

4. The models developed for this work were developed with traffic, crash, and road data from Ohio State. The models are, therefore, appropriate for (rural two lane roads in) Ohio, but other models would have to be developed for other jurisdictions. (In the case of the GLM models, the recalibration procedure developed by Harwood et al. (2000) and tested by Persaud et al. (2001) could be used to apply the models in other jurisdictions.)

5. The source of the models' data is police reports. The accuracy of the data gathered from police reported crashes is a well-known concern. A common bias may be under-reporting: police may either not be called to a crash, or may not file reports on very minor crashes. The percentage of under-reporting is lowest with fatalities, but increases for crashes in which the injuries and damage are not severe. As the logistic regression models deal with the outcomes of crashes and not the crash rates, the effect of this bias is not important, but the accuracy of recording details of the crash may have a significant influence on the models' accuracy.

6. Most of the results were intuitively acceptable, but there were exceptions. For example, the logistic models showed that an increase speed limits and AADT were associated with reduced odds of a FR crash occurring. At this stage, we cannot explain the interplay between fatigued drivers' possibly increased concentration under such conditions and fatigued drivers' possibly increased difficulty coping with the maneuvering and other demands of such conditions. It would be necessary to know each driver's fatigue levels and the details of how the fatigue affected each driver's performance.

<u>GLM models</u>

7. The variables included in the GLM models have the functional form of an exponential. When the functional form does not match the phenomenon it aims to describe, or when important explanatory variables are missing, even well estimated regression constants cannot be trusted to predict the effect on the dependent variable of a change in an explanatory variable (Hauer, 2004). Further research is needed to discover useful

regularities in observed data, and to cast these into the form of model equations. When suitable equations are available, they can be used in other jurisdictions by applying a re-estimation procedure presented by Vogt and Bared (1998) and estimated by Hauer (2001).

8. The analysis made no allowance for possible inter-correlations among the variables, or for possibly important confounding factors. Internal correlation makes it difficult to estimate the safety effects of a single explanatory variable with confidence as the effects on the variable of other variables in the model are not known. It is, therefore, important to interpret the FR crash safety effect of a single variable with caution. Nevertheless, the models can be used with caution to evaluate the implications of implementing specific countermeasures, such as providing a wider surface width or a wider shoulder width on roadway sections.

9. The models have not been validated using a second datasets. The power of the models can only be proven by showing that they can be replicated using a different and independent dataset.

Logistic Regression Analysis

10. It is important to emphasize the fact that, as the logistic regression models developed in this research are based on crash data, they explain the effects that various factors have on a crash given that the crash has occurred.

11. Exposure is a key variable in traffic research (OECD 1997; Fridstrøm et al. 1995), but as exposure was not measured in the logistic regression models, the absolute and relative exposure on roads where the FR and NFR crashes occurred is unknown.

12. The use of categorical variables may lead to a loss of useful information. For example, the categories may be inappropriate and obscure some potential findings. In this study's age categories, for example, the oldest age category was 46+ so there was no information about the oldest drivers, for example those aged 65+.

7 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

The research undertaken in response to the objectives focused on GLM and on logistic regression analysis. The GLM analysis was directed at all three objectives, but did not include environmental and demographic factors. The logistic regression analysis was directed at the first and second objectives.

Ohio State data were used in the modeling. Single vehicle run-off-road crashes were assumed to be FR crashes, and all other crash types were assumed to be NFR crashes.

This Chapter briefly summarizes the study's approach and results (Sections 7.1 to 7.6). Section 7.7 presents the study's conclusions, and Section 7.8 presents the study's recommendations.

7.1 GLM: Modelling Procedure

Six explanatory variables were tested in the GLM models: AADT, posted speed limit, average horizontal curvature, average vertical gradient, pavement surface width, and average outside shoulder width. Four separate models were calibrated in order to understand the effect of the six variables on FR, NFR, SPFR, and SPNFR crashes.

7.2 GLM: Summary of Modelling Results

The choice of the negative binomial rather than the Poisson formulation for predicting FR, NFR, SPFR, and SPNFR crash rate was considered reasonable and appropriate. The goodness of fit results also showed that the models fit the data well.

All six road and traffic variables had a statistically significant effect on FR and NFR crashes, and all but average horizontal curvature had a statistically significant effect on SPFR crashes.

AADT had by far the largest effects in all four models.

AADT had more effect on the non-fatigue crashes (NFR and SPNFR) than on the fatigue crashes (FR and SPFR), but most of the other variables had more effect on the fatigue crashes than on the non-fatigue crashes. Most of the variables also had more effect on crashes that occurred during the sleepy periods (both fatigue and non-fatigue crashes) than on crashes that occurred during all day period.

7.3 GLM: Network Screening Procedure

GLM modelling was also used to develop SPFs. The SPFs were used to screen the Ohio road network for sites where there was a high potential for safety improvement (PSI) with regard to FR crashes. The road and traffic variables of the top 50 FR sites identified were compared with the road and traffic variables of the 50 comparable sites that had the lowest PSI values.

7.4 GLM: Network Screening Results

The 50 sites with the highest PSI values had higher speed limits, greater average curvature, higher average gradients, narrower pavement surface width, and narrower average outside shoulder width than the 50 comparable sites with the lowest PSI values. Average curvature, average gradient, and average outside shoulder width appeared to show the largest differences.

7.5 Logistic Regression Analysis: Procedure

The logistic regression models were developed to identify potential explanatory variables for FR crashes, and to estimate the occurrence of FR crashes. The models predict the odds of a crash being fatigue related.

The logistic regression variables included potential environment (lighting conditions) and

driver demographic variables (sex and age) in addition to traffic volume, posted speed limit, and roadway alignment variables. Ohio data (36,341 crashes on rural two-lane roads) were used in the analysis. Three models were developed: FR crashes (FR), sleepy-period fatigue-related crashes (SP-FR), and non-sleepy-period fatigue-related crashes (NSP-FR).

7.6 Logistic Regression Analysis: Summary of Results

The model fit statistics and null hypothesis tests showed that the model fit was adequate in all three models, and that the predictor variables of each model affected the occurrence of FR, SP-FR, and NSP-FR crashes.

The results for FR and NSP-FR crashes were similar: the most important factors affecting the odds of a crash being fatigue related were alignment. FR crashes that occurred during sleepy periods (SP-FR crashes) were a little different. The odds of a SP-FR crash being related to fatigue were generally higher than the odds for FR and NSP-FR crashes, and dusk and dawn lighting scored higher than alignment.

Speed limit and AADT were analyzed as continuous variables rather than as category variables. The results showed that an increase in either the speed limit or the AADT decreased the odds of a crash occurring. As it would be unwise to conclude that high speed limits and high volumes of traffic are suitable countermeasures for AADT fatigue crashes on two-lane rural roads, the speed limit and AADT results illustrate the importance of giving all the models' results careful consideration and interpretation.

7.7 Study Conclusions

As mentioned in Section 1.2, fatigue is believed to be involved in more traffic crashes than alcohol or prescription drugs (Flemons, 1999). Fatigued driving and crashes related to fatigue are clearly major issues in road safety.

The GLM models and network screening exercise, and the logistic regression analysis met the study's objectives. The models developed were reasonable, appropriate, and fit the data well. Most of the variables were significant in each model.

The GLM and logistic regression models successfully identified road, traffic, environment, and demographic factors that contribute to FR crashes on rural two-lane roads. The GLM models were also successfully used to develop SPFs and to conduct a network screening exercise that identified sites with a high potential for FR safety improvement.

AADT was the major variable identified in the GLM models. The models' detailed results were encouraging. The results confirmed other research, were largely consistent, and were intuitively acceptable. For example, wider pavement surface widths were associated with a decrease in FR crashes suggesting that the extra space gives fatigued drivers a possibly crucial opportunity to correct their driving before a crash occurs. As might also be expected, most variables had more effect on fatigue crashes than on non-fatigue crashes.

The logistic regression analysis identified alignment and then lighting conditions as important in FR and NSP-FR crashes, and lighting conditions and then alignment as important in SP-FR crashes. The logistic regressions models' results for speed limit and AADT were less intuitive and more ambivalent than the other variables' results.

It is interesting to note that and most GLM variables had more effect on crashes that occurred during sleepy periods of the day than on crashes that occurred during all day period. The logistic regression findings were similar: all the logistic regression variables appeared to play a larger role during sleepy periods of the day (SP-FR crashes). These results possibly confirm the influence of circadian rhythms in road safety.

7.8 Recommendations for Further Work

The recommendations for further work flow from the results of the study and also from the limitations of the study. The limitations were discussed in Chapter 6. The limitations were:

1. A limited number of factors were included in the study. Important variables may have been omitted.
2. The definition of FR crashes was restricted to single vehicle run-off-road crashes. The study may have missed many FR crashes of other types.
3. The models developed are restricted to rural two-lane roads.
4. The models developed are based on data only from Ohio State.
5. The models' data are derived from police reports which may be inaccurate and which may under-report the total number of crashes. Inaccurately recorded crash details may influence on the models' accuracy.
6. A few of the models' results were intuitively unacceptable.
7. The exponential functional form of the variables included in the GLM models may not be the best form possible.
8. The GLM approach does not take internal correlations between variables into account.
9. The power of the GLM models cannot be assessed until the models are tested with a different datasets.
10. The logistic regression models are conditional crash models which examine factors associated with crashes that have already occurred.
11. The logistic regression models do not take exposure into account, and cannot compare the exposure associated with the different crash types analyzed.
12. The use of categorical variables in the logistic regression models may lead to a loss of useful information.

While it is recommended that each of the study's limitations should be addressed as appropriate, the following recommendations are emphasized as being particularly worthwhile and urgent. Research should be directed towards:

1. improving the identification of FR crashes. Obstacles include the driver disregarding fatigue when the crash itself alerts him to his situation and the complete loss of information about the driver's state when the crash has fatal consequences.
2. discovering useful regularities in observed crash data and using these develop the form of model equations. New equations may not only take the exponential functional forms that were used in this research.
3. considering surrogate methods to incorporate exposure into logistic regression analysis.
4. including more years of crash data as more years of data may strengthen the results of FR crash studies.

REFERENCES

Abdel-Aty M.A. and Essam R. E. A., "Modeling traffic accident occurrence and involvement", Accident Analysis and Prevention 32 (2000), pp. 633—642.

Agent, K.R., Pigman, J.G. and Stamatiadis, N., 2001. "Countermeasures for fatal crashes on two-lane rural roads". Lexington, KY: University of Kentucky Transportation Center.

Akerstedt, T., (2000). "Consensus Statement: Fatigue and Accidents in Transport Operations". Journal of Sleep Research, Volume 9, pp. 395.

Armour, M., Carter, R., Cinquegrana, C. and Griffith, J. (1988). "Study of single vehicle rural accidents". Interim report. Report No. RN/88/1. Melbourne: Road Traffic Authority.

Balakrishnan, N. (1991). "Handbook of the Logistic Distribution", Marcel Dekker, Inc.

Ben-Akiva, M. and Lerman, S.R., "Discrete Choice Analysis: Theory and Applications to Travel Demand", The MIT Press, Cambridge, Massachusetts (1993).

Caliendoa, C., Guidab, M., and Parisia, A., "A crash-prediction model for multilane roads", Accident Analysis and Prevention 39 (2007), pp. 657-670.

Cameroon, A.C. and Trivedi, P.K., 1998. "Regression Analysis of Count Data". Cambridge University Press,

Chin, H.C. and Quddus, M.A., "Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections", Accident Analysis and Prevention 35 (2003), pp. 253–259.

Chipman, M. L. and Jin, Y., "Drowsy Drivers and Routinely Collected Crash Data", The Canadian Multidisciplinary Road Safety Conference XVII, June 3-6, 2007, Montreal Quebec,

Choueiri, E.M, Lamm, R., Kloeckner, J.H. and Mailaender, T., "Safety Aspects of Individual Design Elements and Their Interactions on Two-Lane Highways: International Perspective," Transportation Research Record 1445, Transportation Research Board, Washington, D.C.; 1994.

Connor, JL., Whitlock G., Norton, RN. and Jackson, RT., "The role of driver sleepiness in car crashes: a systematic review of epidemiological studies". Accident Analysis and Prevention 33(2000), pp. 31-41.

Donelson, A.C., Ramachandran, K., Zhao, K. and Kalinowski, A., "Rates of Occupant Deaths in Vehicle Rollover: The Importance of Fatality Risk Factors", Transportation Research Record 1665, Transportation Research Board, National Research Council, Washington, D.C., 1999, pp. 109–117.

Dissanayake, S., (2003). "Young Drivers and Run-Off-the-Road Crashes". Proceedings of Mid-Continent Transportation Research Symposium.

Duncan, C.S., Khattak, A.J. and Council, F.M., "Applying the Ordered Probit Model to Injury Severity in Truck-passenger Car Rear-end Collisions", Transportation Research Record 1635, Transportation Research Board, National Research Council (1998), pp. 63–71.

Expert Panel on Driver Fatigue and Sleepiness (EPDFS), "Drowsy Driving and Automobile Crashes", Washington DC: National Centre for Sleep Disorders Research/National Highway Traffic Safety Authority, 1997.

Federal Highway Administration (FHWA), "Drowsy Driving & Fatigue", Retrieved October 22, 2007 from < http://www.nysgtsc.state.ny.us/drow-ndx.htm >

Fitzpatrick, K., Anderson, I., Green, P., Krammes, R., and Poggioli, B., (2000). "Evaluation of design consistency methods for two-lane rural highways, executive summary". Report

FHWARD-99-173. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C.

Flemons, W., (1999). Medical director of the Alberta Lung Association's Sleep Centre at Calgary's Foothills Hospital. Sleep Disorders. Retrieved October 21, 2007 from <http://www.sk.lung.ca/ca/articles/sleepdisorders.html>

Folkard, S., "Black times: Temporal determinants of transport safety". Accident Analysis and Prevention 29(1997), pp. 417-430.

Fridstrom, L., J. Ifver, S. Ingebrigtsen, R. Kulmala, and L.K. Thomsen., (1995). "Measuring the Contribution of Randomness, Exposure, Weather and Daylight to the Variation in Road Accident Counts". Accident Analysis and Prevention 27(1995), pp. 1-20.

Garber, N. J. and Hoel, L. A., Third Edition, "Traffic & Highway Engineering". Publication Date: December 2001

Green, W. H. (2003). "Econometric Analysis", fifth edition. Prentice Hall

Hadi, M. A., Aruldhas, J., Chow, L. F., and Wattleworth, J. A., "Estimating safety effects of cross-section design for various highway types, using negative binomial regression", Transportation Research Record 1500 (1995), pp. 169-177.

Harnen, S., Radin Umar, R.S., Wong, S.V., Wan Hashim, W.I., 2006. "Motorcycle accident prediction model for junctions on urban roads in Malaysia", Advances in Transportation Studies an international Journal Section A 8 (2006), pp 31-40.

Hartley, L. R,, Penna, F., Corry, A. and Feyer, A. M. (2000). "Comprehensive Review of Fatigue Research." Report Number 116: Institute for Research in Safety and Transport.

Harwood, D. W., Council, F. M., Hauer, E., Hughes, W. E. and Vogt, A., (2000). "Prediction

of the expected safety performance of rural two-lane highways". FHWA-RD-99-207, Washington, DC.

Haworth, N., (1998a). "Speed, Alcohol, Fatigue, Effects, Brisbane". Paper presented to 7th Biennial Australasian Traffic Education Conference.

Haworth, N., (1998b). "Fatigue and Fatigue Research: The Australian experience: Speed, Alcohol, Fatigue Effects". Paper presented to 7th Biennial Australian Traffic Education Conference.

Hauer, E. "Statistical Road Safety Modeling". Transportation Research Record, Transportation Research Board, National Research Council, vol. 1897 (2004)

Hauer, E. "Re-Estimation of Models for Two-Lane Rural Road Segments". Draft working paper. FHWA, U.S. Department of Transportation, 2001.

Hauer, E. and Bamfo, J., (1997). Two tools for finding what function links the dependent variable to the explanatory variables. In: Proceedings of the ICTCT 1997 Conference, Lund.

Hildebrand, E., Lougheed, P., and Hanson, T., University of New Brunswick Transportation Group, Retrieved June 21, 2007 from <http://www.unb.ca/transpo/documents/ >

Horne, J. A. and Reyner L. A., "Sleep Related Vehicle Accidents", British Medical Journal Vol. 310 (1995).

Hosmer, D. W. and Lemeshow, S., (2000), "Applied Logistic Regression", Second edition, New York; Chichester, Wiley.

House of Representatives Standing Committee on Communications, Transport and the Arts. 2000. "Beyond the Midnight Oil: an inquiry into managing fatigue in transport." Canberra: The Parliament of the Commonwealth of Australia.

http://roadsafetydirectory.com

http://userwww.sfsu.edu/~efc/classes/biol710/logistic/logisticreg.htm

Insurance Research Council. 2001. "Characteristics of auto accidents – an analysis of auto injury claims". Malvern, PA: Insurance Research Council.

Ivan, J. and O'Mara, P., 1997. "Prediction of Traffic Accident Rates Using Poisson Regression". Presented at the 76th Annual Meeting of the Transportation Research Board.

Joksch, H.C., "Velocity change and fatality risk in a crash – a rule of thumb", Accident Analysis and Prevention 25 (1993), pp. 103-104.

Joshua, S. and Garber, N., "Estimating truck accident rate and involvement using linear and Poisson regression models". Transportation Planning and Technology 15 (1990), pp. 41–58.

Khattak A.J. and Council F.M., Effects of Work Zone Presence on Injury and Non-Injury Crashes, Accident Analysis and Prevention 34 (2002), pp. 19-29

Kim, D., Washington, S. and Oh, J., "Modeling crash outcomes: new insights into the effects of covariates on crashes at rural intersections", J. Transport. Eng. 132 (2006) (4), pp. 282–292.

Kim, K. S. and Yamashita, E., "An Analysis of Alcohol Impaired Motorcycle Crashes in Hawaii, 1986 to 1995", Transportation Research Record 1734", Transportation Research Board, National Research Council, Washington, D.C., 2000, pp. 77–85.

Knipling, R.R. and Wang, J.S. (1994). Research Note: Crashes and Fatalities Related to Driver Drowsiness/Fatigue. National Highway Traffic Safety Administration.

Knuiman, M.W., Council, F.M. and Reinfurt, D.W. "Association of median width and highway accident rates (with discussion)", Transportation Research Record, vol. 1401, (1993), pp 70-82.

Krull, K. A., Khattak, A.J. and Council, F.M., "Injury Effects of Rollovers and Events Sequence in Single-Vehicle Crashes," Transportation Research Record 1717, Transportation Research Board, National Research Council, Washington, D.C., 2000, pp. 46–54.

Kulmala, R., (1995), Safety at Rural Three-and Four-arm Junctions: Development and Application of Accident Prediction Models. VTT publications. Espoo: Technical Research Center at Finland.

Lal, S.K., and Craig, A., "A critical review of the psychophysiology of driver fatigue", Biological Psychology 55 (2001) (3), pp. 173–194.

Leggett, L. M. W. (1988), "Truck accidents, fatigue and driving hours in Tasmania (Research Report 63)". Hobart: Transport Tasmania.

Lin, T. D., Jovanis, P.P., and Yang, C. Z., "Modeling the Safety of Truck Driver Service Hours Using Time-Dependent Logistic Regression", Transportation Research Record 1407, Transportation Research Board, National Research Council, Washington, D.C., 1993, pp. 1–10.

Martin, J. L., "Relationship crash rate and hourly traffic flow on interurban motorways", Accident Analysis and Prevention. 34 (2002), pp. 619–629.

Maschner, H. D. G., "Geographic Information Systems in Archaeology", In New Methods, Old Problems: GIS in Modern Archaeological Research, 1996, pp. 1-23.

Maycock, G., "Sleepiness and driving: The experience of UK car drivers", Accident Analysis

and Prevention 29 (1997), pp. 453-462.

Mayo foundation for Medical Education and Research, (1995). "These 24-hour cycles keep you on schedule", Retrieved October 21, 2007 from <http://www.hallym.ac.kr/~neuro/kns/tutor/medical/rhy.html>

McCartt, A.T., Ribner, S.A., Pack, A.I. and Hammer, M.C., "The scope and nature of the drowsy driving problem in New York State", Accident Analysis and Prevention 28 (1996), pp. 511-517.

McCullagh, P. and Nelder, J. A. (1989), "Generalized Linear Models". Second edition, Chapman and Hall, London.

McGinnis, R.G., Wissinger, L.M., Kelly R.T., and Acuna, C.O., "Estimating the Influences of Driver, Highway, and Environmental Factors on Run-off-Road Crashes Using Logistic Regression", TRB Preprint, Annual Meeting of Transportation Research Board, National Research Council, Washington, DC (1999).

Menard, S., (1995). "Applied Logistic Regression Analysis". Sage Publications.Series: Quantitative Applications in the Social Sciences, No. 106.

Miaou, S. P., "The relationship between truck accidents and geometric design of road section: Poisson versus negative binomial regression", Accident Analysis and Prevention 26 (1994), pp. 471-482.

Miaou, S. P. and Lump, H., "Modeling Vehicle Accidents and Highway Geometric Design Relationships", Accident Analysis and Prevention 25 (1993), pp. 689-709.

Miaou, S.P., Hu, P.S., Wright, T., Rathi, A.K. and Davis, S.C., "Relationship between truck accidents and geometric design: a Poisson regression approach", Transportation Research Record, Transportation Research Board, National Research Council, vol. 1376 (1992), pp. 10-18.

Miaou, S. P., Hu, P. S., Wright, T., Davis, S. C. and Rathi, A. K., "Development of relationships between truck accidents and highway geometric design: Phase I. Technical memorandum prepared by the Oak Ridge National Laboratory", Washington, DC: Federal Highway Administration, 1991.

Milton, J. and Mannering, F., 1996. "The Relationship Between Highway Geometrics, Traffic Related Elements and Motor Vehicle", Washington State Dept. of Transportation.

Mountain, L., Maher, M. and Fawaz, B., "The Influence of Trend on Estimates of Accidents at Junctions", Accident Analysis and Prevention 30 (1998), pp. 641–649.

Nassar, S.A., Saccomanno, F.F. and Shortreed, J.H., "Road accident severity analysis: a macro level approach", Canadian Journal of Civil Engineering, National Research Council of Canada 21 (1994), pp. 847–855.

National Highway Traffic Safety Administration. 2003a. "Traffic safety facts, 2001: rural/urban comparison". Report no. HS-809-524. Washington, DC: U.S. Department of Transportation.

National Highway Traffic Safety Administration. (2003b). "Fatality Analysis Reporting System", Washington, DC: U.S. Department of Transportation.

National Highway Traffic Safety Administration. (2002), "Traffic safety facts, 2001." Report no. HS-809-484. Washington, DC: U.S. Department of Transportation.

National Highway Traffic Safety Administration. (1995). "National occupant protection use survey: controlled intersection study". Research Note. Washington, DC: U.S. Department of Transportation.

Organisation for Economic Co-operation and Development (OECD), 1997. Road Transport

Research, Road Safety Principles and Models: Review of Descriptive, Predictive, Risk and Accident Consequence Models. IRRD NO. 892483. OECD Scientific Expert Group, Paris.

Oh, J., Lyon, C., Washington, S., Persaud, B., and Bared, J., "Validation of the FHWA crash models for rural intersections: lessons learned", Transportation Research Record 1840 (2003), pp. 41-49.

O'Donnell, C.J. and Connor, D.H., "Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice", Accident Analysis and Prevention 28 (1996), pp. 739-753.

Pack A.I., Pack A.M., Rodgman E., Cucchiara A., Dinges D. F. and Schwab C. W., "Characteristics of crashes attributed to the driver having fallen asleep", Accident Analysis and Prevention 27 (1995), pp. 769-775.

Persaud, B. and Dzbik L., "Accident prediction models for freeways", Transportation Research Record 1401 (1993), pp. 55-60.

Persaud, B., Retting, R. A. and Lyon, C. A., "Crash reduction following installation of centerline rumble strips on rural two-lane roads", Accident Analysis and Prevention 36 (2004), pp. 1073-1079.

Persaud, B.N., Lord D. and Palmisano J., "Issues of calibration and transferability in developing accident prediction models for urban intersections", Transportation Research Record 1784 (2001), pp. 57-64.

Persaud B., Retting R.A., Lyon C., "Guidelines for the identification of Hazardous Highway Curves", Transportation Research Record 1717 (2000), pp. 14-18.

Persaud, B., Lyon, C. and Nguyen, T., "Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement", Transportation Research Record 1665 (1999), pp.7-12

Poch, M., and Mannering, F., "Negative binomial Analysis of Intersection-Accident Frequencies", Journal of Transportation Engineering 122 (1996), pp. 105-113.

Polus, A., Poe, C.M. and Mason, J.M. (1998), "Review of international design speed practices in roadway geometric design." Transportation Research Circular, Boston Meeting.

Qin, X., Ivan, J. N., Ravishanker, N., "Selecting Exposure Measures in Crash Rate Prediction for Two-lane Highway Segments", Accident Analysis and Prevention 38 (2003), pp. 1–9.

Radin, U. R. S., Mackey, M. G. and Hills, B. L., "Modeling of conspicuity-related motorcycle accidents in Sermbean and Shah Alam, Malaysia", Accident Analysis and Prevention 28 (1996), pp. 325–332.

Rosekind, M. (1999). "Fatigue in transportation: Physiological, performance and safety Issues". Evaluation of US Department of Transportation Efforts in the 1990s to Address Operator Fatigue. Safety Report NTSB/SR–99/01,Washington DC.

RTA, "Fatigue: Fatigue Problem Definition and Countermeasures Summary", 2001. Road and Traffic Authority, NSW, Retrieved October 21, 2007 from <www.rta.nsw.gov.au>.

SAS Institute, 1998a. SAS Institute Inc. SAS Users Manual, SAS Institute Inc, Cary, NC (1998).

SAS Institute, 1998b. SAS Institute Inc. "Logistic Regression Examples Using the SAS System", Cary, NC (1998).

Sawalha, Z., "Statistical Issues in Traffic Accident Modeling". In Proceedings of the 82th Annual Meeting of the Transportation Research Board, January 12-16, Washington, D.C, 2003.

Shankar, V., Mannering, F. and Barfield, W., "Statistical analysis of accident severity on rural freeways". Accident Analysis and Prevention 28 (1996), pp. 391–401.

Shankar, V., Mannering, F. and Barfield, W., "Effect of roadway geometric and environment factors on rural freeway accident frequencies", Accident Analysis and Prevention 27(1995), pp. 371-389

Vogt, A., (1999). "Crash models for rural intersections: four-lane by two-lane stop-controlled and two-lane by two-lane signalized". FHWA-RD-99-128, Washington, DC.

Vogt, A., and Bared, J., (1998). "Accident prediction models for two-lane rural roads: segments and intersections". FHWA-RD-98-133, Washington, DC.

Washington. S. P., Karlaftis, M. G.. and Mannering, F. L., "Statistical and Econometric Methods for Transportation Data Analysis", Chapman & Hall/CRC, 2002

Wong, C. H., (2003), "Contractor Performance Prediction Model for the United Kingdom Construction Contractor: Study of Logistic Regression Approach", Journal of Construction Engineering and Management 130 (5), pp. 691-698

Wylie C.D,, Shultz T., Miller J.C., Mitler M.M., and Mackie R.R. (1996). "Commercial Motor Vehicle Driver Fatigue and Alertness Study". Technical summary. Montreal: Transport Development Centre.

www.safetyanalyst.org

Zegeer, C., Stewart, R., Reinfurt, D., Council, F., Neuman, T., Hamilton, E., Miller, T., and Hunter, W., "Cost-Effective Geometric Improvements for Safety Upgrading of Horizontal Curves", Report No. FHWA-RD-90-021," Federal Highway Administration, 1991.

APPENDIX A. Applying SAS System for Data Analysis

SAS is a software system for data analysis. Basic SAS software provides tools for information storage and retrieval, data modification and programming, report writing, statistical analysis, and file handling.

There are three primary windows in SAS. The 'Program Editor' window is used to enter, edit, and submit SAS programs. The 'Log' window displays messages about the programs to submit. The 'Output' window displays output from procedures.

Only one window at a time is active. The active window is where the cursor is located. Commands are given through the SAS menu bar will be executed in the active window. Windows can be scrolled through using the arrow buttons, and can enlarge a window by clicking on the up arrow in the right hand corner of the window. Windows can be changed by using the mouse to select a window, or choosing 'Windows' from the menu bar at the top of the screen.

A SAS program consists of a series of statements. Each statement must end with a semicolon. The two main steps in preparing a SAS program are the DATA step which creates a SAS data set, and the PROCEDURE step which performs tasks on the data. To enter the SAS program make sure the cursor is in the 'Program Editor' window, and then type.

1. Data Definitions and Options

The Data Definitions and Options are at the top of most SAS programs. The first step is to use SAS to locate the data. In other words, the location of the data on the computer or storage device must be defined. To define the location of the data in SAS, the type of data working with must be known. For reading in a SAS data set, a LIBNAME statement is needed.

SAS Options statement is used to define an environment for the program. It changes the standard settings. Some common options include:

- LS - defines the line size for output.
- OBS - limits the number of observations processed to allow for program testing on a small subset rather than reading in the entire dataset.
- NOCENTER - writes all the output in the log and listing files flush left.

2. The Data Step

The first part of the SAS program is the data step. The data step names the SAS dataset, identifies the location of the data, describes the data values to be read, and computes new variables. The DATA steps included:

A. DATA statement

It is the first statement in the SAS job, which marks the beginning of the Data Step, and gives the name chosen for the SAS data set being created. The data set name must begin with a character or an underscore, can include characters, underscores, or letters, and is limited to eight characters.

B. The INFILE or CARDS statement

This is used to identify the location of the data. If the data in the SAS program is included, put the data

lines immediately after the CARDS statement, then place a semicolon after the final data line.

C. INPUT statement

The INPUT statement describes the data to the SAS system and assigns variable names. Each variable must be defined as either character or numeric. A $ sign after the variable indicates a character variable.

D. Creating new variables

To create a new variable, the new variable name and the expression for calculating it must be specified

3. The Procedure Step

The second part of a SAS program is the procedure step. SAS procedures are computer programs that use the dataset to perform various computations, and print results.

The general form of a SAS procedure is:

PROC NAME DATA=NAME;

... other lines specific to the procedure **NAME** ...

RUN;

If a SAS data set name is not specified, SAS uses the most recently created data set.

The GENMOD procedure fits generalized linear models, as defined by Nelder and Wedderburn (1972). The class of generalized linear models is an extension of traditional linear models that allows the mean of a population to depend on a linear predictor through a nonlinear link function and allows the response probability distribution to be any member of an exponential family of distributions. Many widely used statistical models are generalized linear models. These include classical linear models with normal errors, logistic and probit models for binary data, and log-linear models for multinomial data. Many other useful statistical models can be formulated as generalized linear models by selecting an appropriate link function and response probability distribution.

RUN instructs SAS to run the previously listed data step or procedure. Ending the programs with a run statement ensures that SAS will run all procedures which have been requested.

4. Submitting or Running a SAS Program

To submit a SAS program make sure that the cursor is in the 'Program Editor' window and press <F8> or click on the person running icon to submit a SAS program. Another method is through the menu bar at the top of the screen, by selecting 'Locals', 'Submit'. The program will disappear and the 'Log' window will start showing the SAS statements and notes with information about the programs progress. The output window will come up automatically when the program is finished, and will have the results of the procedures. Printing or scrolling through any of the windows is allowed. If something is wrong with the SAS program, SAS will give red error messages in the LOG file. If the red error messages are received user should go back to the "Program Editor", recall the program by hitting <F4> or choosing "Locals" and "Recall text" from the menu bar, and then correct the codes and rerun the program. Before re-running a program it is often helpful to clear the LOG and OUTPUT windows of their text from the run that generated the error messages. An easy way to do this is to select the window and use <CNTRL>+<E> to clear it.

5 Saving, Retrieving, and Printing the Results

Click the following menu commands from the menu bar at top of the screen with the mouse.

- To Save:

  SAS will save the active window that the cursor is in.

  Click 'FILE', 'SAVE AS' if the file is new or the file is renamed. Otherwise click 'SAVE'.

  Specify the drive and the name of the file. It can be saved on either a floppy disk or a hard drive.

  Click 'OK'.

- To Retrieve a File:

  If a file that has been previously saved is wanted to work with, make sure the 'Program Editor' window is active.

  Click 'FILE', 'OPEN'.

  Select the drive where the file has been saved, choose the file name, and click on the 'OK' box.

  If successful the file will appear in the 'Program Editor' window.

- To Print:

  Printing is allowed from any window. Make sure the window to be printed is active by selecting it with the mouse.

  Click 'FILE', 'PRINT'.

  From the print window, select the proper printer options with mouse.

  Click 'OK'

6. Exiting the Program

Before exiting, save any files needed for future use.

- Select 'File' from the menu bar.
- Select 'Exit'

APPENDIX B. SAS Variables of the OHIO HSIS Data used to develop the Dataset for GLM analysis

LIST OF SAS VARIABLES FOR ALL OHIO HSIS FILES	
SAS VARIABLE NAME	**DESCRIPTION**
CASENO	UNIQUE ACCIDENT CASE NUMBER
VEHNO	VEHICLE NUMBER
CNTYRTE	COUNTY ROUTE
MILEPOST	MILEPOST
BEGMP	BEGINNING MILE POST
ENDMP	END MILE POST
XMILEPST	CROSS ROUTE MILEPOST
XCNTYRTE	CROSS COUNTY ROUTE
CNTY_RTE	COUNTY ROUTE

APPENDIX C. SAS Codes for Dataset Development for GLM Analysis

```
data oh00road;
        set oh00road;
       rtenbr =input(rte_nbr, 5.);
run;
data oh01road;
        set oh01road;
       rtenbr =input(rte_nbr, 5.);
run;
data oh02road;
        set oh02road;
       rtenbr =input(rte_nbr, 5.);
run;
data oh03road;
        set oh03road;
       rtenbr =input(rte_nbr, 5.);
run;
data oh04road;
        set oh04road;
       rtenbr =input(rte_nbr, 5.);
run;
data oh00acc;
 set oh00acc;
   drop FLIP_IND;
run;
data oh01acc;
 set oh01acc;
   drop FLIP_IND;
run;
data oh02acc;
 set oh02acc;
   drop FLIP_IND;
run;
data oh03acc;
 set oh03acc;
   drop FLIP_IND;
run;
data oh04acc;
 set oh04acc;
   drop FLIP_IND;
```

```
run;
/*This program first creates homogeneous segments based upon changes in given variables.
   Then the program can disaggregate these sections into smaller sections, e.g. 0.1 mile
   segments if desired*/
/*Document order of site set. (This can be used later to eliminate duplicates)*/
proc sort data=Oh04road; by rtenbr begmp;
data Oh04road;
 set Oh04road;
n=_n_;
run;
/*Identify homogeneous sections.   Use retain statement to compare current record to
previous record.   If they are different, set flag to indicate start of new segment*/
proc sort; by n;
data reducedsites;
set Oh04road;
length lrtenbr $ 12 laccess $ 5 lfunc_cls $5 lmed_type $ 5 lmed_wid $ 5 ;
retain siteno 0 lrtenbr " laadt 0 lspdlimt 0 laccess " lfunc_cls "
        lmed_type "   lmed_wid " lno_lanes 0 lsurfwidr 0 lrodwycls 0 lendmp 0;
if rtenbr ne lrtenbr or aadt ne laadt or spdlimt ne lspdlimt or
    access ne laccess or func_cls ne lfunc_cls or med_wid ne lmed_wid
    or med_type ne lmed_type or
    no_lanes ne lno_lanes or
    surfwidr ne lsurfwidr or rodwycls ne lrodwycls or begmp ne lendmp then do;
    siteno=siteno + 1;
/*All segments to be merged will have same siteno*/
frec=1;
/*This is flag for first record*/
end;
output;
/*Output the record before reassigning values*/
lrtenbr = rtenbr;
lendmp = endmp;
laadt = aadt;
lspdlimt = spdlimt;
laccess = access;
lfunc_cls = func_cls;
lmed_wid = med_wid;
lmed_type = med_type;
lno_lanes = no_lanes;
lsurf_wid = surf_wid;
lrodwycls = rodwycls;
```

```
RUN;
*Sort in descending order of n to save data contained in first record of
homogeneous segment, especially beginning milepost number;

proc sort data=reducedsites; by descending n;
*This procedure actually puts segments together and assigns the
ending milepost of the last record to the first record in the
homogeneous segment;
data reducedsites; set reducedsites;
retain lpm 0 lsiteno 0;
/*Keep ending milepost and siteno with retain*/
if _n_=1 or siteno=1 then do;
      lsiteno=siteno;
      lpm=endmp;
end;
if siteno ne lsiteno then do;
      lsiteno=siteno;
      lpm=endmp;
end;
if frec=1 then do;
      endmp=lpm;
      seg_lng=endmp-begmp;
      output;
/*output record only when you have reached the end*/
end;
RUN;
*Save homogeneous segments permanently;
Data OhAggSegments;
set reducedsites(drop= lrtenbr laadt lspdlimt laccess lfunc_cls lmed_wid lmed_type    lno_lanes lsurfwidr
lrodwycls lpm lsiteno n frec);
RUN;
data OhAggSegments;
set OhAggSegments;
    newleng=seg_lng;
    bpnew=begmp;
    epnew=endmp;
RUN;

proc sort data=OhAggSegments; by rtenbr bpnew;
data OhAggSegments;
set OhAggSegments(keep = cnty_rte spdlimt rtenbr access divided func_cls
```

```
                                        med_wid med_type no_lanes pavecond rd_width
                                        surf_wid    rodwycls newleng rte_type OUTSHWD1 INSHWD1
OUTSHWD2 INSHWD2
                                        bpnew epnew siteno);
num=_n_;
RUN;
/*Merge OhDisaggSegments data with AADT data for other years*/
data road00ohtemp (rename=( rtenbr =route00 aadt=aadt00));
set oh00road(keep= rtenbr begmp endmp aadt);
run;
data road01ohtemp (rename=( rtenbr =route01 aadt=aadt01));
set oh01road(keep= rtenbr begmp endmp aadt);
run;
data road02ohtemp (rename=( rtenbr =route02 aadt=aadt02));
set oh02road(keep= rtenbr begmp endmp aadt);
run;
data road03ohtemp (rename=( rtenbr =route03 aadt=aadt03));
set oh03road(keep= rtenbr begmp endmp aadt);
run;
data road04ohtemp (rename=( rtenbr =route04 aadt=aadt04));
set oh04road(keep= rtenbr begmp endmp aadt);
run;
proc sql;
      create table matchedaadt00 as
      select *
      from ohAggSegments seg, Road00ohtemp traffic
      where    seg.rtenbr =traffic.route00 and
      ((traffic.begmp between seg.bpnew and seg.epnew) or
       (traffic.endmp between seg.bpnew and seg.epnew) or
       ((traffic.begmp <= seg.bpnew) and (traffic.ENDMP >= seg.epnew)));
quit;
run;
proc sql;
      create table matchedaadt01 as
      select *
      from ohAggSegments seg, Road01ohtemp traffic
      where    seg.rtenbr =traffic.route01 and
      ((traffic.begmp between seg.bpnew and seg.epnew) or
       (traffic.endmp between seg.bpnew and seg.epnew) or
       ((traffic.begmp <= seg.bpnew) and (traffic.ENDMP >= seg.epnew)));
quit;
```

```
run;
proc sql;
      create table matchedaadt02 as
      select *
      from ohAggSegments seg, Road02ohtemp traffic
      where    seg.rtenbr =traffic.route02 and
      ((traffic.begmp between seg.bpnew and seg.epnew) or
       (traffic.endmp between seg.bpnew and seg.epnew) or
       ((traffic.begmp <= seg.bpnew) and (traffic.ENDMP >= seg.epnew)));
quit;
run;
proc sql;
      create table matchedaadt03 as
      select *
      from ohAggSegments seg, Road03ohtemp traffic
      where    seg.rtenbr =traffic.route03 and
      ((traffic.begmp between seg.bpnew and seg.epnew) or
       (traffic.endmp between seg.bpnew and seg.epnew) or
       ((traffic.begmp <= seg.bpnew) and (traffic.ENDMP >= seg.epnew)));
quit;
run;
proc sql;
      create table matchedaadt04 as
      select *
      from ohAggSegments seg, Road04ohtemp traffic
      where    seg.rtenbr =traffic.route04 and
      ((traffic.begmp between seg.bpnew and seg.epnew) or
       (traffic.endmp between seg.bpnew and seg.epnew) or
       ((traffic.begmp <= seg.bpnew) and (traffic.ENDMP >= seg.epnew)));
quit;
run;
data matchedaadt00;
set matchedaadt00;
if ((begmp >= bpnew) and (endmp <= epnew)) then do;
varleng00=endmp-begmp;
aadtvar00=aadt00*varleng00;
end;
if ((begmp <= bpnew) and (endmp >= bpnew) and (endmp <= epnew)) then do;
varleng00=endmp-bpnew;
aadtvar00=aadt00*varleng00;
end;
```

```
if ((begmp >= bpnew) and (begmp <= epnew) and (endmp >= epnew)) then do;
varleng00=epnew-begmp;
aadtvar00=aadt00*varleng00;
end;
if ((begmp <= bpnew) and (endmp >= epnew)) then do;
varleng00=epnew-bpnew;
aadtvar00=aadt00*varleng00;
end;
run;
data matchedaadt01;
set matchedaadt01;
if ((begmp >= bpnew) and (endmp <= epnew)) then do;
varleng01=endmp-begmp;
aadtvar01=aadt01*varleng01;
end;
if ((begmp <= bpnew) and (endmp >= bpnew) and (endmp <= epnew)) then do;
varleng01=endmp-bpnew;
aadtvar01=aadt01*varleng01;
end;
if ((begmp >= bpnew) and (begmp <= epnew) and (endmp >= epnew)) then do;
varleng01=epnew-begmp;
aadtvar01=aadt01*varleng01;
end;
if ((begmp <= bpnew) and (endmp >= epnew)) then do;
varleng01=epnew-bpnew;
aadtvar01=aadt01*varleng01;
end;
run;
data matchedaadt02;
set   matchedaadt02;
if ((begmp >= bpnew) and (endmp <= epnew)) then do;
varleng02=endmp-begmp;
aadtvar02=aadt02*varleng02;
end;
if ((begmp <= bpnew) and (endmp >= bpnew) and (endmp <= epnew)) then do;
varleng02=endmp-bpnew;
aadtvar02=aadt02*varleng02;
end;
if ((begmp >= bpnew) and (begmp <= epnew) and (endmp >= epnew)) then do;
varleng02=epnew-begmp;
aadtvar02=aadt02*varleng02;
```

```
end;
if ((begmp <= bpnew) and (endmp >= epnew)) then do;
varleng02=epnew-bpnew;
aadtvar02=aadt02*varleng02;
end;
run;
data    matchedaadt03;
set matchedaadt03;
if ((begmp >= bpnew) and (endmp <= epnew)) then do;
varleng03=endmp-begmp;
aadtvar03=aadt03*varleng03;
end;
if ((begmp <= bpnew) and (endmp >= bpnew) and (endmp <= epnew)) then do;
varleng03=endmp-begmp;
aadtvar03=aadt03*varleng03;
end;
if ((begmp >= bpnew) and (begmp <= epnew) and (endmp >= epnew)) then do;
varleng03=endmp-begmp;
aadtvar03=aadt03*varleng03;
end;
if ((begmp <= bpnew) and (endmp >= epnew)) then do;
varleng03=endmp-begmp;
aadtvar03=aadt03*varleng03;
end;
run;
data    matchedaadt04;
set matchedaadt04;
if ((begmp >= bpnew) and (endmp <= epnew)) then do;
varleng04=endmp-begmp;
aadtvar04=aadt04*varleng04;
end;
if ((begmp <= bpnew) and (endmp >= bpnew) and (endmp <= epnew)) then do;
varleng04=endmp-begmp;
aadtvar04=aadt04*varleng04;
end;
if ((begmp >= bpnew) and (begmp <= epnew) and (endmp >= epnew)) then do;
varleng04=endmp-begmp;
aadtvar04=aadt04*varleng04;
end;
if ((begmp <= bpnew) and (endmp >= epnew)) then do;
varleng04=endmp-begmp;
```

```
aadtvar04=aadt04*varleng04;
end;
run;
*Calculates average AADT;
proc sort data= matchedaadt00; by num;
PROC MEANS NOPRINT; by num;   VAR varleng00 aadtvar00;
OUTPUT OUT= aadtdata00 sum=varleng00 aadtvar00;
RUN;
data   aadtdata00;
set   aadtdata00(drop=_type_ _freq_);
if varleng00 > 0 then do;
newaadt00=aadtvar00/varleng00;
end;
if varleng00=0 then do;
newaadt00=0;
end;
run;

proc sort data= matchedaadt01; by num;
PROC MEANS NOPRINT; by num;   VAR varleng01 aadtvar01;
OUTPUT OUT=aadtdata01 sum=varleng01 aadtvar01;
RUN;
data aadtdata01;
set aadtdata01(drop=_type_ _freq_);
if varleng01 > 0 then do;
newaadt01=aadtvar01/varleng01;
end;
if varleng01=0 then do;
newaadt01=0;
end;
run;
proc sort data= matchedaadt02; by num;
PROC MEANS NOPRINT; by num;   VAR varleng02 aadtvar02;
OUTPUT OUT=aadtdata02 sum=varleng02 aadtvar02;
RUN;
data aadtdata02;
set aadtdata02(drop=_type_ _freq_);
if varleng02 > 0 then do;
newaadt02=aadtvar02/varleng02;
end;
if varleng02=0 then do;
```

```
newaadt02=0;
end;
run;
proc sort data= matchedaadt03; by num;
PROC MEANS NOPRINT; by num;    VAR varleng03 aadtvar03;
OUTPUT OUT=aadtdata03 sum=varleng03 aadtvar03;
RUN;
data aadtdata03;
set aadtdata03(drop=_type_ _freq_);
if varleng03 > 0 then do;
newaadt03=aadtvar03/varleng03;
end;
if varleng03=0 then do;
newaadt03=0;
end;
run;
proc sort data= matchedaadt04; by num;
PROC MEANS NOPRINT; by num;    VAR varleng04 aadtvar04;
OUTPUT OUT=aadtdata04 sum=varleng04 aadtvar04;
RUN;
data aadtdata04;
set aadtdata04(drop=_type_ _freq_);
if varleng04 > 0 then do;
newaadt04=aadtvar04/varleng04;
end;
if varleng04=0 then do;
newaadt04=0;
end;
run;
*Merge AADT data with disaggregate road segmemt data;
data ohAggSegments; merge ohAggSegments(in=a) aadtdata00(in=b);
by num;
if a;
run;
data ohAggSegments; merge ohAggSegments(in=a) aadtdata01(in=b);
by num;
if a;
run;
data ohAggSegments; merge ohAggSegments(in=a) aadtdata02(in=b);
by num;
if a;
```

```
run;
data ohAggSegments; merge ohAggSegments(in=a) aadtdata03(in=b);
by num;
if a;
run;
data ohAggSegments; merge ohAggSegments(in=a) aadtdata04(in=b);
by num;
if a;
run;
data ohAggSegments;
set ohAggSegments(drop= varleng00 aadtvar00 varleng01 aadtvar01 varleng02 aadtvar02 varleng03 aadtvar03
varleng04 aadtvar04 siteno);
if newaadt00=0 then do;
newaadt00='';
end;
if newaadt01=0 then do;
newaadt01='';
end;
if newaadt02=0 then do;
newaadt02='';
end;
if newaadt03=0 then do;
newaadt03='';
end;
if newaadt04=0 then do;
newaadt04='';
end;
run;
data ohcurv(rename=(rte_nbr=route)); set oh04curv(keep= rte_nbr begmp endmp deg_curv);
run;
proc sql;
      create table matchedcurve as
      select *
      from ohAggSegments seg, ohcurv curve
      where   seg.rtenbr = curve.route and
      ((curve.begmp between seg.bpnew and seg.epnew) or
       (curve.endmp between seg.bpnew and seg.epnew) or
       ((curve.begmp LE seg.bpnew) and (curve.endmp GE seg.epnew)));
quit;
run;
data   matchedcurve;
```

```
set    matchedcurve;
if ((begmp >= bpnew) and (endmp <= epnew)) then do;
curvleng=endmp-begmp;
curvvar=deg_curv*curvleng;
end;
if ((begmp <= bpnew) and (endmp >= bpnew) and (endmp <= epnew)) then do;
curvleng=endmp-bpnew;
curvvar=deg_curv*curvleng;
end;
if ((begmp >= bpnew) and (begmp <= epnew) and (endmp >= epnew)) then do;
curvleng=epnew-begmp;
curvvar=deg_curv*curvleng;
end;
if ((begmp <= bpnew) and (endmp >= epnew)) then do;
curvleng=epnew-bpnew;
curvvar=deg_curv*curvleng;
end;
run;
data    matchedcurve;set    matchedcurve;
  if curvleng=0 then do;
curv_id = 0;
  if curvleng > 0 then do;
curv_id = curvvar/curvleng/curvleng/52.8 ;
run;
data    matchedcurve;set    matchedcurve;
  curv_ind = 0 ; if    curv_id    ge 4 then curv_ind = 1;
run;
*Calculates horizontal curvature variable;
proc sort data=matchedcurve; by num;
PROC MEANS NOPRINT; by num;    VAR curvleng curvvar curv_ind;
OUTPUT OUT=curvdata sum=curvleng curvvar curv_ind;
RUN;
data curvdata;
set curvdata(drop=_type_ _freq_);
if curvleng > 0 then do;
curv_hi=curvvar/curvleng/curvleng/52.8;
end;
if curvleng=0 then do;
curv_hi=0;
end;
run;
```

```
data curvdata;set curvdata;
curv_idt = 0 ;   if curv_ind   >= 1 then curv_idt = 1 ;
run;
*Merge curve data with disaggregate road segmemt data;
data ohAggSegments; merge ohAggSegments(in=a) curvdata(in=b);
by num;
if a;
run;
data ohAggSegments;
set ohAggSegments(drop= curvvar);
if curv_hi = 0 then do;
curv_hi ='';
end;
run;
data ohgrad(rename=(rte_nbr=route1 begmp=begmp1 endmp=endmp1));
set oh04grad(keep= rte_nbr begmp endmp pct_grad dir_grad);
run;
proc sql;
      create table matchedgrad as
      select *
      from ohAggSegments seg, ohgrad grad
      where   seg.rtenbr = grad.route1 and
      ((grad.begmp1 between seg.bpnew and seg.epnew) or
       (grad.endmp1 between seg.bpnew and seg.epnew) or
       ((grad.begmp1 LE seg.bpnew) and (grad.endmp1 GE seg.epnew)));
quit;
run;
data   matchedgrad;
set   matchedgrad;
if ((begmp1 >= bpnew) and (endmp1 <= epnew)) then do;
gradleng=endmp1-begmp1;
gradvar=pct_grad*gradleng;
end;
if ((begmp <= bpnew) and (endmp >= bpnew) and (endmp <= epnew)) then do;
gradleng=endmp1-bpnew;
gradvar=pct_grad*gradleng;
end;
if ((begmp >= bpnew) and (begmp <= epnew) and (endmp >= epnew)) then do;
gradleng=epnew-begmp1;
gradvar=pct_grad*gradleng;
end;
```

```
if ((begmp <= bpnew) and (endmp >= epnew)) then do;
gradleng=epnew-bpnew;
gradvar=pct_grad*gradleng;
end;
run;
*Calculates vertical gradient variable;
proc sort data=matchedgrad; by num;
PROC MEANS NOPRINT; by num;   VAR gradleng gradvar   ;
OUTPUT OUT=graddata sum=gradleng gradvar ;
RUN;
data graddata;
set graddata(drop=_type_ _freq_);
if gradleng > 0 then do;
grad_hi=gradvar/gradleng;
end;
if gradleng=0 then do;
grad_hi=0;
end;
run;
*Merge grade data with disaggregate road segmemt data;
data ohAggSegments; merge ohAggSegments(in=a) graddata(in=b);
by num;
if a;
run;
data ohAggSegments;
set ohAggSegments(drop= gradvar);
if grad_hi = 0 then do;
grad_hi ='';
end;
run;
data gradtemp; set matchedgrad; keep num dir_grad; run;
data ohAggSegments; merge ohAggSegments(in=a) gradtemp(in=b);
by num;
if a;
run;
*Select fatigue accidents only;
data oh00acc;
  set oh00acc;
  if NUMVEHS='1' and onoff_rd='2' or onoff_rd='3' then fatiacc=1; else fatiacc=0;
  run;
data oh00acc;
```

```
 set oh00acc;
 drop street_1 ; if fatiacc=1;
 run;
data oh01acc;
 set oh01acc;
 if NUMVEHS='1' and onoff_rd='2' or onoff_rd='3' then fatiacc=1; else fatiacc=0;
 run;
data oh01acc;
 set oh01acc;
 drop street_1 ; if fatiacc=1;
 run;
data oh02acc;
 set oh02acc;
 if NUMVEHS='1' and onoff_rd='2' or onoff_rd='3' then fatiacc=1; else fatiacc=0;
 run;
data oh02acc;
 set oh02acc;
 drop street_1 ; if fatiacc=1;
 run;
data oh03acc;
 set oh03acc;
 if NUMVEHS='1' and onoff_rd='2' or onoff_rd='3' then fatiacc=1; else fatiacc=0;
 run;
data oh03acc;
 set oh03acc;
 drop street_1 ; if fatiacc=1;
 run;
data oh04acc;
 set oh04acc;
 if NUMVEHS='1' and onoff_rd='2' or onoff_rd='3' then fatiacc=1; else fatiacc=0;
 run;
data oh04acc;
 set oh04acc;
 drop street_1 ; if fatiacc=1;
 run;
*Get a few fields of the road recordset to merge with accidents;
data segments; set    ohAggSegments(keep = cnty_rte rtenbr bpnew epnew num);
run;
data oh00accnew;
set oh00acc;
accyr=00;
```

```
run;
data oh01accnew;
set oh01acc;
accyr=01;
run;
data oh02accnew;
set oh02acc;
accyr=02;
run;
data oh03accnew;
set oh03acc;
accyr=03;
run;
data oh04accnew;
set oh04acc;
accyr=04;
run;
*Put all accident records into one file by appending;
data accidents; set oh00accnew oh01accnew oh02accnew oh03accnew oh04accnew;
run;
*Order the accident records;
proc sort data=accidents; by rte_nbr milepost;
*Number the accident records;
data accidents; set accidents;
accno=_n_;
run;
/*Merge road segments data with accident data it is helpful
if the variables used to connect the datasets have different names*/
proc sql;
      create table matchedaccs as
      select *
      from accidents acc, segments road
      where   acc.cntyrte=road.cnty_rte and
      (acc.milepost between road.bpnew and road.epnew);
quit;
run;
*Delete duplicatess;
proc sort data=matchedaccs;
by caseno milepost;
*First delete all accidents occurring at beginning mile post;
data matchedaccs; set matchedaccs;
```

```
if milepost=bpnew then delete;
*Delete all accidents appearing in multiple segments;
data matchedaccs; set matchedaccs;
*accidents are uniquely defined by accno;
by caseno;
if first.caseno;
run;
*Final check to make sure there are no duplicates--see log for zero
selected;
proc sort data=matchedaccs nodupkey;
by caseno;
run;
*Save dataset permanently;
data matchedaccs;
set matchedaccs;
run;
*Total number of accidents per year to merge with homogeneous segment file;
*Retrieve temporary road file;
data roads; set ohAggSegments;
*Create variables to sum;
data assgnaccidents; set matchedaccs;
tot=1;
if accyr=00 then acc00=1;
if accyr=01 then acc01=1;
if accyr=02 then acc02=1;
if accyr=03 then acc03=1;
if accyr=04 then acc04=1;
run;
*Sum accidents across years;
proc sort data= assgnaccidents; by num;
PROC MEANS NOPRINT; by num;   VAR tot acc00 acc01 acc02 acc03 acc04;
OUTPUT OUT=accfreq sum= tot acc00 acc01 acc02 acc03 acc04;
RUN;
*Merge road segment data with accident data;
data   finalaggbytype; merge   roads(in=a) accfreq(in=b);
by num;
if a;
run;
data   ohfinalagg; set finalaggbytype ;
dirgra="flat" ;
if dir_grad= "-" then dirgra="down";
```

```
if dir_grad= "+" then dirgra="up";
run;
data    ohfinalagg; set    ohfinalagg ;
drop surfwidr med_type _type_ _freq_ dir_grad;
run;
data    ohfinalagg; set    ohfinalagg ;
avg_outsh=( OUTSHWD1+ INSHWD2)/2;
 avg_insh=( OUTSHWD2+ INSHWD1)/2;
run;
proc sort data=ohfinalagg out=ohfinalagg nodupkey;
    by num;
    run;
data    ohfinalagg; set    ohfinalagg;
*Fill in missing values as zeros;
if tot=. then tot=0;
if acc00=. then acc00=0;
if acc01=. then acc01=0;
if acc02=. then acc02=0;
if acc03=. then acc03=0;
if acc04=. then acc04=0;
run;
data    ohfinalagg; set    ohfinalagg;
*Fill in missing values as zeros;
if med_wid=. then med_wid=0;
if curvleng=. then curvleng=0;
if curv_hi=. then curv_hi=0;
if curv_idt=. then curv_idt=0 ;
if gradleng=. then gradleng=0;
if grad_hi=. then grad_hi=0;
run;
data    ohfinalagg; set    ohfinalagg;
avg_aadt = ( newaadt00+newaadt01+newaadt02+newaadt03+newaadt04 )/ 5;
log_aadt = log(avg_aadt);
off = log((epnew-bpnew)*5);
run;
data    ohfinalagg; set    ohfinalagg ;
avg_outsh=( OUTSHWD1+ INSHWD2)/2;
 avg_insh=( OUTSHWD2+ INSHWD1)/2;
run;
data    ohfinalagg; set    ohfinalagg ;
   rcl=curvleng/newleng;
```

```
   rgl=gradleng/newleng;
run;

data   ohfinalagg; set   ohfinalagg;
*Fill in missing values as zeros;
if avg_outsh =. then avg_outsh =0;
if avg_insh =. then avg_insh =0;
if rcl =. then rcl =0;
if rgl =. then rgl =0;
run;
data fati_r2ln;
   set ohfinalagg;
   drop rte_type;
   if rodwycls="08" ;
run;
```

APPENDIX D. SAS Codes for GLM Analysis

```
proc genmod data=Fati_r2ln;
      TITLE 'Fatigue Accidents';
      model tot=log_aadt SPDLIMT curv_hi grad_hi surf_wid avg_outsh
 / dist=NEGBIN offset=off type1 type3 ;
data Fati_r2ln; set Fati_r2ln; run;
proc genmod data=Nf_r2ln;
TITLE 'Non-Fatigue Accidents';
      model tot=log_aadt SPDLIMT curv_hi grad_hi surf_wid avg_outsh
 / dist=NEGBIN offset=off type1 type3 ;
data Nf_r2ln; set Nf_r2ln; run;
proc genmod data=SPFR_r2ln;
      TITLE 'SP Fatigue Accidents';
      model tot=log_aadt SPDLIMT curv_hi grad_hi surf_wid avg_outsh
 / dist=NEGBIN offset=off type1 type3 ;
data SPFR_r2ln; set SPFR_r2ln; run;
proc genmod data= SPNFR_r2ln;
TITLE 'SP Non-Fatigue Accidents';
      model tot=log_aadt SPDLIMT curv_hi grad_hi surf_wid avg_outsh
 / dist=NEGBIN offset=off type1 type3 ;
data SPNFR_r2ln; set SPNFR_r2ln; run;
```

APPENDIX E. SAS Codes for Dataset Development for Logistic Analysis

```
data oh00road;
        set oh00road;
       rtenbr =input(rte_nbr, 5.);
run;
data oh01road;
        set oh01road;
       rtenbr =input(rte_nbr, 5.);
run;
data oh02road;
        set oh02road;
       rtenbr =input(rte_nbr, 5.);
run;
data oh03road;
        set oh03road;
       rtenbr =input(rte_nbr, 5.);
run;
data oh04road;
        set oh04road;
       rtenbr =input(rte_nbr, 5.);
data oh00acc;
 set oh00acc;
   drop FLIP_IND rte_nbr;
run;
data oh01acc;
 set oh01acc;
   drop FLIP_IND rte_nbr;
run;
data oh02acc;
 set oh02acc;
   drop FLIP_IND rte_nbr;
run;
data oh03acc;
 set oh03acc;
   drop FLIP_IND rte_nbr;
run;
data oh04acc;
 set oh04acc;
   drop FLIP_IND rte_nbr;
run;
```

```
data oh00road;
 set oh00road;
   drop NHS_INTR rte_nbr;
run;
data oh01road;
 set oh01road;
   drop NHS_INTR rte_nbr;
run;
data oh02road;
 set oh02road;
   drop NHS_INTR rte_nbr;
run;
data oh03road;
 set oh03road;
   drop NHS_INTR rte_nbr;
run;
data oh04road;
 set oh04road;
   drop NHS_INTR rte_nbr;
run;
data oh00curv(rename=(begmp=begmp1 endmp=endmp1 seg_lng=curv_lng));
   set oh00curv;
   run;
data oh00grad(rename=(begmp=begmp2 endmp=endmp2 seg_lng=grad_lng));
   set oh00grad;
   run;
data oh01curv(rename=(begmp=begmp1 endmp=endmp1 seg_lng=curv_lng));
   set oh01curv;
   run;
data oh01grad(rename=(begmp=begmp2 endmp=endmp2 seg_lng=grad_lng));
   set oh01grad;
   run;
data oh02curv(rename=(begmp=begmp1 endmp=endmp1 seg_lng=curv_lng));
   set oh02curv;
   run;
data oh02grad(rename=(begmp=begmp2 endmp=endmp2 seg_lng=grad_lng));
   set oh02grad;
   run;
data oh03curv(rename=(begmp=begmp1 endmp=endmp1 seg_lng=curv_lng));
   set oh03curv;
   run;
```

```
data oh03grad(rename=(begmp=begmp2 endmp=endmp2 seg_lng=grad_lng));
   set oh03grad;
   run;
   data oh04curv(rename=(begmp=begmp1 endmp=endmp1 seg_lng=curv_lng));
   set oh04curv;
   run;
data oh04grad(rename=(begmp=begmp2 endmp=endmp2 seg_lng=grad_lng));
   set oh04grad;
   run;
proc sql;
     create table merge00 as
     select *
     from oh00acc left join oh00road
     on cntyrte = cnty_rte and
                (milepost between begmp and endmp) ;
quit;
run;
data merge00;
  set merge00;
   drop cnty_rte;
   run;
proc sql;
     create table merge00 as
     select *
     from merge00 left join oh00curv
     on rtenbr = rte_nbr and
                (milepost between begmp1 and endmp1) ;
quit;
run;
data merge00;
  set merge00;
   drop rte_nbr;
   run;
proc sql;
     create table merge00 as
     select *
     from merge00 left join oh00grad
     on rtenbr = rte_nbr and
                (milepost between begmp2 and endmp2) ;
quit;
run;
```

```
proc sort data=merge00 out=merge00 nodupkey;
    by caseno;
    run;
proc sort data=merge00 out=merge00 nodupkey;
    by caseno;
    run;
data oh00occ1; set oh00occ;
 keep   CASENO   VEHNO   SEATPOS   physcond   alcohol_test_status SOB_TST
          BAC drug_test_status   ;
 if VEHNO=1   and SEATPOS=1 ;
run;
data oh00veh1; set oh00veh;
 keep   CASENO   VEHNO   num_occs damsev veh_speed_post_2000 spd_limt drv_age drv_inj drv_sex ;
 if VEHNO=1 ;
run;
data oh00vc; merge oh00occ1 oh00veh1;
  by CASENO;
run;
proc sql;
    create table merge000 as
    select *
    from merge00 natural join oh00vc
  ;
quit;
run;
proc sort data=merge000 out=merge000 nodupkey;
    by caseno;
    run;
proc sort data=merge000 out=merge000 nodupkey;
    by caseno;
    run;
proc sql;
    create table merge01 as
    select *
    from oh01acc left join oh01road
    on cntyrte = cnty_rte and
              (milepost between begmp and endmp) ;
quit;
run;
data merge01;
 set merge01;
```

```
   drop cnty_rte;
   run;
proc sql;
     create table merge01 as
     select *
     from merge01 left join oh01curv
     on rtenbr = rte_nbr and
                (milepost between begmp1 and endmp1) ;
quit;
run;
data merge01;
 set merge01;
   drop rte_nbr;
   run;
proc sql;
     create table merge01 as
     select *
     from merge01 left join oh01grad
     on rtenbr = rte_nbr and
                (milepost between begmp2 and endmp2) ;
quit;
run;
proc sort data=merge01 out=merge01 nodupkey;
     by caseno;
     run;
proc sort data=merge01 out=merge01 nodupkey;
     by caseno;
     run;
data oh01occ1; set oh01occ;
 keep   CASENO   VEHNO   SEATPOS   physcond   alcohol_test_status SOB_TST
          BAC drug_test_status   ;
 if VEHNO=1   and SEATPOS=1 ;
run;
data oh01veh1; set oh01veh;
 keep   CASENO   VEHNO   num_occs damsev veh_speed_post_2000 spd_limt drv_age drv_inj drv_sex ;
 if VEHNO=1 ;
run;
data oh01vc; merge oh01occ1 oh01veh1;
   by CASENO;
run;
proc sql;
```

```
    create table merge001 as
    select *
    from merge01 natural join oh01vc
  ;
quit;
run;
proc sort data=merge001 out=merge001 nodupkey;
    by caseno;
    run;
proc sort data=merge001 out=merge001 nodupkey;
    by caseno;
    run;
proc sql;
    create table merge02 as
    select *
    from oh02acc left join oh02road
    on cntyrte = cnty_rte and
                (milepost between begmp and endmp) ;
quit;
run;
data merge02;
 set merge02;
   drop cnty_rte;
   run;
proc sql;
    create table merge02 as
    select *
    from merge02 left join oh02curv
    on rtenbr = rte_nbr and
                (milepost between begmp1 and endmp1) ;
quit;
run;
data merge02;
 set merge02;
   drop rte_nbr;
   run;
proc sql;
    create table merge02 as
    select *
    from merge02 left join oh02grad
    on rtenbr = rte_nbr and
```

```
                (milepost between begmp2 and endmp2) ;
quit;
run;
proc sort data=merge02 out=merge02 nodupkey;
    by caseno;
    run;
proc sort data=merge02 out=merge02 nodupkey;
    by caseno;
    run;
data oh02occ1; set oh02occ;
 keep   CASENO   VEHNO   SEATPOS   physcond   alcohol_test_status SOB_TST
          BAC drug_test_status   ;
 if VEHNO=1   and SEATPOS=1 ;
run;
data oh02veh1; set oh02veh;
 keep   CASENO   VEHNO   num_occs damsev veh_speed_post_2000 spd_limt drv_age drv_inj drv_sex ;
 if VEHNO=1 ;
run;
data oh02vc; merge oh02occ1 oh02veh1;
  by CASENO;
run;
proc sql;
    create table merge002 as
    select *
    from merge02 natural join oh02vc   ;
quit;
run;
proc sort data=merge002 out=merge002 nodupkey;
    by caseno;
    run;
proc sort data=merge002 out=merge002 nodupkey;
    by caseno;
    run;
proc sql;
    create table merge03 as
    select *
    from oh03acc left join oh03road
    on cntyrte = cnty_rte and
                (milepost between begmp and endmp) ;
quit;
run;
```

```
data merge03;
 set merge03;
  drop cnty_rte;
  run;
proc sql;
    create table merge03 as
    select *
    from merge03 left join oh03curv
    on rtenbr = rte_nbr and
              (milepost between begmp1 and endmp1) ;
quit;
run;
data merge03;
 set merge03;
  drop rte_nbr;
  run;
proc sql;
    create table merge03 as
    select *
    from merge03 left join oh03grad
    on rtenbr = rte_nbr and
              (milepost between begmp2 and endmp2) ;
quit;
run;
proc sort data=merge03 out=merge03 nodupkey;
    by caseno;
    run;
proc sort data=merge03 out=merge03 nodupkey;
    by caseno;
    run;
data oh03occ1; set oh03occ;
 keep  CASENO  VEHNO  SEATPOS  physcond  alcohol_test_status SOB_TST
        BAC drug_test_status  ;
 if VEHNO=1  and SEATPOS=1 ;
run;
data oh03veh1; set oh03veh;
 keep  CASENO  VEHNO  num_occs damsev veh_speed_post_2000 spd_limt drv_age drv_inj drv_sex ;
 if VEHNO=1 ;
run;
data oh03vc; merge oh03occ1 oh03veh1;
  by CASENO;
```

```
run;
proc sql;
    create table merge003 as
    select *
    from merge03 natural join oh03vc;
quit;
run;
proc sort data=merge003 out=merge003 nodupkey;
    by caseno;
    run;
proc sort data=merge003 out=merge003 nodupkey;
    by caseno;
    run;
proc sql;
    create table merge04 as
    select *
    from oh04acc left join oh04road
    on cntyrte = cnty_rte and
                (milepost between begmp and endmp) ;
quit;
run;
data merge04;
 set merge04;
   drop cnty_rte;
   run;
proc sql;
    create table merge04 as
    select *
    from merge04 left join oh04curv
    on rtenbr = rte_nbr and
                (milepost between begmp1 and endmp1) ;
quit;
run;
data merge04;
 set merge04;
   drop rte_nbr;
   run;
proc sql;
    create table merge04 as
    select *
    from merge04 left join oh04grad
```

```
    on rtenbr = rte_nbr and
                (milepost between begmp2 and endmp2) ;
quit;
run;
proc sort data=merge04 out=merge04 nodupkey;
    by caseno;
    run;
proc sort data=merge04 out=merge04 nodupkey;
    by caseno;
    run;
data oh04occ1; set oh04occ;
 keep   CASENO   VEHNO   SEATPOS   physcond   alcohol_test_status SOB_TST
          BAC drug_test_status   ;
 if VEHNO=1   and SEATPOS=1 ;
run;

data oh04veh1; set oh04veh;
 keep   CASENO   VEHNO   num_occs damsev veh_speed_post_2000 spd_limt drv_age drv_inj drv_sex ;
 if VEHNO=1 ;
run;
data oh04vc; merge oh04occ1 oh04veh1;
   by CASENO;
run;
proc sql;
    create table merge004 as
    select *
    from merge04 natural join oh04vc
   ;
quit;
run;

proc sort data=merge004 out=merge004 nodupkey;
    by caseno;
    run;
data agg01234;
 set merge000   merge001   merge002   merge003   merge004 ;
run;
data agg01234;
 set agg01234 ;
 drop JUR_TYPE RTE_NBR POP_GRP ACCYR FED_FACI ANGLE fed_medw
       RTE_SUFX RTE_TYPE FED_SPSY FUNC_CLS PK_LANES PAV_ROUG MED_TYPE
```

```
        PAS_NHS ID_CNTRL OUTSHWD1 INSHWD1 OUTSHWD2 INSHWD2 SURFWIDR
        SURFWIDL FED_ACES SEQ_NBR mile_cls MUN_NAM RURUID ;
run;
data agg01234;
 set agg01234;
 if NUMVEHS='1' and onoff_rd='2' or onoff_rd='3' then fatiacc=1; else fatiacc=0;
 run;
data   logis_agg; set   agg01234;
*Fill in missing values as zeros;
if med_wid=. then med_wid=0;
if curv_lng=. then curv_lng=0;
if deg_curv=. then deg_curv=0;
if grad_lng=. then grad_lng=0;
if pct_grad=. then pct_grad=0;
If rodwycls=99 then rodwycls=03;
If no_lanes="A" then no_lanes=1;
If no_lanes="B" then no_lanes=1;
If LIGHT=6 then LIGHT=1;
If RDSURF=0 then RDSURF=1;
If RODWYCLS=3 then RODWYCLS="03";
   If light=0   or light=" " or light=. then light=1;
   If rd_char1=0   or rd_char1=" " or rd_char1=. then rd_char1=1;
   If weather=0   or weather=" " or weather=. then weather=1;
   If rdsurf=0   or rdsurf=" " or rdsurf=. then rdsurf=1;
   If loc_type=0   or loc_type=" " or loc_type=. then loc_type=7;
   If med_wid=0   or med_wid=" " or med_wid=. then med_wid=0;
   If spdlimt=0   or spdlimt=" " or spdlimt=. then spdlimt=0;
   If rd_width=0   or rd_width=" " or rd_width=. then rd_width=0;
   If aadt=0   or aadt=" " or aadt=. then aadt=0;
   If deg_curv=0   or deg_curv=" " or deg_curv=. then deg_curv=0;
   If pct_grad=0   or pct_grad=" " or pct_grad=. then pct_grad=0;
   If drv_sex ="U" or drv_sex =" " then drv_sex ="M";
run;
data   logis_agg; set logis_agg ;
      nolanes =input(no_lanes, 5.);
run;
data   logis_agg; set logis_agg ;
vetical="flat" ;
if dir_grad= "-" or dir_grad= "+"   then vetical="veti";
run;
data   logis_agg; set logis_agg ;
```

```
horizental="tangent" ;
if deg_curv^= 0 then horizental="hori";
run;
data    logis_agg; set logis_agg ;
hi_curvature=deg_curv/seg_lng/seg_lng/52.8;
run;
data    logis_agg; set logis_agg ;
   if drv_age>=16 and drv_age<=25    then age_group='16-25' ;
   if drv_age>=26 and drv_age<=45    then age_group='26-45' ;
   if drv_age>=46    then    age_group='46+' ;
run;
data log_r2ln;
   set logis_agg;
   drop rte_type;
   if rodwycls="08" ;
run;
data sp_r2ln;
   set log_r2ln;
   drop div_code;
   if hour=2 or hour=3 or hour=4 or hour=14 or hour=15 ;
run;
data nsp_r2ln;
   set log_r2ln;
   drop div_code;
   if hour^=2 and hour^=3 and hour^=4 and hour^=14 and hour^=15 ;
run;
```

APPENDIX F. SAS Codes for Logistic analysis

```
proc logistic data= log_r2ln;
    class light rd_char1 sex age_group ;
    model fatiacc = light rd_char1 sex age_group
                    spd_limt aadt rd_width
                    / selection=stepwise ;
run;
proc logistic data= sp_r2ln;
    class light rd_char1 sex age_group ;
    model fatiacc = light rd_char1 sex age_group
                    spd_limt aadt rd_width
                    / selection=stepwise ;
run;
proc logistic data= nsp_r2ln;
    class light rd_char1 sex age_group ;
    model fatiacc = light rd_char1 sex age_group
                    spd_limt aadt rd_width
                    / selection=stepwise ;
run;
```