DETERMINING THE SAFETY OF URBAN ARTERIAL ROADS

by

Meredith Leigh Campbell

a Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Civil Engineering

April 29, 2004

APPROVED:

Dr. Malcolm H. Ray, Major Advisor

Dr. Frederick L. Hart, Head of Department

Abstract

The purpose of this project was to investigate the safety of urban arterial non-access controlled roads in Worcester, Massachusetts. An investigation into the dependent variable proved inconclusive and the historical accident rate was used. The best functional form for these roads was unclear so both linear and log-linear models were developed. A linear model was developed that predicted the total accident crash rate and log-linear model was developed to predict the same thing. A second linear model was developed to predict the total injury accident crash rate. The models were validated using independent data where the linear total accident crash rate model was found to be the most robust of the three in that both state primary roads and other arterial roads could have crash rates predicted to a better than fifty percent error.

Acknowledgements

I would like to take a moment to acknowledge the help and talents of the following people:

Professor Jason Wilbur

Jennifer Williams and Jeremy St. Pierre

Jennifer Weir

Nancy Sonnefeld

Balgobin Nandram

James Kempton

Elizabeth Cash

Professor Malcolm Ray

Karen and Archie Campbell

Jonathan Graham

This would have been a much harder and longer process without you all. Thank you.

Table of Contents

List of Figures

List of Tables

Notation and Abbreviations

AADT –Average Annual Daily Traffic
AASHTO –American Association of State Highway and Traffic Officials
ADT –Average Daily Traffic
AIC -Akaike's information criterion
ANOVA –Analysis of Variance
DOF –Degrees of Freedom
GLM –Generalized Linear Model
PDO –Property damage only
SAS –Statistical Analysis Software
TWLTL –Two way left turn lane

Allaccess –variable representing the total number of access points per segment including minor roads, parking lots and driveways
Benches –variable representing the number of benches on the segment
Commercial –variable representing the percentage of commercial land use per segment
Crest –variable for the maximum observed crest per segment in percent
Curve –variable indicating the number of curves on each segment
Curves –variable indicating any horizontal curvature for each segment
Density –variable representing the hazard density per segment
Drivepark –variable representing the total number of driveways and parking lots per segment
Driveways –variable representing the number of driveways per segment
Fence –variable indicating the number of fences on each segment
Grade –variable for the maximum observed grade per segment in percent
Hazards –variable indicating the total number of roadside hazards per segment
Heavyveh –variable indicating the percentage of heavy vehicles in the volume per segment
Hydrants –variable representing the number of fire hydrants on each segment
Industrial –variable representing the percentage of industrial land use per segment
Length –variable for the length in feet for each segment
Lighting –variable for the percentage of street lighting per segment
Maccess –variabel representing the number of minor road access points per segment
Markings –variable representing the quality of the pavement markings for each segment
Median –variable representing the presence of a curbed median on each segment
Ospole –variable for the number of overhead sign poles per segment
Other/electrical –both representing the number of electrical boxes on each segment
Parking –variable for the percentage of on-street parallel parking per segment
Parkinglots –variable for the number of parking lot entrances per segment
Pavement –variable representing the quality of the pavement
Pmeter –variable representing the number of parking meters per segment
Pole –variable for the total number of poles on each segment, including telephone poles, light poles, sign poles
Residential –variable for the percentage of residential land use per segment
SD –variable representing any sight distance problems on each segment

Spole –variable representing the number of sign poles per segment
Trees –variable of the number of trees per segment
Upole –variable representing the number of utility poles per segment
Vol –variable representing the average daily traffic for each segment
Widtha –variable representing the average lane width per segment
Widthm –variable representing the median width per segment
Widthsida –variable representing the average sidewalk width per segment

1 Introduction

Road safety is important to all of society. Even though people seldom consciously think about road safety, almost everyone uses the road network in one capacity or another and expect to survive the experience without injury. More than that, people don't even consider the event something "to survive" and consider traveling on the roads to be a basic part of life. Since there is such a large volume of road users, safety is important. Everything from cars and trucks, to public transportation and pedestrians needs the transportation network to be safe and efficient.

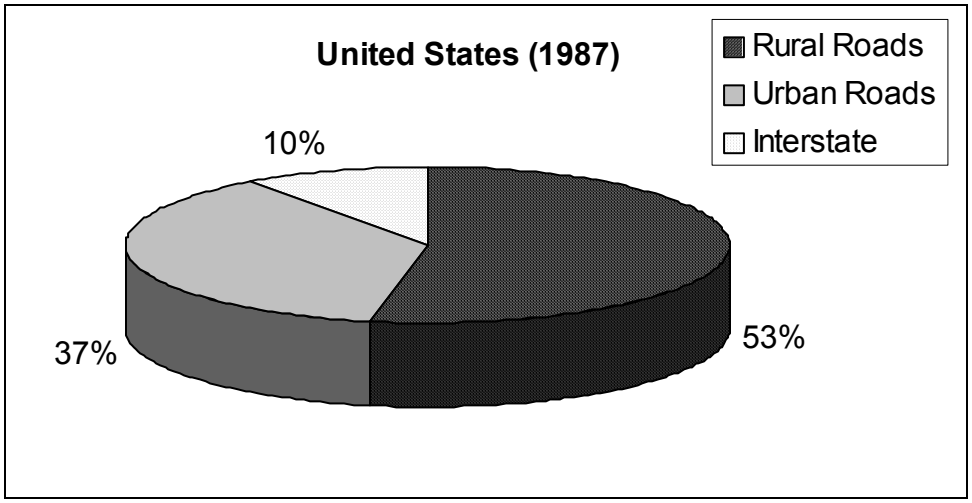


Figure 1: Distribution of Fatalities for Different Road Categories in the United States

Crashes can occur on any road at any time when a vehicle comes in conflict with a fixed or moving object. The majority of accidents occur on "two-lane rural roads … which are the locations of 50 to 60 percent of all severe accidents in Europe and the United States (Lamm, 9.1)." Rural roads have the majority of crashes occurring on them, so the majority of safety research has been focused on those roads. That still leaves approximately 40 to 50 percent of crashes occurring on urban roads and interstates (See Figure 1). Patrons of those roads also deserve to be treated to safe roads.

When looking at the numbers of fatalities and injuries that occur annually on the roadway system in the United States, the safety issue becomes even more evident. In 1998, in the Unites States alone there were 41,171 fatalities that occurred on the roadway system. There were even more injuries, almost 3.2 million injures (See Figure 2). With approximately half of these occurring in urban areas that is a staggeringly large number of accidents that safety improvements can strive to eliminate.

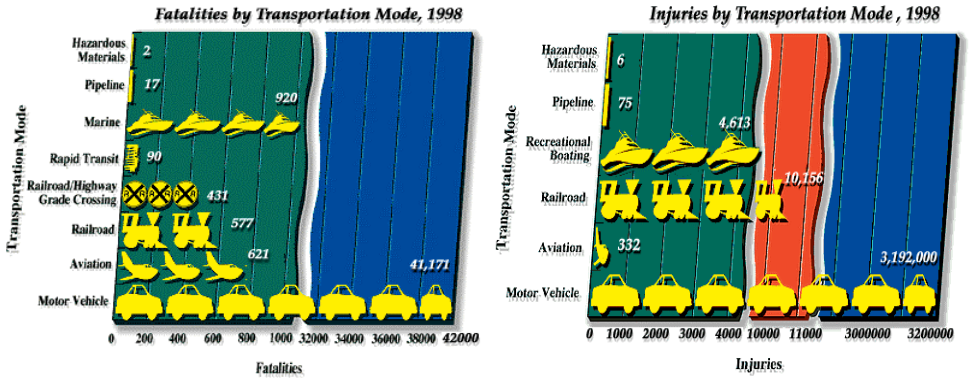


Figure 2:Fatalities and Injuries by Transportation Mode in the United States (1998) (Pedestrian Safety Roadshow)

The calculated costs of accidents come from wage and productivity losses, medical expenses, administrative expenses, vehicle damage, and employer costs. In 1993 the cost of a death due to traffic accidents was calculated to be $900,000, a disabling injury was calculated to be $32,000 and a property-damage only (PDO) accident was calculated to cost $5,800 (Poch and Mannering 105). These values, however, underestimate the cost of accidents by not including the value of a "person's natural desire to live longer or to protect the quality of one's life" (Poch and Mannering 105). This desire is difficult to place a monetary value on and in 1995 the willingness to pay for this was estimated at $3,000,000 (Poch and Mannering 105). Even if some percentage of these accidents can be prevented, millions of dollars could be saved each year.

The safety of urban roads has not yet been fully examined due to the complex nature of the issues and the lack of resources available to devote to the problem. The main factors exerted on driving behavior include human factors, physical features of the site, traffic, legal issues, environment, and the vehicle (Choueiri et al 34), all which contribute to the complex mix of causes of traffic accidents.

Urban roads can be divided by more than just location in terms of population centers, but by the type of traffic using the roads. Table 1 shows the typical distribution of travel volume and length of roadways of the functional systems for urban areas. Road systems developed for urban areas usually fall within the percentage ranges shown. This table shows that the majority of travel in urban areas occur on the arterial roads. These arterial roads account for up to 25% of the urban roadway length indicating that the majority of travel occurs on a minority of roads. Accidents may not be exactly linearly distributed between these types of roads, but the most efficient way to improve the overall safety of the road system is to focus on the areas with the most traffic. Fortunately, this area of arterial roads has the least number of actual miles, making improvements to this area effect the majority of drivers.

Table 1:Typical Distribution of Urban Functional Systems

	Range	
Systems	Travel Volume (%)	Length (%)
Principal arterial system	40-65	5-10
Principal arterial plus minor arterial street system	65-80	15-25
Collector Road	5-10	5-10
Local Road System	10-30	65-80

(Greeenbook, Exhibit 1-7, 12)

1.1 Problem Statement

Quantifying the safety of urban and suburban roads and streets has not attracted the same attention as two-lane rural roads. Since two-lane rural roads have been examined and analyzed in depth, determining the safety of urban and suburban arterials is the next area to be attacked. The creation of a method that can quantify the safety of urban arterials would enable transportation planners and managers to determine the safety of their particular network and help prioritize road creation and improvement projects. Currently the agencies that are responsible for all the road systems do not have quantifiable tools for considering safety in their decisions. Often when difficult choices need to be made, priority is given to factors such as cost, operational impacts, environmental impacts and experience, but not necessarily safety improvement. The purpose of this research is to help predict the safety performance of various elements considered in planning, design, and operation of non-limited-access urban arterials. By monitoring accident rates at a specific site, traffic safety engineers and researchers hope to be able to detect when or if safety has deteriorated. An accurate prediction of the number of accidents, or accident rate, occurring at a particular site is invaluable in the assessment of the effectiveness of an improvement program (Higle & Witkowski 24). An accurate way to help prioritize improvement projects will allow the limited dollars to be used in such a way as to make the most of them and the most possible improvement.

Safety is often defined as the accident rate of a road section. "Vehicle accidents are complex events involving the interactions of five major factors: drivers, traffic, road, vehicles, and environment (e.g., weather and lighting conditions)" (Miaou, 7). Developing accident prediction models is a way to summarize these complicated

interactive effects and try to explain the variation between sites from one time to another. Once a model is found that represents the relationship between all factors, it can be used to aid in finding cost-effective methods to reduce accident frequency/severity over the long term. Traffic and safety engineers would like to control all of these major factors, but are limited to what they can actually influence, which puts limits on how effective prediction models can be. Driver behavior is a complex issue that has been attempted to be modeled, but to no great success and is therefore usually left out of prediction models. Environmental conditions cannot be controlled and vehicles are available today in greater number of types and quality causing many areas where uncertainty can occur in prediction models. This leaves only roadway and traffic characteristics that can be controlled by highway engineers and used with any level of certainty in prediction models.

This project will develop an accident prediction model for the safety of urban, non-access controlled, arterial roadways. This will involve looking at variation in accident frequency due to both systematic variations due to differences in sites and random variation. Systematic variation can be explained as the variation of long-term means among different sites and time intervals while random variation can only be explained as the accident variation without physical explanation. The random variation is, however, assumed to follow probability laws and relatively homogeneous sites are often characterized by a probabilistic distribution. Researchers typically use normal distributions, Poisson distributions or negative binomial distributions. Variation also enters modeling because not all the needed information is readily available and the available sample size is finite. There is also the issue that the accident rate associated

with a particular site is itself a random variable, which cannot be predicted with absolute certainty (Higle & Witkowski 24). The variables this project will examine as possible regressor variables are limited to ones that are either already available, or easily obtainable without complex collection procedures, which would restrict the use of any developed models.

A standard practice for identifying unsafe locations is based on historical data where a site is classified as hazardous if accident history exceeds a specified level usually defined as a certain accident rate or number of accidents per year (Higle & Witkowski 24). A common method used in practice is to identify a site as hazardous if its accident rate exceeds the mean accident rate over all sites in the region plus a multiple of the standard deviation (Higle & Witkowski 24). But, due to the random variations that are inherent in accident phenomena, historical accident data do not always accurately reflect long-term accident characteristics making this an inaccurate method for identifying hazardous sites (Higle & Witkowski 24). A better method for identifying hazardous locations includes looking at factors other than just historical accident data. The more factors used the more accurate identification as a hazardous site can be. In short, arterial roads in Worcester, Massachusetts will be examined for their traffic, land use, access, alignment, hazards and other characteristics that can affect the causes of accidents and models will be developed to predict the safety of urban arterial roads.

Chapter 2 gives background information related to the types of roads under consideration and some background on the mathematical theory. Chapter 3 gives an overview of the methods used to complete this project while chapter 4 covers what data was collected and how that was done. Chapter 5 consists of the majority of the

mathematical analysis while chapter 6 gives the results of that analysis with an overview of the three models developed in this project. The validation of the three models is covered in chapter 7 and chapter 8 gives the conclusion that can be draw from this work.

2 Background Information

For an accident prediction model, there are several areas where some background information would be useful. These areas encompass topics relating to roadway and traffic concern as well as those that are related solely to modeling.

2.1 Functional Classification

Functional classification is the grouping of highways by the type of service they provide and was developed to help with transportation planning (Greenbook 1). The classification system recognizes that individual roads do not serve travel independently; rather, travel involves movement through a network of roads, which can be separated by use (Greenbook 4). Roads are classified in the United States according to the combination of mobility and access on each roadway. The type of classification determines and aids in the design and maintenance of the road networks. The major divisions between access and mobility necessitate the differences in the functional classes (Greenbook 6). The higher the access function of a road, the lower its mobility function becomes, similarly the higher the mobility function the lower the access function; this can be seen in Figure 3. Limited access on arterials enhances their primary function of mobility while full access on local roads promotes accessibility to individual land parcels.

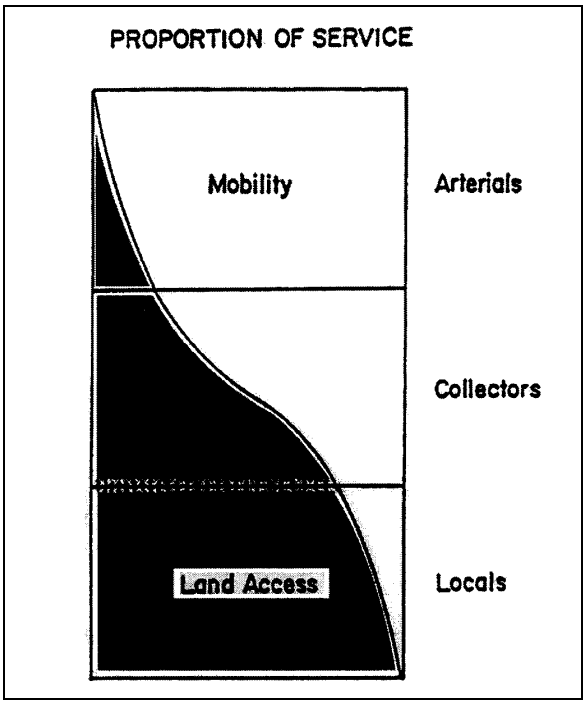


Figure 3: Relationship of Functionally Classified Systems in Serving Traffic Mobility and Land Access

(AASHTO Greenbook Exhibit 1-5)

Highways and streets are described as rural or urban roads, depending on their location. This differentiation is due to fundamental differences in characteristics between urban and rural areas specifically in land use and population density, which significantly influence travel patterns (Garber & Hoel 658). After the primary classification, highways are then classified under the following categories: arterials, collectors, and local roads. Local roadways emphasize the access function. Arterials emphasize mobility for through movements over long distances, while collectors offer approximately balanced service for both mobility and access.

2.1.1 Urban Roads

Urban roads are facilities located in urban areas, which are designated by state and local officials. Areas designated as urban can vary slightly by state though they are usually classified as having populations of 5,000 or more (Garber & Hoel 658). Urban

locations can be further divided into areas with population of 50,000 or more, *urbanized areas*, and areas with populations between 5,000 and 50,000, *small urban areas* (Garber & Hoel 658). Urban areas have a high intensity of land use and large amounts of travel, which makes the placement of urban roads more critical than those in rural areas, since urban roads have less space in which to be built. The high density of roads and traffic makes the safety of these roads critical. Figure 4 shows the basic layout of an urban network.

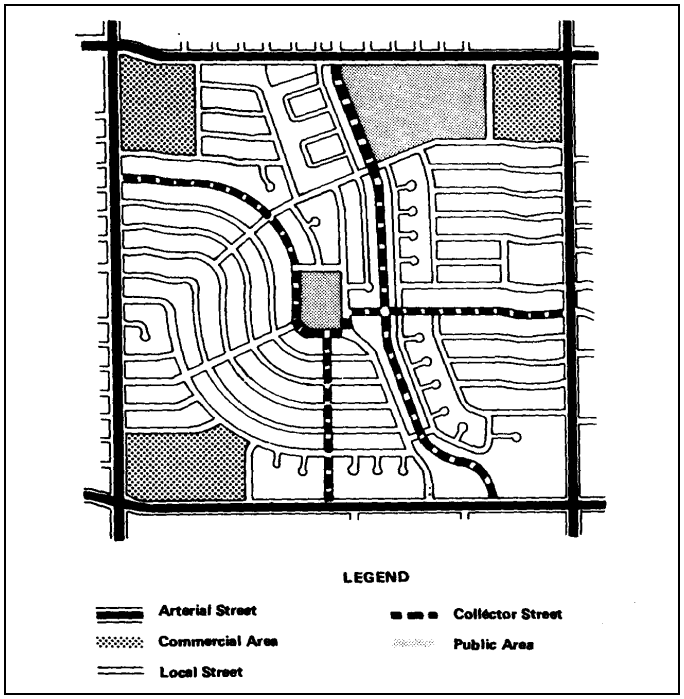


Figure 4: Schematic of the Functional Classes of Urban Roads

(Garber and Hoel 659)

2.1.1.1 Urban Arterial System

The urban arterial system is divided into principal arterials and minor arterials. Urban principal arterials serve the major activity centers, which consist of the highest traffic volume corridors, which carry the longest trips. They carry a high proportion of the total vehicle-miles of travel within the urban areas, even though they amount to a

relatively small percentage of the total network (Greenbook 11). Principal arterials tend to bypass the central business districts and carry most of the trips entering and leaving cities. All controlled access facilities are within this system, though access control is not necessarily a condition. Principal arterials can also be further divided into subclasses based mainly on access control: (1) interstates with full access control and grade-separated interchanges, (2) expressways which have controlled access but may also include at-grade interchanges and (3) and other principal arterials which have little or no access control. (Garber & Hoel 659).

Streets that interconnect with and augment the urban primary arterials are classified as urban minor arterials. This system places more emphasis on access and offers lower mobility than the primary arterials. Although minor arterials "may serve as local bus routes and may connect communities within the urban areas, they do not normally go through identifiable neighborhoods" (Garber & Hoel 659, Greenbook 11). Despite the differences that exist between principal arterials and minor arterials, they are all classified as high mobility and low access facilities.

2.1.1.2 Urban Collector System and Local Road System

Urban collector streets' main purpose is to gather traffic from local streets in residential areas or central business districts and channel it into the arterial system. Collectors, therefore, go through residential and commercial areas and ease traffic circulation through neighborhoods and business districts. Collectors can penetrate residential neighborhoods, distributing trips from the arterials through the area to their ultimate destinations.

The urban local road system includes all other streets in urban areas that have not been included in the previous systems. The main purpose of these streets is to provide access to abutting land and furthermore to allow traffic on that land access to the collector system (Garber & Hoel 660). The local roads are intended to serve multiple types of traffic, including pedestrians and cyclists, and due to the many users through traffic is discouraged to improve safety for the slower ones (Lamm 3.1). This system has the lowest level of mobility, but the highest level of accessibility.

2.1.2 Rural roads

Rural roads consist of all other roads not located in an urban area. They function by connecting separate cities together instead of connecting parts of cities together as is commonly found in urban roads (Garber & Hoel 660). Arterial highways in rural network provide direct service between cities and larger towns, while collectors serve smaller towns connecting them to the arterial network, gathering traffic from the local roads, which serve individual farms and other uses. This network can be viewed in Figure 5. Similar to the urban network, the rural network is divided into arterial, collector and local roads.

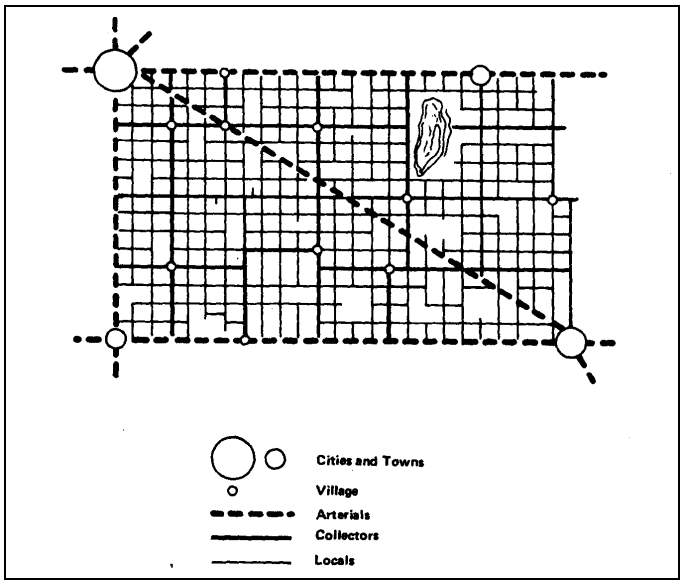


Figure 5: Schematic of the Functional Classes of Rural Roads

(AASHTO Greenbook Exhibit 1-3)

2.1.2.1 Rural Arterial System

The rural arterial system is divided into principle arterials and minor arterials. The principle arterials are composed of most of the interstate and account for most statewide trips. Freeways are a special type of arterial consisting of divided highways with full access control and no at-grade crossings (Garber & Hoel 660). This class of highway includes the heavily traveled routes that warrant multilane improvements and most of the existing rural freeways (Greenbook 8). The minor arterials assist in connecting cities and towns and all the rural arterials are characterized by uninterrupted, high-speed flow. Due to the large traffic volume on these roads much time has been spent researching the safety of this part of the road network.

2.1.2.2 Rural Collector System and Local Road System

Highways classified as rural collectors primarily carry traffic within individual counties. Major collector roads mostly carry traffic to and from large cities that are not directly served by the arterial system, and also carry the majority of the intra-county

traffic (Garber & Hoel 660). The rural minor collectors bring traffic from local roads and transport it to the arterial systems. Collectors are all characterized by more moderate speeds than arterials, and a larger amount of accessibility, though some can have access control.

The rural local road system contains all the roads still remaining within the rural classification. These roads serve trips of short distances and provide direct access to individual residences (Garber & Hoel 661). Conversely, the system also links the individual properties to the collector system. Like all local roads, rural local roads are characterized by low speeds and high access.

2.2 Roadway Alignment

A roadway's alignment is composed of its horizontal and vertical orientation. Vertical alignment includes tangent grades and sag, or crest, vertical curves. Horizontal alignment, similarly, consists of level tangents and circular curves. These elements all contribute to the safety of the road design.

Many studies have been conducted to investigate the effects of various alignment designs on safety including those by Lamm, Hadi, and Gibreel (Lamm et al) (Hadi et al 169) (Gibreel et al 305). Many elements have been found to affect safety through all aspects of alignment design. Studies have also indicated that improvements to highway alignment could significantly reduce the number of crashes that occur on those roadways (Gibreel et al 305) (Poe & Mason) (Miaou et al A). But, only quantitative relationships can adequately show the relationship between design elements and crash rates allowing highway planners and designers to use the information to make informed decisions about better designs.

2.2.1 Cross Section

Much of the research on cross-section design safety has been devoted to two-way two-lane rural highways. Figure 6 shows the major components in a divided cross-section design. The cross slope, lane width, shoulder width and type are the elements given the most focus during the design process.

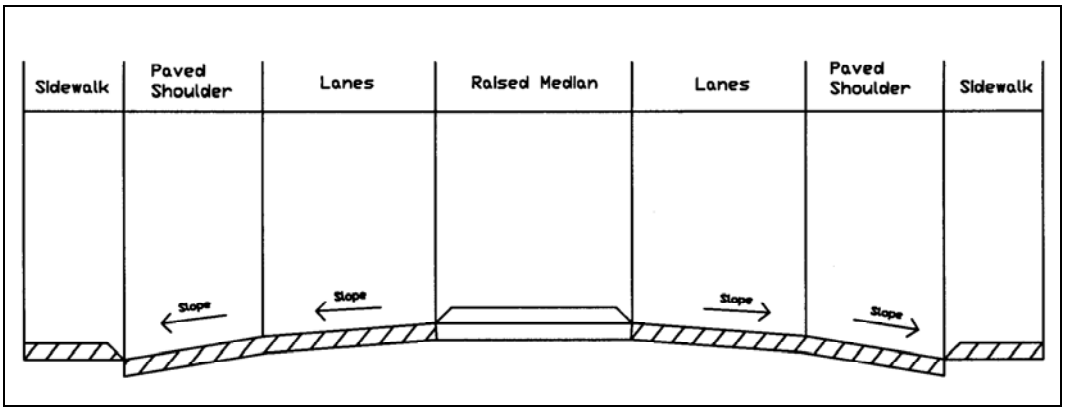


Figure 6: Cross-section of a Divided Roadway

2.3 Cross slope

Undivided roads have a crown or high point in the middle with a downward slope towards both edges, though unidirectional slopes may also be used. The primary purpose of having a cross slope is to facilitate drainage. A steep crown is desirable to make the water flow as quickly as possible away from the main traveled path, but too large of a slope can cause vehicles to drift towards the lower edge of the road (Greenbook 313). The two elements need to be balanced in order to get the most benefit from the crown before the negative consequences come into play. American Association of State Highway and Transportation Officials (AASHTO) has produced a generalized set of guidelines to help designers in choosing the proper amount of cross slope to use on road designs. Accepted cross slope rates range from 1.5 to 2 percent for two lane roads. As additional lanes are added the cross slope rate may be increased by 0.5 to 1 percent.

Slopes larger than two percent are not desired on high-speed roads due to the fact that high crowns can cause trucks with high centers of gravity to sway when traveling at high speeds (Greenbook 313). In areas of high rainfall, cross slopes can be extended to 2.5 percent to handle the large volume of water (Greenbook 314).

2.3.1.1 Lane width

The lane width of roads can greatly influence the safety and comfort of driving. Lane widths generally range between nine and twelve feet where the minimum width is limited by the width of the design vehicle for the road. The maximum width for lanes is limited by the amount of space needed where drivers could perceive a lane where one does not actually exist. The recommended lane width for all new roads by AASHTO is twelve feet (Greenbook 316). Increasing lane width to the maximum value can reduce crash rates for urban freeways and undivided highways (Hadi et al 176). In some situations such as low-speed facilities, urban areas with restrictive development and right-of-way, and low volume roads in rural and residential areas, smaller lane widths are permitted. Russia and European countries have developed an empirical relationship between pavement width and accidents $N = \frac{1}{0.173W - 0.21}$ where N is the number of accidents per million-vehicle kilometers and W is the pavement width in meters. This shows that accident rate decreases with an increase in pavement width (Gibreel et al 308). The above relationship helps support the idea that lane widths affect roadway safety.

2.3.1.2 Shoulder Types and Width

Shoulders are the area of the road intended for stopped vehicles, emergency vehicles and structural support of the roadway. Shoulders can vary in width and type,

surfaced or un-surfaced. Surfaced shoulders use asphalt or concrete pavement, gravel, shells, and crushed rock as surfacing material while un-surfaced shoulders are typically dirt and grass. In urban situations, parking lanes can help to provide some of the same services as shoulders on rural roadways. Widths range from two feet wide on minor rural roads to twelve feet on major roads with most shoulders ranging between six and eight feet (Greenbook 318). Research has shown that increasing the outside shoulder width to between ten and twelve feet helps to decrease accident rates (Hadi et al 176).

Choueiri et al found that there is a tendency for accident rates to decrease with increasing overall pavement width up to 7.5 meters (25 feet) on two-lane roads (Choueiri et al 37). This was confirmed by many studies in countries including the United States, Germany, Canada, and the former United Soviet Socialist Republic (Choueiri et al 37). Though the accident rate decreased, the accident cost rate, an indication of severity tended to go up with increased pavement widths (Choueiri et al 37). This is due to the fact that roads with wide lanes and shoulders tend to have higher speeds and the accidents that occur on them tend to be very severe. This shows why the individual lane and shoulder widths, as well as the overall pavement width of the road, are important.

Some roads, especially in urban areas have shoulders that are used primarily for parking. This allows space for parallel parking, but increases the number of roadside hazards that can be struck by moving vehicles. The problem of hazards versus need for parking in commercial urban areas needs to be balanced to prevent problems occurring from the presence of parked vehicles. This balance is mostly necessary in locations where the road has been divided to allow for higher speeds, where the parked vehicles permit for increased pedestrian presence.

2.3.1.3 Curbs

The type and location of curbs can affect driver behavior, especially their feelings of comfort. Curbs can make drivers more comfortable by illuminating the edge of the road. Curbs are primarily intended for drainage and delineation of road and sidewalks. They consist of a vertical or raised portion to physically create a barrier between spaces with different purposes such as roads for vehicle travel and sidewalks for pedestrian travel. Curbs are used on all types of low speed urban highways, though caution needs to be applied when placing curbs on high-speed roads (Greenbook 323). Caution is needed because curbs can cause problems when they are struck at high speeds causing vehicles to flip. The positive benefits of curbs, for delineation and directional control of water, need to be balanced with their adverse affects on safety for high-speed vehicles.

2.3.2 Horizontal Alignment

Horizontal alignment describes the variation in placement of horizontal design elements of the roadway, which consists of level tangents separated by curves. Horizontal curves can consist of simple curves, single circular arcs or compound curves of two circular arcs on the same side of a common tangent (Easa 1). A simple curve is bordered on both sides by tangents and consists of a single circular curve. Compound curves consist of two or more curves in a row, which all turn in the same direction and any two successive curves have a common tangent point (Garber & Hoel 701). Reverse curves consist of two simple curves of equal radii turning in opposite directions with a common tangent point. Reverse curves are generally used to alter the alignment of a highway (Garber & Hoel 706). Designers try to avoid reverse curves whenever possible, in order to avoid the sudden radical change in alignment which can cause the driver to

have problems staying in their own lane (Garber & Hoel 707). Spiral curves are also known as transition curves and gradually increases or decreases the radial force as a vehicle is entering or departing from a circular curve (Garber & Hoel 707).

A large number of accidents tend to occur at horizontal curves. A study by Choueiri et al showed that a negative relationship between radius of curve and accident rate exists, meaning the smaller the radius the more accidents occurred (Choueiri et al 44). To combat this safety issue, when there is space available, large radii should be used on horizontal curves. Once radii became greater than 400 to 500 meters (1,650 feet), the marginal increase in safety per increase in radius is very low (Choueiri et al 44).

Horizontal alignment uses design speed as an overall design control and uses friction, superelevation and curvature to set specific limits. The limits are based on mechanical relationships, but the values used in design are adjusted due to practical limits determined empirically over the range of values allowed (Greenbook 131). A design speed, superelevation, and friction factor have to be chosen and then the minimum radii can be determined by $R = \frac{u^2}{15(e + f_s)}$ where R is the minimum radius (ft), u is the design speed (mph), e is the superelevation, and fs is the coefficient of side friction.

Superelevation is an "inclination of the roadway towards the center of the curve" (Garber & Hoel 67) and is regulated by AASHTO with maximum values being limited by design speed and environmental factors. In areas with snow and ice the super elevation is restricted to less than eight percent, though in other areas it can be as high as ten or twelve percent (Greenbook 141). The relationship between geometric design, specifically horizontal design and operating speed has been shown in studies for all types of roadways. Relationships between geometric design and operating speed on two-lane

rural highways show that horizontal curvature is a significant effect on operating speed (Poe & Mason 18). High-speed geometric design is based on design values for geometric elements that promote speed consistency and safety (Poe & Mason 18). Low-speed design tries to provide access and accommodate mixed types of users such as bicyclists and pedestrians with the goal of maintaining lower speeds to achieve the functionality of the road and improve overall safety (Poe & Mason 18).

Due to the relationship between horizontal alignment and operating and design speeds, many researchers have attempted to create a quantifiable relationship between the two. Lamm and Glennon independently examined this relationship in depth. Both groups developed models for predicting the 85^{th} percentile speeds of vehicles using degree of curvature (degrees/100 ft) as a variable.

V85=94.37-1.83DC (Lamm's group)

V85=93.8-2.59DC (Glennon's group) (Poe & Mason 19)

Both models displayed very similar relationships with only minor differences. The constant reflects the differences in the maximum speeds allowed on the tangent or straight sections of roads and then an adjustment is made based on the specific curve. Lamm and Choueiri's work in the late 1980's confirmed the importance of the radius of curve (degree of curve) by concluding that it is the most influential parameter in determining accident rates on horizontal curves (Gibreel et al 309). The probability of accidents is higher on curves than on tangents since the road is changing causing the driver to do more work allowing room for more mistakes and can be especially dangerous when high-speed roads have sharp curves that abruptly slow traffic making the situation ripe for an accident.

2.3.3 Vertical Alignment

Vertical alignment consists of straight sections of grades, or tangents connected by vertical curves. The curves consist of single parabolic arcs (sag or crest) or compound curves (unsymmetrical curves) of two parabolic arcs with a common tangent (Easa 1). Design of vertical alignment, therefore, consists of choosing the proper grade and the layout of the curve. The proper grade is important since vehicles traveling upward tend to loose speed due to the downward force from the weight of the vehicle unless the driver accelerates (Garber & Hoel 56). Trucks and buses are especially affected by long grades, on upgrades speed reduction can be extreme and on downgrades the brakes may not be strong enough to slow and stop heavy vehicles. This is a key concern on higher speed roads (45 mph and up), but is less of a concern on slower speed roads. The sharpness of the grade will also affect this, with larger grades having a more significant effect on traveling vehicles.

The selection of maximum grades for a highway depends on the design speed, and a general heuristic is that grades of 4 to 5 percent have little to no effect on passenger cars (Garber & Hoel 675). Table 2 shows the maximum allowable grades for urban arterials as recommended by AASHTO. Similar tables exist for urban and rural collectors and local roads with the allowable grades increasing slightly as roads increase in accessability and decrease in mobility. Maximum grades are specified by design speed and terrain type.

Table 2: Maximum Grades for Urban Arterials

	US Customary Units						
	Maximum Grade (%) for Specified Design Speed (mph)						
Type of Terrain	30	35	40	45	50	55	60
Level	8	7	7	6	6	5	5
Rolling	9	8	8	7	7	6	6
Mountainous	11	10	10	9	9	8	8

(AASHTO Exhibit 7-10)

Some studies have examined the point when grade starts playing a significant role in increasing accident rates. A study done in 1973 with data from the United Kingdom, the former Soviet Union and Germany found a direct relationship between accident rate and grade. $N = 0.265 + 0.105G + 0.023G^2$ where N equals the number of accidents and G is the percent of grade. This shows that accident rates increase with an increase in grade (Gibreel 309). A later study in 1994 concluded that accident rate slightly increases with increases in grade up to six percent and sharply increase at grades higher than six percent indicating that for rolling and mountainous terrain, the grade plays a large role in effecting accidents (Choueiri et al 44). Minimum grades can also be an important issue. They are based on the need to provide adequate drainage especially when there are curbs present, which prevent free drainage from all parts of the roadway (Garber & Hoel 676). If the minimum grade is not large enough, water can collect on the pavement and contribute to the road's deterioration and increase accidents by causing vehicles to hydroplane.

Vertical curves are supposed to provide a gradual change from one grade to the next for a smooth overall ride and are mostly parabolic in shape and can be classified as crest or sag curves (Garber & Hoel 676). To design a vertical curve, the criteria to consider includes the minimum stopping sight distance for crest curves, headlight sight distance for sag curves, drainage, comfort and appearance for both types of curve.

Headlight glare and minimum sight distance work in a similar fashion, by providing minimum allowable lengths for the curves. Available sight distance should be designed to be equal or greater than the required sight distance to make certain that all the design requirements are met. Headlight glare conditions are most important on sag vertical curves where on-coming traffic can blind the driver if the curve is designed improperly. Driver comfort is also most important in sag vertical curve conditions where gravitational and vertical centripetal forces are acting in opposite directions, so the rate of change of grade needs to be kept within "tolerable limits" (AASHTO Greenbook 269). The appearance consideration is that long curves have a more pleasing appearance than short ones, which can give the appearance of a sudden break in the profile (AASHTO Greenbook 270). Appearance and comfort are only given a passing consideration, as most curves that are designed for the minimum sight distance will already be appropriate for comfort and appearance.

2.4 Access Control

The function of a highway system is to provide both mobility and access. Arterial roadways can be designed with various levels of both accessibility and mobility. Arterials often have infrequent access points and barriers to prevent crossing, as found in the interstate system or principal arterials, or they can be designed with low access control with many direct access points for all land uses as in the minor arterials. Improving safety is an important goal of access control management. To help in evaluating the possible benefits, models to predict crashes based on road geometry and access control characteristics are being developed.

One of the major indications of access control management is the presence of medians and islands on the roads and at intersections. A common access control technique involves the use of medians and refuge islands to increase safety by decreasing the number of possible vehicle or pedestrian conflicts. The definition of a median is "the portion of a highway separating opposing directions of the traveled way"(Green Book, 341). This definition does not, however, state what the function of a median is or how it is to be constructed. There are a variety of different median types in use where some are combined with barriers designed to prevent out-of-control vehicles from crossing into opposing vehicles and wider medians relying on their width to prevent opposing vehicle crashes. Medians can be divided into three major types: raised, depressed or flush, and installed for several different reasons.

2.4.1 Median Purpose

Medians are an effective method for increasing safety and vehicle capacity on arterials and are generally considered to improve pedestrian safety. The main goals of a median include a) separating opposing vehicles b) providing vehicles with a safe clear zone to avoid other moving vehicles and reduce roadside object collisions and c) providing a refuge for turning or crossing vehicles and pedestrians (Knuiman et al 71). Medians can be designed for one or more of these general goals. One way for reaching these goals is for medians to provide an additional lane for thigh speed traffic by creating left turn bays and removing the turning vehicles from blocking the traffic flow. Similarly, medians will protect entering vehicles that want to cross one or both directions of traffic. Medians on a divided highway can provide a recovery area for out-of-control vehicles, by allowing space for the vehicle to regain control before crossing into the

opposing traffic. A side benefit of medians on arterials is that they can provide a landscaping area, as long as vegetation is frangible and will not cause fixed object collisions. Despite these opportunities for medians to protect vehicles and pedestrians, their safety benefits are largely unknown and theoretical since the true effects of medians are difficult to quantify.

Similar to medians, refuge islands are designed to provide a place of safety for pedestrians who cannot safely cross the entire roadway at one time due to changing traffic signals, oncoming traffic, or the pedestrian's own capabilities. They are particularly useful at locations where heavy volumes of traffic make crossing difficult especially on multilane roadways, large or irregularly shaped intersections and at signalized intersections (Bowman & Vecellio a 180). However many studies done on the effect of medians on improving pedestrian safety have been called into question due to the researchers disregard of changing pedestrian and vehicular volumes throughout the time period of the study (Bowman & Vecellio a 183).

The before and after studies of pedestrian accidents in areas with median installations often do not take into account the increased number of pedestrians when a median or island is installed. Larger numbers of pedestrian accidents at a specific location may not be alarming if the accident rate is calculated, but getting realistic pedestrian counts is difficult and rarely done. Therefore, Bowman and Vecellio's findings of higher accident rates for undivided arterials than for arterials with raised or two-way-left-turn-lane may be due to larger volumes of pedestrians being attracted to the areas with undivided cross sections than the median treatment being effective. Medians and refuge islands are both techniques intended to increase pedestrian safety, but the

actual effect on pedestrian safety is unclear and, like medians, difficult to quantify especially as most studies have focused on the safety benefits to motorized vehicles.

2.4.2 Median types

There are three major types of medians, raised, depressed, and flush. Depressed medians are generally used on freeways to help create more efficient drainage and snow removal. According to AASHTO's Policy on Geometric Design of Highways and Streets, depressed medians should have side slopes of 1V:6H, but 1V:4H also may be adequate (Green Book 341). Figure 7 shows the layout of typical depressed medians. This type of median separates the opposing traffic, but may cause problems in providing a safe clear zone between the two directions. This can be due to the depression intended to aid with drainage not being properly maintained and vegetation growing up. Also, if the slopes are built too steep a vehicle could roll over while in the median.

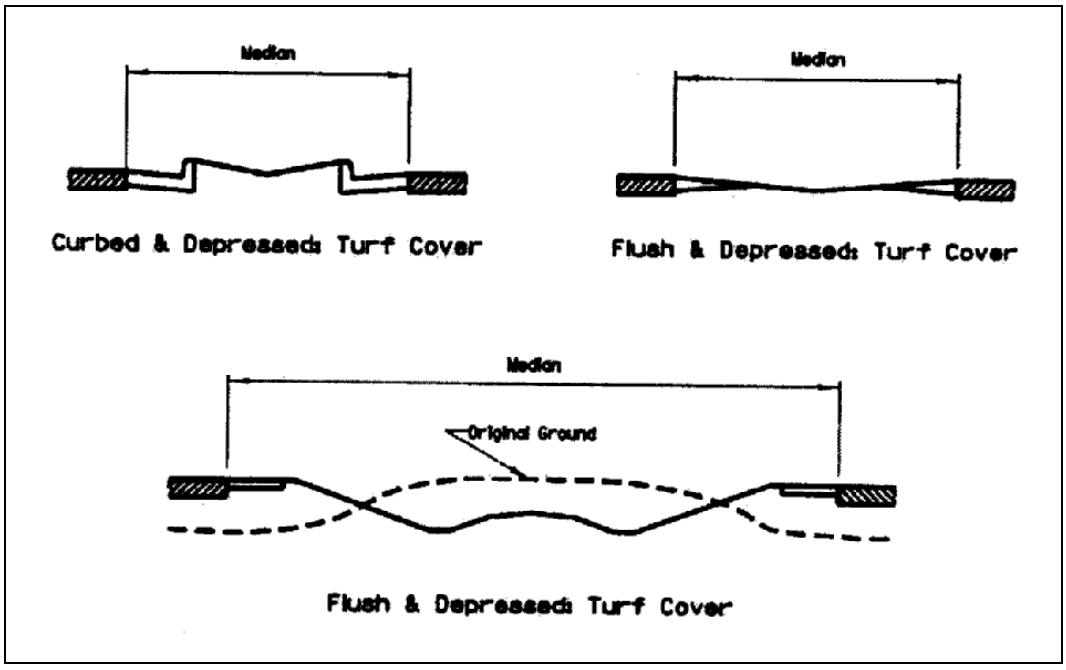


Figure 7: Depressed Median

Exhibit 7-7 AASHTO's Green Book

Raised medians, on the other hand, are seldom used in freeway situations any more. On freeways, raised medians cause problems for out-of-control vehicles. The slope, while separating the traffic flows, does not allow for the out-of-control vehicles to use the median as a place of refuge and avoid vehicles and objects. The out-of-control vehicle cannot climb the slope and the high slope tends to cause the vehicles to roll over and land back in the traffic stream that was just left.

However, raised medians of a different style have an application on arterial streets where it is desirable to regulate left-turn movements, by limiting left turns and U-turns except at designated points. Separating the traffic in arterial streets also increases the comfort level of the driver and increases the traffic speed. In this situation, the term raised median implies the use of a curb and ability to be used as a pedestrian refuge as seen in Figure 8. In order to be officially called a pedestrian refuge, medians must be at least 4 feet wide, though 6 feet is needed for multiple pedestrians, bicyclists and wheelchairs.

Figure 8: Raised Curb Median

Raised curb medians were the predominant treatment first used in urban areas. They were found to be effective in controlling left turn movements and separating opposing traffic flows as well as providing pedestrian refuge. Table 3 shows a compiled list of the advantages and disadvantages of raised medians. Use of raised medians increases traffic flow and speed limits while reducing the number of mid block collisions by limiting the number of conflict points. However, there is often an increase in crashes at intersections and sometimes an increased number of fixed object collisions. Increasing congestion, limited right-of-way, high construction cost, and the need for more left turn opportunities resulted in the increasing use of flush medians, specifically two-way left-turn lanes in urban locations where previously a raised curb median would have been installed (Bowman & Vecellio a 181).

Table 3: Advantages and Disadvantages of Raised Medians

Advantages	Disadvantages
1. Discourages new strip development and encourages large planned development	1. Reduces operational flexibility for emergency vehicles and others
2. Allows better control of land use by local government	2. Increases left turn volume at major intersections and median openings
3. Reduced number of conflicting vehicle maneuvers at driveways	3. Increases travel time for vehicles desiring to turn left where median openings are not provided
4. Safer on major arterials with high (>60) number of driveways per mile (>37 driveways per km)	4. Reduces capacity at signalized intersections
5. Increases traffic flow	5. Possible increase of accidents at intersections and median openings
6. Desirable for large pedestrian volumes	6. Usually increases fixed object accidents
7. Permits circuitous flow of traffic in grid patterns	7. Requires motorists to organize their trip making to minimize the need for U-turns and use the arterial only for relatively long through movements
8. Allows greater speed limits on through road	8. To minimize delay requires inter-parcel access, which may not be under government control or would be expensive to purchase and construct
9. Safer than TWLTL in 4 lane sections	9. Restricts direct access to adjoining property
10. Safer than TWLTL in 6 lane sections but depends on number of signals/mile, driveways/mile, ADT and approaches/mile	10. Installation costs are higher
11. Encourages access roads and parallel street development	11. Can create on over concentration of turns at median openings
12. Reduces accidents in mid-block areas	12. Indirect routing may be required for some vehicles
13. Reduces total driveway maneuvers on the major roadway	13. When accidentally stuck, curb may cause driver to lose control of the vehicle
14. Low maintenance cost of raised medians, depending on final design	14. A median width of 25 ft (7.6 m) is needed to accommodate U-turns
15. Studies have shown that delay per left turning vehicle does not increase, up to the studied volume of 3700 vph	
16. Curbs discourage arbitrary and deliberate crossings of the median	
17. Reduces number of possible median conflict points	
18.Provides separation between opposing traffic flows	

Table 3: Advantages and Disadvantages of Raised Medians Continued

Advantages	Disadvantages
19. Provides a median refuge area for pedestrians	
20. With raised grass medians, an open space is provided for aesthetics	

Bowman & Vecellio

Two-way left-turn lanes are a type of flush or traversable median, which is a median treatment type that is delineated but does not physically restrict traffic movements. Delineation comes from marking the pavement with appropriate stripping. Common types of flush medians are narrow divider strips, alternating left turn lanes and two-way left-turn lanes, which are collectively referred to as painted medians (Bowman & Vecellio a 180). Two-way left-turn lanes (see Figure 9) are intended to remove left turning vehicles from the main traffic throughways and to provide a storage area until a large enough gap in traffic is available to complete the turning movement.

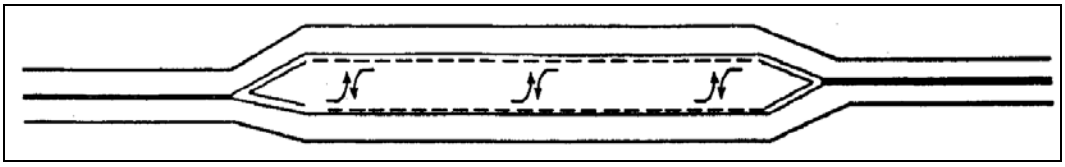


Figure 9: TWLTL

Garber & Hoel 164

A compiled list of advantages and disadvantages that come from installing two-way left-turn lanes can be seen in Table 4. Two-way left-turn lanes help to improve safety by removing the turning vehicles from the through-traffic lanes, but at the same time maximizing access for the turning vehicles. This is a beneficial solution because emergency vehicles do not run into access problems and the two-way left-turn lanes eliminates island fixed objects, which occur with raised medians. Problems can occur, however, with conflicting turning movements, visibility problems and safety for pedestrians. Visibility problems range from problems seeing he turning vehicles to

problems, especially at night, in determining the location of the two-way left-turn lane, while pedestrians loose their island refuge and have a further lane of traffic to cross.

Table 4: Advantages and Disadvantages of TWLTL

Advantages	Disadvantages
1. Left turning vehicles are removed from through traffic while maximum left turning access to side streets and driveways is still provided	1. There are conflicting vehicle maneuvers at driveways
2. Delay to left turning vehicles and others is often reduced	2. Poor operation of roadway if stopping sight distance is less than AASHTO minimum design
3. Operational flexibility for emergency vehicles and others is enhanced	3. No pedestrian refuge areas for pedestrians free from moving vehicles
4. When less than 60 commercial driveways per mile (37 driveways per km) are permitted to be constructed two-way left turn lanes appear to be safer	4. Operate poorly under high volume of through traffic
5. Roads with two-way left turn lanes are operationally safer than roadways with no separate left turn lanes in the median	5. Should not be used when access is required on only one side of the street
6. Detours can be easily implemented when required by maintenance in adjacent lanes	6. Visibility problem of painted median especially with snow and rain or when pavement markers outlive their design life
7. Provides spatial separation between opposing traffic flows	7. A safety problem when they are used as a passing lane
8. Eliminates the median island fixed object	8. High maintenance cost of keeping the pavement striped and raised pavement markers in proper operating condition
9. Provides temporary refuge for disabled vehicles	9. Must continually instruct the public on proper use and operation
10. Can be used as a reversible lane during peak hours	10. Delays to left turning vehicles increase dramatically when two way through volume reaches 2800 vpd
11. Permits direct access to adjoining properties	11. Limits operating speed to a maximum rate 45 mph (73 km/hr)
	12. Does not guarantee unidirectional use at high volume intersections
	13. Are not aesthetically pleasing for some people
	14. Allows numerous potential traffic conflict points

Bowman & Vecellio

Another type of flush median that has attempted to eliminate some of these problems is the alternating left turn lane which provides left turn opportunities for one direction at a time with both directions have turning capabilities over limited sections of the roadway (Bowman & Vecellio a 181). Alternating lanes have similar properties to two-way left-turn lanes, but eliminate possible conflicts by turning vehicles at the price of eliminating some access. This type of median works well in small urban areas especially where only one side of the road is developed otherwise the access restrictions can create more problems.

2.4.3 Median Width

Median width is defined as the width separating the traveled ways and includes the median width as well as the inside shoulder width. This is an important distinction, especially with traversable medians, because shoulder width provides some of the same services as a median, recovery room specifically, and may sometimes be difficult to distinguish especially for unpaved shoulders next to grass medians. It has been suggested that median widths should be at least 60 feet wide on rural highways and as narrow as 10 feet on urban highways if a barrier is used, but these are just heuristics and few studies have provided quantitative measures on the effect of median width on frequency and severity of accidents (Knuiman et al 70). Little guidance is given for median widths even by AASHTO. AASHTO's guidelines give a general range of median widths ranging from four to eighty feet or more, with no apparent upper limit. In urban arterial situations, a minimum width of four feet is used under the assumption that "a median 4 ft wide is better than none" (Green Book 478). When left turn lanes are desired, the median should be at least eighteen feet wide allowing room for the lane and a separator, though

in restricted locations a twelve foot median may be used (Green Book 478). Overall, the median must be wide enough to give the motorist the perception of safety for whatever movements are being completed, turning, crossing or straight movements (Knuiman et al 79). While minimal guidelines are given by AASHTO on the widths of medians, there is no agreed upon way to quantify what widths should be used to increase or even to ensure safety of either vehicles or pedestrians. The following sections go into further detail about the effects medians have on safety.

2.4.4 Effects of Medians on safety

Medians have long been recognized as an effective method of increasing vehicle safety and capacity on urban arterials. But, a summary of quantitative results for flush medians on highways has only shown that wider medians have lower accident rates. There is not a fixed amount of safety gained per increase in width. This unknown quantity of safety is reflected in the limited amount of guidelines for median widths. Since the safety benefit of medians is unknown, the best width to maximize safety is equally unknown.

Knuiman et al looked at the effect of median width on frequency and severity of accidents on homogenous highway sections with a traversable median (Knuiman et al 70). A homogenous section in this case means that the geometric and cross-section variables (lane width, pavement type, shoulder width, shoulder type, number of lanes) are constant. The aims of Knuiman et al's modeling process were to obtain standard errors and confidence intervals for estimated accident rates and to determine whether the observed reduction in crude accident rates for wider medians persisted after adjusting for other roadside characteristics.

Using a log-linear regression model, Knuiman *et al* included variables such as functional classification, posted speed limit, access control (none, full, partial) curvature, average daily traffic and section length in their models. Many of the variables considered were correlated with median width, which made the fitting of the interactions between median width and other variables difficult. The estimated effects of median width obtained from the fitted models may, therefore, be conservative due to the inclusion of variables correlated with the width (Knuiman *et al* 73). Knuiman *et al* found that there is little reduction in accident rates for medians up to twenty-five feet and decline in rates is most apparent for median widths beyond twenty to thirty feet with the decreasing trend leveling off somewhere between sixty to eighty feet (Knuiman *et al* 76).

While not giving exact numbers, Knuiman *et al* did manage to give a better range of median widths to use than do earlier assumptions. They found that the decrease in accident rates tapers off after sixty to eighty feet, showing that building medians larger than eighty feet will not be cost effective in reducing accidents. A few more accidents may be prevented by larger medians, but not to any noticeable degree. Also shown was that the minimum width should really be approximately twenty-five to thirty feet which is where observable decreases in accident rates can be seen. The study concluded "accident rates decrease with increasing median width, even when other confounding variables are controlled for" (Knuiman et al 77). What was not found with the decreasing accident rates was a concurrent decrease in the severity of accidents. Median width affected as many of the severe crashes as the less severe ones, and primarily lowered multi-vehicle crashes but had no effect on single vehicle run-off-the-road crashes (Knuiman *et al* 79).

So while a more effective median width can be chosen, there are still many other confounding variables that affect safety of vehicles.

2.4.5 Comparison of Median treatment safety

Urban locations primarily use raised curb medians or two-way left-turn lanes. Studies looking at the relative safety between the two have discovered conflicting results. Some researchers have found no difference in the accident rates of the two treatment types, some found two-way left-turn lanes to have higher rates and still other researchers found raised medians to have the higher accident rates. When examined individually, the installation of a median whether raised or painted typically resulted in a lowering of accident rates and improvement of safety (Bowman & Vecellio a 182). Both median types showed typical reduction in total number of vehicle accidents in the 25 to 35-percentage range (Bowman & Vecellio a 186) and both resulted in a reduction in accident severity (Bowman & Vecellio a 187).

Brown and Tarko have developed prediction models for total number of crashes, number of property-damage only crashes and number of fatal and injury crashes with the prime interest of seeing if controlling access does improve safety. Brown and Tarko chose to make crash frequencies proportional to traffic volume, despite this not being an exact fit, the data showed this to have an insignificant effect on the models (Brown and Tarko 71). Brown and Tarko found more access points to results in a higher crash rate, the presence of an outside shoulder reduces crashes, the presence of traffic signals to increase rates, and medians with no opening to decrease accident rates (Brown and Tarko 72). Brown and Tarko concluded that in general access control has a beneficial effect on safety.

Bonneson and McCoy also developed models for predicting the safety of urban arterial streets focusing on use of specific median types (Bonneson & McCoy 33). They created three median specific models for raised medians, two-way left-turn lanes and undivided cross sections. For arterial streets the independent variables included in the accident prediction models include traffic demand, road length, driveway density, median type, number of lanes, and adjacent land use. Bonneson and McCoy found several trends from their modeling, including raised-curb median treatments having the lowest accident rate, two-way left-turn lanes slightly higher and undivided segments the highest rates (Bonneson & McCoy 35). Land use was also show to be important with business and office land use locations having consistently higher accident rates than residential and industrial areas. Despite being unable to yet agree on the safer median treatment between raised medians and two-way left-turn lanes, most researchers agree that either treatment will reduce accident rates compared with an undivided cross section, so that proper use of access control methods does result in safer roads.

2.5 Intersection Accidents

A major theory behind intersection accidents is that the number of accidents at an intersection is proportional to the sum of flows that enter the intersection (Hauer et al 49). This is sometimes referred to as the traffic intensity or the total number of vehicles entering an intersection per year and is often one of the most important factors in predicting injury accidents (Lau & May 63). Several problems exist with this type of thinking including that problems occur when looking at specific accident types, it is an overly simplistic version of events and is very dependent on correlation. Another theory is that accidents relate to the products of conflicting flows (Hauer et al 49). Hauer et al

found that accidents tend to be related to the product of flows with each flow raised to a power of less than 1 (Hauer et al 49). This cross street traffic, traffic from the minor road is an indication of how many possible conflicts could exist at the intersection. Accidents between vehicles proceeding in the same direction have to be estimated separately from accidents between vehicles (turning, left) in multiple approaches. Customary categorization of accidents by initial impact (rear end, turning movement, sideswipe, etc) is not very informative (Hauer et al 56). It cannot be assumed that classification of an accident as an angle accident implies that vehicles were traveling at right angles to each other. To be specific the categories need to clearly show the relationship between the vehicles involved in the accident. This becomes an important issue when categorizing accidents.

Important factors when developing models that deal exclusively with intersection accidents include traffic intensity, percent of cross street traffic, intersection type, signal type, number of lanes on the main and side streets, and left turning arrangements (Lau & May 65). At the time of Lau and May's work the current intersection models in California only used traffic intensity and intersection type to predict accidents (Lau & May 65). Other factors such as turning movement counts and conflict analysis may help in creating prediction models, but these types of data are more time intensive and difficult to collect and are not readily available for use in developing prediction models.

Hauer et al find that intersection accidents are not proportional to the sum of entering volumes. Accident rates should not be calculated on the "basis of the sum of entering volumes to compare the safety of two different intersections" (Hauer et al 57).

Another issue with junction models is that they are usually limited to major intersections with roads of collector or arterial classification. There are many minor junctions that exist where knowledge of the traffic flows on the minor roads are unknown and the collection of such data would prohibit the usefulness of such a model. Separate models of minor junctions are not possible without data collected just for that purpose (Mountain 705). The separation and delineation made between link sections and minor and major intersections make the combination of the three important and the effect of one on the other significant.

2.6 Modeling Types and Issues Related to Modeling

Mathematical modeling is a technique to create a quantifiable method to predict the occurrence of certain events. An accident prediction model is an equation that expresses accident frequency as a function of traffic flow and other road characteristics. Many models have been created to calibrate relationships between shoulder width, lane width and shoulder type on two-lane rural highways and several studies have looked at the effects of median width and type. Hadi et al looked at roads in Florida separated by location, access type and number of lanes (Hadi et al 170). Many issues have been brought to light due to issues relating to both modeling and the nature of traffic accidents. Several of the more important issues comprise the following sections.

2.6.1 Generalized Linear Modeling

Generalized linear modeling (GLM) is the most straight forward method used to develop mathematical models. A GLM is usually made up of three components: a random component, a systematic component, and a link function that connects the other two to produce a linear predictor (Lord & Persaud, 103). In generalized linear modeling

an important assumption is that random error occurs only in the dependent variable and that the explanatory variables are known without error (Maher & Summersgill 293). This is an important assumption to keep in mind since not all the necessary variables contributing to car accidents are known without error. For geometric and control variables such as number of lanes and presence of a median, the variables are known without error, but not so for all the traffic characteristic variables such as volume and percentage of heavy vehicles. Ideally traffic flow should be the average annual daily traffic (AADT) over the whole time period under consideration, but data often comes from a "snapshot" of a single day from the study period and some time not even that (Maher & Summersgill 293). Since volume studies are very time consuming, they are not performed on a regular basis and are adjusted based on state factors.

The GLM is flexible in the choice of probability distribution for the random component, making this kind of model effective for traffic safety where number of accidents and other variables follow a Poisson or negative binomial distribution and further variables follow a normal distribution. In the past, models have been developed that follow all of these distributions depending on what exactly is being studied.

2.6.2 Linear Modeling

There have been many studies which have the goal of establishing relationships between traffic accidents and road geometry, as well as determining the effect of road and intersection design on the frequency of accidents (Maher & Summersgill 281). The majority of studies have historically used conventional analysis, linear regression, which assumes that the dependent variable is continuous and normally distributed with a

constant error variance. Most often the regression coefficients are found by the traditional method of least squares (ordinary least squares).

This method results in point estimators, β, that have minimum variance. Analysis of variance (ANOVA) approach is typically used and separates the sum of squares and degrees of freedom associated with the dependent variable. The mean squared error, MSE, can be found on the ANOVA table and is an unbiased estimator of variance (σ^2). The variance of the error terms (ε_i) is also an indication of the variance of the probability distributions of the dependent variable.

The variance is used to calculate the coefficient of determination, R^2, which represents the proportion of variability explained by the regression function. The coefficient of determination is the most common method for determining the quality of the model in question and ranges between zero and one. An R^2 value near zero indicates that there is not a strong linear relationship between the dependent and independent variables. A value of R^2 near one indicates a strong linear fit where the model explains the variability in the data. The use of R^2 should be used with caution to ensure the correct interpretation and be accompanied by the examination of scatter plots (Garber and Ehrhart 78). R^2 is only a useful parameter when looking at linear regression models; it does not apply to anything other than a normal distribution. A low value may not just mean that the model is a bad fit for the data, but that there is not a linear relationship between the examined variables and another functional form (logarithmic, exponential) or distribution (Poisson, negative binomial) should be used.

Some traffic engineers believe that the coefficients of accident prediction models can not be properly estimated by ordinary least-squares or weighted least-squares

regression methods due to the non-negative, discrete nature of accident counts and the fact that variance of the number of accidents increases, but not linearly, as traffic flow increases (Lord & Persaud, 103). In approximately the last ten to twenty years there has been a tacit agreement among modelers that conventional normal or lognormal regression models don't have the necessary statistical properties to describe vehicle accidents. A major problem with linear/multilinear modeling is that it may predict negative accidents, which is not a possibility in real life (A Miaou et al 12). A location with no accidents can occur, but not a location with negative ones. The relationships between accidents and related factors do not always reflect linear behavior causing multi-linear regression to be inappropriate for analyzing the causes of accidents (Saccomanno & Buyco 24). Instead, as modeling programs have become more accessible, sophisticated and user friendly, transportation professionals have begun to estimate model coefficients by using maximum-likelihood methods to calibrate generalized linear models. The use of other types of distributions has also become more popular. The favored choice of models appears to be the Poisson and negative binomial distributions. Another natural choice of function due to the nature of accidents is the exponential function, which has been widely used by statisticians and econometricians (Miaou, 8).

2.6.2.1 Model Fit

Once a model has been developed, it needs to be shown to work for the application for which it has been applied. The quality of the model must also be obtained.

The coefficient of determination (R^2) has traditionally been used over the past approximately thirty years as a criterion to determine how well the developed models fit

the observed data (Miaou 6). R^2 has been used to determine overall quality and usability of a model. "The R^2 statistic is a measure of the percentage of unconditional variance of the dependent variable explained by the available covariates" (Miaou, 13). For any given data set the R^2 value of the developed model has a minimum lower bound of zero and an maximum upper bound of one. So a model with a coefficient of determination of 0.85 would be considered good while a model with a coefficient of 0.36 would be considered as a poor candidate. An R^2 value of 0.7 or less is often considered the breaking point and models with lower values are typically not recommended for use (Miaou 6). The R^2 is often used to indicate the model fit to the data but also as a way to compare models. When comparing two or more models that predict the same thing, whether vehicle speed or accident rates, often models can look very different from each other with different variables and coefficients. Using the R^2 values to compare the relative quality of models from different studies helps by standardizing the model quality and simplifying the comparison process. The decision to try and add variables to the model can also be formed from the R^2 value. Using a constant upper bound of one, many researchers look at $\left(1-R^2\right)$ as a measure of potential improvement that can be gained by including additional covariates (Miaou 6). Increasing the number of variables is not, however, always the best move.

The adjusted coefficient of determination, or R_a^2, is a modified measure that allows the total number of degrees of freedom (DOF) in the model to be reflected in R^2. R_a^2 is used in model's developing phase to decide which explanatory variables should be included. The model with the largest R_a^2 value is typically considered the best. The reason for using the adjusted coefficient is that it includes information about the degrees

of freedom in the model. Including more variables in a model may slightly improve the R^2 value, but if the increase in the coefficient is not large enough, the loss of degrees of freedom can counteract the minimal benefits. This adjusts for the fact that more variables is not always better. Both the coefficient of determination and the adjusted coefficients are most commonly used for models with normal distributions and can loose some or all of their true meaning if applied to non-normal distributions (Bonneson & McCoy 31). Miaou *et al*. found that the R^2 statistic is only meaningful in measuring the goodness-of-fit for "normal linear regression models with additive mean functions" (Miaou 13). Accident prediction models are non-normal and typically non linear. Miaou *et al.* showed by example that R^2 is not always an appropriate way to make decisions about quality and goodness-of-fit for accident models. Since the use of these coefficients is relatively simple (larger value equals better quality) the temptation to use coefficients of determination with non-normal distributions must be avoided.

Another major pitfall of coefficients of determination comes with the use of binary response models. The upper bound for a perfect model can be less than one, implying that a model with a low value of R^2 does not mean the fit is poor. Brűde and Larsson showed that the R^2 value of "Poisson regression models is dependent on the mean level of the dependent variable (i.e., the mean level of accident frequency)" (Miaou 6). It was shown that the higher mean accident levels would result in higher R^2 values regardless of the quality of the model. This is a reason why R^2 values of accident prediction models for urban areas have typically been reported higher than those for rural areas, based solely on the higher accident rates (Miaou 6). This also implies that R^2 values should not be the only method chosen for comparing goodness-of-fit of models

when they are from different studies especially when different locations, accident types, or time periods are involved (Miaou 7).

There are many statistical tests and criteria that are available for evaluating the quality of the goodness-of-fit of a model and several should be used in conjunction to determine the quality for accident prediction models. A good check of model fit is the statistical significance of the variable coefficients, which can be found by looking at the standard error and 95 percent confidence intervals for each coefficient (Bonneson & McCoy 30). Checking that the individual variables are significant and that with 95 percent confidence their coefficients won't become zero helps to ensure the quality of the model.

Other well-known statistics to measure the quality of the fit between the observed Y_i and the fitted values $\hat{\mu}_i$ are the scaled deviance (SD) and the Pearson χ^2 statistic.

$$SD = \sum_i 2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right]$$

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}$$

When there is perfect agreement these statistics are zero, otherwise they are positive. The scaled deviance is based on the log likelihood function and the estimation of parameter estimates are obtained through the maximum likelihood and is the more commonly used of the two statistics (Maher & Summersgill 283). This statistic follows the χ^2 distribution with n-p-1 degrees of freedom, where n is the number of observations, and p is the number of model variables. This statistic is asymptotic to the χ^2 distribution for large sample sizes and exact for normally distributed error structures (Bonneson &

McCoy 30). However, this statistic is not well defined in terms of minimum sample size and non-normal distributions (Bonneson & McCoy 30). This is a statistic that people tend to take at face value, but since it is not well defined for non-normal distributions, care should be taken to ensure that it is applied mainly to linear models, but if it is used for non-normal distribution models, that it is not the only qualification for goodness of the model.

Other model fit techniques include the Cumulative Residuals Method (CURE), which investigates the quality of fit by plotting the cumulative residuals for each independent variable. This is a graphic method that allows the fit of the function to the data to be observed (Lord & Persaud 106). An advantage of this and other graphical methods is that CURE is not dependent on the number of observations as other techniques are which allows models developed from any sample size to be assessed with this method (Lord & Persaud 106).

Akaike's information criterion, AIC, can be used for multivariate models to predict the fit of a model based on the expected log likelihood (Garber and Ehrhart 78). It is based on the Kullback-liebler information criterion, which measures the distance between the true model and the hypothesized model (Garber and Ehrhart 78). $ACI = -2\ln(L) + 2k$ where L is the Gaussian likelihood of the model and K is the number of free parameters in the model. In terms of sum of square of the errors $ACI = n\ln\left(\frac{SSE}{n-k}\right) + 2k$ where n is the number of model residuals, $SSE = \sum (y_i - \hat{y}_i)^2$ y_i is the observations $\hat{y}$ =model estimates. The first term measures badness of fit or bias and the second measures complexity of the model. The goal for

selecting the model is to minimize the criterion and select the best fit with the least complexity (Garber and Ehrhart 78).

The dispersion parameter, σ^2, can also be used to measure fit by assessing the amount of variation in the observed data. A dispersion parameter near one indicates that the assumed error structure is approximately equivalent to that found in the data (Bonneson & McCoy 31).

2.6.3 Bernoulli Random Variables

A Bernoulli random variable, named after the Swiss mathematician James Bernoulli, can take on only two values (e.g. 0/1, on/off, yes/no, present/not present, success/failure) with respective probabilities of 1-p and p (Ross 144).

$p(1) = p$

$p(0) = 1 - p$

$p(x) = 0$ if $x \neq 0$ or $x \neq 1$

A Bernoulli trial consists of selecting and testing one item from a finite set of items and seeing which value it has (Petruccelli et al 136). The probability of success in a Bernoulli trial is always nonnegative and at most unity.

An indicator variable is used to designate whether or not an event occurred or if a characteristic is present. If A is an event, then the indicator random variable I_A takes on the value of 1 if A occurs and the value of zero if A does not occur.

$I_A(z) = 1$, if $z \in A$

$I_A(z) = 0$, otherwise (Rice 34)

Indicator random variables are, therefore, a special case of Bernoulli random variables with only probabilities of zero or one. Both Bernoulli random variables and the

more specific indicator variables are commonly used in traffic models. For instance in a model that is predicting the 85th percentile speed of a vehicle an indicator variables could be used to show the presence of horizontal curves where a zero would mean a straight road and a value of one would men that one or more curves were present.

2.6.4 Binomial Distribution

There are n independent experiments or trials performed in a binomial distribution where each trial results in a "success" with the same probability p or a "failure" with the same probability 1-p. "The total number of successes, X, is a binomial random variable with parameters n and p" (Rice 34). K is the number of successes that occur throughout the entire experimental program. Each experiment is constructed from independent Bernoulli trials.

A classic example used in binomial distributions is the situation of tossing a coin multiple times. A coin is tossed 10 times (i.e., n, the number of trials, equals 10) and the total number of tails is recorded (i.e., k, the number of successes, equals the number of tails observed). The probability that X=k or $p(k)$ can be found by the following method:

$$P(X) = p(k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ where } k = 0,1,2,\ldots.$$

The distribution for tossing a coin 10 times is shown in Figure 10 as a binomial distribution. "There are $\binom{n}{k}$ ways to assign k successes to n trials" (Rice 34). The combinatorial notation $\binom{n}{k}$ can also be written in the following way: $\frac{n!}{k!(n-k)!}$ (Petruccelli et al 167). This allows the entire probability distribution to be

shown by: $p(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$. The mean of the binomial distribution is $\mu = np$, the variance is $\sigma^2 = np(1-p)$, and the standard deviation is $\sigma = \sqrt{np(1-p)}$ (Petruccelli et al 1168).

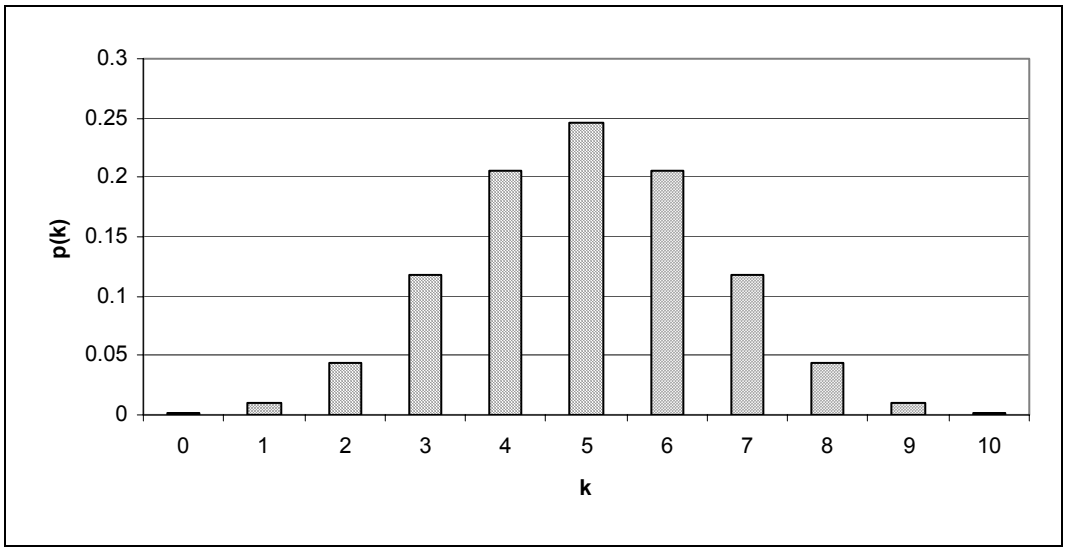


Figure 10: Binomial Frequency Function n=10. p=0.5

The binomial distribution can consist of Bernoulli trials and other types of situations. In the Bernoulli trial, there are only two options, but binomial distributions can be used when there are more than two optional answers. For instance, a die typically has six sides. This can be used in binomial distributions in many different ways. For example, a success could be considered rolling an even number (2, 4, or 6). Therefore there are multiple chances for a success to happen, but there is still only the two options of "success"(even number) and "failure" (odd number). There are three key assumptions in binomial distributions: (1) each trial is independent, (2) each trial results in only one of two possible outcomes, and (3) the probability of a success in each trial is constant (Montgomery and Runger 74). The binomial distribution is used extensively in statistical and probability applications. In spite of the need for the individual trials to be

independent, certain continuous problems can be modeled using this distribution. For example, time and space problems, which are generally continuous, may be modeled by discretizing time into finite intervals with only two possibilities within each interval. Then what happens in each time (or space) interval becomes a trial (Ang & Tang 109).

2.6.5 Log-Linear Models

Log-linear models assume that the effect of variables on the accident rate is multiplicative rather than additive as in linear models (Knuiman et al 72). Estimated rates from log-linear models cannot be negative, which fit accident rates in that you can have zero accidents or a positive number of accidents, but negative accidents do not exist. "Zegeer *et al* considered both additive and multiplicative (log-linear) models and concluded that the multiplicative models provided a better fit to the data" (Knuiman et al 72). Knuiman et al assumed a negative-binomial variance function for the accident count per section so $Var(Y) = E(Y) + k * [E(Y)]^2$ where k is the same for all section and Var(Y) and E(Y) are the variance and expected value respectively.

This has the form of $\log(\lambda) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$ where $\lambda = R = E(R) = \left[\frac{E(Y)}{ADT * 365 * T * L} \right] * 10^8$ and X_i is the indicator variable for categorical roadway characteristics or actual values for quantitative roadway characteristics. Loglinear models are where the predictive variable is really the log of the variable. Advantages of using loglinear models include having continuous and categorical variables. "A loglinear approach allows the statistical significance of partial and marginal association to be tested for a given combination of categorical factors" (Saccomanno & Buyco 25). Multiplicative models also assume that the effects of

individual variables work together and that they do not act independently from one another, so that combinations of characteristics rather than individual ones better explain events.

2.6.6 Poisson Modeling

The majority of studies, historically, have used conventional regression analysis, which assumes that the dependent variable is continuous and normally distributed with a constant variance. Early modeling work used multiple linear regression modeling with assumed normally distributed errors, but as work progressed the nature of traffic accidents showed that it is better to assume a Poisson distribution for the frequency of accidents. The assumption of a normal distribution is not correct when applied to crashes, which are discrete, non-negative variables whose variance depends on its mean (Hadi et al 169). Beginning in the early 1990's, researchers started to try to over come some of the problems associated with linear regression. Poisson regression models, widely used in modeling accident and mortality data in epidemiology, began to be applied to traffic accidents (A Miaou et al 12). Poisson regression and negative binomial regression have both been used to combat the incorrect assumptions of normality for accident counts. The Poisson model "although representing a significant advance in accurate and reliable modeling capability, is not without its weaknesses and technical difficulties which must be overcome if it is to be used effectively" (Maher & Summersgill 282).

Poisson regression is a nonlinear approach to modeling where the response variable is a count, or a discreet event, with large outcomes being rare events (Neter et al 609). The Poisson distribution model was named for the French mathematician S. D.

Poisson who lived from 1781 to 1840 (Petruccelli et al 147). He introduced the concept in a book regarding the application of probability theory to lawsuits and criminal trials (Ross 154). The book was designed as a contribution to judicial practices and contains "so much preliminary material of a purely mathematical and probabilistic nature that it must be regarded as a textbook on probability with illustrations from the courts of law" (Haight 113). The following are examples of random variables that usually obey the Poisson probability laws:

- The number of people in a community living to 90 years of age,
- The number of customers entering a post office on a given day, or
- The number of α-particles discharged from radioactive material over a given time.

Count data has been analyzed by ordinary linear regression and the advantage of using Poisson regression comes from the fact that the distribution is tailored to the discrete and often highly skewed distribution of the dependent variables.

In a Poisson distribution, there are two main sources of variability; the differences in mean accident frequency among similar segments and randomness in accident frequency. In spite of similarity between roadway segments, each has its own unique mean accident frequency (m), where the distribution of m within a group of similar segments can be described by a probability density function with mean E(m) and variance Var(M) (Bonneson & McCoy 29). This distribution has been adequately described by the gamma density function (Bonneson & McCoy 29). If accident occurrence at a segment is Poisson distributed then the distribution of accidents around the E(m) of a group of segments can be described by the negative binomial distribution.

Poisson regression models discreet events ($Y_i = 0,1,2,...$) where a large number of occurrences is rare. The dependent variable follows a Poisson distribution where

$$f(Y) = \frac{\mu^Y \exp(-\mu)}{Y!} \quad Y_i = 0,1,2,...$$

$f(Y)$ is the probability that the outcome is Y

Y!=Y(Y-1)(Y-2)…3*2*1

While Y can take on only nonnegative, integer values, μ can be any positive number. As can be see in Figure 11, where μ =1.75, the probabilities for the Poisson distribution are graphed. The probability mass function is defined for an infinite set of possible values of Y, though there will be a finite upper bound on the values of Y that are actually observed (Petruccelli et al 147). Despite there being an upper bound on the observed values of Y, the Poisson distribution allows for modeling of random phenomena without having to know the maximum value that the random variable can take (Petruccelli et al 148).

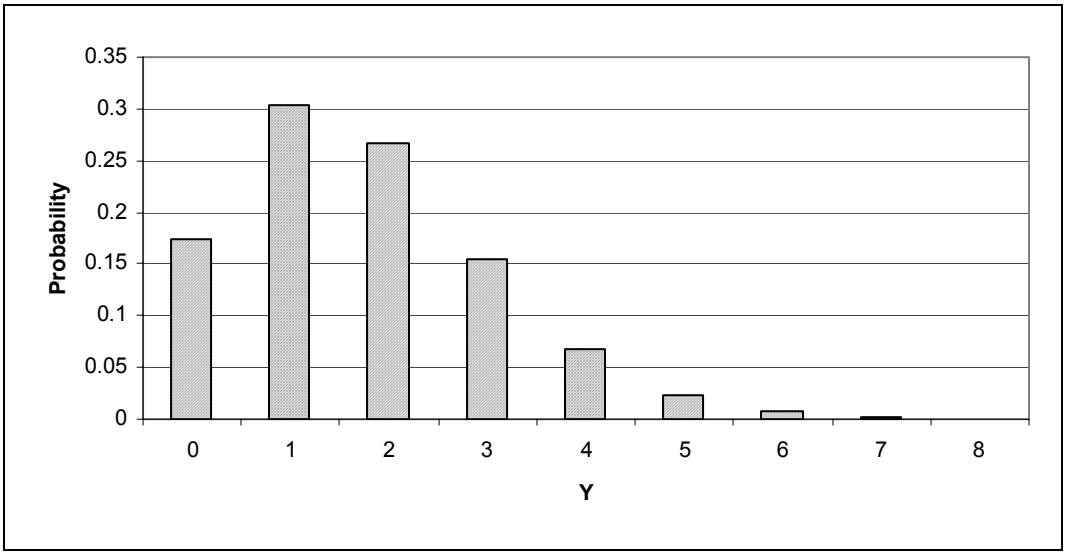


Figure 11: Probability Mass Function of a Poisson distribution with μ =1.75

As μ gets larger, the mode moves away from zero causing the distribution to resemble more and more that of a normal distribution (Allison 218). A unique feature of the Poisson distribution is that the mean is equal to the variance.

$$E\{Y\} = \mu$$

$$\sigma^2\{Y\} = \mu$$

The parameter μ depends on the explanatory variables and it is standard to let μ be a log-linear function of the X variables $\log \mu_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}$.

In the above model form it is assumed that the counts were collected over a certain period of time. The Poisson distribution can also be applied when the dependent variable is collected over different lengths of time or space for different individuals. In ordinary regression analysis, the individual event count could be simply divided by the length or time interval. That will not work in "Poisson regression because a division by time implies that the resulting model no longer has a Poisson distribution" (Allison 228) and the observed number of accidents at a site is assumed to be Poisson distributed about a mean of μ_i, which is assumed to be proportional to the length of the observation period T_i (Maher & Summersgill 282). When this situation arises, the probability distribution can be adapted by t the number of units of time or space to which the Y value corresponds.

$$f(Y) = \frac{(t\mu)^Y \exp(-t\mu)}{Y!} \quad Y_i = 0,1,2,...$$

The Poisson regression model can be stated as $Y_i = E\{Y\} + \varepsilon_i \quad i = 1,2,...n$. The mean response for the i[th] case, μ, is assumed to be a function of the set of predictor variables $X_1,...,X_{p-1}$. $\mu(X_i, \beta)$ denotes the function that relates the mean response μ_i to

X_i, the values of the predictor variables for case i and β the values of the regression coefficients (Neter et al 610). There are several commonly used functions for Poisson regression including:

$$\mu_i = \mu(X_i, \beta) = X_i'\beta$$

$$\mu_i = \mu(X_i, \beta) = \exp(X_i'\beta)$$

$$\mu_i = \mu(X_i, \beta) = \log_e(X_i'\beta)$$

In all of the cases the mean response μ_i is a nonnegative value. The distribution of the error terms ε_i is a function of the distribution of the response variable which is Poisson distributed. The Poisson model can be stated as: Y_i are independent Poisson random variables with expected values μ_i where $\mu_i = \mu(X_i, \beta)$.

Poisson distributions model the probability of discrete events by $P(Y) = \frac{e^{-\mu}\mu^Y}{Y!}$. The Poisson distribution can be derived as the limit of a binomial distribution as the number of trials, n, approaches infinity and the probability of success on each trial, p, approaches zero in such a way that $np = \lambda$ (Rice 39). Where Y is the number of events in a chosen period and μ is the mean number of events in the chosen period. The Poisson regression model assumes that the mean number of events is a function of regressor variables. To estimate crash frequencies, they are assumed to be Poisson distributed by $P(Y = Y_i) = \frac{e^{-\mu(X_i\beta)}[\mu_i(X_i, \beta)]^{Y_i}}{Y_i!}$. Yi equals the number of crashes at road section 'i' for a chosen time period. β is the vector of parameters to be estimated $\mu_i(X_i, \beta)$is the mean number of crashes on section 'i' which is a function of a set of regressor variables X. Xi

is the vector of regressor variables for segment i. The function $\mu_i(X_i, \beta)$, which relates the distribution mean to regressor variables, is the link function $\mu_i(X_i, \beta) = e^{X_i \beta}$. The regressor or explanatory variables are items such as traffic glows and geometric characteristics. The vector **X**, containing the explanatory variables has 1 as its first term so that the first term in vector **β** is the interceptor or constant. When sites are lengths of road rather than junctions it is usually assumed that μ_i is also proportional to the length L_i as well as the time period, so that λ_i is in terms of accidents per kilometer per year.

One of the main problems is the phenomenon of overdispersion where the assumption of a pure Poisson error structure can be seen to be inadequate. The negative binomial model is often chosen to overcome this issue as an extension to the Poisson model. Often, however, variances greater than the mean are observed due in part to not including all the relevant variables in the model (Knuiman et al 72). When variances greater than the mean are observed, it is called overdispersion.

2.6.6.1 Overdispersion

It is important for models to try and explain the variation in accidents between sites. A model should have terms for the relevant flows, then explanatory variables for physical characteristics and control variables. But final models still are often in the technical sense inadequate, with the explanatory variables not providing complete explanation of the variability between sites. The major reasons for this are that there are (a) unobserved explanatory variables, (b) there are errors in the explanatory variables, and (c) the model was mis-specified (Maher & Summersgill 288). Overdispersion is the term used to describe this problem of not fully explaining the variability in the model and

is a problem often associated with Poisson regression. This occurs when variances greater than the mean are observed which can be due in part to not including all the relevant variables in the model (Knuiman et at 72).

Overdispersion occurs because there is no random disturbance term in the equation $\log \mu_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}$ that would allow for omitted explanatory variables (Allison 223). This is because a disturbance term would produce larger variances in the dependent variable. Overdispersion does not produce a bias in the regression coefficients, but it will cause underestimation of standard errors and overestimation of chi-square test statistics, which can cause a model to be regarded more highly than it should. Also, implied by overdispersion is that the "conventional maximum likelihood estimates are not efficient, meaning that other methods can produce coefficients with less sampling variation" (Allison 223). If the lack of efficiency is ignored, it is relatively simple to correct the standard errors and test statistics for overdispersion. "Take the ratio of the goodness-of-fit chi-square to its degrees of freedom, and call the result C. Divide the chi-square statistic by C. Multiply the standard error of each coefficient by the square root of C." (Allison 223)

The deviance and the Pearson chi-square are both goodness-of-fit chi-square values and the theory of quasi-likelihood estimation proposes the use of the Pearson chi-square statistic (Allison 223). Adjustment for overdispersion can greatly affect the significance of the regression coefficients. Comeeron and Trivedi have suggested a test involving simple least-squares regression to test the significance of the overdispersion coefficient (Hadi et al 171).

Statistical Analysis System (SAS) can control for overdispersion by using either of the above methods: the deviance or the Pearson chi-square value. To do this automatically, SAS has the options of PSCALE (for Pearson) and DSCALE (for deviance) as options in the MODEL statement. This produces the corrected standard deviations without the uncorrected ones being present in the output.

There are several ways in which a basic Poisson model can be modified to correct for overdispersion. One that has been suggested is the quasi-Poisson (QP) model where the variance of Y_i is given by $k^2\mu$. The parameter k^2 can be estimated by any of the statistics $\frac{SD}{(N-p)}$, $\frac{X^2}{(N-p)}$, and $\frac{SD}{E(SD)}$ (Maher & Summersgill 288). The parameters estimated are identical to those of a pure Poisson model with the difference occurring in the magnitude of the standard errors, which are inflated by a factor of k. Due to this, some model variables would no longer be found to be significant. In terms of significance thee types of models perform badly when the percent of fitted values less than 0.5 gets over 60 percent (Maher & Summersgill 288).

2.6.6.2 Maximum Likelihood

The maximum likelihood method is commonly used to estimate regression coefficients.

$$L(\beta)=\prod_{i=1}^{n} f_i(Y_i)=\prod_{i=1}^{n}\frac{[\mu(X_i,\beta)]^{Y_i}\exp[-\mu(X_i,\beta)]}{Y!}=\frac{\left\{\prod_{i=1}^{n}[\mu(X_i,\beta)]^{Y_i}\exp\left[-\sum_{i=1}^{n}\mu(X_i,\beta)\right]\right\}}{\prod_{i=1}^{n}Y_i!}$$

A functional form is chosen and the maximum likelihood estimates of the regression coefficients are produced. Numerical search procedures, iteratively reweighed least

squares and statistical software can be used to obtain the maximum likelihood estimates (Neter et al 610)

2.6.6.3 Test of Fit

A formal test of the fit of the response function is based on the model deviance $DEV(X_0, X_1, ...X_{p-1})$. If n is large then the deviance follows an approximate chi-square distribution with n-p degrees of freedom (Neter et al 595). If $DEV(X_0, X_1, ...X_{p-1}) \leq \chi^2(1-\alpha; n-p)$ then H_o is concluded

If $DEV(X_0, X_1, ...X_{p-1}) > \chi^2(1-\alpha; n-p)$ then H_a is concluded

Where H_o is the model is a satisfactory fit for the type of model chosen.

2.6.6.4 Deviance Residuals

A large ratio of deviance to degrees of freedom suggests that a problem with the model exists. A large deviance relative to the degrees of freedom exemplifies the problem of overdispersion (Allision 222).

Residual analysis helps to show if models follow the model assumptions. This type of analysis is most useful when using a normal distribution and must be modified when being applied to different distributions. Instead of just residual analysis, the deviance residual is more useful when dealing with Poisson distributions. The deviance residual for case i, dev_i is defined as

$$dev_i = \pm\left[2Y_i \log_e\left(\frac{Y_i}{\hat{\mu}_i}\right) - 2(Y_i - \hat{\mu}_i)\right]^{1/2}$$

and the overall deviance is defined as

$$DEV\left(X_0, X_1, \ldots X_{p-1}\right) = 2\left[\sum_{i=1}^{n} Y_i \log_e\left(\frac{Y_i}{\hat{\mu}_i}\right) - \sum_{i=1}^{n}\left(Y_i - \hat{\mu}_i\right)\right]$$

where $\hat{\mu}_i$ is the fitted value for the i[th] case (Neter et al 611). The sign of the deviance residual is selected according to whether $Y_i - \hat{\mu}_i$ is positive or negative.

A graphic display of the deviance residuals that helps to identify outlying residuals is the index plot. Index plots and half-normal probability plots are useful in identifying outliers and checking model fit (Neter et al 611).

Inferences for a Poisson regression model can be carried out. The mean response for predictor variables X_h can be estimated by substituting X_h into $\hat{\mu} = \mu\left(X, b\right)$. Estimation of probabilities of certain outcomes for given predictor variables can also be obtained by substituting $\hat{\mu}_h$ into $f\left(Y\right) = \frac{\mu^Y \exp\left(-\mu\right)}{Y!}$. Interval estimation of individual regression coefficients can be carried out by using the large-sample estimated standard deviations furnished by regression programs (Neter et al 612).

2.6.7 Geometric Distribution

The geometric distribution is constructed from independent Bernoulli trials, but instead of a fixed number of trials, trials are conducted until a success is obtained. A success occurs with probability p, and X is defined as the total number of trials up to and including the first success. "In order that X=k there must be k-1 failures followed by a success" (Rice 36). $p\left(k\right) = P\left(X = k\right) = \left(1 - p\right)^{k-1} p$, $k = 1,2,3,\ldots$

Figure 12 shows an example of a geometric probability mass function. The distribution acquires its name from the fact that the probabilities decrease in a geometric progression (Montgomery and Runger 78).

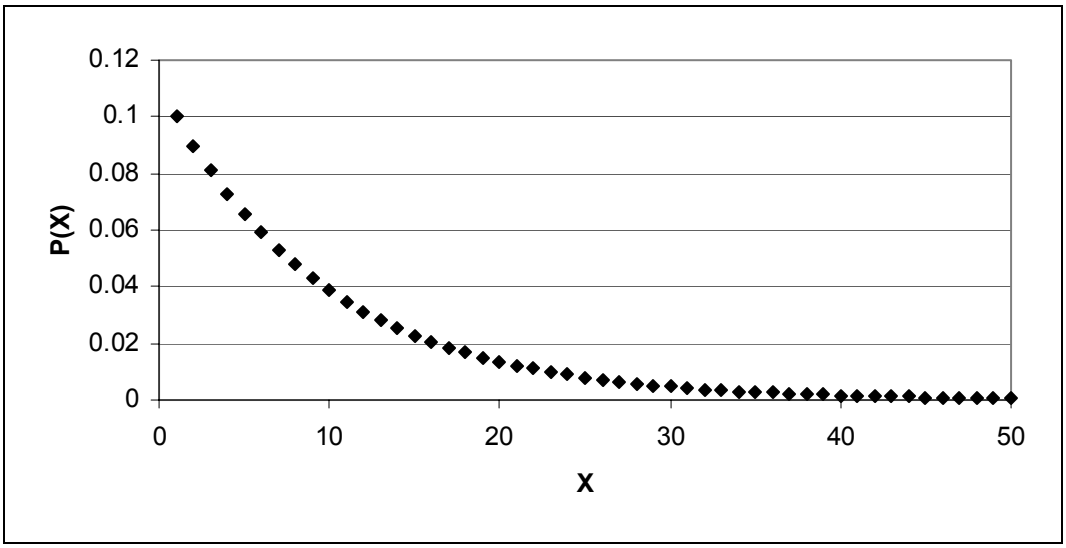


Figure 12: Probability Mass Function of a Geometric Random Variable with p=0.1

2.6.8 Negative Binomial Regression

The negative binomial distribution is a natural extension from the Poisson distribution, which accounts for the excess variability that is sometimes observed in accident prediction model. This distribution has gained favor for use in transportation studies, being used to help overcome the problems that occur with Poisson modeling, specifically the variance is allowed to be different from the mean in negative binomial regression (Hadi et al 171). Both models are related to the Bernoulli sequence (Ang & Tang). The negative binomial model can be considered a more generalized distribution for count data than the Poisson model due to a disturbance term that helps to overcome the overdispersion problems that Poisson modeling is prone to (Allison 226). The beta coefficients in the model were estimated by the method of quasi-likelihood (Knuiman et al 72). Maximum likelihood estimation is also an efficient way to estimate parameters in negative binomial regression. $\log \lambda_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik} + \sigma\varepsilon_i$ The dependent variable Y is assumed to follow a Poisson distribution with the expected value λ_i conditional on ε_i (Allison 226). The expected value of ε_i is assumed to follow a

standard gamma distribution. It then follows that the unconditional distribution of Y_i follows a negative binomial distribution (Allison 226).

The negative binomial distribution is based on a negative binomial random variable where the number of successes is fixed and the number of trials is random. This is different from the binomial distribution, where the number of trials is fixed (Devore 111). There are several conditions that need to be satisfied for an experiment with a negative binomial random variable and distribution. These include the following:

1. The experiment consists of independent trials,
2. Each trial can result in a success or a failure,
3. The probability of success is constant from trial to trial, and
4. The experiment continues until a total of r successes have been observed, where r is a specified positive integer (Devore 111).

The random variable of interest is X = the number of failures which precede the r^{th} success. X has possible values of 0,1,2… The probability mass function for the negative binomial distribution can be written as $P(X=k)=\binom{k-1}{r-1}p^r(1-p)^{k-r}$ where $k=r,r+1,...$. Figure 13 shows the probability mass function of a negative binomial random variable. "Suppose that a sequence of independent trials is performed until there are r successes in all; let X denote the total number of trials. To find P(X=k), we can argue in the following way: Any particular such sequence has probability $p^r(1-p)^{k-r}$, from the independence assumption. The last trial is a success, and the remaining r-1 successes can be assigned to the remaining k-1 trials in $\binom{k-1}{r-1}$ ways" (Rice 37). If the r^{th} occurrence happens at the k^{th} trial, there will be exactly r-1 occurrences of the event in

the prior n-1 trials and at the k^{th} trial, the event also occurs (Ang & Tang 113). 'X' is usually defined as the total number of trials in the distribution, but is sometimes defined as the total number of failures in the distribution (Rice 38). The way of writing the probability mass function allows for the relationship between the binomial distribution and the negative binomial distribution. Both distributions consist of a sequence of independent trials.

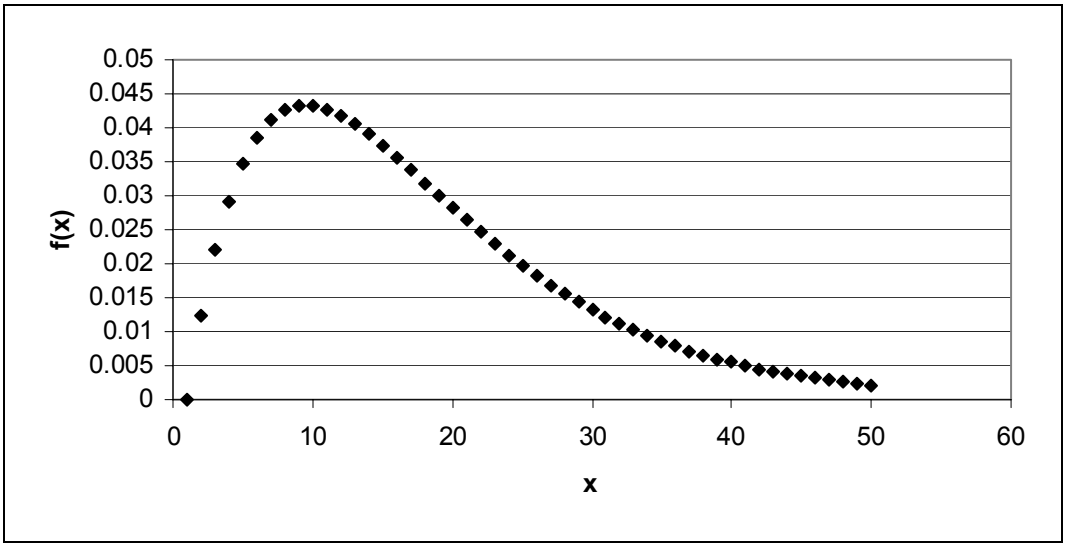


Figure 13: Probability Mass Function of a Negative Binomial Random Variable with k=1/9 and r=2

Since the mean does not have to be equal to the variance in a negative binomial distribution, it follows that the mean does not equal the variance. The mean for a negative binomial random variable is equal to $\mu = E(x) = r/p$. The variance is equal to $\sigma^2 = V(x) = r(1-p)/p^2$ (Montgomery and Runger 82).

Brown and Tarko have used negative binomial regression models with the following form $Y = k * LEN * YRS * AADT^{\gamma} * \exp\left(\sum \beta_i * X_i\right)$ where Y=expected number of total, fatal injury or PDO crashes, k=intercept coefficient, LEN = length of the segment, YRS =number of years of accident data, AADT =average annual daily traffic, γ,

β are model parameters, and X_i are variables representing segment characteristics. The models found all employed the same parameters of access density, indicator variable for outside shoulder, indicator variable that a TWLTL is present, indicator variable if median has no openings, and proportion of access points that are signalized (Brown and Tarko).

2.6.8.1 Goodness of fit

Hadi *et al* found overdispersion to be significant for all the highway types they investigated and chose negative binomial regression to estimate the model parameters (Hadi et al 172). All Poisson and negative binomial models used by Hadi failed to pass the chi-squared goodness of fit test at the 0.05 percent confidence level. Hadi *et al* found similar results reported by other researchers. The chi-squared goodness of fit test is not truly suitable for non linear problems, which includes models following a Poisson or negative binomial distribution (Hadi et al 172). Due to the goodness of fit test not being truly applicable, other criteria have been suggested for determining model acceptance including the following:

- The signs of all parameter coefficients are as expected,
- AIC is the lowest possible value, and
- Each individual parameter is accepted when tested with appropriate statistical methods (Hadi et al 172).

2.6.9 Variable Selection

In addition to choosing the correct model distribution, there needs to be methods for choosing the correct variables to include in a regression model. Most studies that evaluate the effects of road safety measure are observational studies, non-experimental, in which the treatment being studied is not assigned at random. There are many such variables that exist, some of which can be evaluated and some which cannot. "A

confounding variable is any exogenous (i.e., not influenced by the road safety measure itself) variable affecting the number of accidents or injuries whose effects, if not estimated, can be mixed up with effects of the measure being evaluated" (Elvik, 631). "Controlling, or not controlling, for confounding factors may profoundly affect study results" (Elvik, 635), some of this must be done in the early stages of the study when first selecting variables to gather information on, and some can be done in the later stages of modeling.

Several different methods are available to select the variables once they have been included in the study. To determine which variables to include in the model with non-normal distributions, Hadi *et al* prefer the Akaike's information criterion (AIC). AIC=-2*ML+2*K; K is the number of free parameters in the model and ML is the maximum log likelihood(Hadi et al 171). The smaller the AIC value is the better the model (Hadi et al 171).

The development of a model is typically obtained by including additional terms one at a time and testing their significance by the drop in scaled deviance or by the t-ratio (ratio of the estimated coefficient to its standard error) (Maher & Summersgill 283). The drop in scaled deviance should be compared with a χ^2 distribution with as many degrees of freedom as there are extra parameters in the model (Maher & Summersgill 283). A well fitting model or adequate model, the value of the scaled deviance and x^2 should come from a χ^2 distribution with $(N-p)$ degrees of freedom where N is the number of observations and p is the number of parameters which have been estimated (Maher & Summersgill 283).

A formal method for testing that an individual parameter should be included in the regression model exists. Individual parameters, regression coefficients from the β-vector, can be tested to see if the null hypothesis that a given parameter β_j is zero is true. The method used by Hadi *et al* was based on the standard errors of coefficients $\chi^2 = \frac{b_j^2}{(SE_j)^2}$ where b_i is the estimate of β_j and SE_j is the standard error of the coefficient β_j. A chi-square test with one degree of freedom was used to test the hypothesis (Hadi et al 171). This test allows for enough evidence to exist to show either that a β_j is equal to zero, that the corresponding X –variable should not be included in the model, or that β_j is not equal to zero and the corresponding X-variable should be included in the model.

An important part of determining if a variable should be included is that the coefficient should have the expected sign and the t-statistic should show that the variable is significant (A Miaou et al 13). The level of statistical significance needs to be carefully considered. Maher and Summersgill did not accept variables at less than five percent level and did not reject any variables at the one percent level or better without careful thought (Maher & Summersgill 284). A level of significance of five to ten percent is commonly used, depending on the study parameters. The stability of the model should also be considered. When variables are associated with one another then introducing one will tend to strongly affect model parameters. Care should be taken to minimize the correlation between variables that are likely to appear in the models. It is also important that the effect of the variables is understandable and makes sense. Mainly the sign of the parameter should make sense in the context of the study. If the volume is a variable and the sign is negative, that would mean the more traffic, the fewer accidents

and that is not typically the case. The size of the effect and ease of measurement is important in that variables which have a large effect on accidents in relation to their range and were straight forward to measure are preferred for ease of duplication (Maher & Summersgill 285).

2.6.9.1 Variable Transformations

Transformations on certain variables can improve their statistical power for identifying possible relationships. Typically curve radius and grade are variables that are transformed (Fitzpatrick et al (2001) 20). Fitzpatrick *et al* (2001) kept grades at +/-4 percent or essentially flat and constant between all sites so were not used as a variable. Common transformations for curve radius are square root of radius and inverse of radius (Fitzpatrick et al (2001) 20).

During data analysis, modifications of variables may occur. In Fitzpatrick *et al* (2001) access density was originally modeled as a continuous variable but analyses showed that access density was not significant. Further investigation was done due to the preliminary work. A break point was identified for a reasonable division and access density was changed to a class or indicator variable with classes of low density (<12 points/km) and high density (>12 points/km) (Fitzpatrick et al (2001) 20). Another modification that was done by Fitzpatrick et al (2001) was changing median type from three classes (raised, TWLTL, none) to two classes (presence or absence of median). Transforming variables whether by a mathematical change such as a square root, or by content change, by changing a continuous variable into an indicator variable, is done to increase the statistical power of both the individual variable, but more importantly that of the model as a whole.

2.6.9.2 Multicollinearity

Focusing on a specific group of roads gives some variables a limited range of possible values. Due to the limited range, some variables may be correlated with others and in some cases can be explained and expected. In some circumstances the limited range in variables can create apparent relationships that may not be valid and can significantly affect the results of regression analysis (Fitzpatrick et al (2001) 20). "Using Statistical Analysis System (SAS) and the proc CORR command, those variable pairs with multicollinearity problems were identified. The value of 0.05 for alpha was used." (Fitzpatrick et al (2001) 20). To help minimize the effects of multicollinearity, Fitzpatrick et al (2001) averaged inside and outside lane widths to create one lane width variable, similarly inside and outside super-elevation rates were averaged to create one value for each curve (Fitzpatrick et al (2001) 20). The correlation between variables means that the variation in the data explained by one is replicated by the other and that there is no statistical gain from including both in the final model. To have the best possible model, it would be advantageous that the included variables explain different part of the variation within the data set.

2.6.9.3 Outliers

In addition to knowing what type of data to include, it is important to know what type of data to not include. Outliers are data points that were collected using the same methods as all the other points, but do not fall within the same range as the remainder of the data. Points that are outliers are often summarily discarded. This is a problem, because the only points that should be discarded are if there is a known error that occurs with the measurements, otherwise the points may be showing a valid trend in the data that

there is not enough other data to strongly support, or the point could be different due to lack of an additional explanatory variable.

In addition to outliers, influential points also need special consideration. These are points that do not deviate significantly from the rest, but by including them in the model they have a stronger influence on the model than other points do. Schurr *et al* began the modeling process by identifying "influential study sites" or outliers that would strongly influence the model (Schurr et al 63). The sites so identified were removed from the data set before the model was built. The blanket removal of outlying points from a data set needs to be carefully considered and have valid reasoning behind it, else the model will not be a good reflection of the truth. In the collection process, data can be discarded due to instrumental errors or incomplete data points. But once the model building process is begun, none of the data points should be removed from the data set. This could cause relationships that are not truly present to be seen and conversely cause relationships that are present to be overlooked.

2.6.10 Uncertainty of Predictions

Once the model has been fitted and the parameter estimates found, the amount of uncertainty attached to predictions from the model needs to be considered. The parameter coefficients are only estimates of the true values and as such each has standard errors. Uncertainty in the coefficients leads to uncertainty in the linear predictor and finally to uncertainty in the prediction value. The uncertainty of the prediction, measured by its error variance can be approximated by $Var(\hat{\lambda}) \approx Var(\hat{\eta})\hat{\lambda}^2$ where $\hat{\eta} = \hat{\beta}^T x$ (Maher & Summersgill 290). The uncertainty of the estimate to the true mean λ consists of the

regression effect (uncertainty in λ) and overdispersion (uncertainty in λ about λ, where $Var(\lambda) = Var(X:\hat{\lambda}) + Var(\hat{\lambda})$) (Maher & Summersgill 290).

Quasi-Poisson model: $Var(\lambda) = (k^2 - 1)\hat{\lambda} + Var(\hat{\eta})\hat{\lambda}^2$

Negative Binomial model: $Var(\lambda) = \hat{\lambda}^2\left[\frac{1}{\alpha} + Var(\hat{\eta})\left(1 + \frac{1}{\alpha}\right)\right]$

The predicted error variances of the negative binomial and quasi-Poisson models are very different especially for extreme values. While the choice of model has little effect on the form of the fitted model, it can greatly affect the estimate of the uncertainty of the model (Maher & Summersgill 290).

2.6.11 Trend

Accident counts can show trends due to transitory changes in factors such as flow, weather, economy, and accident reporting practices. Accident models that account for these types of trends should provide better estimates of safety than the more traditional models in identifying hazardous locations and evaluating treatments (Lord & Persaud, 102). There are three main categories of proposed methods to deal with trend: marginal models (MM), transition models (TM), and random-effects models (REM). These three procedures all have different limitations:

- Temporal correlation in the data is ignored (REM & MM),
- Model type may not be appropriate for accident prediction models (REM & TM), or
- Too complicated for average modelers (TM & MM).

The generalized estimating equations (GEE) procedure overcomes these limitations (Lord & Persaud, 102). Lord and Persaud found when comparing generalized linear models and GEE with and without trend that the temporal correlation contributes to

approximately half of the standard errors (Lord & Persaud, 105). The standard errors for the GEE models were roughly twice those of the GLM models. If time trend is not of interest, the dispersion parameter was found to be slightly higher for the GEE than the GLM procedure (Lord & Persaud, 105). Using time trend also allows for potentially dangerous trends to be identified and investigated earlier.

3 Methodology

In order to see what previous research methods have been used, existing methods for the determination of safety of two lane rural roads will be reviewed. This will include a literature search and review of existing techniques. Different techniques will be examined and reviewed for their applicability to urban arterial streets and roads. Work on urban roads will also be assessed to see if it can be applied to urban arterials and to see what types of analysis tools were considered to be reliable.

Miaou dissected the modeling process into five major tasks which are required to develop accident prediction models: (1) find a good probability function to describe the random variation, (2) determine an appropriate functional form and parameterization to describe the effects of multiple variables, (3) select the right variables, (4) obtain estimates of the regression parameters and (5) assess the quality of the model, ways to improve it, and to ensure the model fits the required specifications (Miaou, 8). Sample size is always a crucial point of throughout the modeling process. By nature, sample sizes are limited and minimum sizes need to be chosen to ensure that the best possible model can be developed. The impact of omitted variables should be considered, as well as the potential for variables that were not considered. In addition to considering all possible variables the chosen sites used to create the models should be fairly homogeneous to help eliminate the unforeseen variations.

After a thorough examination of existing research, data will be collected. This will occur by one or more of the following methods, including receiving data from local or regional agencies and gathering data from roads neighboring Worcester Polytechnic Institute. Many different variables need to be considered and then either rejected or

accepted as explaining a significant amount of variation in the final model. Two major types of variable data area needed: geometric and non-geometric data.

Non-geometric data includes information regarding the traffic characteristics and vehicle crashes. This includes traffic flow (AADT), vehicle distribution (trucks, passenger vehicles, vulnerable road users (pedestrians and cyclists)), speed limit, one/two way traffic, surrounding land use, bus stops, parking conditions, and accident number and type.

Geometric data is also needed to help fit the model to the specific location where it is being applied. The geometric data includes segment length, number of lanes, number of minor crossings/side roads, sidewalks (access point frequency), road width, number of driveways (two-way total)/km, number of bus stops (two-way total), crosswalk frequency, type of median (none, TWLTL, raised), traffic islands, type of land use (residential, business, and other (industrial)), and percentage of segment length on which parking is allowed. Some of the variables will be used directly as numerical input values, but some will be used as an indicator variable.

One specific issue that has to be determined is what defines a section length. One rule of thumb is that signalized intersections are natural delineators of road sections since major changes in volume occur at those locations. Traffic signals imply that there is considerable traffic on both roads and the mixing of traffic streams can create an issue in regards to what causes an accident. It could be that the junction is not safe due to the combination of the two different road geometries and usage, but not that the design of the roadway itself is unsafe. The mixture of traffic streams makes it difficult to assign an

accident to only one of the intersecting roads causing discrepancies in the accident data. Another group identified road sections by the type of median.

Once all the data has been acquired, it has to be assembled in order and placed into models. The most common method is to use generalized linear modeling techniques. With linear modeling techniques it has to be assumed that the distribution of accidents follows a pattern (discrete, nonnegative and rare) and is not just a random occurrence. The two most widely used distributions are the Poisson distribution and negative binomial distribution. There are positive and negative aspects to using either major type of distribution. Poisson distribution is easier to use than the negative binomial one, but problems can arise due to the phenomenon of "overdispersion." Overdispersion is when the observed variance is actually greater than the mean and causes standard errors to be underestimated (Greibe, 275). Negative binomial distribution is more difficult to implement, but allows for a greater variance in the data, which eliminates the overdispersion issue.

Separate models can be determined for a combination of all accidents, including property-damage-only accidents, all injury and/or fatality accidents and for specific types of accidents that it may be important to look at more closely (single vehicle accidents, rear-end accidents, crossing accidents, and turning accidents).

Once the model has been developed, it needs to be verified showing it to be an accurate representation of accidents falling into the study's characteristics (size of roadway and AADT). Statistical methods will be used to show that model is a good fit for the data used to develop it. The final step includes using the developed model to compare the predicted results with the actual accident records. A technique known as

bootstrapping allows for the use of part of a database for model development and part of the data base for model verification, which allows for this comparison otherwise a new data set can be used. If the difference in the model's results and the accident records is statistically insignificant then the model is a good representation of the urban arterial roadways that fall into the study's criteria.

4 Data Collection

Data are needed to develop a model for predicting accidents on any road type. "Accuracy of prediction models depends on the details of the information base on which the models are built" (Lau & May 62) which indicates that the better and more accurate the data collection, the better the prediction models will be. The following sections describe the types of data that were collected and how the data were obtained. The site of the road sections used was mostly random in nature. Due to using only sites in a single geographic area, the findings of this study should only be interpreted as explaining the relationships in this study sample and only extrapolated to similar areas (Tarris et al). A goal of the study by Schurr *et al* was to minimize uncertainty in the final results by reducing the number of extraneous variables, which could influence operating speeds, the variable they were most interested in. Only sites with pavement of fair or better were chosen to eliminate the pavement influence. If there were roadside elements near the curve site such as bridges, guardrails, intersections within 1000 feet of the point of curvature on the approach the curve, the site was not used (Schurr et al 62). For this reason, each possible variable was carefully collected so that its importance could be considered and if necessary, used to eliminate outlying data points from the study. An important issue was to keep data collection simple, so if the data was available it was used, otherwise if collection was simple, counting or easy to measure, it was collected in the field. If data collection was difficult or time consuming, such as new volume counts and turning movement counts, then it was not considered a viable variable.

Roadways included in this study were urban arterial roads, consisting mainly of state routes. Belmont Street and Highland Street are both part of Route 9. Chandler

Street is part of Route 122, while Park Avenue is part of Route 12. These roads were chosen in part due to their geographical location of spanning Worcester from east to west. Figure 14 shows the roads used in the study to create the prediction models. The map also displays the boundaries of the City of Worcester and most of the arterial roadways throughout the city.

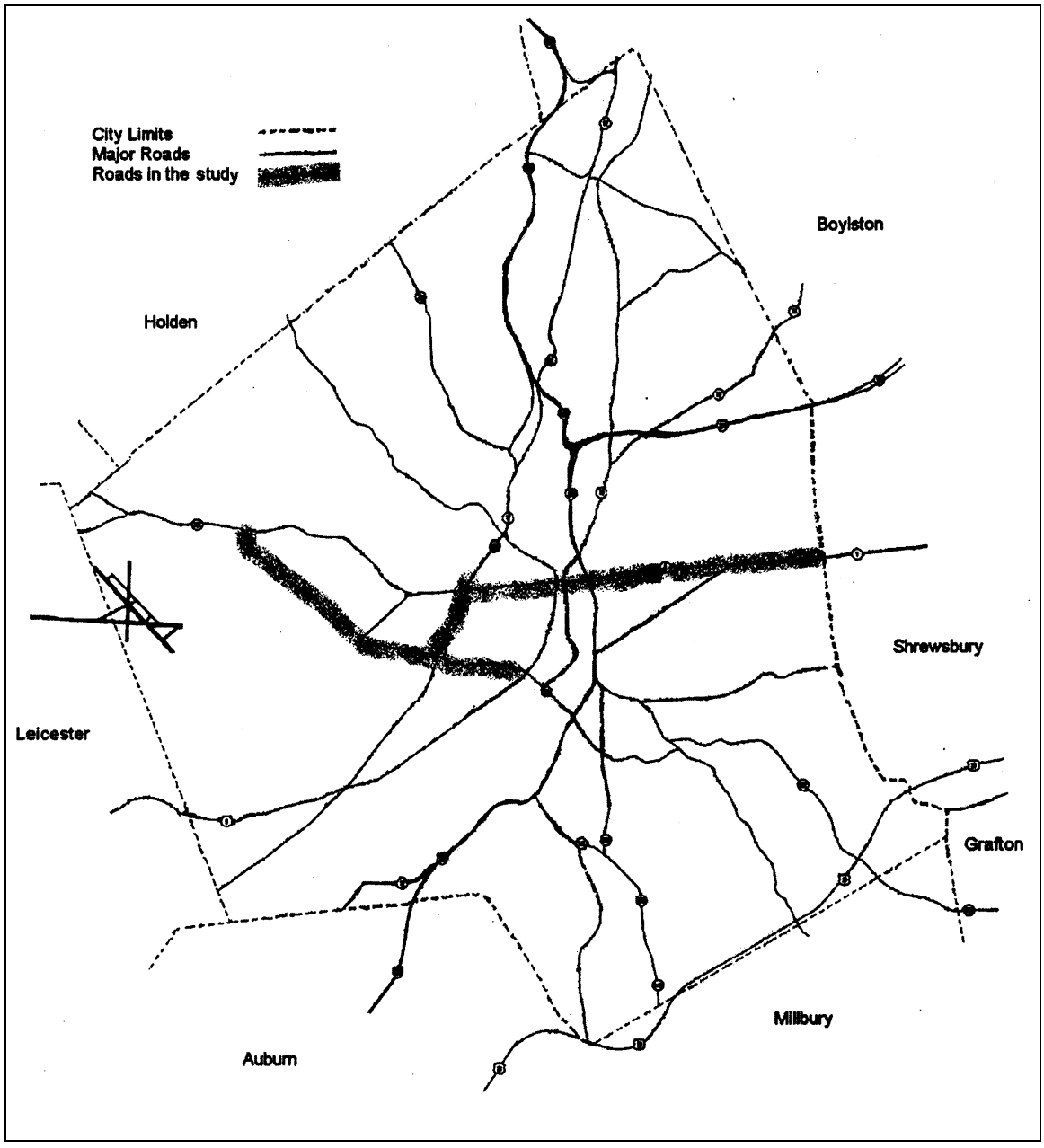


Figure 14: Worcester City Limits Displaying the Study's Road Sections

4.1 On-Site Data

A form was developed in order to assist in the collection of geometric data. This form covers the data that needed to be collected from each site, consisting mainly of geometric, land use, and roadside data. This can be seen in Figure 15.

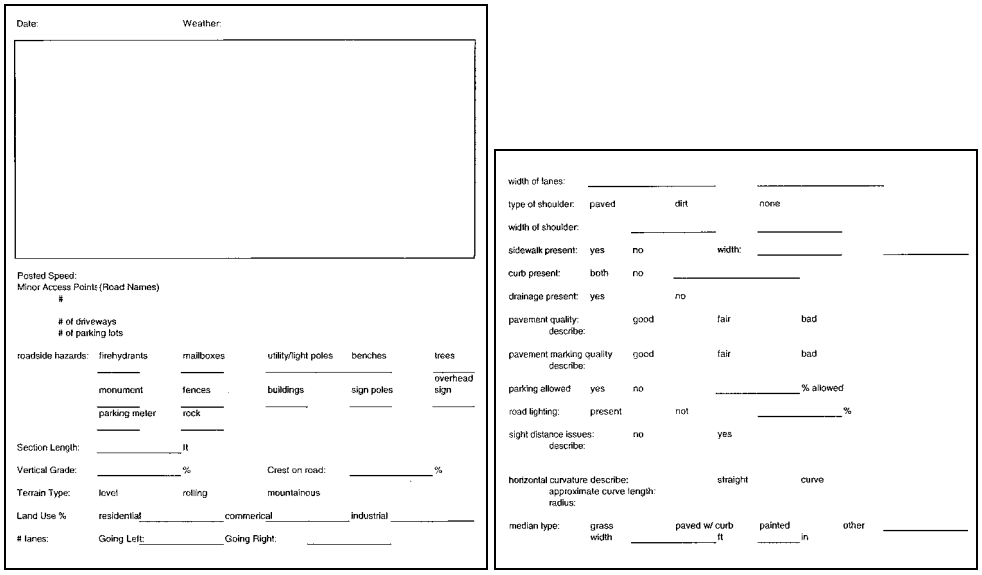


Figure 15: Data Collection Form

4.1.1 Speed Limit

The posted speed limit was gathered to help give an indication of how fast drivers should be going on the road. The posted speed limit also gives an expectation of how the traffic should be flowing. When there is not a posted speed limit in Worcester, the city follows Massachusetts State Law, Chapter 90, Section 17 (www.state.ma.us/legis/laws/mgl/90-17.htm). If a vehicle is on a divided roadway outside of thickly settled areas or business districts, it can travel at 50 mph. If a vehicle is on any other road outside of a thickly settled area or business district, it can travel at 40 mph. Inside thickly settled areas or business districts, vehicles can travel at 30 mph and in school zones are limited to 20 mph. These general rules are superceded by posted speed limits. Most of the road segments examined in this study did not have posted

speed limits. Only ten segments had posted speed limits and the remainder of the segments had their speeds inferred from the Massachusetts State Law or surrounding sections with posted speeds. Speeds throughout the study area range from 25 mph to 40 mph.

4.1.2 Length

Section length plays an important role in predicting accidents. Accidents are usually transformed into accident rates, where the number of accidents is normalized by time, traffic volume and length and then the accident rate is used as the dependent variable. Determining whether accidents are distributed linearly by segment length and traffic volume is key to that assumption. If accidents are not linearly distributed than the use of accident rates is not appropriate. Segment length is also important in that the longer the segment is the more crashes are expected to occur on it. The relationship between accidents and segment length may be linear or exponential in nature, but intuitively the longer a segment the more area where an accident can occur.

Due to the various ways segment length can play a role with crashes and accident rates the way roads are divided into sections is very important. There are two main schools of thought. In rural conditions, where most prior roadway research has been done, segments are divided by changes in geometry, such as changes in lane width or shoulder width or changes in paving materials. In urban locations, segments tend to be defined by intersections. The segment length may include intersections with local roads, while intersections with collectors or arterials indicate the end of the segment (Brown and Tarko 71). The definition of Brown and Tarko's segment length is more appropriate in this situation than the definition used in rural locations. Major intersections with traffic

signals on urban arterials show that there is a significant change in traffic conditions at that point. That change of conditions between one segment and the next is important to recognize. Major intersections also provide a very exact way to identify the segments without the possibility of mistaking the ends of the segment. The segment lengths in this study ranged in length between 226 ft to 5,245 ft with an average segment having a length of 1,346 ft. The variation between residential and commercial land use areas helps to explain the variation in length of the segments.

4.1.3 Access Control

Access points on urban arterial streets consist of major intersections (i.e., intersection with traffic signals), minor intersections (i.e., without traffic signals) and entry points such as driveways and parking lots. The number of access points gives an indication of how many places there are were vehicles could get into turning conflicts and possibly crashes. Brown and Tarko's study used access density as a variable to characterize conflict points and driveway accidents. According to studies in Indiana, driveway accidents compose between 14 and 33 percent of all accidents in cities (Brown and Tarko 68). It included driveways, signalized and un-signalized roads (Brown and Tarko 70). Access density is one way to use the data, but that assumes that the access points are linearly related to the segment length. Using the data as a continuous count variable or as a density variable are both possibilities for variables for predicting accidents. Access points need to be examined to be certain that there is a linear relationship between access points and segment length before using density as a variable in an accident prediction model. Some studies have used access density as a qualitative variable listing the density into groups of high, medium, and low density. This may be an

effective method if access density as a continuous variable is insignificant in an accident prediction model. In this study, the road segments were divided by major intersection, so that there are only minor intersections, driveways, and parking lots that make up the access points. The three classes were recorded separately so that each can be examined individually for any relationships to accident occurrence.

Figure 16: Examples of Minor Access Points

This study defined minor access points as public roadways that intersect the road segment but do not have any signalized control. There may, however, be stop or yield controls present. The occurrence of minor access points ranged from zero to thirteen per segment with an average of four points per segment. Driveway counts varied dramatically between zero and sixty-six per segment with an average of eight driveways per segment. Figure 16 shows an example of a driveway access point and a minor road access point. Some of this variation is due to the fact that some of the road segments were located in fully residential areas and some were located in commercial areas. Parking lot counts varied due to similar reasons as driveways with a range of zero to

thirty-three with an average value of seven per segment. The land use surrounding the segment strongly influences the division between driveways and parking lots and the number of access points is important in showing locations where vehicles can enter the traffic stream.

4.1.4 Vertical Alignment

Vertical alignment has an important role in helping to determine safe design criteria, specifically maximum grade allowances. Vertical grades affect the ability of some vehicles, especially large trucks and buses, to safely traverse some roads. The grades found on the road segments ranged from less than one percent up to a maximum of 10.9 percent grade. As can be seen Table 5 in from AASHTO's Green book, the maximum grade observed falls under the maximum for its design speed of 30 mph in mountainous terrain. Most of the grades observed fall well below the maximum allowable values recommended by AASHTO.

Table 5: Maximum Grades for Urban Arterials

	Maximum Grade (%) for Specified Design Speed (mph)				
Type of Terrain	30	35	40	45	50
Level	8	7	7	6	6
Rolling	9	8	8	7	7
Mountainous	11	10	10	9	9

From Exhibit 7-10 AASHTO's Greenbook

4.1.5 Land Use

Land use gives an indication of the type of traffic that is expected to use the roadway. Residential areas tend to have drivers who are familiar with the roadway and expect turning vehicles and pedestrians throughout the area. Commercial areas, on the

other hand, lend themselves to fewer places for turning, with more parking lots than driveways, while also having pedestrians, the drivers will not be as familiar with the roads and traffic patterns in commercial areas. Examples of residential and commercial land use can be seen in Figure 17. The other main alterative for land use is industrial use. The residential category indicates land use from both single-family dwellings to apartment complexes. Commercial areas are associated with customer trips that occur throughout the business day. Industrial use refers to land where non-professional employees make the majority of trips with the trips taking place during shift changes (Bonneson & McCoy 28). Large trucks are associated with both commercial and industrial areas, which have very different dimensions from passenger vehicles and roads with high percentages of trucks need to be designed to accommodate the larger dimensions.

Figure 17: Examples of Commercial and Residential Land Use

Land use can vary drastically along the length of an arterial, but also can vary significantly between each side of the road. When there were multiple uses along a segment Bowman and Vecellio assigned a type on the basis of observed activity at the time of the field survey (Bowman & Veccellio b 170). Similarly, when Bonneson and McCoy observed varied land use, the most dominant type would be chosen (Bonneson &

McCoy 28). This method of picking one type of land use has the result of eliminating the variation of use throughout the segment, but this variation may strongly influence the travel patterns. With this in mind, land use was categorized by the percentage of land use between all three possible types in each segment; residential, commercial, and industrial. This allows for the possibility of having multiple land uses in a single road segment and does not disregard the differences. If multiple land use does not have a strong influence on the prediction model, the dominant type of use can still be identified and used as a variable in the prediction model. The sections that were used in this study were divided mainly between residential and commercial areas. There was only one segment that had any industrial land use. Overall, approximately 25 percent of the land examined was residential and 75 percent was commercial.

4.1.6 Medians

Medians have always been important in terms of roadway safety. Experts have agreed that the use of medians increases safety, but that affect has not been quantified. Safety experts have also disputed the type of median that provides the best safety measure. The undisputed fact remains, however, that median treatments do have an effect on vehicular safety. An example of a common median treatment in Worcester can be seen in Figure 18 that of a raised and curbed median.

Figure 18: Raised Median from the Study Area

Three major types of median treatments were included; raised median, two-way left turn lanes (TWLTL), and undivided treatment. Due to the area chosen for data collection (i.e., Worcester, MA) there were not any TWLTL available in the study area. There were a few segments that had raised median treatments consisting of curbs surrounding grass or pavement, but most had undivided treatments. The lack of variability in the existing conditions will not allow for a full exploration of this issue with the data available but a partial one may be possible. The width of a median has also been shown to play an important part in the safety of a roadway. Due again to the small number of available sites with suitable treatment, there is not a large enough variability among the sites with raised medians to show effects on safety due to median width. The four sites identified as having raised median treatments had widths ranging from 5.5 feet to eight feet.

4.1.7 Cross-Sectional Alignment

Cross section alignment plays an important role in helping drivers to feel that they are using a safe road especially when referring to lane and shoulder widths. When lanes are narrow, drivers feel crowded by passing vehicles and are more prone to feeling

uncomfortable. Increasing lane widths up to the AASHTO standard of 12 feet helps to alleviate that discomfort. In studies, the number of accidents has been shown to decrease as the lane width increases up to the standard width. For this reason the lane widths were all recorded, to see first if the roadways are being built according to the AASHTO recommendations, and secondly to see if the road sections that are built with 12-foot lanes have fewer accidents than road segments that are smaller. For the same reasons the number of lanes was recorded. Most of the segments had one or two lanes going in each direction, with a few exceptions of three lanes and one case of four lanes. The widths similarly varied depending on the section being examined. There was an overall average lane width of 12.5 feet, which is due to the fact that many of the roads with one lane in each direction were twenty feet wide. These lanes are not truly twenty feet wide but there is no distinction between the parking lane and the traveling lane leading to this large lane width. If a segment had on-street parallel parking, the parking area was included in the lane width measurement because the lanes were not well delineated and some times no vehicles were present at the time of the on-site investigation to mark the parking lane.

Similar to number of lanes and lane width is the effect of shoulder width. Shoulder widths have been examined in great detail in many studies to determine their safety benefits. For that reason the type of shoulders and their widths were recorded. Possible shoulder types include paved shoulders and dirt/grass shoulders. However, in urban settings, roadway shoulders are not a requirement and due to space constrictions are seldom used. This was found to be the case in the sections reviewed during this study. No segments were found to possess actual shoulders, and a variable that has been

thoroughly studied and found to be an important factor in rural settings has little impact in an urban location.

A different variable exists that is seldom found in rural settings and is frequent in urban settings that of sidewalks. Sidewalks provide a place for pedestrians to safely walk along busy roads without intruding on the traveled way. Since wider lanes make drivers safer and feel safer, it has been suggested that the same could hold true for pedestrians feeling safer on wider sidewalks. Therefore both the presence of sidewalks and their width were noted at the physical inspection of each site (See Figure 19). The width of sidewalk was recorded for both sides of the road if it was present, but if a sidewalk was present on at least one side of the road, it was concluded to be present along the entire length. It was found, by this definition of a sidewalk on at least one side of the road, that every road segment reviewed had a sidewalk with widths ranging from five to 12.5 feet. The minimum width of sidewalks should be determined by the necessary width needed to accommodate people with disabilities and strollers. The maximum width is determined by space availability and convention. An average sidewalk width of nine feet was found in the study area in Worcester.

Figure 19: Example of a Sidewalk in a Residential Area

Drainage becomes an important consideration when there is not a large amount of land available for building roads. Water on road surfaces can become a hazard, especially with large rainfall amounts and during winter months when hydroplaning and black ice are of major concern. To investigate whether or not drainage could be a cause of accidents, its presence was noted for each segment. That was accomplished by recording if there were curbs present on the side of the road to help direct water flow and by recording the presence of any drainage structures, such as catch basins or manholes. For each segment in the study, a curb was found to exist on both sides of the road, and drainage structures were present along the entire study length. Figure 20 shows an example of what drainage structures were found throughout all of the roadway segments.

Figure 20: Example of Roadside Drainage

Another feature that assists with drainage is the crest of the road, which helps to direct water away from the main travel path and into the catch basins. The crest was measured along the road segments to see if there was adequate provision for this issue. The values found for the amount of cross slope on the roadway ranged from 0.3 to 6.8

percent with an average of four percent. There were four sections where the cross slope exceeded 6 percent, the maximum recommended value by AASHTO and two cases were the cross-slope was less than the recommended 1.5 percent minimum. This could indicate problems with drainage and may also indicate an increase in accidents on segments that do not meet AASHTO's recommendations.

4.1.8 Roadside Hazards

Roadside hazards provide opportunities for vehicles to hit objects located on the roadside. The more hazards that exist on a given road, the more opportunities are present for a vehicle to collide with those objects. During the on-site inspection, the number and type of roadside hazards were recorded. This was done for the possibility that a relationship exists between either the total amount of hazards or a specific type or combination of hazards. The types of hazards recorded included fire hydrants, mailboxes, light poles, utility poles, benches, trees, monuments, fences, buildings, sign poles, overhead sign poles, parking meters, rocks and electrical boxes (See Figure 21). The number of hazards ranged from ten to 338 per segment with an average of 79 hazards per segment. This is also an area where a rate, or a density, may be a more appropriate representation of the hazards, so the possibility of normalizing the roadside hazards by length may have a better effect for predicting accidents. Either a continuous variable of number of hazards per segment or a qualitative variable of hazards per mile could be used as a variable in the accident prediction model. Some researches have used hazard density as an indicator variable, separating section into high medium and low-density locations, which is another way that the data could possibly be used.

Figure 21: Examples of Roadside Hazards

4.1.9 Horizontal Alignment and Sight Distance

Like vertical alignment and cross sectional alignment, horizontal alignment can have a significant effect on accidents. Horizontal curvature is often a controlling factor for safe speeds on roadways and for the comfort of drivers. If a curve is too sharp for a given design speed, it can cause discomfort for drivers and passengers even if the car can safely travel around the curve. Horizontal alignment can also cause sight distance problems in high-speed areas. There were fifteen curves identified throughout the study segments. Of these curves none were identified as having a radius that was inappropriate for the design speed of the segment. Of the sight distance problems identified throughout the segments, only one was due to the horizontal alignment. The other two were due to vertical alignment that blocked the sight of the traffic signals, but in both cases signs and other warning devices were present to help eliminate the problems. The only horizontal curve that caused possible sight distance problems was like the other sites, marked with signs, specifically chevrons, and at the posted speed limit would be safe.

4.1.10 Other On-Site Data

Several other pieces of information were collected in the hopes that one or more of them may be identified as having a significant influence on accident occurrence.

Pavement quality was identified as something that could cause accidents to occur. Data for this issue was collected at each segment and the pavement was identified to be in good, fair or poor condition. A pavement was classified as a good pavement if there were very few disturbances in the surface of the pavement. A few cracks or patching would qualify a pavement as good. A fair pavement would have significant amounts of cracking and rutting. Bad pavement would have to have visible potholes, large ruts or other serious problems. Problems that can occur to negatively effect pavement quality include rutting and cracking and can be seen in Figure 22. At the sites used in the study, all the pavements fell into either the good or the fair category. This was to be expected due to the usage patterns of the roads investigated. Urban arterial roads have heavy volumes of traffic and poor conditions can cause large congestion problems quickly. Poor conditions on arterial roads are avoided by having significant amounts of repair on the roads.

Figure 22: Examples of Problems in Pavement Quality

Pavement marking, like pavement quality, was theorized to have an influence on accidents on urban arterial roadways. Again, like pavement quality, pavement markings were categorized as good, fair or poor quality. A good pavement marking was all present and able to been easily seen, while a fair marking was starting to fade in places. A bad pavement marking, on the other hand, was very faded and in places not even visible. The majority of pavement markings qualified for fair or good status with only five segments having bad pavement markings. The greater variation in quality is because the lifetime of pavement markings is significantly shorter than that of the pavement, allowing the pavement to still be in good condition while the markings have worn away. Figure 23 shows two locations of pavement markings with the left hand side representing a bad marking and the right hand side representing a fair pavement marking.

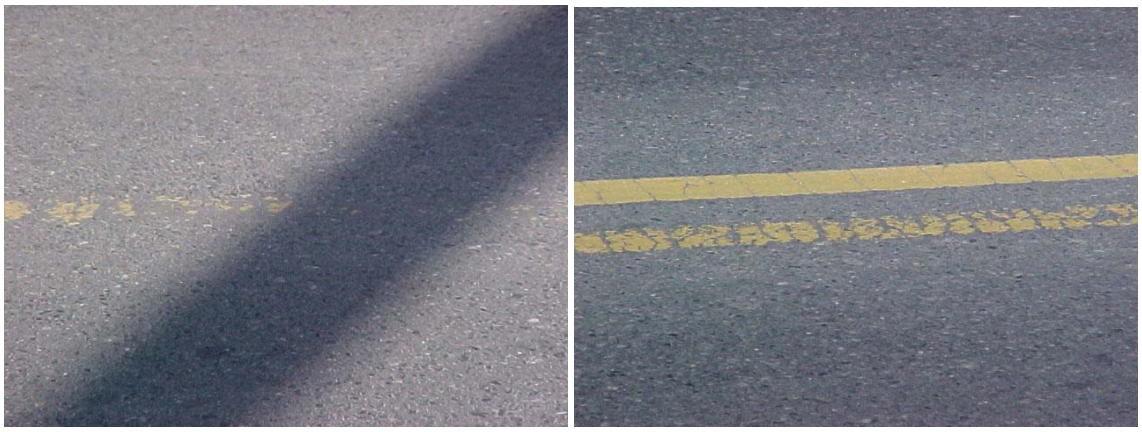


Figure 23: Exampled of Pavement Markings

Lighting is an issue of major concern on rural roads. Due to its importance in that type of road, the amount of roadway lighting was recorded. The urban setting, however, makes lighting a much less prominent issue. Of all the segments in the study, only one did not have roadway lighting along its entire length, and that segment was only approximately 20 percent unlit. Due to the high volume and high speed of urban arterials and their position in important areas of cities with many turning possibilities, the urban

arterials are usually well lit. This has the effect of lighting not playing such a large role for urban arterials as they do in rural locations and possibly urban collectors and local streets.

Another possible variable for consideration is the amount of on street parking. The amount of on street parallel parking gives an indication of the type of expected traffic on the roads. Areas that do not allow on-street parking tend to have higher volumes and higher speeds. Conversely, areas with a large amount of street parking will have slower speeds, but may still have high volumes. Some segments examined had no on-street parking while other segments had 100 percent on-street parking. Twelve segments, in fact, allowed no parking at all. The average amount of parking was 40 percent for each segment. One segment even had a small section of perpendicular parking.

4.2 Off-site Data

Some data was also needed that could not be collected at the individual sites. This was used to supplement the geometric, land use and roadside data by identifying accident and traffic conditions.

4.2.1 Volume Data

Average daily traffic (ADT) and average annual daily traffic (AADT) are used to indicate traffic conditions or congestion levels of a road section. ADT plays an important role in determining the safety of a roadway by helping to characterize the types of accidents that are likely to occur on roads. It is also important because the more traffic on a road the more possibilities exist for conflicts and crashes. Studies performed on two-lane rural highways in the former Soviet Union show that the number of accidents

increased in proportion to the traffic volume (Gibreel et al 309). In Sweden single vehicle accident rates decreased as traffic volume increase, and the accident rate of multiple vehicle accidents increased as traffic volume increased (Gibreel et al 309). Depending on the type of accident being reviewed ADT can have varying effects. The study on Swedish accidents shows this. The more traffic present the more multi-vehicle accidents occur. In the same way as there is more traffic, it is less likely that only a single vehicle will be involved in an accident. This shows why it is important to consider the ADT when looking at accidents in general and at specific types of accidents such as multi-vehicle crashes or single-vehicle run-off-the-road crashes. Hadi *et al* also found that crash frequency increases with higher ADT for all highways types investigated during their study, including two-way two-lane and four-lane undivided urban highways and divided urban highways (Hadi et al 173).

The number of lanes varies from one road section to another, especially in urban areas and the differences in number of lanes can sometimes have a large effect on the ADT. In a study on truck accidents and geometric design, ADT was generalized by considering the AADT per lane (A Miaou et al 15). This was done to help make the volume more representative of the actual road conditions. Using just the volume numbers can be misrepresentative when some roads have only two lanes and others have more. By using just the AADT by lane, comparisons between road segments with differing geometric characteristics can be more easily completed. The above are reasons why it is important to have information available on the ADT in order to develop an accurate prediction model.

Due to time constraints, the ADT needed to be gathered from existing data and could not be gathered specifically for this study over the exact roadway segments. Counts were gathered from several different sources, including the Worcester Department of Public Works, Traffic Engineering Division, the Central Massachusetts Regional Planning Commission (CMRPC) and the Massachusetts Highway Department (MHD). The data from CMRPC consisted of un-factored ADT's throughout Worcester that were gathered by public and private companies. The data from the Worcester Department of Public Works, Traffic Engineering Division was in the original raw data listed by hour. The data from the MHD was already factored and given by year for locations that have had multiple counts over several years. The un-factored data was multiplied by a weekday monthly factor that was obtained from the MHD website. Factoring allows for a more accurate value for the ADT.

The study period covers three years from 2000 to 2002. The most accurate way to deal with volume data would to have volume counts for each of the three years. This however, is unpractical in that the data was not available and counts are not conducted annually through out the study area. Due to those facts the most recent and available data was used and if necessary projected to the center of the study time period. An average growth rate of 2 percent per year was used as the value used by the Worcester Traffic Engineering Division. The ADT's of the road sections ranged from 11,000 vehicles per day to 47,000 vehicles per day with an average ADT of 25,000 vehicles per day.

4.2.2 Heavy Vehicles

The percentage of heavy vehicles can be very influential on the number of accidents. Heavy vehicles have different characteristics than smaller vehicles (Figure

24). The major differences are that heavy vehicles take longer to speed up and slow down, need larger turning radii, on long upgrades they can slow down considerably, and on long downgrades their brakes may not be able to stop the vehicle. This is mostly a concern over the long distances in rural locations, but in the idea that what is important in one region can be important in another the data was gathered. The data came from the Central Massachusetts Regional Planning Commission (CMRPC) and is taken from their list of peak period turning movement counts. When both an morning and evening period was listed, an average of the two was used for the data point. The amount of heavy vehicles ranged from 0.4 to 3.1 percent with an average of 1.7 percent of traffic being heavy vehicles.

Figure 24: Example of a Heavy Vehicle

4.2.3 Crash Data

The other main type of off-site data gathered was the number of observed crashes. The crashes were complied from the Worcester accident database, which lists all reported accidents in the city of Worcester. Three years of crash data was used from 2000 to 2002. Accident data can be separated in many ways: accident type, location and time period. "It might be asked whether data could, or should, be disaggregated so that each year/site combination provides a unit of data" (Maher & Summersgill 292). This can

make a difference in modeling overdispersion because one cause of overdispersion is the influence of variables not included in the model that remain the same from year to year which can be thought of as a site effect (Maher & Summersgill 292). Using each year/site as a data point does not allow for the errors to all be seen as independent as the errors in the same site in different years are likely to be highly correlated (Maher & Summersgill 292). But using each year/site as a data point allows for more data points to be used when considering the data. Use of multiple observations from each intersection could cause the "gamma error term in the negative binomial model could be correlated from one observation to the next, which is a violation of the error-term independence assumption made to derive the model" (Poch and Mannering 111). This results in a loss of estimation efficiency (standard errors of coefficients will become larger) and could lead to wrong conclusions regarding coefficient estimates (Poch and Mannering 111). The way the accident data was recorded allows for this possibility if it is found to be necessary. If at all possible it is better to avoid the problems associated with correlation of the data points.

The accidents were recorded by which segment they occurred on. Further separating the accidents was categorizing them by occurring on the main part of the segment or occurring on the major intersection of the segment. The major intersection of the segment was defined as the intersection occurring at the end of the road segment. The beginning of the roadway segment was the end with the lowest street number and the end of the segment had the highest street numbers. Hadi *et al* performed separate analyses for non-intersection or mid-block crashes and all crashes, which include intersections, interchanges and railway crossing crashes (Hadi et al 171). The accidents

were recorded in such a way that separate analyses for mid-block and all crashes can be done. The crashes were also recorded by type of crash; fatal, injury, and property-damage only (PDO) crashes. Throughout the study period there were 2,842 reported crashes, but there was only one fatal crash on the roads in the study. There were also a total of 1,930 PDO crashes. It is believed that the reporting level for injury accidents is between eighty to ninety percent and that for PDO accidents it is around fifty percent or less of the accidents being reported (Lau & May 58). Since fatality crashes are rare and PDO's are often not reported, Lau and May suggest that injury accidents are the best category for using to develop prediction models (Lau & May 58). The reporting levels for accidents however are not likely to change suddenly, so that if the number of reported PDO accidents is used it represents an unknown but constant percentage of the true number of accidents and therefore is acceptable to use for predictive purposes.

The separation of the observed crashes allows for the possibility of multiple prediction models being developed. Other researchers, including Brown and Tarko, have been able to create prediction models for total number of crashes, fatal crashes, injury crashes and PDO crashes (Brown & Tarko). Due to the nature of the data collected models for total number of crashes, total injury and fatal crashes and PDO crashes can possibly be developed for the data from Worcester. The classification of the accidents was kept simple with just injury, fatal and PDO as options. Lau and May kept to this classification in their intersection crash study and advantages of this include easy comprehension of the type of accident and there can be a simple translation to monetary terms (Lau & May 58). A major disadvantage of this technique is that it is an inadequate way to reflect the overall collision process and the concept of collisions. Further

classification, however, is difficult since the main descriptive terms, sideswipe and angled collision, usually describe more than one situation. An angled collision can be caused when a vehicle is turning left or right or slides sideways, all very different situations described with the same phrase. Further classification, can get complicated very quickly with many possible types of collisions and become a time consuming and tedious process. The accidents can be used in the model to predict the total number of accidents or a more common way is to predict an accident rate. Accident rates can normalize the number of accidents by time, ADT and length. Knuiman *et al* calculated accident rate per 100 million vehicle miles traveled which they calculated by:

$$R = \frac{Y}{ADT * 365 * T * L}$$ (Knuiman et al 71)

where: R= the observed accident rate

Y= the observed number of accidents

ADT= the average daily traffic in vehicles per day

T=the number of years of crash data

L=the section length

Another way to construct accident rates is to use the rate per million of entering vehicles (RMEVs) which is the number of accidents per million vehicles entering the study location.

$$RMEV = \frac{A * 1{,}000{,}000}{V}$$ (Garber & Hoel, 139)

where: RMEV=accident rate per million entering vehicles

A=total number of accidents or number of accidents by type occurring in 1 year at the study location

V=Average daily traffic (ADT) * 365

This type of rate is commonly used to measure accident rates at intersections. Garber and Hoel also developed a rate per 100 million vehicle miles (RMVM) which is the number of accidents per 100 million vehicle miles of travel over the study section (Garber & Hoel).

$$RMVM = \frac{A * 1{,}000{,}000}{VMT}$$ (Garber & Hoel, 140)

where: RMVM= number of accidents per 100 million vehicle miles of travel

A=total number of accidents or number of accidents by type during a given period at the study period

VMT=total vehicle miles of travel during the given period

=ADT*(days in study period)*(length of road)

The number of accidents compared to volume over a roadway segment is small, so that multiplying by a large factor helps in the analysis. The accident rate can correspond to different accidents depending on the desired parameters. Rates for serious injury accidents, all injury accidents, PDO accidents, multi-vehicle accidents, head-on accidents, sideswipe opposite direction accidents, single vehicle accidents, single vehicle rollover accidents and any other type that in depth study is desired for can be calculated and each analyzed individually using regression modeling.

5 Analysis

The analysis procedure began with trying to identify the exact form the dependent variable will take. Traditionally, this would be an accident rate and it was investigated and was found to be the best variable to be used as the dependent variable. Then once the dependent variable was determined a prediction model was developed.

5.1 Accident Rate Analysis

Most traffic and safety engineers take a great deal of their information about a road's safety from its calculated accident rate. An accident rate is a mathematical representation of the relationship between the major factors that influence accidents. This rate allows comparison between different sites, by normalizing the number of accidents on the road by time, length, and volume. If one road has many accidents and a very large volume it can have a lower accident rate and therefore be deemed safer than another road with fewer numbers of accidents but a greatly smaller ADT. Accident rates are usually expressed as a ratio of the number of accidents divided by the amount of travel for a comparable mix of mitigating factors. The amount of travel or exposure measures the number of opportunities available for each accident to occur (Saccomanno & Buyco 23). "The relationship between accidents and traffic flow, the most common measure of exposure, has been shown actually to follow a nonlinear relationship, in which accident counts usually increase at a decreasing rate as traffic flow increases" (Lord 17). Due to this relationship between accident rate and assumed level of safety, the mathematical relationships that go into accident rates were investigated, including total number of accidents per segment, ADT, time period of the study and segment length. One significant issue that occurs when looking at traditional accident rates is that the

numerator and denominator in accident rates are both random quantities that can contribute to the overall uncertainty about accident rates. Accident counts have been found to be an inaccurate estimation of safety since they are usually random and independent events (Lord 18). Since more exact data is not available, these inexact figures must be used.

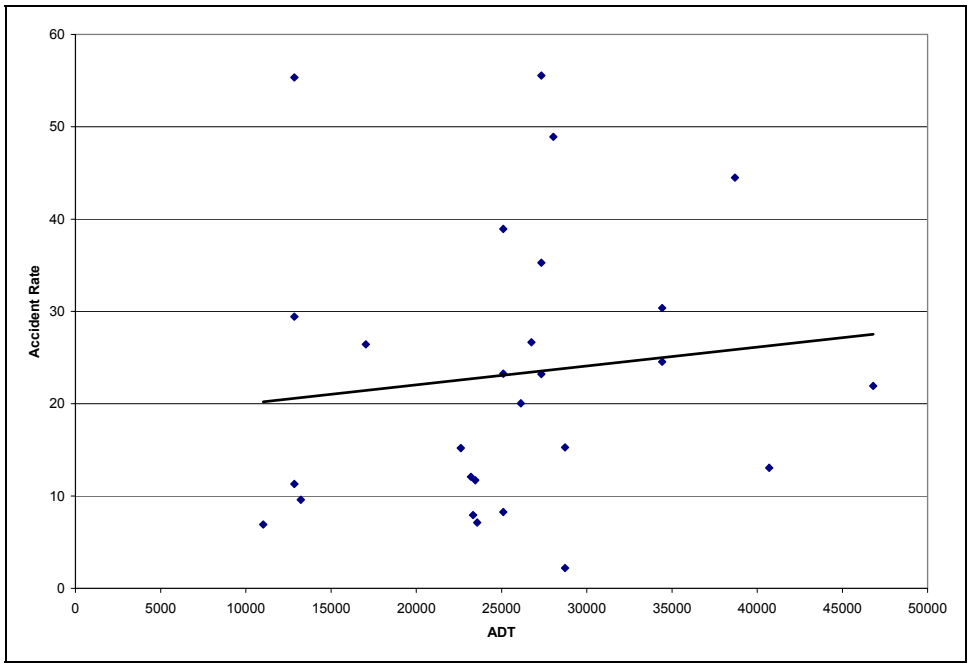


Figure 25: Accident Rate vs. ADT with Linear Trend Line

Figure 25 shows the relationship between accident rate in accidents per million vehicle miles and ADT. The trend line helps to show that as volume increases the number of accidents increase. This is a linear trend line to give the general impression of how the data is represented. The large amounts of scatter make a more specific relationship difficult to assess with the Worcester data.

5.1.1 Linear Accident Rate Analysis

Working towards the goal of finding each segment's accident rate, the first thing that was examined was the linear relationship between the traditional variables involved, specifically the relationship between total number of accidents, volume, and length.

5.1.1.1 Accident Rate and Volume

Volume versus total number of accidents per segment was the first relationship examined. A linear model predicting the total number of accidents for the entire study period per segment by volume was developed: $Acc = 54.67658 + 0.00199\,ADT$. The parameter estimate for volume (ADT) is positive which means that the higher the volume becomes, the more accidents there will be. This is to be expected because the more vehicles that are present on the road the more possibilities exist for conflicts between the different movements of the vehicles. A problem with this model is that if there is no traffic (ADT=0) the model still predicts accidents. Numerically this is not a problem, but in practice if there are no vehicles on the road, no traffic accidents can take place on the road. The Analysis of Variance (ANOVA) table given below gives the highlights of the model. The coefficient of determination is only 0.0853, showing that volume alone is not a good representation of the variability in the data.

Table 6: ANOVA Table for Total Number of Accidents and Volume

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	1	7894.04451	7894.04451	2.33	0.1393
Error	25	84631	3385.2463		
Corrected Total	26	92525			
Root MSE	58.18286		R-Square	0.0853	
Dependent Mean	105.25926		Adj. R-Sq	0.0487	
Coeff Var	55.27577				

One way to see what is happening with a linear regression is to plot the regression line in relationship to the points from which it was formed. This allows the viewer to see if there are any outlying points that are affecting the regression line or if there are any patterns that could be taking place. Including confidence bands on this plot also allows for an observer to see where points should be falling in order for the regression line to be a valid reflection of what is occurring. Figure 26 shows the regression line, the actual points, and the 95 percent confidence bands. The 95 percent confidence bands present with the regression line show the location of where with 95 percent confidence the true regression line of this relationship lies. The use of only volume does not seem to be the best idea for a relationship, as most of the points, showing the actual data, fall far outside of the confidence bands.

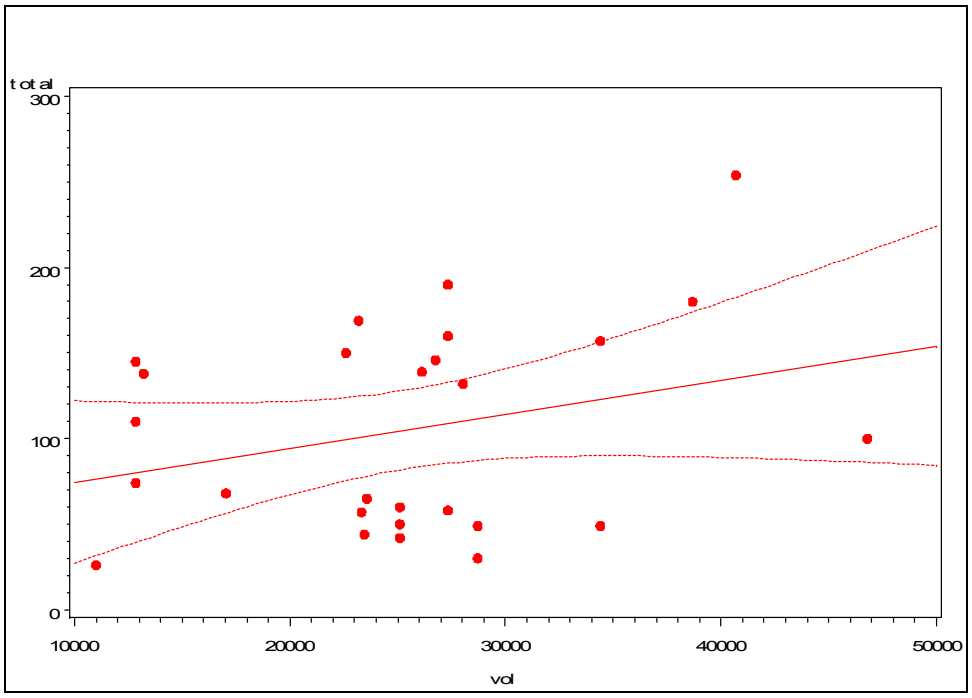


Figure 26: Confidence bands for Regression of Total Number of Accidents and Volume

The assumptions of any model need to be tested in order to determine if the model is an appropriate way to look at the relationships in question. An assumption of linear

regression is that the variables follow a normal distribution. The plot of the predicted values versus the residuals is a good way to see if any deviation from normality exists. By examining Figure 27, there does not appear to be a strong deviation from normality (i.e. there points do not form a pattern) and the variance appears to be fairly constant (i.e. the points lie within a constant band around zero) with the model using only total number of accidents and volume. Constant variance is another assumption in linear modeling. There is a certain amount of symmetry in the residuals with half falling above and half falling below the zero line. No obvious outliers can be identified by lying far from the majority of the points, which are all good indications of following the model assumptions.

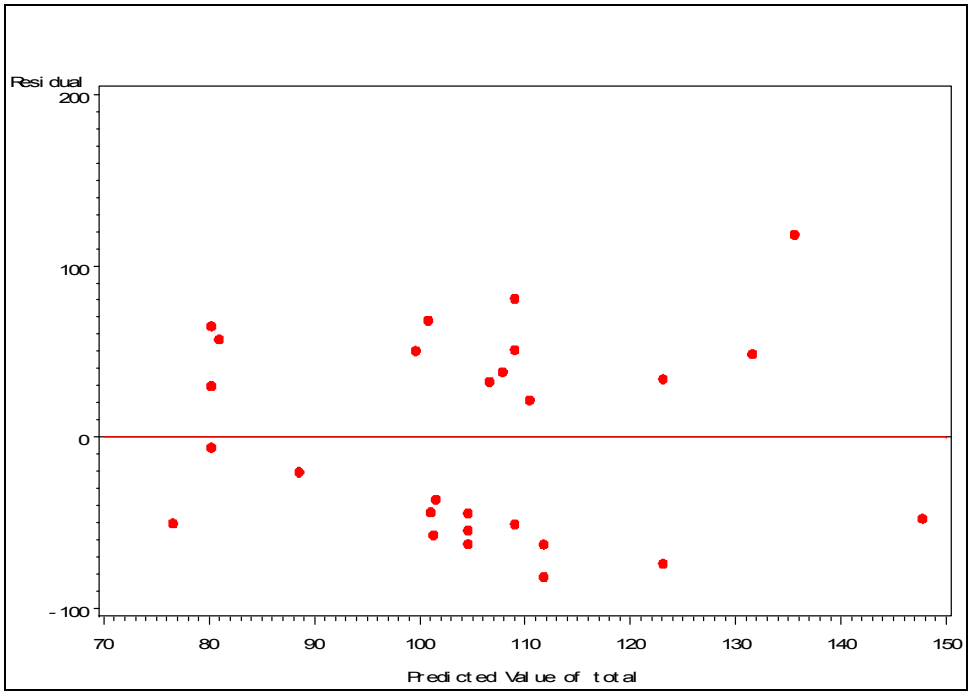


Figure 27: Predicted Values vs. Residuals for Total Number of Accidents and Volume

The normal probability plot in Figure 28 also shows that there is not a significant deviation from normality. The solid line is the normal probability distribution. The dashed line is the distribution from the data set and the histogram is also from the data

set. The model has a distribution with a flatter and lower peak value and a slightly wider base than the normal distribution. These minor departures could also be due to the small sample size used for this investigation. A departure from normality would mean that a model of this functional form would be inappropriate for the given data. Since the dash line follows the solid line closely, normality is assumed.

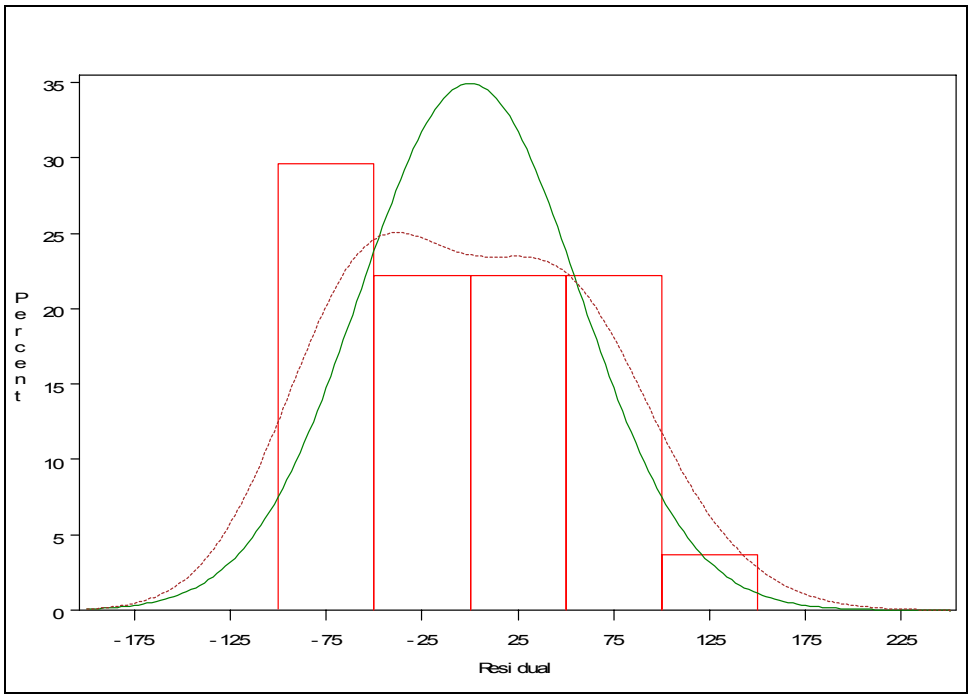


Figure 28: Normal Probability Plot for Total Number of Accidents and Volume

The investigation in the linear relationship between total number of accidents and annual daily traffic shows that while the relationship most likely is linear, there is some minor deviations from normality. Also found was that while there may be a relationship between total number of accidents and ADT, volume does not explain much of the variation that occurs in accident data.

5.1.1.2 Accident Rate and Length

Similarly to the investigation of volume versus total number of accidents, segment length versus total number of accidents per segment was examined with linear regression. A model of the form $Acc = 54.67658 + 0.00199Len$ was found. The parameter estimate for segment length is positive which means that the longer the segment is the more accidents there should be. This like the volume study is an intuitive conclusion as the longer the segment is, the more possibilities for vehicle conflicts. A problem that exists with this model is that if a segment has no length, that there are still accidents occurring. This is impossible in reality.

The analysis of variance table below shows some of the important statistics relating to this model. The coefficient of determination, most often used to compare models is equal to 0.0674 in this case. This shows that the use of length as an explanatory variable can explain only 6.74 percent of the variation in the data and also implies that there are most likely other variables that can explain some of the variation. This combination explains even less of the variation in the data than did the total number of accidents versus volume.

Table 7: ANOVA Table for Total Number of Accidents and Segment Length

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	1	6237.13397	6237.13397	1.81	0.1909
Error	25	86288	3451.52205		
Corrected Total	26	92525			
Root MSE	58.74966		R-Square	0.0674	
Dependent Mean	105.25926		Adj. R-Sq	0.0301	
Coeff Var	55.81424				

Again, looking at a plot with the regression line, the actual points and 95% confidence bands, length alone is not a good indication of total accidents (see Figure 29). As with volume, most of the points fall outside of the confidence bands. This helps to show that a better model is most likely needed to explain the majority of the variation in accident data.

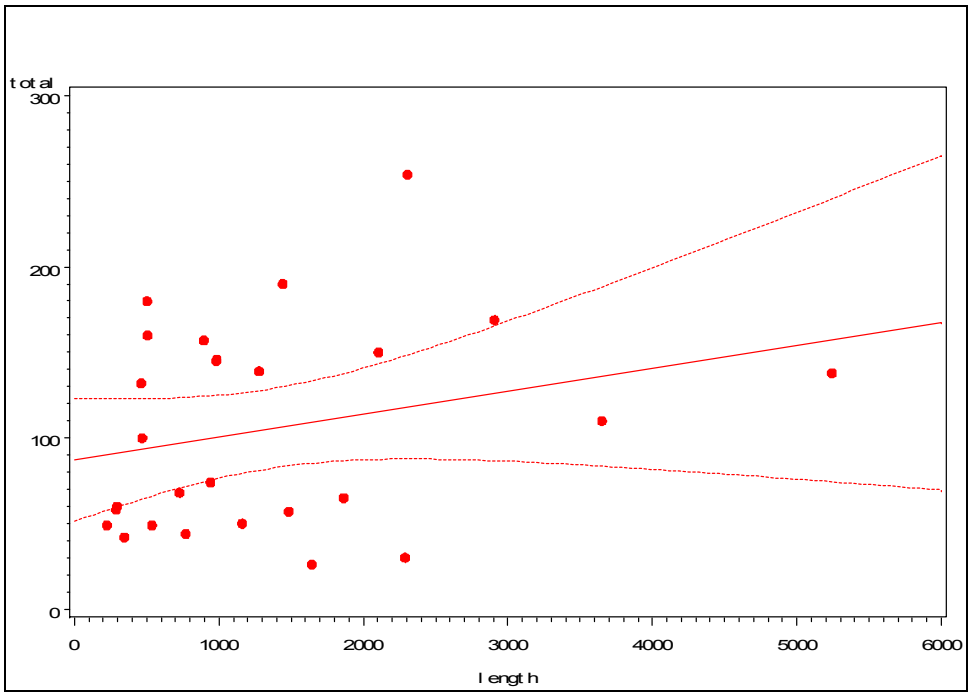


Figure 29: Confidence Bands for Regression of Total Number of Accidents and Segment Length

Checking the model assumptions, as with total number of accidents and volume, there does not appear to be a strong deviation from normality in the predicted versus residual plot in Figure 30. One point appears to be located further away than the others, but not enough to be called an outlier. There appears to be a constant variance, as the points lie in a mostly constant band around zero, which is one of the assumptions for linear modeling. The clustering of the points on the left side of the graph has to do with the selection of the data points rather than with systematic departures from the basic

assumptions. These observations indicate that linear modeling is an acceptable way to look at this relationship.

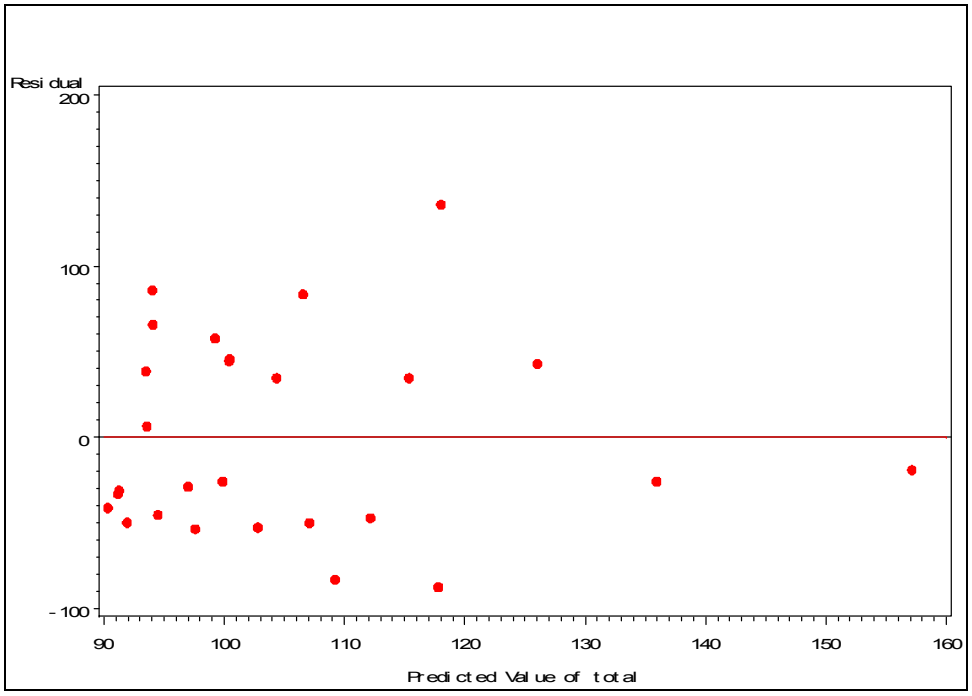


Figure 30: Predicted Values vs. Residual for Total Number of Accidents and Segment Length

The probability plot below does not appear to have a strong deviation from normality. The solid line is the normal distribution. The dashed line is the distribution of the residuals and the histogram is of the residuals. The peak of the distribution from the model is further towards the left than the normal distribution as is the base of the distribution. Since there are only minor departures from normality, the plot shows that the data most likely follows a normal distribution, meaning that a linear relationship is present and the model assumptions hold true.

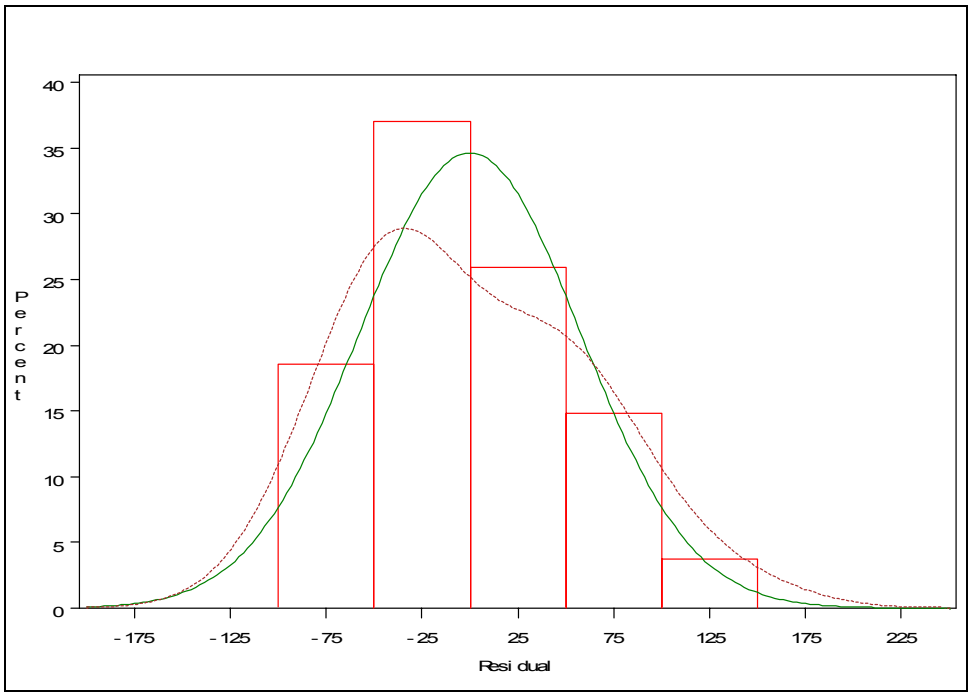


Figure 31: Normal Probability Plot for Total Number of Accidents and Segment Length

The normal quantile plot also reveals a small departure from normality, but this departure could be explained by the use of other explanatory variables (See Figure 32). The solid line shows where the data points would be for perfect normality and the dotted line shows where the data is actually located. This small amount of deviation is not a large concern, but with a larger data set, could prove to be showing that the data is not truly linear.

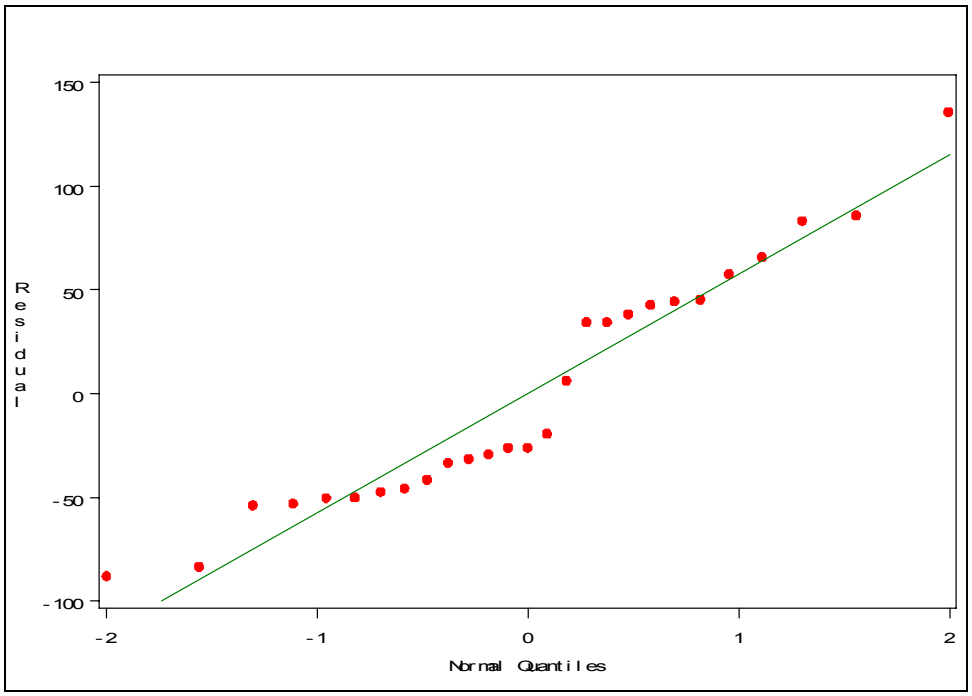


Figure 32: Normal Quantile Plot for Total Number of Accidents and Segment Length

There is a small possibility that the total number of accidents and segment length do not have a linear relationship, but there is no doubt that segment length alone does not describe an adequate amount of the variation in the crash data. The relationship may be linear, but the models formed by both segment length and traffic volume alone, do not correctly represent what happens in actual situations. The fact that according to the two above models developed, accidents can occur when there is no traffic volume on the road or not length to the segment is worrisome. This means that further steps must be taken in looking at accident rates.

5.1.1.3 Accident Rate with Length and Volume

As both volume versus total number of accidents and segment length versus total number of accidents appear to follow a normal distribution but do not explain a large amount of the variation in the data a model was developed that combined the two explanatory variables in one model. The parameter estimate for length is positive which

means that the longer the segment is the more accidents there should be. The coefficient for ADT is also positive which means that the more traffic the more accidents occur. These are the expected values for the sign of each of the two coefficients. By combining these two variables into one equation, much more of the variability in the model is explained. Individually, just using length explained 6.74 percent of the variation and just using volume as an explanatory variable explained 8.53 percent of the variation in the model. Using both variables in the model increased the variation explained to 25.4 percent, which is more than the individual amounts combined. Key numbers, including the coefficient of determination can be seen in Table 8.

Table 8: AVOVA Table for Accidents, Segment Length and Volume

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	2	23545	11773	4.10	0.0295
Error	24	68980	2874.15623		
Corrected Total	26	92525			
Root MSE	53.61116		R-Square	0.2545	
Dependent Mean	105.25926		Adj. R-Sq	0.1923	
Coeff Var	50.93249				

The regression procedure found the following formula to be representative of the given data. $Acc = -7.48971 + 0.02302Len + 0.00321ADT$. Both predictor variables, *Len* and *ADT*, have the expected positive sign, but the intercept term is problematic. The negative intercept shows that if there was no volume and no segment length there would be negative accidents. This is not possible in reality, so this cannot be used to show the relationships between the total number of accidents, segment length and traffic volume. The significance of each of the three parts of the equation can be tested using statistical methods, which show that while the parameters for segment length and ADT are

significant to greater than five percent, the intercept term is not significant and does not help in explaining the variation in the data as shown in Table 9.

Table 9: Parameter Estimates for Accidents, Segment Length and Volume

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	-7.48971	41.80557	-0.18	0.8593
Length	1	0.02302	0.00987	2.33	0.0283
Vol	1	0.00321	0.00131	2.45	0.0218

To check that the model assumptions are met, the predicted values versus the residual values were examined in Figure 33. This residual plot shows that the there is not a substantial departure from normality in the data. There is no discernable pattern in the points and they are evenly distributed between positive and negative values. A constant variance can be seen, by the points being distributed in two constant bands, above and below zero. One point falls slightly further away then the rest at –110 but this remains close enough to not be considered an outlying point and not be considered a departure from a constant variance. This plot allows for the linear modeling assumptions to be met, and for linear regression to be an adequate representation of this particular data.

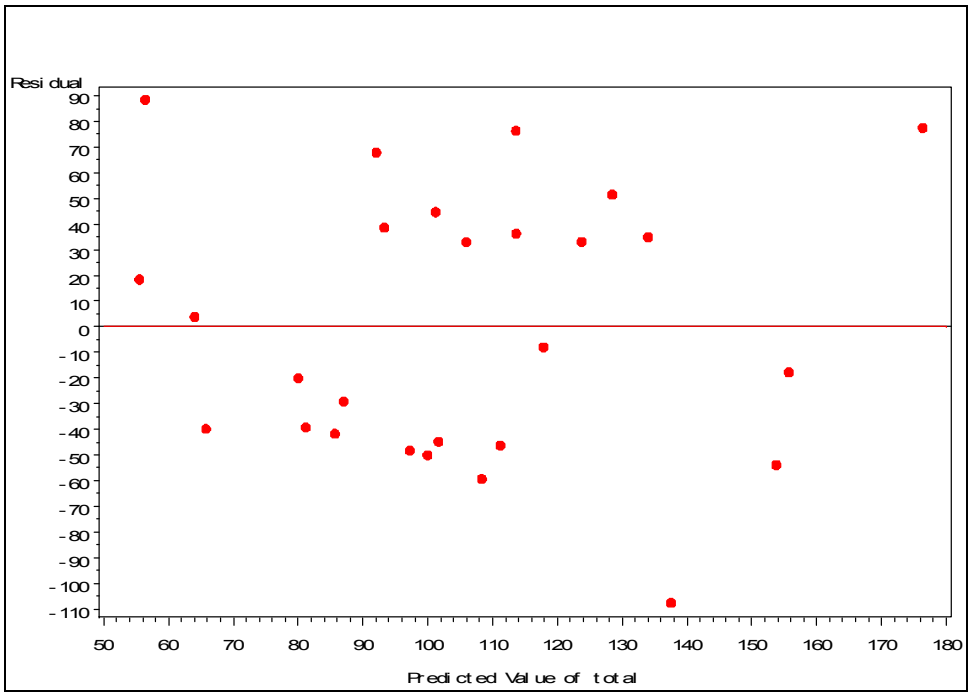


Figure 33: Predicted Values vs. Residuals for Accidents, Segment Length and Volume

Similarly, the boxplot of the residuals shows that they are evenly distributed by the plot being symmetric (See Figure 34). The symmetry helps to confirm that the choice of a linear distribution was appropriate. This also helps to show that no one point is a major outlier and affecting the overall model. There is a slightly larger variation of residuals on the negative side.

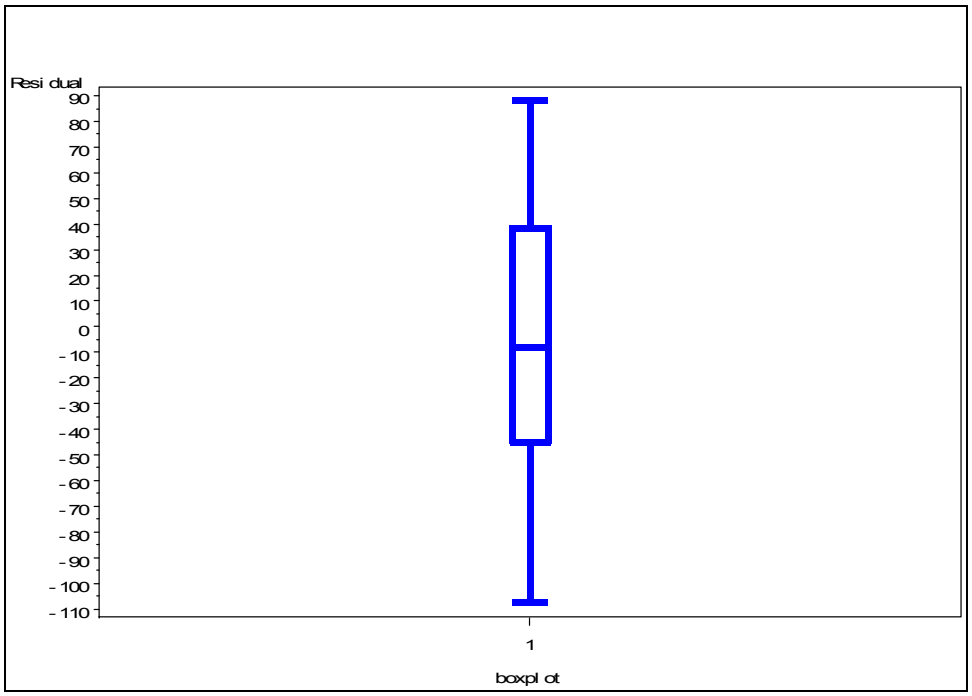


Figure 34: Boxplot of Residuals for Accidents, Segment Length and Volume

The normal quantile plot, shown in Figure 35, demonstrates that there may be some minor deviations from the normal distribution. The solid line represents normality and the dotted line represents the actual data. There is a minor pattern that may be explained by a sinusoidal wave, or could be natural variation in the given data set. The departure from normality, however, is not enough to cause the linear relationship to be entirely disregarded. But due to previous investigations there is a non-linear relationship between accident rate and especially traffic volume. That is what is most likely causing the data to not fully follow a normal distribution, but due to the small data set, the non-linear relationship discussed by Lord (Lord 17) cannot be fully duplicated.

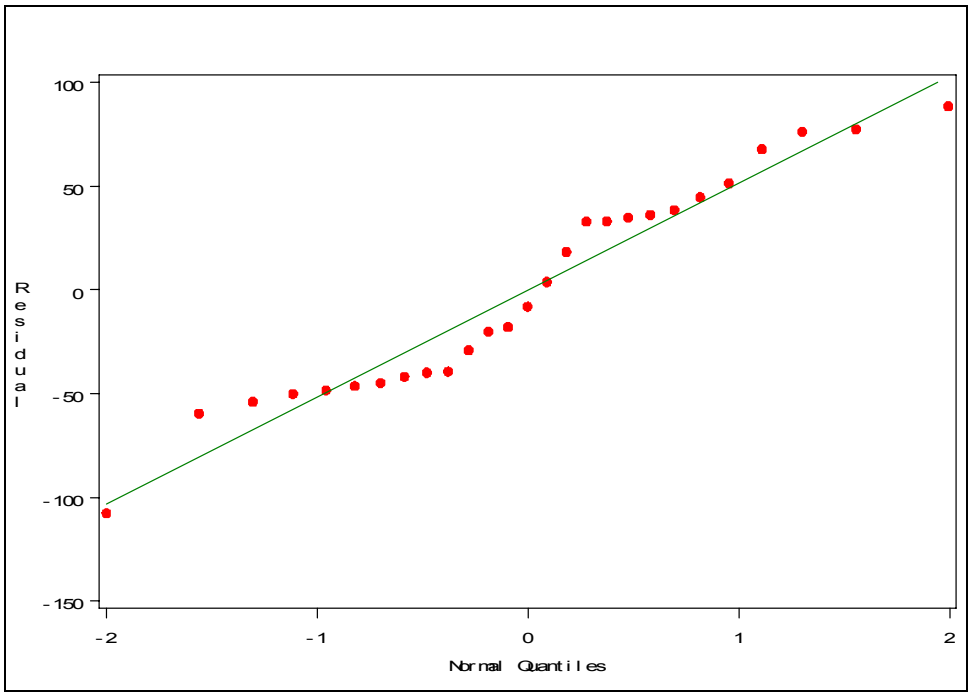


Figure 35: Normal Quantile Plot for Accidents, Segment Length and Volume

5.1.2 Accident Rate with Non-Linear Distributions

Due to the uncertainty about the relationship between length, volume and total accidents, these variables were examined under a Poisson distribution and a negative binomial distribution. The reason for exploring other distributions can from the issue that traffic accidents themselves are non-negative discrete counts that do not follow a normal distribution. Therefore distributions that consider count data as their basis were reviewed as possibly being more appropriate for predicting the number of accidents.

5.1.2.1 Accident Rates with Poisson Distribution

The model developed that used the Poisson distribution showed a large amount of overdispersion, which is an indication that the mean is very different from the variance. This violates a very basic model assumption. The deviance divided by the degrees of freedom shows this quality. This value was 27.7869. A value of one indicates that there is not a problem of overdispersion; the larger the value, the greater the variance and mean

differ. This can be seen in Table 10. This is also an indication that the data does not adequately fit this functional type of model.

Table 10: Criteria for Assessing Goodness of Fit for Accident Rates using a Poisson Distribution

Criterion	DF	Value	Value/DF
Deviance	24	666.8856	27.7869
Scaled Deviance	24	24.0000	1.0000
Pearson Chi-Square	24	644.1454	26.8394
Scaled Pearson X2	24	23.1816	0.9659
Log Likelihood		377.8728	

The model using the Poisson distribution is as follows:

$$Totalaccidents = 3.5914 + 0.0002length + 0.0000vol\ .$$

All of the variables are significant to greater than 95 percent. This can be seen in Table 11. The confidence limits also show that there is a possibility that the coefficients for both segment length and volume can be zero, which is a questionable result: having a coefficient of zero means that the variable in question does not affect the number of accidents that occur. Based on observation, the idea that volume and segment length have no effect on the number of accidents that occur is ludicrous. Since the model assumptions do not hold true this relationship is invalid.

Table 11: Analysis of Parameter Estimates for Accident Rates using a Poisson Distribution

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr> ChiSq
Intercept	1	3.5914	0.4115	2.7848	4.3980	76.16	<0.0001
Length	1	0.0002	0.0001	0.0000	0.0004	5.55	0.0184
Volume	1	0.0000	0.0000	0.0000	0.0001	6.04	0.0140
Scale	0	5.2713	0.0000	5.2713	5.2713		

5.1.2.2 Accident Rate with Negative Binomial Distribution

Using the negative binomial distribution to model length, volume and total accidents allows for the problems of overdispersion to be overcome. The model is almost identical to that which follows the Poisson distribution, but the problem of overdispersion

is almost completely overcome. $Totalaccidents = 3.6605 + 0.0002length + 0.0000vol$ The coefficients are very similar, but the elimination of the overdispersion problem, makes the data better fit this distribution. The deviance divided by the degrees of freedom value is 1.1713, which is a very low value, making this a very good model for these variables (See Table 12). A value of 1.0 would show that there is no problem of the variance being greater than it is allowed to be.

Table 12: Criteria for Assessing Goodness of Fit for Accident Rates using a Negative Binomial Distribution

Criterion	DF	Value	Value/DF
Deviance	24	28.1119	1.1713
Scaled Deviance	24	24.0000	1.0000
Pearson Chi-Square	24	24.9062	1.0378
Scaled Pearson X2	24	21.2632	0.8860
Log Likelihood		9199.8261	

The variables are almost significant to the 95 percentile, with volume being 3.99 percent and length being 5.61 percent. This can be seen in Table 13. Again as with the model developed using the Poisson distribution, the 95 percent confidence limits show that the coefficients for both segment length and volume have a chance of being zero, but as zero is at the lower limit of the confidence band is not a likely situation. Both models, using the Poisson distribution and the negative binomial distribution, however, do not provide a good method for constructing an accident rate.

Table 13: Analysis of Parameter Estimates for Accident Rages using a Negative binomial Distribution

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr> ChiSq
Intercept	1	3.6685	0.4185	2.8483	4.4886	76.85	<0.0001
Length	1	0.0002	0.0001	-0.0000	0.0004	3.65	0.0561
Volume	1	0.0000	0.0000	0.0000	0.0001	4.22	0.0399
Dispersion	1	0.2546	0.0755	0.1424	0.4551		

5.1.2.3 Accident Rate with Natural Logarithm

In hopes that the accident rate can be reconstructed, a model using the natural logarithm of volume, length and total number of accidents was developed. This was done assuming the variables all followed a normal distribution. By using $N = A(ADT)^{B}(length)^{C}$ as the base model where N equals the total number of accidents, and length equals the segment length, if the coefficients are found to be equal to approximately positive one (i.e. B=C=1), then that will show that the traditional formula for accident rates is valid. Since $\frac{N}{ADT * Length} = rate$ to equate the accident rate to the model $N = rate * ADT * Length$ where the coefficient A is equal to the accident rate and B and C should be approximately positive one. For ease of modeling, the following is what was actually modeled: $\ln(N) = \ln(A) + B\ln(vol) + C\ln(length)$. The model then gives values for each of the predictive variable's coefficients. The model that resulted from this is the following: $\ln(totalaccidents) = -4.43339 + 0.66394\ln(vol) + 0.32276\ln(length)$. The coefficient of determination of this model is 0.1915, which means that 19.15 percent of the variation in the variables is explained by this model. This model does not explain all of the variation that occurs in the data, but the rest can hopefully be explained by additional variables. See Table 14 for more detailed numerical analysis.

Table 14: ANOVA Table for Accident Rates with Natural Logarithm

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	2	1.94447	0.97224	2.84	0.0780
Error	24	8.20810	0.34200		
Corrected Total	26	10.15257			
Root MSE	0.58481		R-Square	0.1915	
Dependent Mean	4.48268		Adj. R-Sq	0.1242	
Coeff Var	13.04601				

As with other investigations above, the significance of the coefficients was examined. The natural logarithm of segment length and volume are significant to more than 90 percent, which is a common cut off point for including variables in a regression model. The parameter estimates and their F-values for the significance tests can be seen in Table 15.

Table 15: Parameter Estimates for Accident Rates with Natural Logarithm

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	-4.43339	4.01313	-1.10	0.2802
ln(Length)	1	0.66394	0.34040	1.95	0.0629
ln(Vol)	1	0.32276	0.15509	2.08	0.0483

This investigation results in having $N = e^{-4.43} ADT^{0.664} length^{0.323}$. Here the coefficients for B and C are not equal to positive one, but closer to positive one half. These were not the expected values, which implies that the traditional accident rate formula is not applicable to at minimum this data set and at maximum all accident data. The above investigations show that the traditional relationships used to calculate accident rates are not applicable to this data and another way of determining the accident rate or risk of an accident occurring must be found.

5.1.3 Accident Risk Analysis

The goal of the accident rate analysis is to be able to determine the safety of different road segments based on roadway and traffic characteristics. To compare segments, an accident rate tends to be more helpful than just an accident count. The rate that is being search for is the accident "risk" or the probability that a vehicle on a segment will be involved in an accident. The risk should be different for each road segment. Based on the above work, Poisson regression had severe overdispersion problems, so the negative binomial distribution was examined to try to overcome those problems. Use of the negative binomial distribution and natural logarithm did not appear to adequately describe how the accident data related to ADT and segment length. The earlier linear regression was also not helpful in describing the relationships between volume, length and number of accidents.

The preliminary problem is determining the risk of an accident on an individual segment. This has traditionally been accomplished by using an accident rate. The above analysis has shown that with this data, this is not an adequate way to describe the accidents that occur on the segments. Instead, an accident risk will be used. This is the probability of an accident occurring to an individual vehicle on the segment. Each occurrence of an accident is an independent action. There are a known number of accidents that occur on each segment over the three year time period. There are also a know number of trials, or possibilities of accidents over the three year time period, which is the total number of vehicles that have passed through the segment which is calculated by an accurate estimation of the volume by multiplying the ADT by 365 days per year by three years.

With a known number of trials and known number of successes, or accidents, the best way to determine the actual risk of an individual vehicle being in an accident is through the binomial distribution. The binomial distribution is often used to find the probability of an event with a given number of trials and successes. The binomial distribution deals with independent events, which is true with accident occurrences. The risk of an accident is equal for any passing vehicle and each vehicle has an equal chance of being in a crash.

The traffic volume ranges from 11,000 to 47,000 vehicles per day. Time is constant over all the segments, with each segment lasting three full years. This allows the number of trials per segment to vary from twelve to fifty-one million vehicles. The number of accidents per segment similarly has a large amount of variation between 26 and 254 accidents per segment. The binomial distribution's probability mass function is $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$. There are n trials and k successes. Since this is a distribution, there are infinite possibilities for what the actual probability is. However, the best point estimate, which will be used to identify the risk of an accident occurring for an individual vehicle, is $\frac{k}{n}$. The best point estimate allows for the most likely probability on each segment to be used as the accident risk for each road segment.

The risk for an accident to occur varies according to the roadway segment. These risk range between $9.537*10^{-7}$ and $1.03*10^{-5}$. After further consideration the accident risk was normalized by length converting it back into the more traditional accident rate.

5.2 Accident Risk Prediction Model Development

The first step in the model development was reducing the number of variables to a workable number. The combinations of the variables can be made to produce the best possible model.

5.2.1 Primary Elimination of Variables

Since the data set has a relatively small number of data points, and there exist a potentially large number of variables, some of them need to be eliminated early on in the development process. The primary elimination was to look at groups of variables and remove the ones that do not help explain variation in the data. The fifty-six primary variables were divided up into groups, which have similar characteristics. The variables were divided into six major groups to try and to an initial elimination of variables that do not have a large influence on the data. The groups consist of hazard variables, cross-section variables, traffic characteristic variables, horizontal and vertical alignment variables, access variables and the remaining variables. Each group is examined individually to see if there are any variables that can quickly be eliminated to help lower the number of possible variable to consider for the final model to a more workable size.

5.2.1.1 Variables Relating to Roadside Hazards

There are many of variables that relate to the number and type of roadside hazard. It was decided to try and determine which were the most influential and important of these variables to include in a prediction model that includes the influence of more than just the roadside hazards. Using a selection process of the adjusted coefficient of determination, the variables were compared in multiple combinations to determine the optimum combination. The adjusted coefficient of determination adjusts R^2 by dividing

each sum of squares by its associated degrees of freedom. The adjusted coefficient may actually become smaller when additional X variables are introduced into a model, because any decrease in the error sum of squares may be more than offset by the loss of a degree of freedom in the denominator (Neter et al 231). This is what makes comparisons by adjusted coefficient of determination fit better than comparisons by just the coefficient of determination.

Due to the goal of finding hazard variables of most interest, more possible models other than the model with the greatest adjusted coefficient of determination were examined. The top models sorted by adjusted coefficient of determination were examined to show which variables were used most often in these models. All seventeen possible hazard variables were included in the top models, but as the reasoning for looking at these was to eliminate some possible variables, an in depth look at the variation of the use of the variables was done. The variables *hydrant* (number of fire hydrants on the segment) and *benches* (number of benches on the segment) were included in all the top models. Variables *upole* (number of utility poles on each segment), *building* (number of buildings on each segment), *ospole* (number of overhead sign poles on each segment), and *hazards*, representing the total number of hazards were found in more than eighty percent of the top models. The other variables that were used in more than fifteen percent of the models were *electrical* (number of electrical/traffic control boxes), *pmeter* (number of parking meters), *fence* (number of fences), *trees* (number of trees), *pole* (number of telephone poles, light poles, and sign poles), *spole* (number of sign poles) and *density* (the number of hazards per mile). Some of the variables that were

excluded from further consideration include the counts of mailboxes, stone monuments, rocks, and light poles on each segment.

Table 16: ANOVA Table for the Best Model using only Hazard Variables

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	6	3898.85678	649.80946	6.35	0.0007
Error	20	2048.15673	102.40784		
Corrected Total	26	5947.01352			
Root MSE	10.11968		R-Square	0.6556	
Dependent Mean	23.14741		Adj. R-Sq	0.5523	
Coeff Var	43.71840				

The model that had the largest adjusted coefficient of determination for hazards included just six variables: *hydrant*, *upole*, *benches*, *building*, *ospole*, and *hazards*. *Hydrant* is the total number of fire hydrants on the segment while *benches* is the total number of benches observed on the road segment. *Upole* is the number of utility poles on the road segment while *ospole* is the total number of overhead sign poles observed on the segment. *Hazards* is the variable that represents the total number of roadside hazards observed and building represents the number of buildings throughout the segment. The adjusted coefficient of determination for this model is 0.5523; meaning that 55 percent of the variation in the model can be explained by this model and the coefficient of determination is 0.6556. These and other informative numbers can be seen in Table 16. The coefficients for the different variables may not be what were actually expected (hazards and hydrant had negative coefficients), but the model is not of what was of primary interest in this situation (See Table 17). The model was mainly to show what hazard variables are of main interest.

Table 17: Parameter Estimates for the Best Model using only Hazard Variables

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	26.47713	5.26992	5.02	<0.0001
Hydrant	1	-4.14853	1.91712	-2.16	0.0427
Upole	1	0.79928	0.43373	1.84	0.0802
Benches	1	7.28857	3.69220	1.97	0.0624
Building	1	0.69990	0.30342	2.31	0.0319
Ospole	1	5.74825	1.46561	3.92	0.0008
Hazards	1	-0.23315	0.12022	-1.94	0.0667

Some further analysis was done primarily to confirm that that best model from this group followed the basic model assumptions. Figure 36 shows the distribution of the residuals for this model. This figure shows that the residuals are basically evenly distributed about zero with approximately half falling above and below zero. Normally distributed residuals are a sign that the data fits the normal probability model.

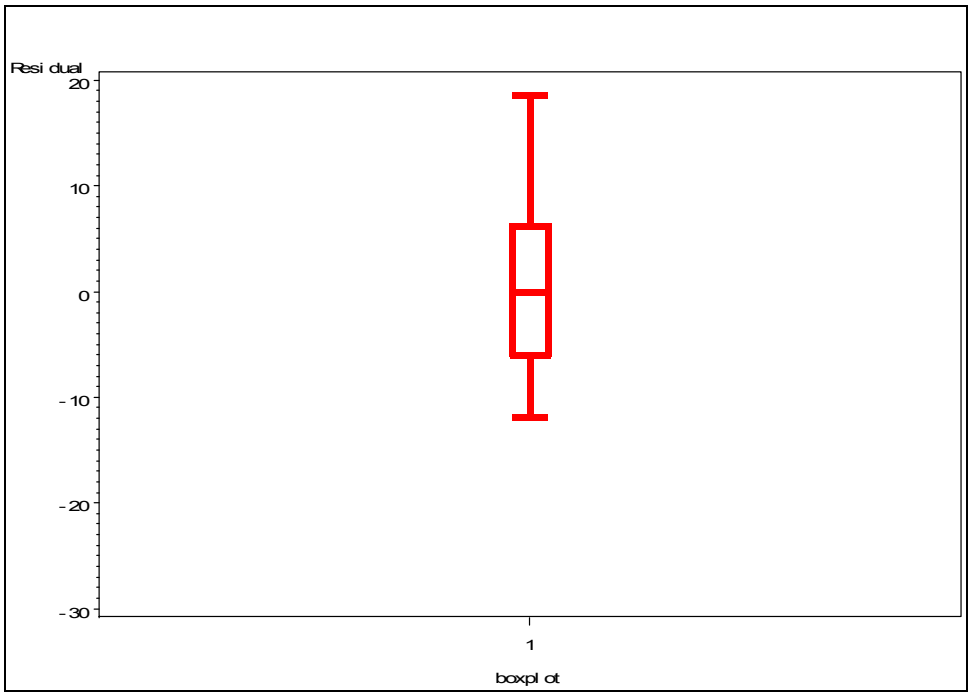


Figure 36: Boxplot of Residuals for the Best Model using only Hazard Variables

An assumption when dealing with multiple linear regression is that the data follows a normal distribution and the variance is constant. The graph in Figure 37 shows the studentized residuals versus the predicted values for this model. This conveys that

there is a constant variance in this model. The studentized residual plot helps to show that there are no severe outlying data points. A heuristic for outliers is that if they are greater than four in the studentized residual plot then the point could be considered an outlier. None of data points follows that heuristic.

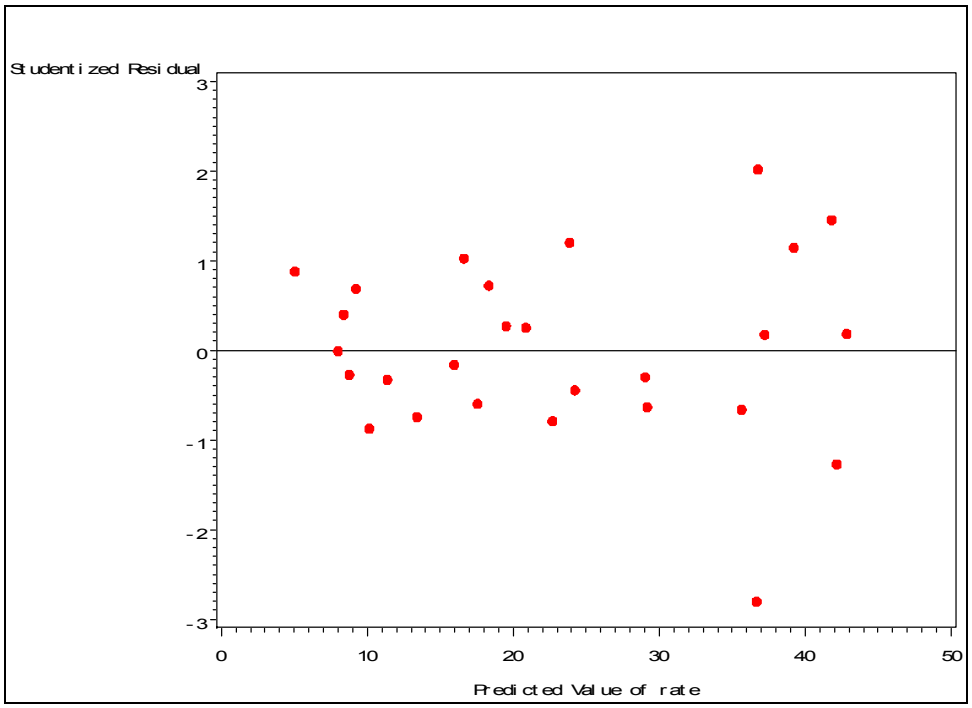


Figure 37: Residuals and Studentized Residuals vs. Predicted Values for the Best Model using only Hazard Variables

Another way to visually check that the data follows a normal distribution is to look at the normal probability plot (See Figure 38). The solid line is the normal probability distribution, while the dashed line represents the distribution that can be developed using the data from the model. The two lines almost exactly line up, showing that using the normal probability distribution was a good assumption for this data.

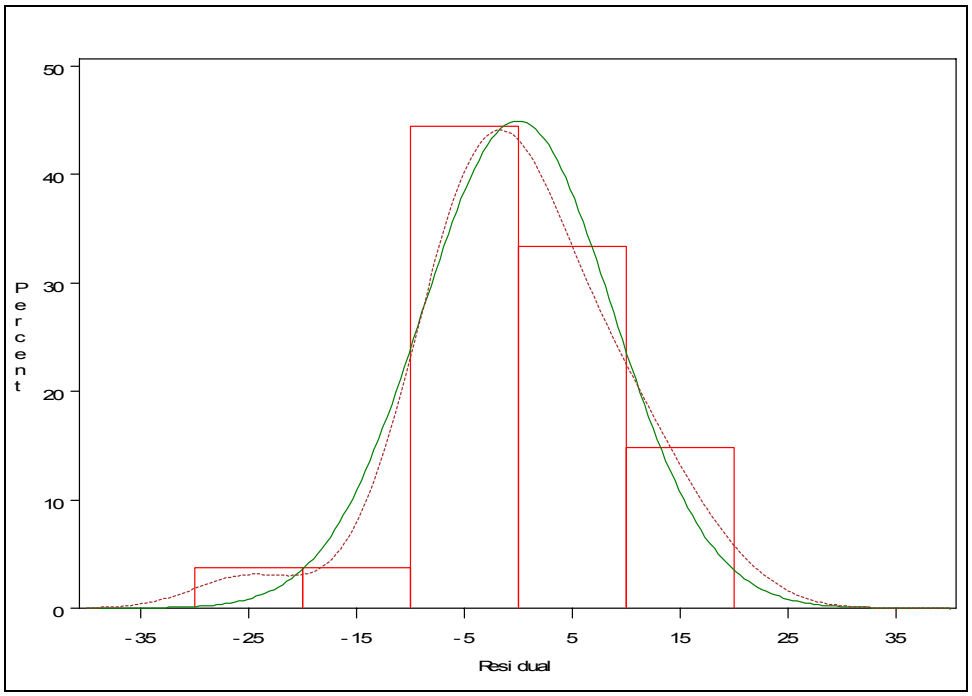


Figure 38: Normal Probability Plot for the Best Model using only Hazard Variables

Similarly the normal quantile plot is effective in showing when the data does not follow a normal distribution. When the assumption is correct, the residuals fall along the straight line. If the assumption is wrong, the residuals will not fall along the straight line, but may follow a different pattern. Figure 39 show that the residuals fall along the straight line, showing that the assumption of normality is correct with using the hazard variables regressed against the rate variable.

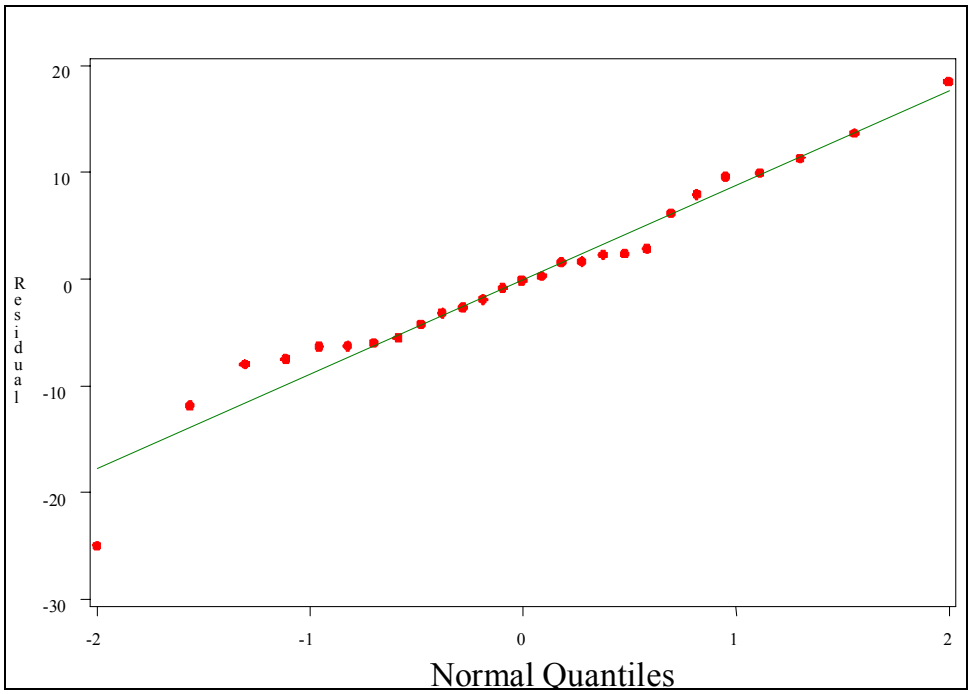


Figure 39: Normal Quantile Plot for the Best Model using only Hazard Variables

The best model using only hazard variables does follow all the assumptions of linear regression. This shows that this is so far a good choice of distributions for this data set and allows the four variables to be removed from further consideration since hazard variable models are normal in distribution.

5.2.1.2 Variables Relating to Cross-Section Alignment

There are also many variables that relate to the different elements that compose cross-sectional alignment. Of the nineteen identified variables, it was felt that some of them would not have strong influences on accident rates. It was decided to try and eliminate the least influential of these variables. Using a selection process of the adjusted coefficient of determination, the variables were compared in multiple combinations to determine the optimum combination.

The top models, sorted by adjusted coefficient of determination, were examined to show which variables were used most often in these models. Only eighteen of the

nineteen possible variables were present in the top models. The missing variable was *perpendicular*, which represents the amount of perpendicular parking on each road segment, however, this only occurred on one segment so was not expected to be influential. A further examination was made of the remaining eighteen variables. The variables of *widthsr* (width of the right shooulder), *widthsidl* (width of the left sidewalk), and *widthl2* (width of the second lane in the left direction) were included in more than 80 percent of the top models. Variables that appeared in more than fifteen percent of the top models were retained for inclusion in further model development.

Some of the variables that were excluded from further consideration include the percentage of parking, the number of lanes going in the right direction, and the width of the second and third lanes going in the right direction. By eliminating these variables, there is a more reasonable number of variables that are related to cross-sectional alignment to include in further model development.

Table 18: ANOVA Table for the Best Model using Cross-Section Variables

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	5	2474.03957	494.80791	2.99	0.0342
Error	21	3472.97395	165.37971		
Corrected Total	26	5947.01352			
Root MSE	12.86000		R-Square	0.4160	
Dependent Mean	23.14741		Adj. R-Sq	0.2770	
Coeff Var	55.55700				

The model that had the largest adjusted coefficient of determination for hazards included just five variables: *crest*, *llanes*, *widtha*, *widthsr*, and *widthsidl*. *Crest* is the maximum recorded value of the crest on each segment while *llanes* is the total number of lanes in the left direction on the road segment. *Widtha* is the average width of the lanes

on each road segment while *widthsr* is the width of the should on the right side of the road. It is interesting that this variable was shone to be such a significant one, since there was only one segment with a recorded shoulder. *Widthsidl* is the variable that represents the width of the left hand sidewalk. The adjusted coefficient of determination for this model is 0.2770, meaning that 27 percent of the variation in the model can be explained by this model and the coefficient of determination is 0.4160. These and other informative numbers can be seen in Table 18. The coefficients for the different variables may not be what were actually expected (*llanes* has a negative coefficient meaning that the more lanes in the left direction there are the fewer accidents occur), but the model is not of primary interest in this situation (See Table 19). The model was mainly to show what cross-section variables are of primary concern.

Table 19: Parameter Estimates for the Best Model using Cross-Section Variables

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	80.02889	32.53926	2.46	0.0227
crest	1	2.20930	1.77988	1.24	0.2282
llanes	1	-11.57391	6.94379	-1.67	0.1104
Widtha	1	-4.74875	1.58690	-2.99	0.0069
widthsr	1	-3.87422	2.01539	-1.92	0.0682
Widthsidl	1	2.77757	1.29274	2.15	0.0435

Some further analysis was done primarily to confirm that that best model from this group followed the basic model assumptions. Figure 40 shows the distribution of the residuals for this model in a boxplot. This figure shows that the residuals are basically evenly distributed about zero with approximately half falling above and below zero. Normally distributed residuals are a sign that the data fits the normal probability model.

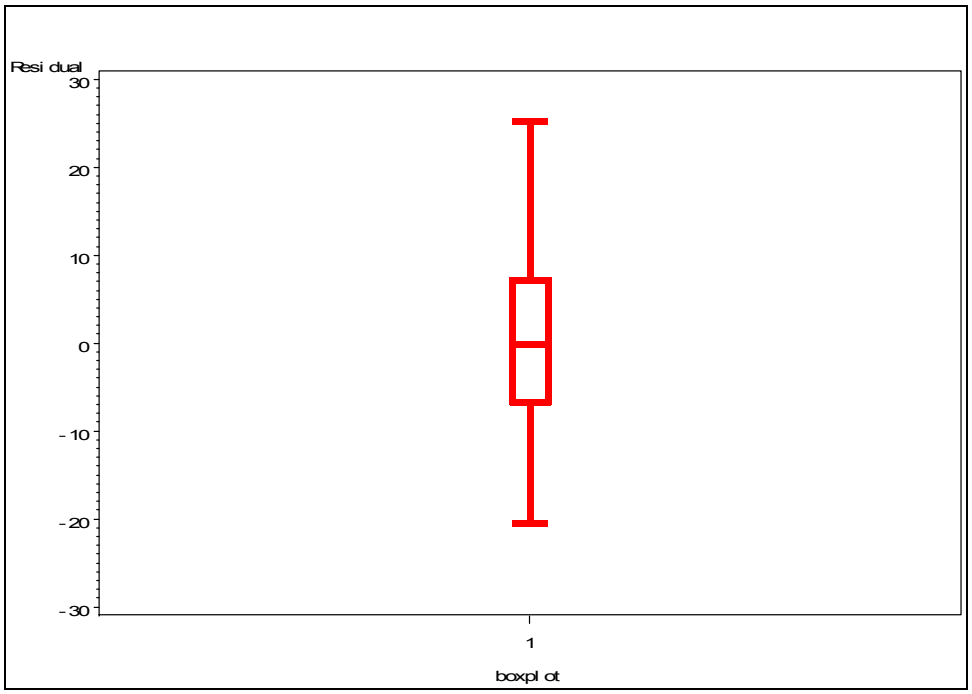


Figure 40: Boxplot of Residuals for the Best Model using Cross-Section Variables

An assumption when dealing with multiple linear regression is that the data follows a normal distribution and the variance is constant. The graph in Figure 41 shows the studentized residuals versus the predicted values for this model and shows that there is a constant variance in this model. The studentized residual plot also helps to show that there are no severe outlying data points. A heuristic for outliers is that if they are greater than four in the studentized residual plot then the point could be considered an outlier. None of data points follows that heuristic.

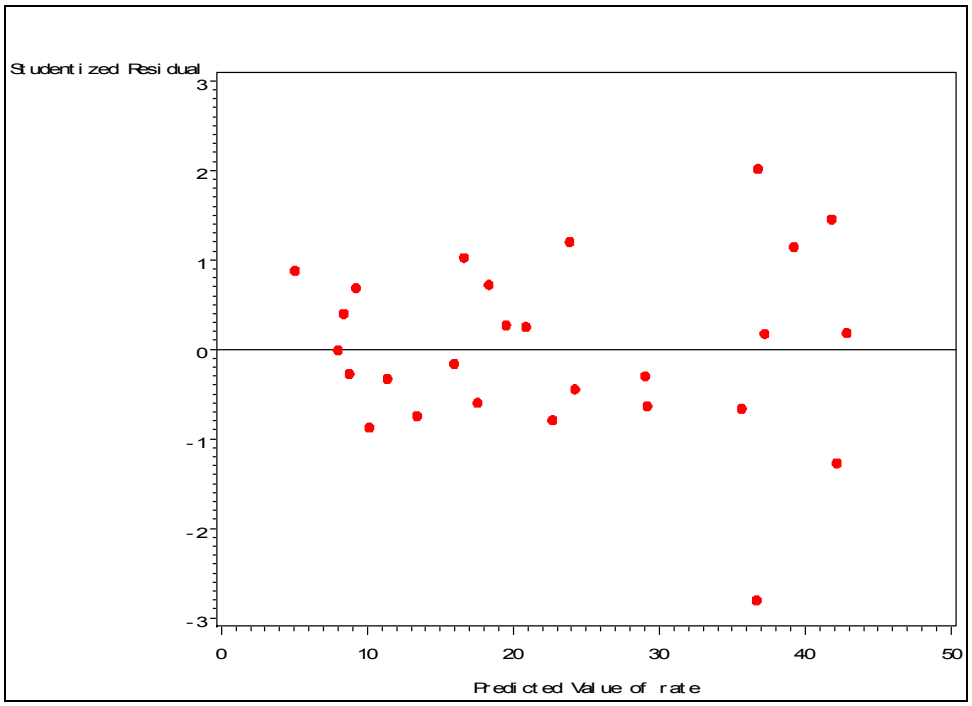


Figure 41: Studentized Residuals vs. Predicted Values for the Best Model using Cross-Section Variables

Another way to visually check that the data follows a normal distribution is to look at the normal probability plot (See Figure 42). The solid line is the normal probability distribution, while the dashed line represents the distribution that can be developed using the data from the model. The two lines almost exactly line up with the model's distribution peaking to the left of the normal distribution, showing that using the normal probability distribution was a good assumption for this data.

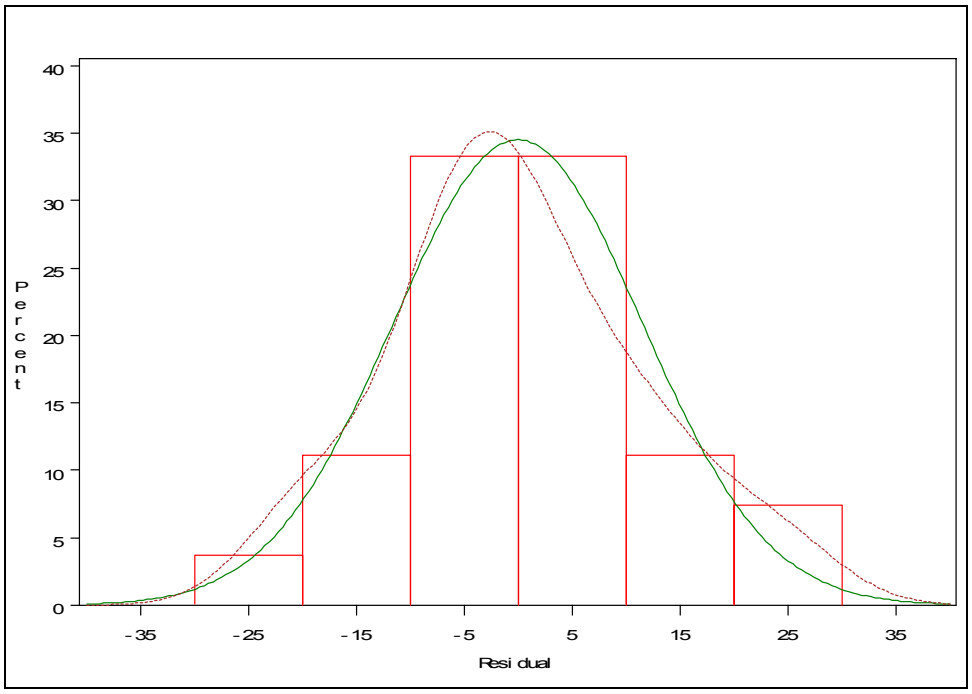


Figure 42: Normal Probability Plot for the Best Model using Cross-Section Variables

Similarly the normal quantile plot is effective in showing when the data does not follow a normal distribution. When the assumption is correct, the residuals fall along the straight line. If the assumption is wrong, the residuals will not fall along the straight line, but may follow a different pattern. Figure 43 shows that the residuals fall along the straight line, showing that the assumption of normality is correct with using the hazard variables regressed against the rate variable.

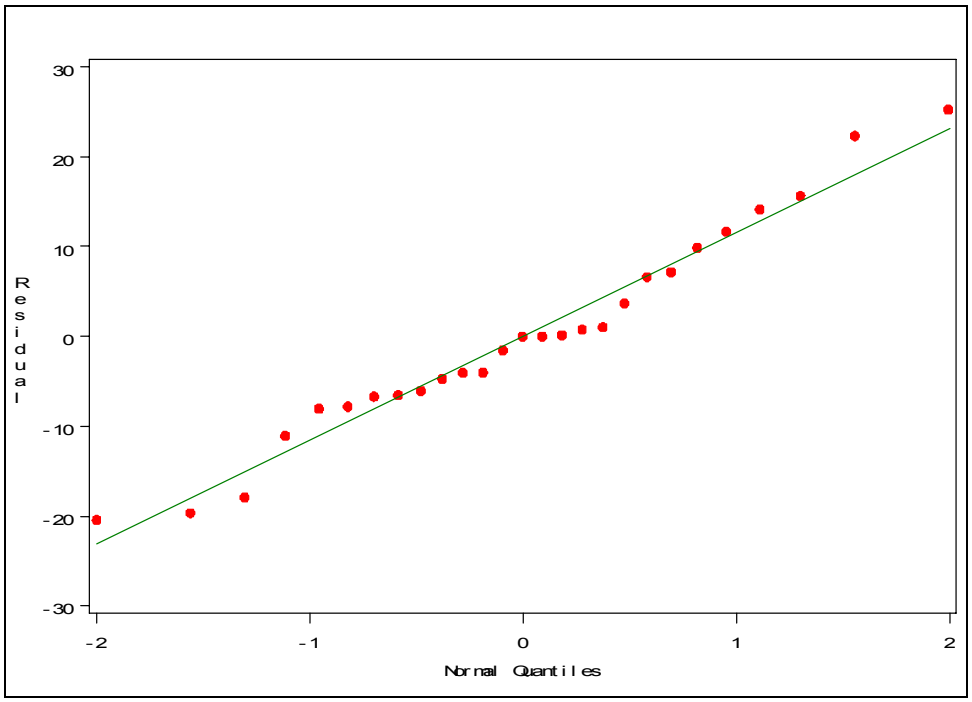


Figure 43: Normal Quantile Plot for the Best Model using Cross-Section Variables

The best model using cross-sectional alignment variables follows all the assumptions of linear regression. This shows that this is an acceptable choice of distributions for this data set.

5.2.1.3 Variables Relating to Traffic Characteristics

There are two variables that relate to traffic characteristics. It was decided to try and determine if both would be important in a prediction model. Again, using a selection process of the adjusted coefficient of determination, the variables were compared in together individually to determine if they should be combined or kept separate.

Due to the goal of finding the traffic characteristics of most interest, all three models including the one with the greatest adjusted coefficient of determination were examined. The two possible traffic characteristic variables examined were *vol* and *heavyveh*. *Vol* is the annual daily traffic of each roadway segment and *heavyveh* is the

percentage of volume that is composed by heavy vehicles. The top model consisted both of the traffic characteristic variables.

Table 20: ANOVA Table for the Best Model using Traffic Characteristics

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	2	1536.79919	768.39959	4.18	0.0277
Error	24	4410.21433	183.75893		
Corrected Total	26	5947.01352			
Root MSE	13.55577		R-Square	0.2584	
Dependent Mean	23.14741		Adj. R-Sq	0.1966	
Coeff Var	58.52681				

The adjusted coefficient of determination for this model is 0.1966, meaning that 19 percent of the variation in the model can be explained by this model and the coefficient of determination is 0.2584. These and other informative numbers can be seen in Table 20. The coefficients for the variable may not be significant to the desired amount of $\alpha = 0.01$, with volume being significant to a 0.12 level, but the model is not of what was of primary interest in this situation (See Table 21). The model was mainly to show which traffic characteristics are of major importance.

Table 21: Parameter Estimates for the Best Model using Traffic Characteristics

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	-4.91602	11.50883	-0.43	0.6731
Vol	1	0.00051943	0.00032341	1.61	0.1213
heavyveh	1	7.65513	2.72203	2.81	0.0096

Some analysis was done to confirm that that the model from this group of variables followed the basic model assumptions. Figure 44shows the distribution of the residuals for this model, which shows that the residuals are basically evenly distributed about zero with approximately half falling above and below zero. There is a small lack of symmetry in that there is a larger variance on the positive side for the residuals, but

this is not large enough to cause serious concern. Normally distributed residuals are a sign that the data fits the normal probability model.

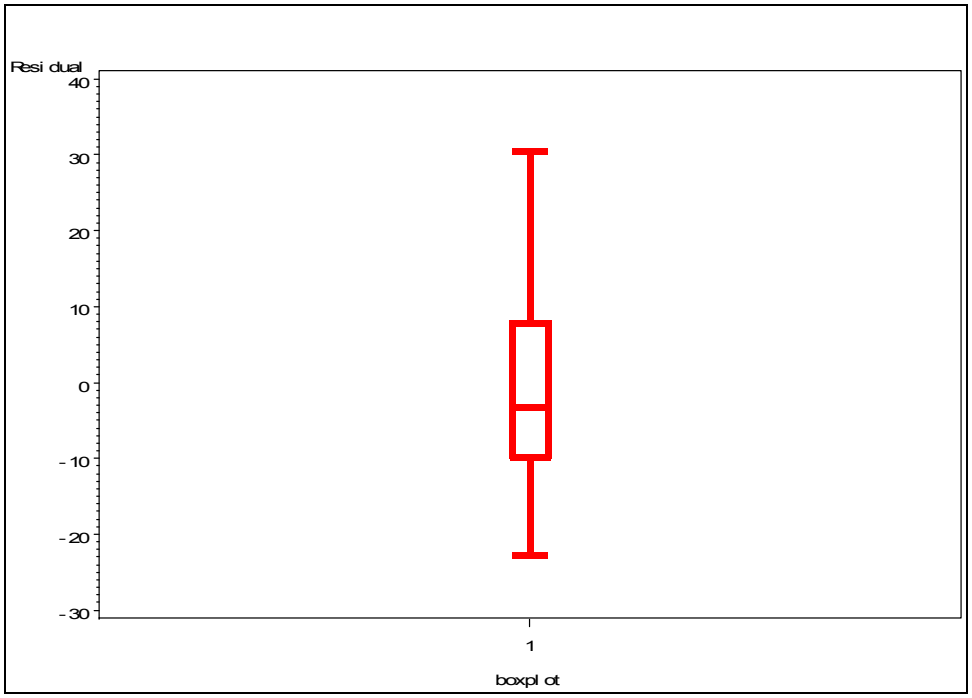


Figure 44: Boxplot of Residuals for the Best Model using Traffic Characteristics

An assumption when dealing with multiple linear regression is that the data follows a normal distribution and the variance is constant. The graph in Figure 45 shows the studentized residuals versus the predicted values for the traffic characteristics model and conveys the basic principle that there is a mostly constant variance in this model. This can be seen by the even distribution of the residuals around zero and by the lack of a pattern in the locations. A heuristic for outliers is that if they are greater than four in the studentized residual plot then the point could be considered an outlier. Based on this rule of thumb there are no outlying points in this data set.

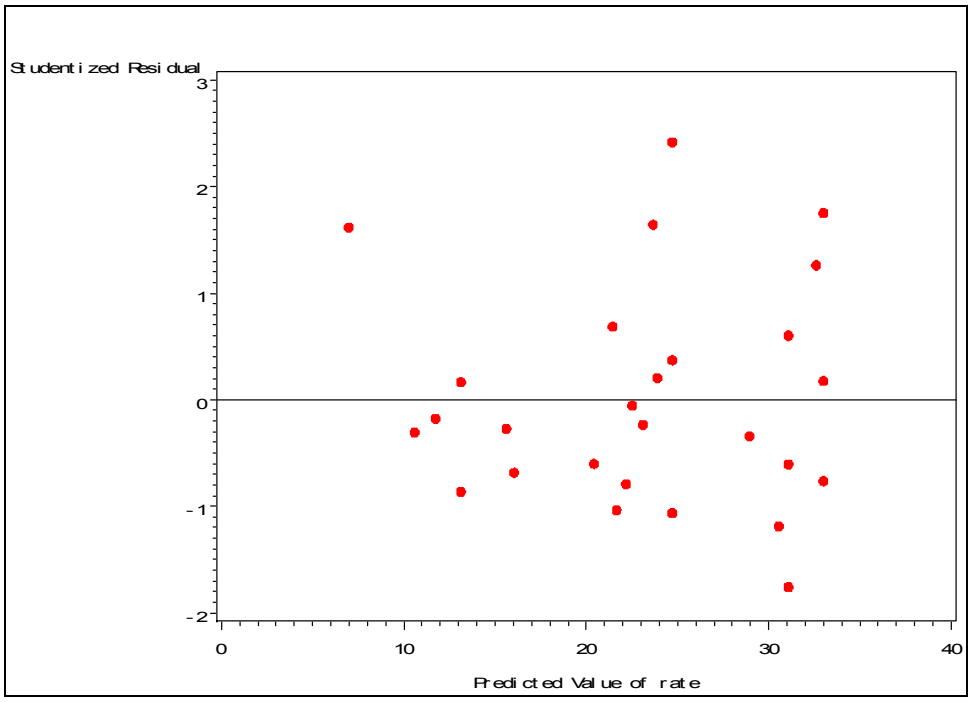


Figure 45: Studentized Residuals vs. Predicted Values for the Best Model using Traffic Characteristics

Another way to visually check that the data follows a normal distribution is to look at the normal probability plot (See Figure 46). The solid line is the normal probability distribution, while the dashed line represents the distribution that can be developed using the data from the model. The two lines match closely; deviating only on the right side of the plot, showing that using the normal probability distribution was a good assumption for this data.

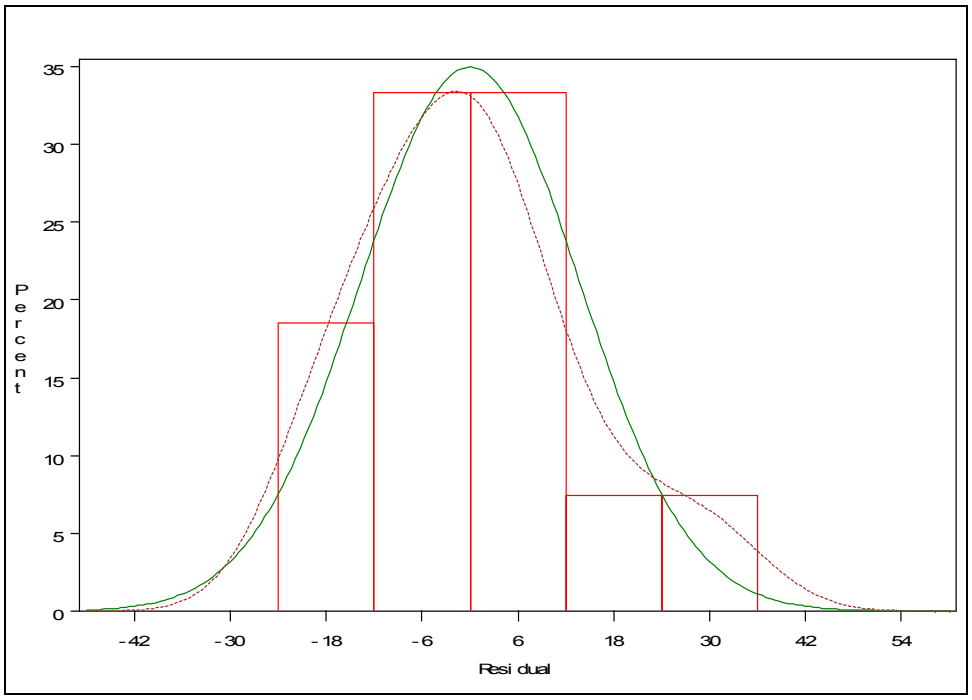


Figure 46: Normal Probability Plot for the Best Model using Traffic Characteristics

Similarly the normal quantile plot is effective in showing when the data does not follow a normal distribution. When the assumption is correct, the residuals fall along the straight line. If the assumption is wrong, the residuals will not fall along the straight line, but may follow a different pattern. Figure 47 shows that the residuals almost all fall along the straight line, showing that the assumption of normality is correct with using the hazard variables regressed against the rate variable.

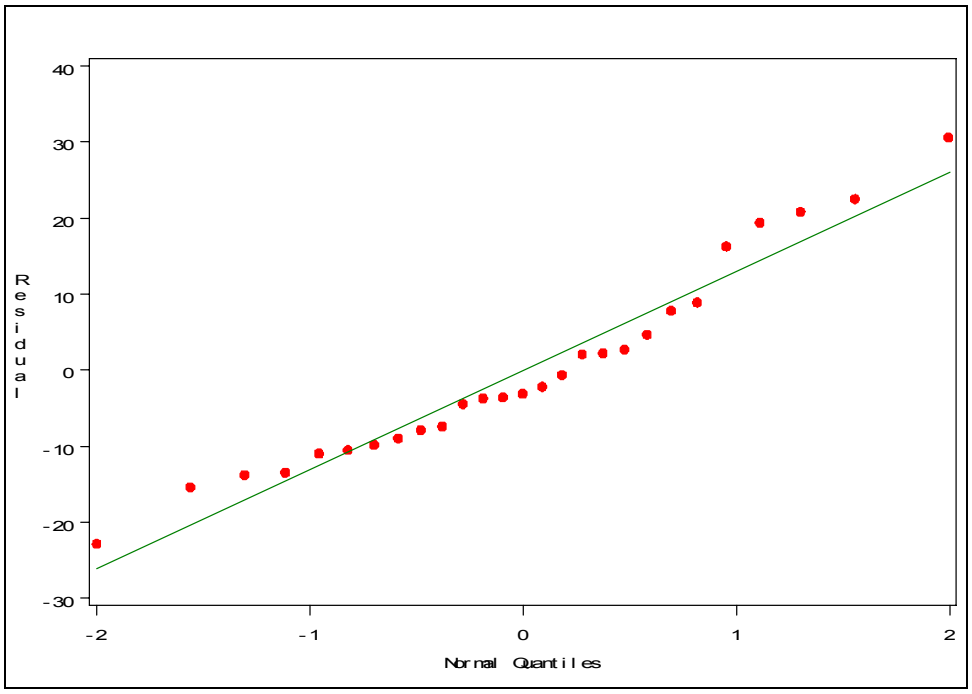


Figure 47: Normal Quantile Plot for the Best Model using Traffic Characteristics

While not being able to eliminate any of the traffic characteristic variables, the model using them follows all the assumptions of linear regression. This continues to shows that a normal distribution is a good choice for this data.

5.2.1.4 Variables Relating to Horizontal and Vertical Alignment

There are five variables that relate to horizontal and vertical alignment. Using a selection process of the adjusted coefficient of determination, the variables were compared in multiple combinations to determine the premier combination. The five possible horizontal and vertical alignment variables examined were *length*, *SD*, *curve*, *type*, and *grade*. *Length* is the overall length of the segment, while *SD* represents the presence of a stopping sight distance problem. *Curve* is an indication of how many horizontal curves there are in the roadway segment. If this variable proves to be insignificant during the model development process it may be converted to a simple indicator variable showing that the segment is either straight or curved. *type* indicates

what the terrain is classified as with zero representing level terrain, one representing rolling terrain and two representing mountainous terrain. *Grade* indicates the maximum grade observed on the roadway segment.

The goal of examining this group is to find which variables are of most interest in further model development. The models with the highest adjusted coefficient of determination were examined to see which of the alignment variables occurred most often. All of the possible alignment variables were included in the top models, but as the reasoning for looking at these was to eliminate some possible variables, a further examination of the use of the variables was done. The findings of this review show that each variables was used the same number of times as the other variables in the top models with each variable appearing in over fifty percent of the top models sorted by adjusted coefficient of determination. This shows that there is not enough of a difference between the variable that would support dropping any of them at this time.

Table 22: ANOVA Table for the Best Model using Alignment Variables

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	3	2807.71384	935.90461	6.86	0.0018
Error	23	3139.29968	136.49126		
Corrected Total	26	5947.01352			
Root MSE	11.68295		R-Square	0.4721	
Dependent Mean	23.14741		Adj. R-Sq	0.4033	
Coeff Var	50.47195				

The model with the largest adjusted coefficient of determination included the variables *length*, *SD* and *curve*. The adjusted coefficient of determination for this model is 0.4033; meaning that 40 percent of the variation in the model can be explained by this model and the coefficient of determination is 0.4721. These and other informative

numbers can be seen in Table 22. This model was examined in further depth than the others, to ensure that the model assumptions are being followed. The coefficients for the variable may not be significant to the desired amount of $\alpha = 0.01$, with *SD* being significant to a 0.28 level, but the model is not of what was of primary interest in this situation (See Table 23). The model was mainly to show which variables relating to horizontal and vertical alignment are of greatest interest in further modeling development.

Table 23: Parameter Estimates for the Best Model using Alignment Variables

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	34.67169	3.79513	9.14	<0.0001
Length	1	-0.01237	0.00307	-4.03	0.0005
SD	1	8.03475	7.31378	1.10	0.2833
curve	1	7.63052	3.56643	2.14	0.0432

Some analysis was done to confirm that that the model from this group of variables followed the basic model assumptions. Figure 48 shows the distribution of the residuals in a boxplot for this model, which shows that the residuals are basically evenly distributed about zero with approximately half falling above and below zero. Symmetric residuals are a sign that the data follows the normal probability model.

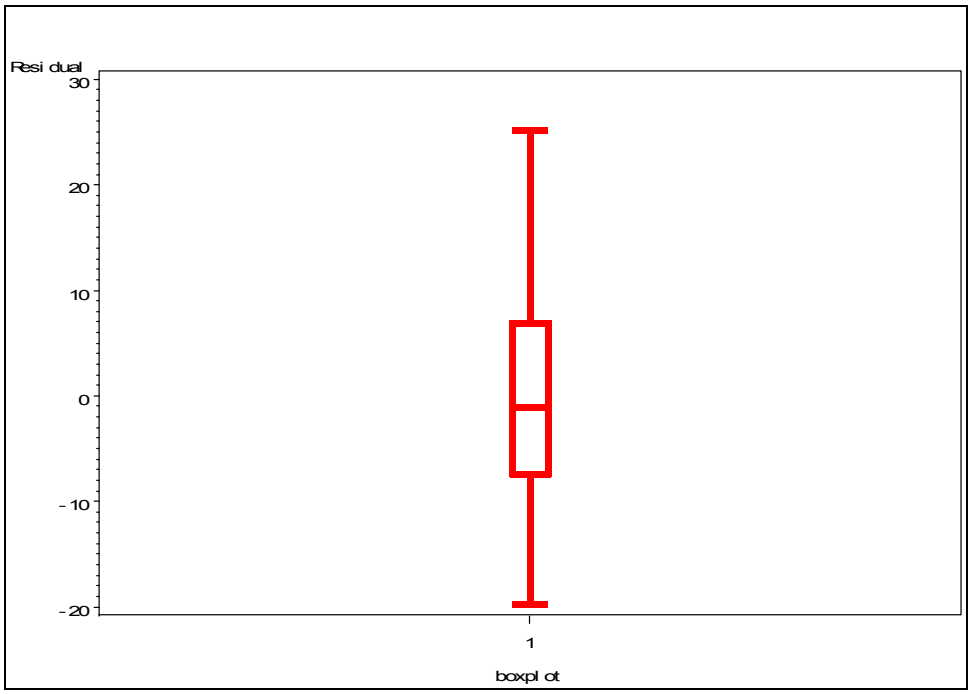


Figure 48: Boxplot of Residuals for the Best Model using Alignment Variables

An assumption when dealing with multiple linear regression is that the data follows a normal distribution and the variance is constant. The graph in Figure 49 shows the studentized residuals versus the predicted values for the alignment model and conveys the basic principle that there is a mostly constant variance in this model. This can be seen by the even distribution of the residuals around zero and by the lack of a pattern in the locations. A heuristic for outliers is that if they are greater than four in the studentized residual plot then the point could be considered an outlier. Based on this rule of thumb there are no outlying points in this data set. There is a slight bias towards positive residuals, but this is not strong enough to imply that the data does not follow a normal distribution.

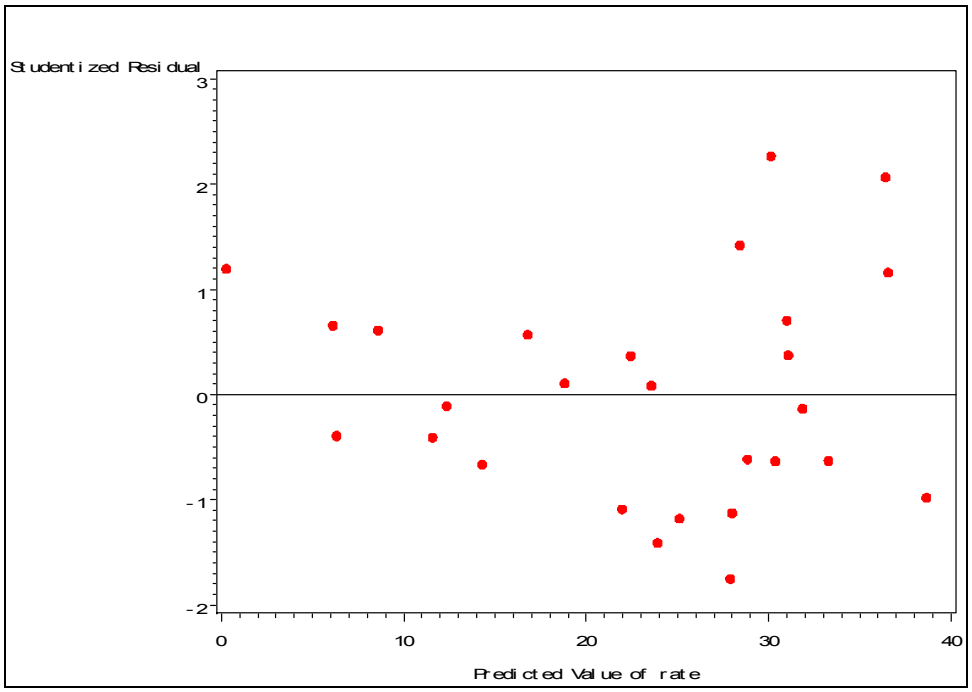


Figure 49: Studentized Residuals vs. Predicted Values for the Best Model using Alignment Variables

Another way to visually check that the data follows a normal distribution is to look at the normal probability plot (See Figure 50). The solid line is the normal probability distribution; while the dashed line represents the distribution that can be developed using the model developed with just horizontal and vertical alignment variables. The two lines match closely, deviating only slightly with the model having a lower and flatter peak than the normal distribution, showing that using the normal probability distribution was a good assumption for this data.

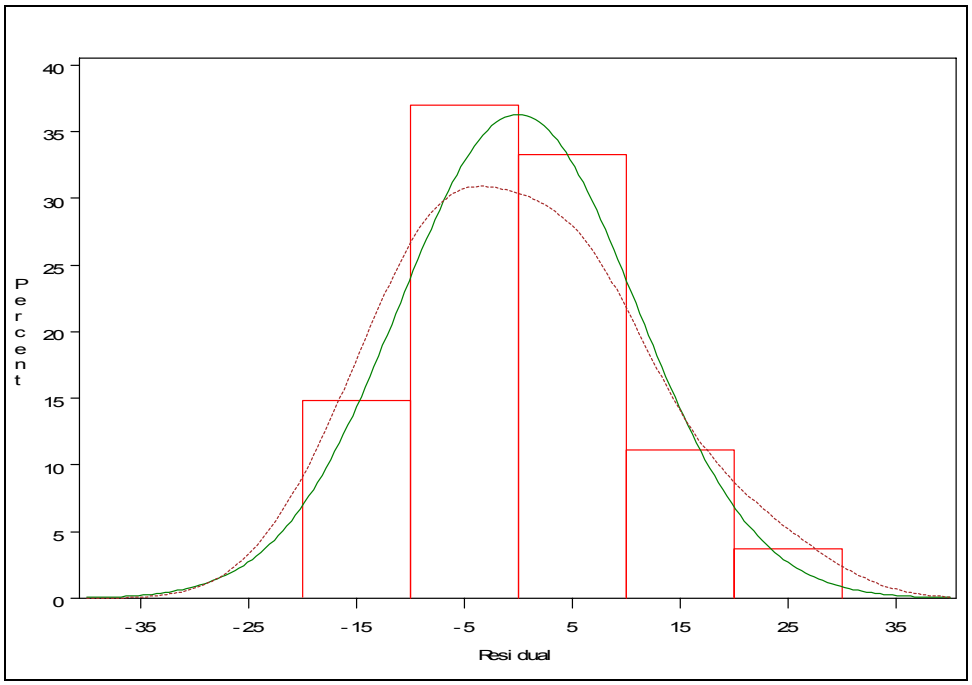


Figure 50: Normal Probability Plot for the Best Model using Alignment Variables

Similarly the normal quantile plot is effective in showing when the data does not follow a normal distribution, which does not apply in this situation. Figure 51 shows that the residuals almost all fall along the straight line, showing that the assumption of normality is correct with using the hazard variables regressed against the rate variable.

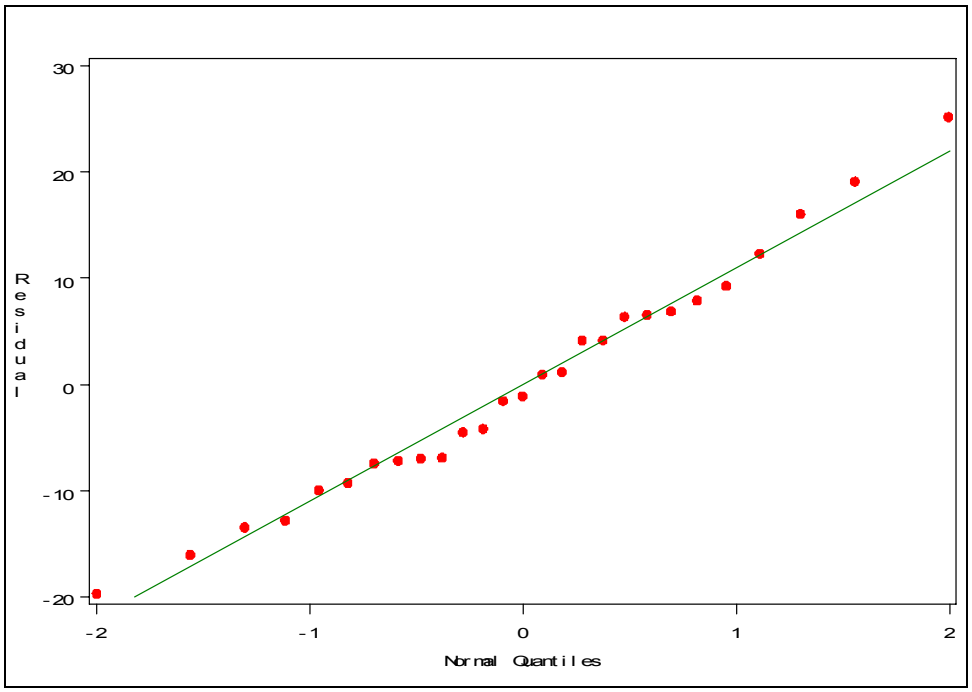


Figure 51: Normal Quantile Plot for the Best Model using Alignment Variables

While not being able to eliminate any of the horizontal and vertical alignment variables, the model using those variables follows all the assumptions of linear regression. Not being able to eliminate any of the variables also leads to the assumption that these may all prove to be important variables for safety purposes.

5.2.1.5 Variables Relating to Access Control

There are several variables that relate to the number and type of access control. It was decided to try and determine which were the most influential and important of these variables to include in a prediction model that includes the influence of more than just access control. Using a selection process of the adjusted coefficient of determination, the variables were compared in multiple combinations to determine the optimum combination.

Due to the goal of finding access control variables of most interest, more possible models other than the model with the greatest adjusted coefficient of determination were

examined. The top models sorted by adjusted coefficient of determination were examined to show which variables were used most often in these models. All of the five possible access control variables were included in the top models, but as the reasoning for looking at these was to eliminate some possible variables, an in depth look at the variation of the use of the variables was done. The variables considered were *maccess* (the number of minor street access points on each segment), *driveways* (the number of driveways on each segment), *parkinglots* (the number of parking lots on each segment), *drivepark* (the total number of driveways and parking lots on each segment), and *allaccess* (the total number of access points on each segment). Out of the top twenty-three models, each variable was used either nine or ten times. So each access control variable was present in over forty percent of the top models. This prevents any of the access control variables from being immediately eliminated from the list of potential variables.

Table 24: ANOVA Table for the Best Model using only Access Variables

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	1	1073.64374		5.51	0.0272
Error	25	4873.36978			
Corrected Total	26	5947.01352			
Root MSE	13.96191		R-Square	0.1805	
Dependent Mean	23.14741		Adj. R-Sq	0.1478	
Coeff Var	60.31736				

The model that had the largest adjusted coefficient of determination for hazards included just one variable: *allaccess*. *Allaccess* is a continuous variable that represents the total number of access points on each roadway segment. The access points include minor roads, driveways and parking lots. The adjusted coefficient of determination for

this model is 0.1478; meaning that 14 percent of the variation in the model can be explained by this model and the coefficient of determination is 0.1805. These and other informative numbers can be seen in Table 24. The coefficients for the variable may not be what were actually expected, *allaccess* has a negative coefficient meaning that the more access points present the fewer accidents occur, but the model is not of what was of primary interest in this situation (See Table 25). The model was mainly to show what access control variables are of main interest.

Table 25: Parameter Estimates for the Best Model using only Access Variables

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	29.14389	3.70789	7.86	<0.0001
Allaccess	1	-0.32124	0.13688	-2.35	0.0272

Further analysis was done primarily to confirm that that best model from this group followed the basic model assumptions. Figure 52 shows the distribution of the residuals for this model. This figure shows that the residuals are basically evenly distributed about zero with approximately half falling above and below zero. Normally distributed residuals are a sign that the data fits the normal probability model.

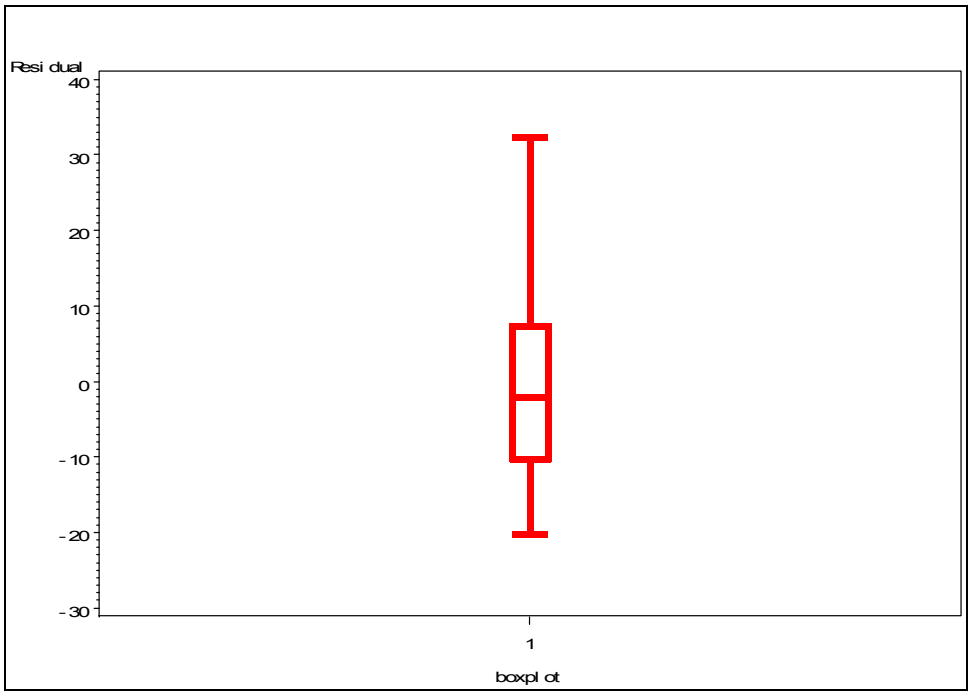


Figure 52: Boxplot of Residuals for the Best Model using only Access Variables

An assumption when dealing with multiple linear regression is that the data follows a normal distribution and the variance is constant. The graph in Figure 53 shows the studentized residuals versus the predicted values for the best access model and conveys the basic principle that there is a mostly constant variance in this model. A heuristic for outliers is that if they are greater than four in the studentized residual plot then the point could be considered an outlier. Based on this rule of thumb there are no outlying points in this data set.

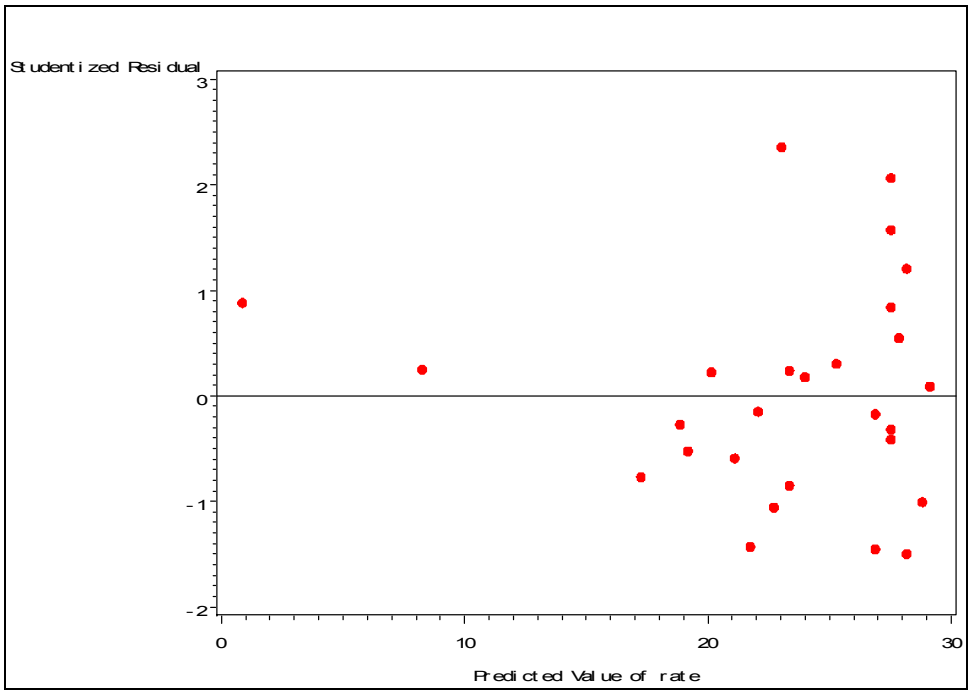


Figure 53: Studentized Residuals vs. Predicted Values for the Best Model using only Access Variables

Another way to visually check that the data follows a normal distribution is to look at the normal probability plot (See Figure 54). The solid line is the normal probability distribution, while the dashed line represents the distribution that can be developed using the data from the model. The two lines match closely; showing that using the normal probability distribution was a good assumption for this data.

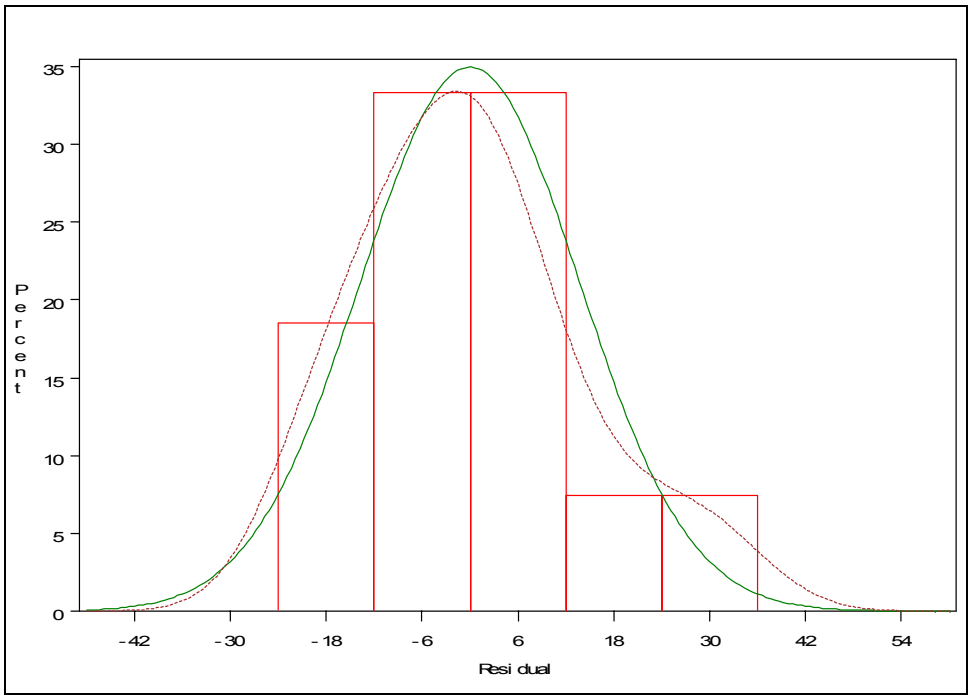


Figure 54: Normal Probability Plot for the Best Model using only Access Variables

Similarly the normal quantile plot is effective in showing when the data does not follow a normal distribution. When the assumption is correct, the residuals fall along the straight line. If the assumption is wrong, the residuals will not fall along the straight line, but may follow a different pattern. Figure 55 shows that the residuals follow the straight line, showing that the assumption of normality is correct with using the hazard variables regressed against the rate variable.

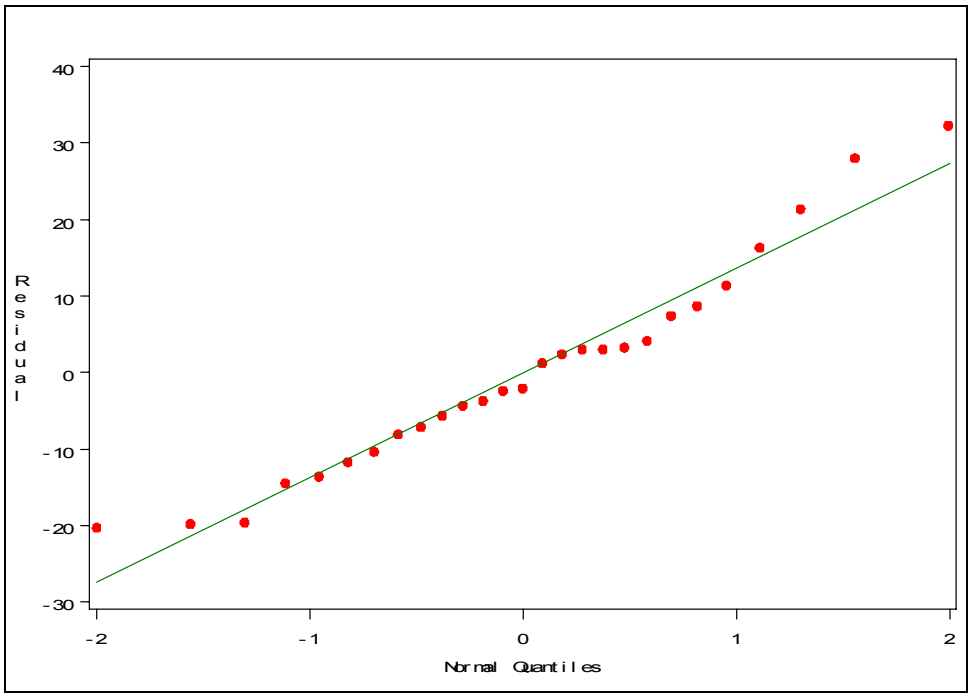


Figure 55: Normal Quantile Plot for the Best Model using only Access Variables

The best model using only access control variables follows all the assumptions of linear regression. This shows that this is a good choice of distributions for this data set.

5.2.1.6 Variables Relating to All Other Characteristics

There are several variables that have not found a home in any of the earlier categories. It was decided to put any remaining variables in a group and determine which were the most influential and important of these variables to include in a prediction model. Using a selection process of the adjusted coefficient of determination, the variables were compared in multiple combinations to determine the optimum combination.

There were four variables that did not fit into any of the other categories which include *markings*, *lanelength*, *pavement*, and *lighting*. *Markings* is the variable that considers the condition of the pavement markings on each segment. These can be classified as good, fair or poor depending on their quality. Similarly, *pavement* is the

variable that considers the condition of the pavement and again it can be classified as good, fair or poor. *Lighting* represents the percentage of each roadway segment that has lighting, this is important as lack of lighting is often a cause of accidents. *Lanelength* is the variable that represents the total miles of lanes on each segment. This helps to normalize segments that have different lengths and different numbers of lanes.

Due to the goal of finding the variables of most interest, the top models were sorted by adjusted coefficient of determination and examined to show which variables were used most often in these models. All of the possible variables were included in the top models, but as the reasoning for looking at these was to eliminate some possible variables, an in depth look at the variation of the use of the variables was done. The top models were compared to see how often the variables appeared in each. There was no clear division with one or more of the variables not appearing in the top models. Each variable was present in over fifty percent of the top models. This prevents any of the variables from being eliminated from the list of potential variables.

Table 26: ANOVA Table for the Model using Other Variables

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	2	2130.96238	1065.48119	6.70	0.0049
Error	24	3816.05114	159.00213		
Corrected Total	26	5947.01352			
Root MSE	12.60960		R-Square	0.3583	
Dependent Mean	23.14741		Adj. R-Sq	0.3049	
Coeff Var	54.47524				

The model that had the largest adjusted coefficient of determination for hazards included just two variables: *markings* and *lanelength*. The adjusted coefficient of determination for this model is 0.3049; meaning that 30 percent of the variation in the

model can be explained by this model and the coefficient of determination is 0.3583. These and other informative numbers can be seen in Table 26. The coefficients for the variable may not be what were actually expected, *allaccess* has a negative coefficient meaning that the more access points present the fewer accidents occur, but the model is not of what was of primary interest in this situation (See Table 27). The model was mainly to show what access control variables are of main interest.

Table 27: Parameter Estimates for the Model using Other Variables

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	39.47000	5.79125	6.82	<0.0001
Markings	1	-4.23065	3.80893	-1.11	0.2777
lanelength	1	-14.62690	4.17076	-3.51	0.0018

Some further analysis was done primarily to confirm that that best model from this group followed the basic model assumptions. Figure 56 shows the distribution of the residuals for this model. This figure shows that the residuals are evenly distributed about zero with approximately half falling above and below zero. Normally distributed residuals are a sign that the data fits the normal probability model.

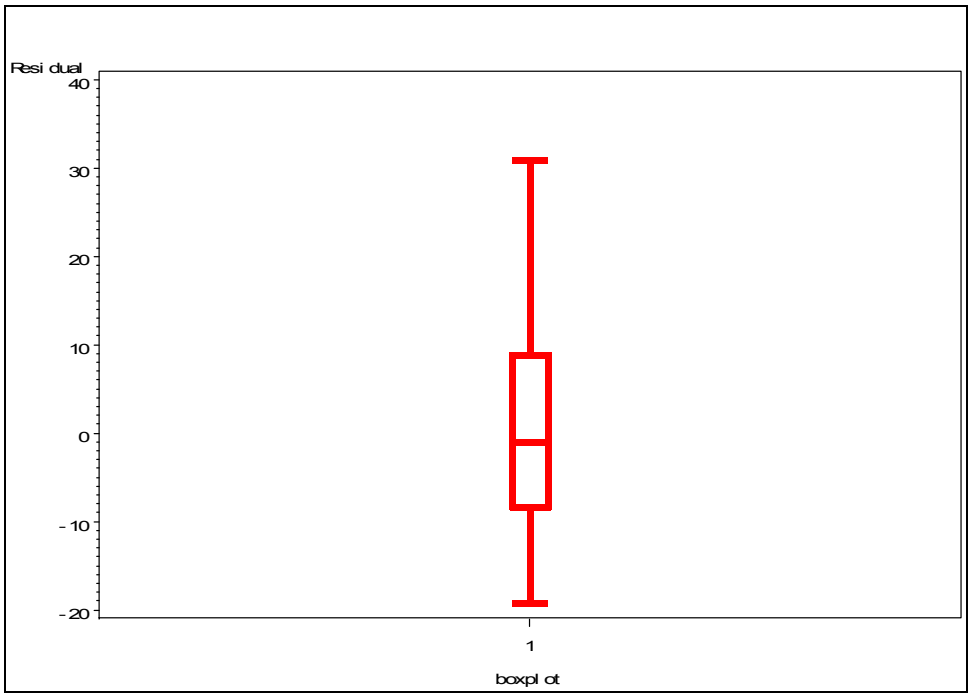


Figure 56: Boxplot of Residuals for the Model using Other Variables

An assumption when dealing with multiple linear regression is that the data follows a normal distribution and the variance is constant. The graph in Figure 57 shows the studentized residuals versus the predicted values for the best other model and conveys the basic principle that there is a mostly constant variance in this model. There is a slight unevenness with the positive residuals having a larger variance, but this is not large enough to be of any concern. A heuristic for outliers is that if they are greater than four in the studentized residual plot then the point could be considered an outlier. Based on this rule of thumb there are no outlying points in this data set.

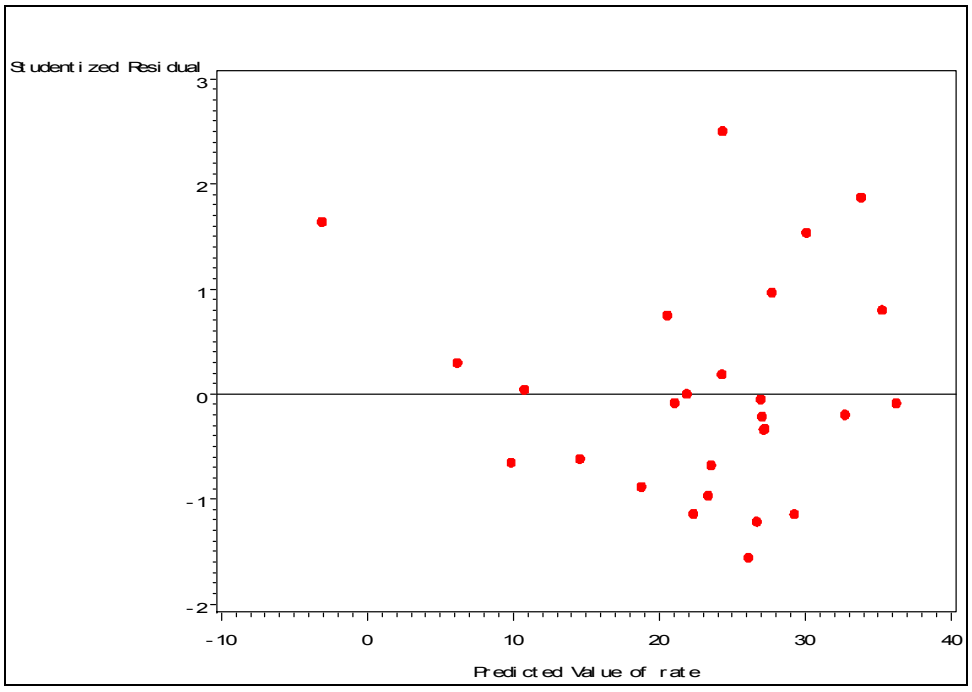


Figure 57: Studentized Residuals vs. Predicted Values for the Model using Other Variables

Another way to visually check that the data follows a normal distribution is to look at the normal probability plot (See Figure 58). The solid line is the normal probability distribution, while the dashed line represents the distribution that can be developed using the data from the model. The two lines match closely; showing that using the normal probability distribution was a good assumption for this data.

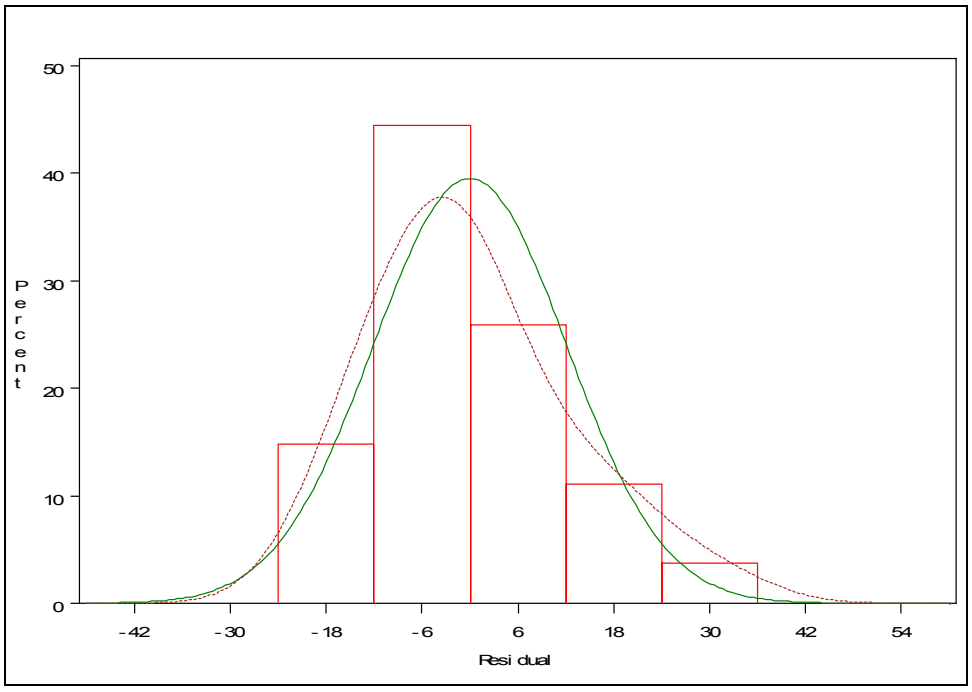


Figure 58: Normal Probability Plot for the Model using Other Variables

Similarly the normal quantile plot is effective in showing when the data does not follow a normal distribution. When the assumption is correct, the residuals fall along the straight line. If the assumption is wrong, the residuals will not fall along the straight line, but may follow a different pattern. Figure 59 shows that the residuals closely follow the straight line, showing that the assumption of normality is correct with using the hazard variables regressed against the rate variable.

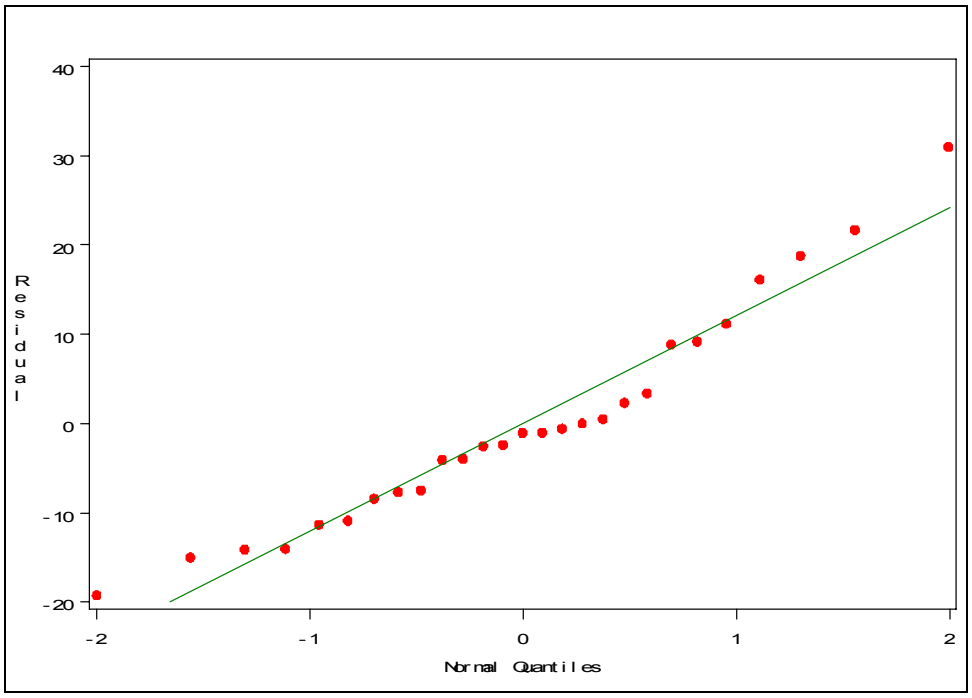


Figure 59: Normal Quantile Plot for the Model using Other Variables

The best model using other variables follows all the assumptions of linear regression. This shows that this is a good choice of distributions for this data set.

5.2.1.7 Summary of Primary Variable Elimination

The primary elimination was intended to be a rough elimination of variables that do not have a strong effect on predicting crashes. The variables eliminated at this stage deal mainly with roadside hazards and geometric alignment. This is too be expected since these are the areas with the largest number of possible variables. The variables that were eliminated include the number of mailboxes, the number of stone monuments, the number of rocks, the number of light poles, the percent of perpendicular parking, the percent of parallel parking, the number of lanes going in the right direction, the width of the second and third lanes in the right direction. The first of these can be eliminated based on the fact that they were not used often or found to be significant and that they are accounted for in the overall variable that accounts for all the roadside hazards present on

the road segment. The number of light poles again is counted in the variable pole, which is a count of all the poles on the segment. The percent of perpendicular parking was a variable that was expected to have little or no effect with predicting crashed due to the fact that perpendicular parking was only found to exist on one road segment and is an unusual style of parking on urban streets. The information in the other variables relating to the number of lanes traveling in the right direction and the width of the second and third lanes traveling in that direction is also duplicated in other variables that remain for further consideration. The total number of lanes and the average lane width take these variables into account. This primary elimination however did allow for some variables to be eliminated from further consideration and it allowed for information to be gathered relating to how the different variables relate to each other and to the crashes that occurred over the arterial segments.

5.2.2 Secondary Variable Elimination

The first round of variable elimination allowed for eight variables to be discarded at this stage of the model development. This reduction brought the total number of possible variables down to forty-eight which can be seen in Table 28. The variables were divided into two groups that could be run together and the most common variables examined, in the same way as the primary variable elimination method. There were still too many variables to be run in one modeling attempt, so a secondary elimination process was undertaken.

Looking at variables that could be combined into one overall variable and looking at correlations between similar variables was the basis of the second elimination method. By looking at correlations, it can be seen if variables are describing the same variation in

the data. A high correlation value means that the variables in question describe the same variation in the data and are highly correlated, while a low correlation value means that the variables do not describe the same variation in the data.

Table 28: Variables Remaining after the Primary Elimination

Variables:			
ospole	drivepark	length	llanes
upole	allaccess	grade	widthl3
vol	benches	SD	median
Pmeter	hydrant	curve	widthm
maccess	building	curves	widthr1
Fence	other/electrical	crest	widthsida
Spole	hazards	widthl2	lane
residential	density	widthl1	widtha
commercial	driveways	markings	widthsidr
pole	heavyveh	widthsidl	widthsr
parkinglots	trees	pavement	lighting
lanelength	industrial	type	parking

Six variables describe the access on each roadway segment. The correlation between these variables was reviewed to try and eliminated some of them from further investigations. The variable of *allaccess* was considered the basic variable in that as it is a count of all access points on a road segment, it should explain the majority of the variation in the data. Two of the other access variables, *driveway*s and *drivepark*, have high correlation coefficients with 0.9041 and 0.9854 respectively (See Table 29) allowing them to be removed from further consideration. Since the data variation can be almost equally described by another variable, they are not needed for further model development. It was also determined on further reflection that the variable *density* should be eliminated since it is the number of hazards per mile for each segment. It is a compiled variable that takes into account the total number of roadside hazards and the segment length. Since it is made up of variables that are already included in the model development it can be left out of further development.

Table 29: Pearson Correlation Coefficients for Access Variables

	maccess	parkinglots	driveways	drivepark	allaccess	density
maccess	1	0.2549	0.5775	0.5713	0.7027	-0.29
parkinglots	0.2549	1	0.2493	0.6381	0.606	0.208
driveways	0.5775	0.2493	1	0.9047	0.9041	-0.06
drivepark	0.5713	0.6381	0.9047	1	0.9854	0.042
allaccess	0.7027	0.606	0.9041	0.9854	1	-0.02
density	-0.288	0.2081	-0.0628	0.0416	-0.0238	1

There were three variables that describe the width of the existing sidewalks: a variable for the 'left' sidewalk width, the 'right' sidewalk width and the average sidewalk width. The correlation between the three variables was examined to see if they were describing the same variation in the data. The Pearson correlation coefficients can be seen in Table 30. There is a strong correlation between the variables of *widthsida* and *widthsidl* with a coefficient of 0.9635. Strong correlation also exists between *widthsida* and *widthsidr* with a coefficient value of 0.9670. These coefficients show that there is a high correlation between the variables in question and that these variables are describing almost the same variation in the base data. Since the variables are describing the same variation, they are not all needed to be in the final model. This allows for both *widthsidr* and *widthsidl* to be eliminated from further models with *widthsida* covering the same data variation.

Table 30: Pearson Correlation Coefficients for Sidewalk Widths

	widthsida	widthsidl	widthsidr
widthsida	1.0000	0.9635	0.9670
widthsidl	0.9635	1.0000	0.8644
widthsidr	0.9670	0.8644	1.0000

Similarly to the variables describing sidewalk width above, there are three variables that explain the number of lanes that exist on each roadway segment: *llanes*, *rlanes*, and *lane*. These describe the total number of lanes in the 'left' direction, the total

number of lanes in the 'right' direction and the total number of lanes on the segment. The Pearson correlation coefficients (as seen in Table 31) were examined in the hope that two of the variables could be eliminated, having the variation in the data that they explain be covered by the joint variable of *lane* which can be described as *llanes* + *rlanes*. The correlation between lane and the other two variables were greater than 95 percent allowing both *llanes* and *rlanes* to be removed from further consideration.

Table 31: Pearson Correlation Coefficients for Lane Variables

	rlanes	llanes	lanes
rlanes	1.0000	0.8637	0.9675
llanes	0.8637	1.0000	0.9631
lane	0.9675	0.9631	1.0000

There are variables that describe the width of the different lanes in addition to the variables that describe the number of lanes on each road segment. The correlation coefficients can be seen in Table 32. In this set *widtha* was the variable assumed to be the base, since it contained the information from the other variables by being an average width of all the lanes. Using this assumption of a base variable, it was determined that two other variables are highly correlated with *widtha*, that of *widthl1* and *widthr1*, the widths of the centermost lane going in both directions. They were correlated with Pearson coefficients of 0.9048 and 0.9176 respectively. This allows the two variables to be eliminated from further use in the final model development. The variables of *widthl2* and *widthl3* were also looked at because their values are included in the average width variable, which means that including them and the average width lets that information be double counted in the final model development. Due to this repetition of the data the two variables were also removed from further consideration.

Table 32: Pearson Correlation Coefficients for Lane Width Variables

	widthl1	widthl2	widthl3	widthr1	widtha
widthl1	1.0000	-0.6562	-0.2717	0.8329	0.9048
widthl2	-0.6562	1.0000	0.0760	-0.7353	-0.5517
widthl3	-0.2717	0.0760	1.0000	-0.3112	-0.4125
widthr1	0.8329	-0.7353	-0.3112	1.0000	0.9176
widtha	0.9048	-0.5517	-0.4125	0.9176	1.0000

In terms of cross section variables there are two that describe the presence of a median, by use of an indicator variable, or its width, by the use of a continuous variable. The two variables show a very high correlation with each other, allowing the base variable to be kept for further model development (See Table 33 for correlation coefficients). It was decided to use the presence of a median as the more important of the two variables. This was done because on the range of segments examined there was not a large amount of variation in the median widths observed, with variation existing only from 5.5 to 8 feet. Then the indicator variable was used as the base variable and the continuous variable was removed from further development.

Table 33: Pearson Correlation Coefficients for Median Variables

	median	widthm
medain	1.000	0.987
widhtm	0.987	1.000

There are a lot of possible variables that can be used to describe roadside hazards. In order to eliminate some of them, first all the variables that describe a pole were examined. These included variables that describe overhead sign poles, utility poles, and sign poles. *Pole* was used as the base variable since it consists of all the other pole variables added together. The correlation between *pole* and *spole* is very high with a Pearson's coefficient of 0.9822, which means the *pole* variable describes the same variation, as does the *spole* variable, letting *spole* be removed from further consideration.

This can be seen in Table 34 with the correlation coefficients for the pole variables. There is also a fairly high correlation between *upole* and *pole* with a coefficient of 0.7643. Though this is a slightly lower correlation that would be discarded without any thought, it was deemed large enough to allow the variable to be discarded and get the total number of variable to be used in further model development to become smaller.

Table 34: Pearson Correlation Coefficients for Pole Variables

	upole	spole	ospole	pole
upole	1.0000	0.6580	-0.0631	0.7643
spole	0.6580	1.0000	0.0331	0.9822
ospole	-0.0631	0.0331	1.0000	0.0313
pole	0.7643	0.9822	0.0313	1.0000

The other variables that represent roadside hazards were also looked at for possible correlations. *Hazards* was used as the base variable, which represents the total number of hazards on each road segment. This comparison took place in several steps to make looking at the correlation matrixes easier. Table 35 shows the first set of correlations that show a large correlation between *hazards* and *hydrants*, *buildings* and *trees*. All three of these correlation coefficients are greater than 0.8 allowing the variables to be removed from further evaluations. The variable *electrical* was also removed from further consideration based on the fact that only three segments have the variable and it does not appear to be significant in the amount of variation in the data that it can explain. So in an effort to reduce the total number of variables *electrical* was discarded.

Table 35: Pearson Correlation Coefficients for Hazards (1)

	hazards	hydrant	building	electrical	trees
hazards	1.0000	0.8382	0.9296	0.0652	0.8603
hydrant	0.8382	1.0000	0.7352	-0.0322	0.7388
building	0.9296	0.7352	1.0000	0.0512	0.7464
electrical	0.0652	-0.0322	0.0512	1.0000	0.0670
trees	0.8603	0.7388	0.7464	0.0670	1.0000

Looking at the second matrix of correlation coefficients in Table 36, there is only one variable that has a strong correlation to the base variable of *hazards*. The variable *pole* has a correlation coefficient of 0.9528 meaning that most of the variation in the data that is explained by the variable *pole* is also explained by the variable *hazards*, allowing *pole* to be disregarded.

Table 36: Pearson Correlation Coefficients for Hazards (2)

	hazards	benches	pole	fence	pmeter	ospole
hazards	1.0000	0.1209	0.9528	0.5242	-0.0365	-0.0769
benches	0.1209	1.0000	0.0730	0.1602	-0.0543	-0.1335
pole	0.9528	0.0730	1.0000	0.5581	-0.1427	0.0313
fence	0.5242	0.1602	0.5581	1.0000	-0.1518	-0.0505
pmeter	-0.0365	-0.0543	-0.1427	-0.1518	1.0000	-0.0982
ospole	-0.0769	-0.1335	0.0313	-0.0505	-0.0982	1.0000

There are two variables that describe vertical alignment that of *grade* and *type*. *Grade* is a continuous variable giving the maximum vertical grade observed on the road segment. *Type* classifies the segments according to level, rolling, or mountainous terrain, so both variables give similar information. The correlation matrix between the two variables was examined and the coefficient was found to be 0.8888 (See Table 37). This is large enough to allow one of the variables to be removed from further examination. The variable of *grade* was kept as the base variable on the understanding that in this case, the divisions of the *type* variable may not be the best possible and that the maximum grade would be more useful.

Table 37: Pearson Correlation Coefficients for Vertical Alignment

	grade	type
grade	1.0000	0.8888
type	0.8888	1.0000

Similar to the variables relating to vertical alignment, there are two variables that describe a segments horizontal alignment. Curve and curves are a continuous and indicator variable respectively that represent either the number of horizontal curves or the presence of one or more horizontal curves. The coefficient between *curve* and *curves* is 0.7906, meaning that 79 percent of the variation in the data is explained by the two variables (See Table 38). This allows one of the two to be eliminated from further evaluation. It was determined that the presence of horizontal curvature was more important than the actual number of horizontal curves that where present on each road segment. The variable *curve* was removed from further consideration.

Table 38: Pearson Correlation Coefficients for Horizontal Alignment

	curve	curves
curve	1.0000	0.7906
curves	0.7906	1.0000

There are three variables that describe land use on each road segment. The variable that represents the percentage of industrial land use was eliminated from further consideration by several reasons. It did not appear in the top models when half the variables were run together to look at the top models. Another reason for discarding this variable was that only one road segment had industrial land use, so for the areas under consideration in this study, industrial land use is not a large percentage so should not have a large effect on the overall prediction model. The correlation between the remaining variables that describe residential and commercial land use was very high with a coefficient of –0.9997. Table 39 shows the full correlation matrix for the land use

variables. The negative sign in this case means that the two variables are present in opposite conditions, when one segment shows ninety percent residential use, commercial use will then conversely be ten percent. Despite being negatively correlated, the two variables are still strongly correlated meaning that one of them can be removed from further evaluation. It was decided to leave the variable representing the percentage of residential land use for use in further model developments.

Table 39: Pearson Correlation Coefficients for Land Use Variables

	commercial	residential
commercial	1.0000	-0.9997
residential	-0.9997	1.0000

One final variable was eliminated from further evaluation during the secondary variable elimination stage. This variable, *widthsr*, is the width of the shoulder on the road segment and was eliminated since shoulders only occurred on one road segment, it was determined that the variable did not carry enough information that could be used to make further conclusions about the data. The secondary variable elimination stage allowed for many variables to be eliminated and the total number to be used for further model development brought down to a manageable twenty-five.

5.2.3 Linear Model Groups

After the primary and secondary variable elimination methods were used, three models were contenders for accident prediction models. There were three sets of variables ranging from 24 to 26 variables. A model selection criterion of the highest adjusted coefficient of determination was used to choose the most significant model from the three variable groups.

5.2.3.1 Variable Group One

The first group was run with the remaining 24 variables after the primary and secondary elimination methods had been used to bring the total number of variables down to a workable number. The adjusted R-square selection method was used in that the best models were sorted by the largest adjusted R-square values, but the coefficient of determination was also give for comparison purposes. See Table 40 for the list of possible variables.

Table 40: Variable Group One

Variable	
ospole	Length
vol	Grade
pmeter	SD
maccess	Curves
fence	Crest
residential	Markings
parkinglots	Pavement
allaccess	Median
benches	Widthsida
hazards	Lane
heavyveh	Widtha
parking	lighting

The best model that was developed from the top group of variables included 19 variables with a coefficient of determination of 0.9451 and an adjusted coefficient of 0.7961, both values are extremely good. The analysis of variance table seen below shows important values relating to this model, including the F-statistic value and the P-statistic value which indicate that the overall model is significant to a greater than 0.05 percent.

Table 41: ANOVA Table for First Model from Variable Group One

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	19	5620.52175	295.81693	6.34	0.0093
Error	7	326.49177	46.64168		
Corrected Total	26	5947.01352			
Root MSE	6.82947		R-Square	0.9451	
Dependent Mean	23.14741		Adj. R-Sq	0.7961	
Coeff Var	29.50426				

The parameter estimates and standard errors can be seen in Table 42. All but four of the variables are significant to greater than 0.1 percent. And twelve variables are significant to greater than 0.05 percent which leaves only three variables significant between 0.1 and 0.05 percent. This shows that most of the included variables are important to the model. It is desirable, however, to have a model where all of the variables are significant. As the model currently stands this is not the case and the model is cumbersome with so many variables being included.

Table 42: Parameter Estimates for First Model from Variable Group One

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	-129.33712	69.27231	-1.87	0.1041
Benches	1	-14.91338	3.69498	-4.04	0.0050
Fence	1	2.73710	1.02511	2.67	0.0320
Ospole	1	3.30055	1.70866	1.93	0.0947
Pmeter	1	-0.51043	0.43612	-1.1	0.2801
parkinglots	1	-1.75103	0.55093	-3.18	0.0155
allaccess	1	0.85415	0.32706	2.61	0.0348
Vol	1	0.00287	0.00051511	5.57	0.0008
Length	1	-0.02017	0.00453	-4.45	0.0030
Grade	1	-3.91493	1.53683	-2.55	0.0382
SD	1	18.37744	10.87777	1.69	0.1350
Curves	1	17.41656	5.78500	3.01	0.0196
Crest	1	10.54793	1.87318	5.63	0.0008
Widtha	1	-6.51575	1.85164	-3.52	0.0097
widthsida	1	5.11999	1.47432	3.47	0.0104
Parking	1	0.20599	0.09052	2.28	0.0570
Median	1	14.13734	7.94326	1.78	0.1183

Table 42: Parameter Estimates for First Model from Variable Group One Continued

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Lane	1	-16.19383	4.96841	-3.26	0.0139
markings	1	6.55813	4.31383	1.52	0.1723
Lighting	1	1.45846	0.66942	2.18	0.0658

In an attempt to have a more workable model and one where the variables are significant, further work was done. By looking at the individual variable's significance and coefficient of partial determination, variables were removed from the model. An alpha level of 0.10 was set and a coefficient of partial determination level was set at 150. This criterion must be met to be kept for further model development. The coefficient of partial determination "measures the marginal contribution of one X variable when all others are already included in the model" (Neter et al 274). If this contribution is small and the variable insignificant then the variable was removed from further development.

The graphical diagnostics showed this model to follow a normal distribution and the overall model was significant. Despite these attributes, four variables were not significant enough and had low coefficients of partial determination so were eliminated. Based on significance less than 0.1 and coefficients of partial determination less than 150, the variables *pmeter*, *SD*, *median* and *markings* were be eliminated to produce a better model.

The model was rerun with the remaining fifteen variables and overall was again significant. The coefficients of determination and P-value can be seen in Table 43. But, once more, not all the individual variables were significant. Six more variables were identified for removal.

Table 43: ANOVA Table for Second Model from First Variable Group

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	15	4967.41170	331.16078	3.72	0.0167
Error	11	979.60182	89.05471		
Corrected Total	26	5947.01352			
Root MSE	9.43688		R-Square	0.8353	
Dependent Mean	23.14741		Adj. R-Sq	0.6107	
Coeff Var	40.76863				

This process was repeated three more times until all the remaining variables were significant to better than $\alpha = 0.10$. This resulted in all but two variables being removed from the model. The variables that remained were *ospole* and *length*. So now all the variables in the model and the model as a whole were significant as can be seen Table 44 in by the F-statistic. Unfortunately, the coefficient of determination was lowered as more variables were eliminated to such a level that the model no longer explains an acceptable amount of the variation in the data. With $R^2 = 0.4095$ not even half the variation is explained so that the model is not effective at predicting an accident rate.

Table 44: ANOVA Table for Best Model from Variable Group One

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	2	2435.48284	1217.74142	4.26	0.0063
Error	24	3511.53068	146.31378		
Corrected Total	26	5947.01352			
Root MSE	12.09602		R-Square	0.4095	
Dependent Mean	23.14741		Adj. R-Sq	0.3603	
Coeff Var	52.25649				

$Rate = 24.6167 + 2.9903 ospole - 0.0069 length$ The coefficients are mostly the expected signs and even with a 90 percent confidence level do not become zero. The parameter estimate for ospole is positive indicating that the more overhead sign poles on the road segment the higher the accident rate becomes. The coefficient's sign for the

length parameter by intuition would be positive meaning that the longer the segment the more accidents but turned out to be negative implying that the longer segments have lower accident rates. This is due to the division of road segments by major signalized intersections where the shorter the road segment the closer together the signalized intersections are which is where there are large numbers of conflicts and accidents are more likely to occur.

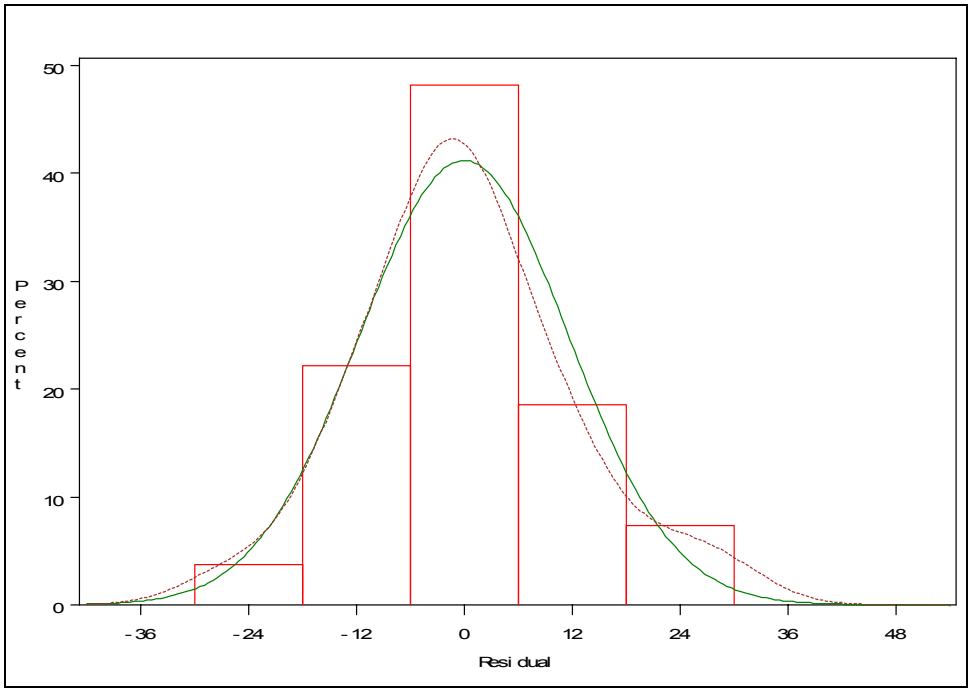


Figure 60: Normal Probability Plot for Best Model from Variable Group One

In spite of the fact that the model does not violate any of the assumptions and follows a normal distribution as seen in Figure 60 this model does not perform well. The coefficient of determination is low and only two variables are included in the model. This model could possibly be used to compare whether or not a road segment has an accident rate extremely different from other similar segments, but even that would not produce reliable results or be helpful in determining what is causing an accident problem on a segment.

5.2.3.2 Variable Group Two

Variable group two consists of twenty-six variables that can be seen in Table 45. The difference between group one and group two are the two variables of *pole* and *lanelength*. These two variables are compilations from other variables that are also in the group of variables, which is why they were excluded from variable group one. *Pole* and *lanelength* were accidentally left into the calculations, but the resulting adjusted coefficient of determination and the coefficient of determination were very high, so the top model was left in for consideration. The top model from this set of variables included twenty-five of the possible twenty-six variables and had a coefficient of determination of 0.9997 and an adjusted coefficient of determination of 0.9925, both of which are extremely high values.

Table 45: Variable Group Two

Variables	
ospole	length
vol	grade
pmeter	SD
maccess	curves
fence	crest
residential	markings
pole	pavement
parkinglots	median
lanelength	widthsida
allaccess	lane
benches	widtha
hazards	lighting
heavyveh	parking

In addition to the high coefficients, the overall model is significant to greater than 0.1 percent with a P-value of 0.0669. The parameter estimates can be seen in Table 46. Nine of the variables are not significant to greater than 0.1 percent. Nine variables are also significant to greater than 0.05 percent, leaving eight that are significant between 0.1

and 0.05 percent. The graphical diagnostics show that the normal distribution and model assumptions are not violated, but despite that there is some concern since many of the variables have a possibility that their parameters could be zero, so this is not the best possible model. Since there are so many variables in this model, it is very cumbersome to use and since so many of the variables are not significant in this model, further work will be done looking for the best model.

Table 46: Initial Model from Variable Group Two

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	-432.53610	31.46908	-13.74	0.0462
Benches	1	-10.06556	1.73497	-5.80	0.1087
Fence	1	6.62433	0.90468	7.32	0.0864
Ospole	1	9.89939	0.58810	16.83	0.0378
Hazards	1	1.13725	0.08916	12.76	0.0498
Pole	1	-1.16616	0.11494	-10.15	0.0625
Maccess	1	2.16695	0.77721	2.79	0.2192
Parkinglots	1	-4.74461	0.21713	-21.85	0.0291
Allaccess	1	0.78614	0.18859	4.17	0.1499
Vol	1	0.00233	0.00034406	6.76	0.0935
Heavyveh	1	-11.00644	1.16075	-9.48	0.0669
Lanelength	1	11.43583	4.04092	2.83	0.2162
Residential	1	-0.33445	0.06253	-5.35	0.1177
Length	1	-0.04953	0.00679	-7.29	0.0868
Grade	1	-2.70225	0.84165	-3.21	0.1922
SD	1	10.49614	4.06884	2.58	0.2354
Curves	1	18.03077	1.34017	13.45	0.0472
Crest	1	16.58164	0.89836	18.46	0.0345
Widtha	1	-10.77932	0.65018	-16.58	0.0384
widhtsida	1	7.88668	0.77867	10.13	0.0627
Parking	1	0.02956	0.01891	1.56	0.3624
Median	1	18.23474	4.30093	4.24	0.1475
Lane	1	-18.03378	1.67547	-10.76	0.0590
Pavement	1	-43.05961	5.03997	-8.54	0.0742
Markings	1	21.79006	1.50729	14.46	0.0440
lighting	1	4.68225	0.33764	13.87	0.0458

The second variation of a model from variable group two consisted of sixteen variables. This model had a coefficient of determination of 0.8199, an adjusted

coefficient of 0.5317, and overall was significant with a P-statistic of 0.049. This model has all the indications of a good predictor. The overall model is significant, only three individual variables are insignificant and none of the model assumptions were violated. The normal probability plot in Figure 61 shows how closely this model follows the normal distribution.

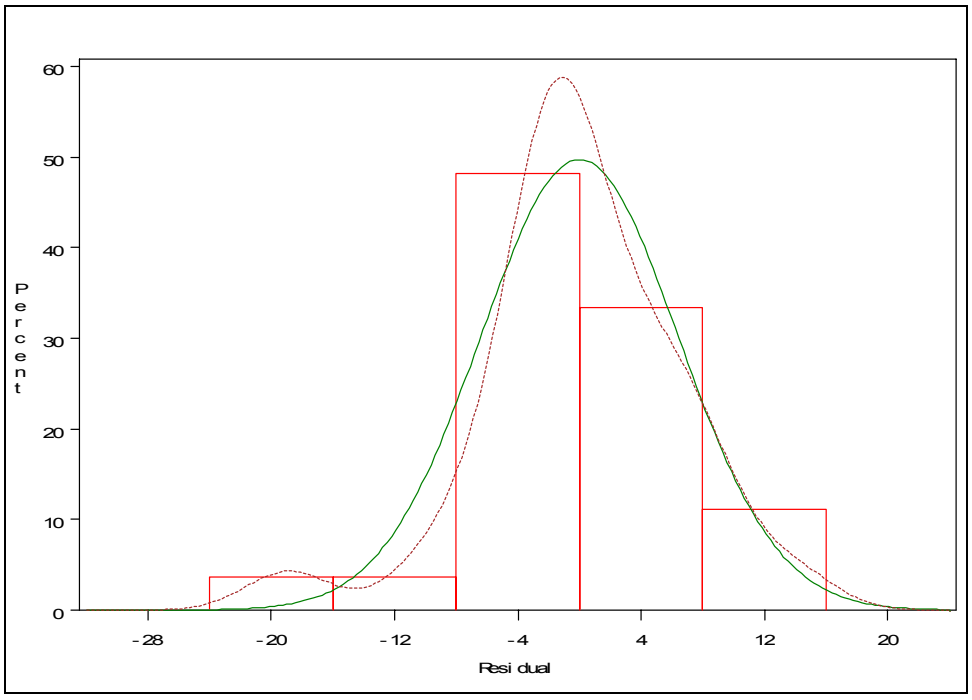


Figure 61: Normal Probability Plot from the Second Model from Variable Group Two

Since this model was so close to working, the three insignificant variables were removed and the model was rerun in the hope that this would be a final model. Unfortunately, this was not to be. The model was run with thirteen variables, and only one variable remained significant. A model with only one variable, besides not doing a good job at predicting an accident rate, will not be useful in finding areas where the road segment differs from other similar section and needs improvement. The lack of significant variables makes this stream of models unacceptable for a final model.

5.2.4 Variable Group Three

Variable group three consists of one more variable than does group one with the addition of the variable *lanelength*. This can be seen in Table 47. This is because in the model elimination process, this variable as a combination of other variables slipped passed the elimination process. By keeping this variable in the group of possible variables, the adjusted coefficient of determination of the primary model increased from 0.7961 to 0.8114.

Table 47: Variable Group Three

Variables	
ospole	length
vol	grade
pmeter	SD
maccess	curves
fence	crest
residential	markings
parkinglots	pavement
lanelength	median
allaccess	widthsida
benches	lane
hazards	widtha
heavyveh	lighting
	parking

The overall model is also significant to greater than 0.05 percent. That and other important numbers can be seen in the ANOVA table below.

Table 48: ANOVA Table for First Model from Variable Group Three

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	17	5558.71886	7.58	6.34	0.002
Error	9	388.29466	43.14385		
Corrected Total	26	5947.01352			
Root MSE	6.56840		R-Square	0.9349	
Dependent Mean	23.14741		Adj. R-Sq	0.8114	
Coeff Var	28.37639				

The parameter estimates of the seventeen variables included in this model are mostly significant and can be seen in Table 49. Only two are significant to less than 0.1 percent and eleven are significant to more than 0.05 percent, leaving four variables than are significant to between 0.05 and 0.1 percent. This appears to be a good start of a model with most of the variables being significant.

Table 49: Parameter Estimates from First Model from Variable Group Three

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	-128.80349	60.38898	-2.13	0.0617
Benches	1	-13.45230	3.96662	-3.39	0.0080
Ospole	1	4.03357	1.61127	2.50	0.0337
Pmeter	1	-0.70446	0.32812	-2.15	0.0603
Maccess	1	2.00412	0.88028	2.28	0.0488
parkinglots	1	-1.57689	0.46397	-3.40	0.0079
lanelength	1	-10.73818	5.26643	-2.04	0.0719
Vol	1	0.00170	0.00054720	3.11	0.0124
residential	1	-0.19403	0.09163	-2.12	0.0633
Grade	1	-0.74894	0.59604	-1.26	0.2406
Curves	1	16.46455	4.77198	3.45	0.0073
Crest	1	7.89294	1.69974	4.64	0.0012
Widtha	1	-5.32279	2.05407	-2.59	0.0291
widthsida	1	2.49877	1.12959	2.21	0.0543
Parking	1	0.21567	0.08139	2.65	0.0265
Lane	1	-11.11592	4.40139	-2.53	0.0325
markings	1	7.03538	4.409001	1.72	0.1195
Lighting	1	1.68026	0.67519	2.49	0.0345

The coefficients of partial determination are also relatively high, which is a good indication of the quality of the parts of the model. As with any model the model assumptions must be reviewed to ensure that the data and the model do not violate any of the assumption. Looking at both the diagnostic graphs, it can be seen that the model assumptions are not violated. The residuals versus the fitted values give a good impression if the model fits the assumptions by showing that there is a constant variance and symmetry about zero, implying that the model follows the normal distribution. This is seen in Figure 62.

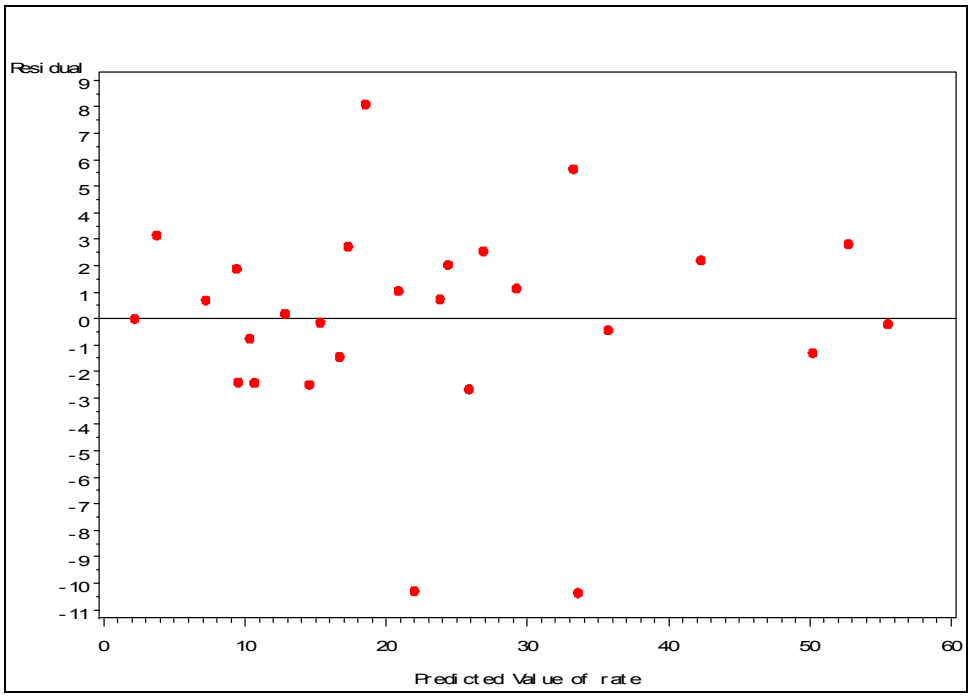


Figure 62: Residuals versus Fitted Values for first Model from Variable group Three

The box plot of the residuals also helps to show this by showing the symmetry in the residuals. In this particular instance there is a small lack in symmetry as there is a greater variation of values on the positive side as can be seen in Figure 63. There are also several points that fall outside of the range of the majority. This would lead to questions

of outlying points expect that there are no points that appear to quality as outliers when looking at the residual scatter plots, so that this is not a cause for concern.

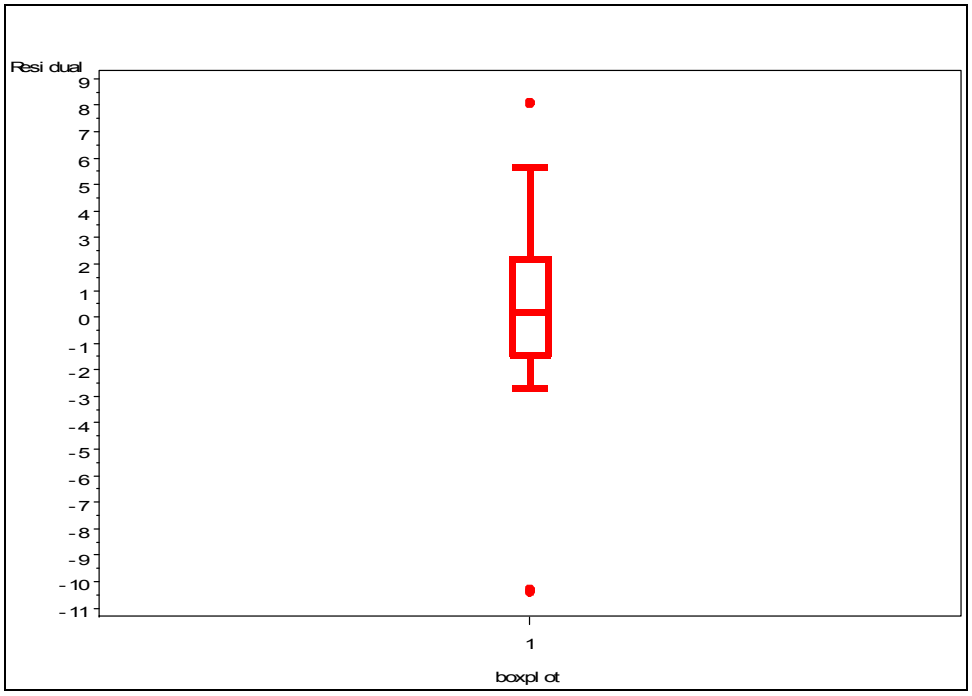


Figure 63: Boxplot for first Model from Variable group Three

Since there were two variables that were insignificant in the model, they were removed and the model was run again. The coefficient of determination and the adjusted coefficient decreased a small amount from 0.9349 and 0.8114 to 0.8882 and 0.7358 respectively, but the overall model was still significant. The new version of the model had fifteen variables, but sadly the previous removal of two insignificant variables caused an avalanche reaction of more variables being insignificant. Now seven variables became insignificant to the model. The model diagnostics still showed that the type of model was appropriate, but the variable parameters being insignificant over rules the positive aspects.

Once again the model was rerun with the insignificant variables removed. This created a model with eight variables and a coefficient of determination of 0.7788. The overall model is highly significant with a P-statistic of 0.0001 (See Table 50).

Table 50: ANOVA Table from Second Model from Variable Group Three

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	8	4631.70060	578.96258	7.92	0.0001
Error	18	1315.31292	73.07294		
Corrected Total	26	5947.01352			
Root MSE	8.54827		R-Square	0.7788	
Dependent Mean	23.14741		Adj. R-Sq	0.6805	
Coeff Var	36.92971				

$$Rate = -48.49 - 1.41benches + 5.07ospole - 1.69parkinglots - 0.38residential + 5.4curves + 3.51crest + 0.13parking + 0.59lighting$$

This time three variables were shown to be insignificant those of benches, curves, and *lighting*. The fact that the number of benches was shown to be insignificant was not unexpected and the percentage of lighting on the segment is also not surprising since most urban arterials have some amount of lighting many with 100 percent lighting. The presence of horizontal curves being found to be insignificant is less expected since horizontal curvature is typically an area where many accidents occur in rural areas.

Again, the insignificant variables were removed and the model was rerun. This time, however, the overall model was shown to be significant and all the remaining variables were shown to be significant. The coefficient of determination was 0.7301 and the adjusted coefficient was 0.6658, both of which are only slightly lower than those of the previous model. The parameter estimates and their standard errors can be seen in Table 51. The only parameter estimate that is not significant is that of the model's intercept. The 95 percent confident interval for the intercept is –4.534 to 21.463 which

does mean that there is a possibility that the intercept is zero. This however is not such a problem that the intercept could be zero as it would be if a parameter estimate for the variable was zero. If the variable's parameter was zero it would mean that the variable possibly should not be included in the model at all, but the intercept gives a value when the variables do not affect the model and a zero value is acceptable.

Table 51: Parameter Estimates for Significant model from Variable Group Three

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	8.46399	6.25048	1.35	0.1901
Ospole	1	5.49178	1.26238	4.35	0.0003
parkinglots	1	-1.56312	0.30367	-5.15	<0.0001
residential	1	-0.30680	0.05085	-6.03	<0.0001
Crest	1	3.32415	1.18981	2.79	0.0109
Parking	1	0.16131	0.05364	3.01	0.0067

The graphical diagnostics show that the model does not violate any of the model assumptions. The residuals versus the fitted values show that there is a constant variance and no points appear to be strong outliers as can be seen in Figure 64. The box plot of the residuals shows a slight tendency for the model to predict accident rates that are lower than those that are actually experienced by the road segments. This can be seen in Figure 65. This is, however, not a large tendency and is not cause for any concern.

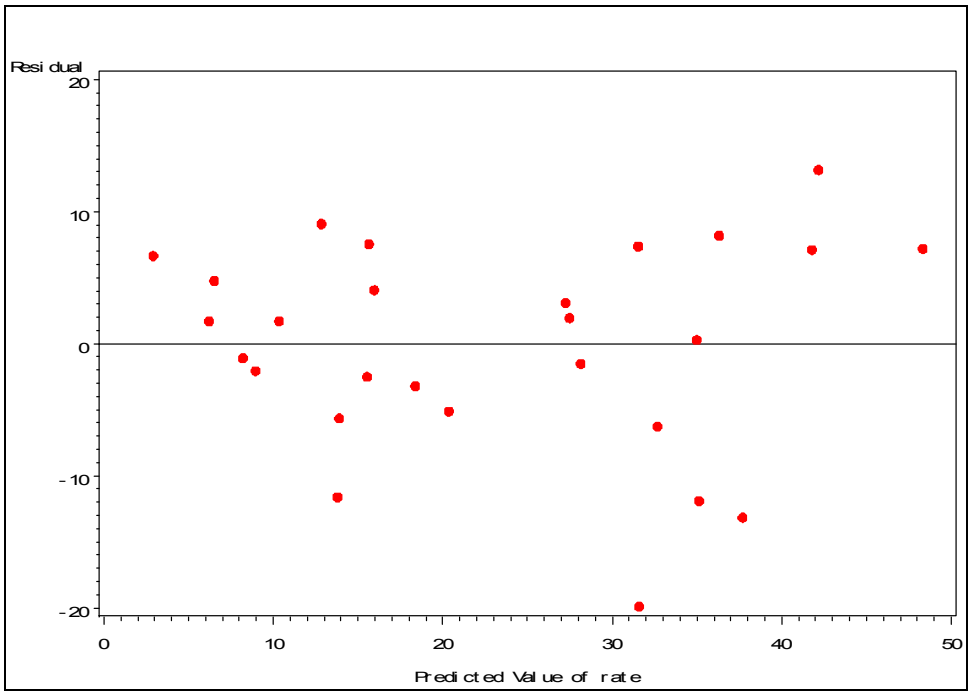


Figure 64: Residuals versus Fitted Values for Significant Model

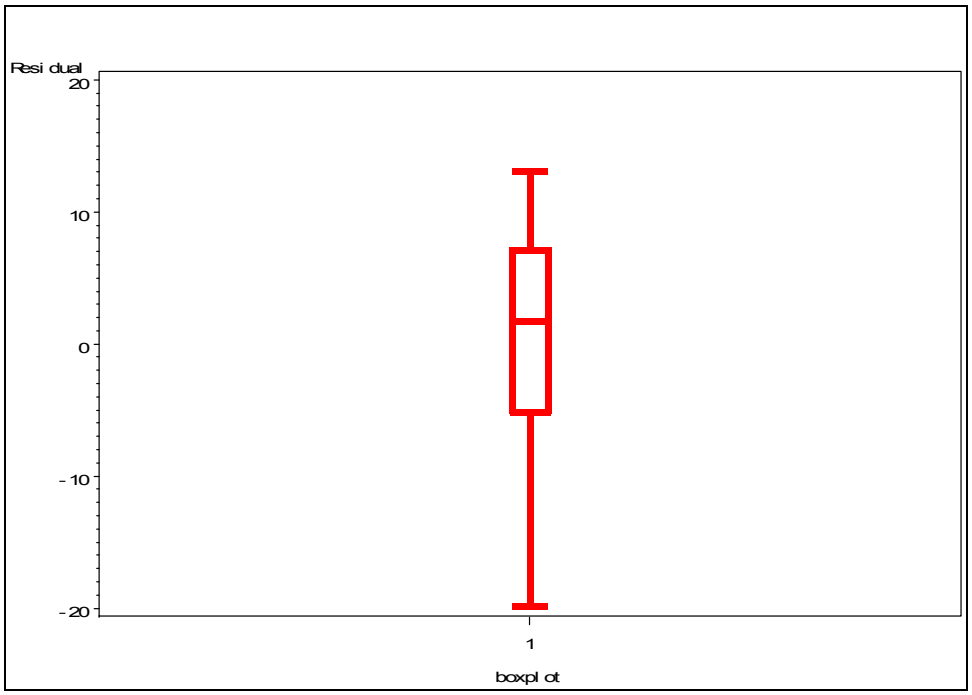


Figure 65: Boxplot for Significant Model

The normal probability plot shows that the model closely follows a normal distribution with only very minor deviations. Figure 66 shows that with the model's distribution falling a little lower than that of the normal distribution. The maximum

value falls along the same plane and minor variations appear on the left hand side of the graph.

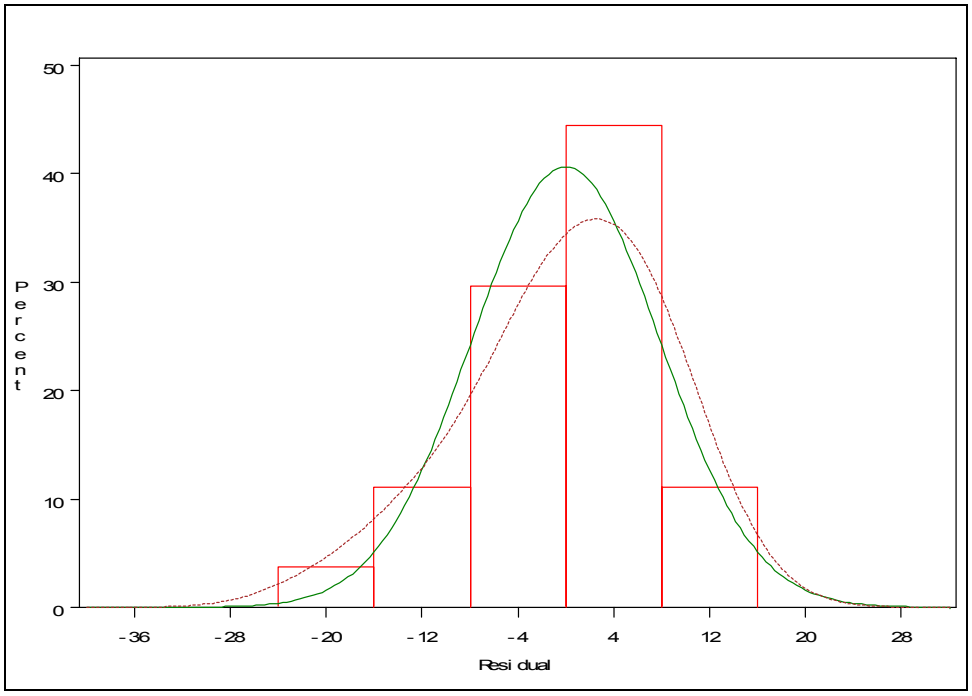


Figure 66: Normal Probability Plot for Significant t Model

This model is composed of only five variables, which will allow for road segments to compare their accident rates to that of other segments with similar characteristics to give a base line to determine if a road segment has an abnormally high accident rate. Since the number of variables is on the low side, it does make identifying locations were improvements could be made more difficult. To try and improve this quality in the model, the last three variables that were removed at one time from the model were removed one at a time to see the effect each one has on the overall model.

The variable representing the total number of benches on the segment was the first to be removed. This was for several reasons including primarily that it had the lowest significance between itself and lighting and curves. Another reason was that so few segments had benches and it was more likely representing the presence of

pedestrians and the use of residential land type helps to represent the major types of pedestrian use that would be seen on the segment. The model without benches had a very similar coefficient of determination to the model with eight variables changing from 0.7788 to 0.7759, but had a better adjusted coefficient changing from 0.6805 to 0.6933. This improvement in the adjusted coefficient of determination helps to show that more variables do not always create a better model. In this instance, it was better to remove the variable benches rather than keep it in the model.

This new version of the model was overall significant, but the remaining two variables curves and lighting still proved to be insignificant as can be seen in Table 52 showing the parameter estimates. The coefficients of partial determination held out the same information, identifying lighting and curves as variables that should be removed from the model. Looking at the 95 percent confidence intervals for the parameter estimates also identified lighting and curves as the only two variables that could possibly have parameters with zero value coefficients making them the only variables that maybe should not be included in the model.

Table 52 : Parameter Estiamates for 7 Variable Model

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	-43.92377	45.93690	-0.96	0.3510
Ospole	1	5.01303	1.26393	3.97	0.0008
parkinglots	1	-1.65882	0.29631	-5.60	<0.0001
residential	1	-0.31326	0.05289	-5.92	<0.0001
Curves	1	5.83270	4.46161	1.31	0.2067
Crest	1	3.32708	1.14025	2.92	0.0088
Parking	1	0.12963	0.05576	2.32	0.0313
Lighting	1	0.54114	0.45446	1.19	0.2484

The graphical diagnostics continue to show that these models do not violate the model assumptions. The plot of the residuals versus the fitted values in Figure 67 show

the constant error variance and show that there is a fairly even distribution around zero, with a slight tendency toward larger negative residuals but not a strong one.

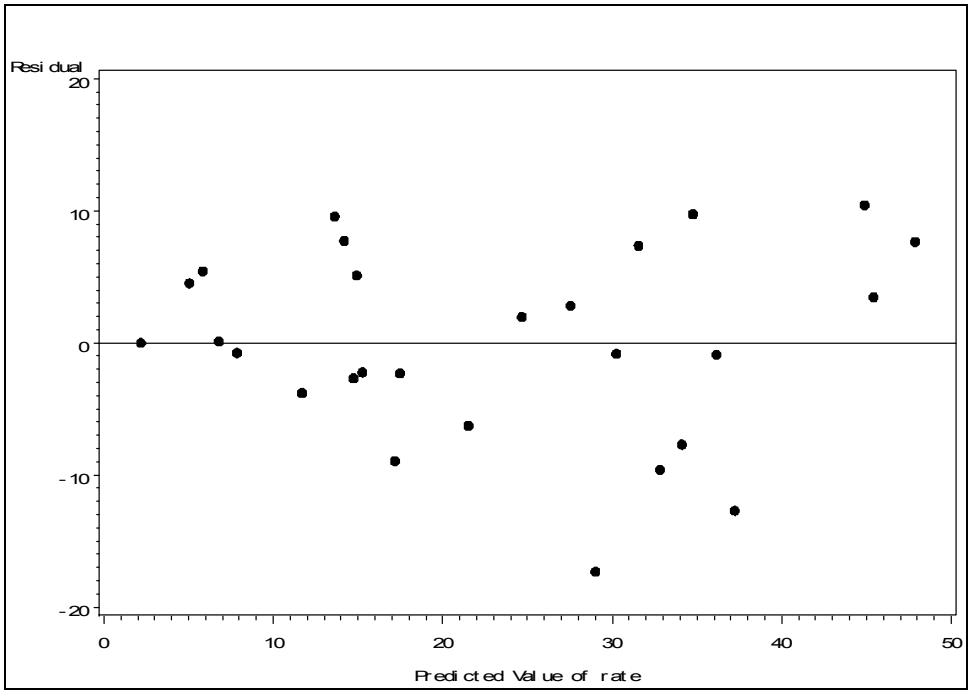


Figure 67: Residuals versus Fitted Values for 7 Variable Model

The normal probability plot shows that there is very little difference between a normal distribution and the distribution that occurs in the residuals which indicates an almost exact normal distribution of the residuals. This can be seen in Figure 68.

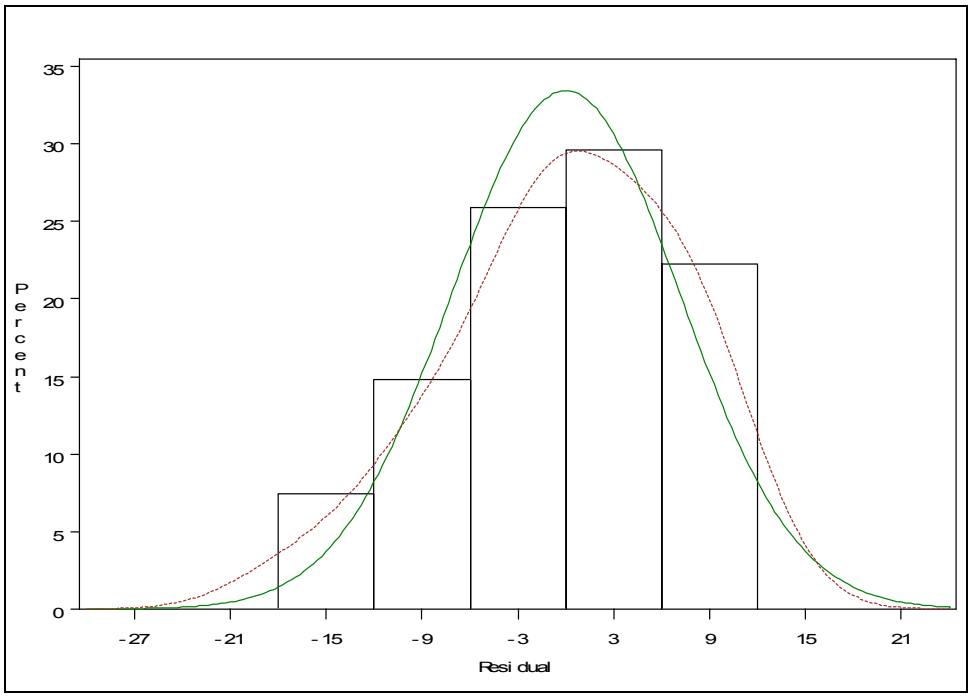


Figure 68: Normal Probability Plot for 7 Variable Model

To check that *lighting* was the better of the remaining variables to remove, the model was run with the variable *lighting* and without the variable *curves*. When this happened, the coefficient of determination was slightly lower than then model with both *curves* and *lighting* in it with a value of 0.7557 versus 0.7759. The adjusted coefficient of determination was also slightly lower at 0.6824 as opposed to 0.6933. The overall model was still significant and the variable lighting was still insignificant.

Since a six variable model with lighting was still insignificant, a six variable model without *lighting* but with *curves* was explored. In the seven variable model curves was of higher significance than was lighting, so this model was expected to perform better. The coefficient of determination is again slightly lower than that of the model with eight variables changing from 0.7788 to 0.7557. The adjusted coefficient of determination, however, is again larger than that of the eight variable model going from 0.6805 to 0.6869. The overall model exhibits full significance with the variable *curves*

remaining insignificant in this version of the model. The alpha level for significance was set at 0.10 and the value from curves is only 0.136 which is only slightly above the limit set. All of the coefficients of partial determination indicate that the variables should remain in the model, so that there is some debate that could occur on whether or not curves should be removed. Since the presence of horizontal curves historically plays a large role in identifying potential accident locations it would be informative if it were left in as a variable in the model. In looking at the 95 percent confidence levels for the parameter estimates, again, the only questionable estimate where the value could be zero is for the one variable that does not reach the full significance that was indicated.

$$Rate = 10.29 + 4.92 ospole - 1.65 parkinglots - 0.33 residentid + 6.87 curves + 3.30 crest + 0.13 parking$$

The graphical diagnostics show that there is no problem perceived with this model violating the linear model assumptions. The plot of the residuals versus the fitted values in Figure 69 shows a very constant error variance and an even distribution between positive and negative residuals. No extreme points are observed on the graph that would imply an outlying point.

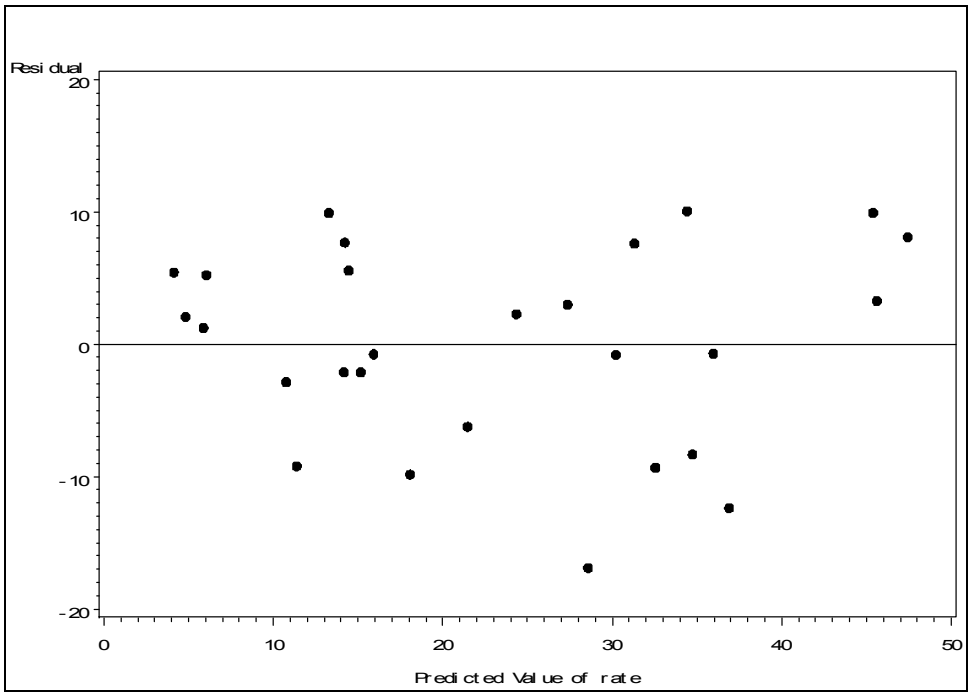


Figure 69: Residuals versus Fitted Values for 6 Variable Model with Curves

Only slight departures from normality can be observed in Figure 70 of the normal probability plot. The distribution for the model has a slightly lower maximum value, but other wise is very similar.

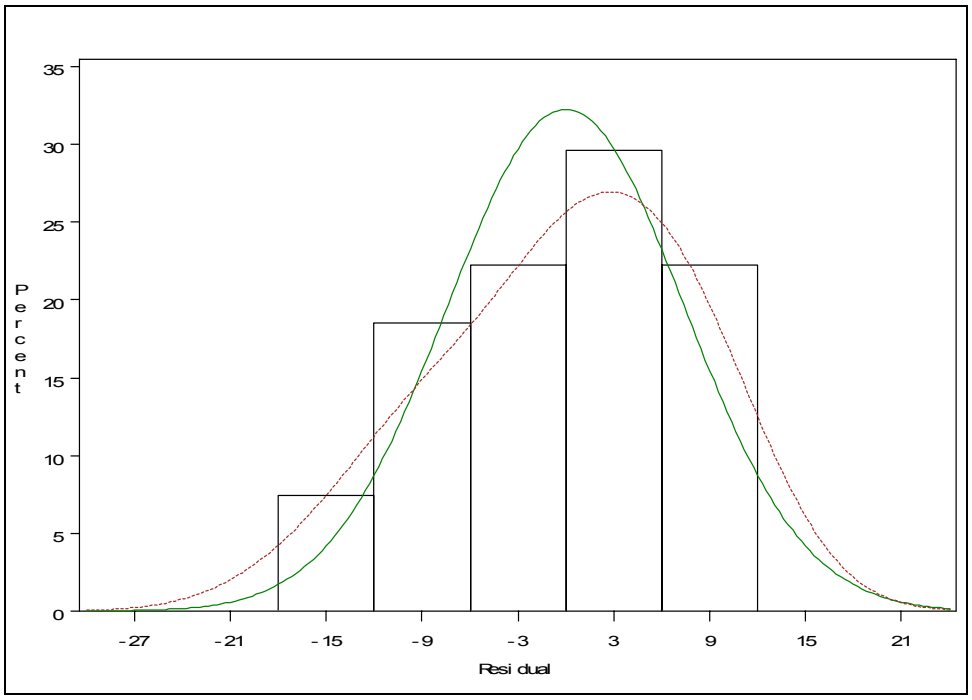


Figure 70: Normal Probability Plot for 6 Variable Model with Curves

This last version of the model from variable group three with six variables including the number of overhead sign poles, the number of parking lots, the percentage of residential land use, an indication of horizontal curves, the largest crest value and the percentage of on-street parking, was the best model in terms of having an acceptable coefficient of determination and adjusted coefficient while also being overall significant and having variables that are significant under statistical testing.

5.2.4.1 Linear Model Summary

In the search for the best possible model to predict the total accident rate, two viable contenders were developed. Variable group one and group three yielded models where the overall model was significant and the individual variables were significant. The coefficients of determination and the adjusted coefficients can be seen in Table 53 to establish the better model.

Table 53: Comparison of Final Linear Accident Rate Models

Variable Group	# of Variables	R^2	R_a^2
1	2	0.4095	0.3603
3	6	0.7591	0.6869

As can be seen in the above table the better of the two models comes from variable group three. This model has a higher coefficient of determination and a higher adjusted coefficient. The overall significance of model is also greater than the model from variable group one. Since the coefficients of the model from group three are higher it is the better choice of a model to predict the total accident rate. The higher coefficients mean that that model can explain more of the variation in the data. Comparison by the coefficients of determination is possible because the models were developed at the same time from the same data set. If they had been created at different times with different data sets, more care would need to be taken instead of this straightforward comparison.

5.2.5 Multiplicative Model Development Process

An additive model silently assumes that the effect of different roadside characteristics are separate and don't effect each other. This is not the best assumption so a multiplicative model was attempted where the roadside characteristics would work with each other to predict the accident rate. The same method was used as when looking for the best risk and accident rate compilation in section 3.4. The first attempt used the variables that were determined to have some significance from the additive model development. The variables that appeared in the top additive models were considered for the multiplicative model. The problem that developed from this automatic transference of variables, is that any variable that had a zero value, whether it was an indicator variable or just a value of zero, did not work well with the multiplicative methodology.

To do the multiplicative model, the log of each variable was taken. So that what is actually modeled is the log of the variable.

The characteristic of not being able to take the logarithm of zero caused many of the variables to be unable to be just transferred from the additive model variable set. Several variables were eliminated totally due to their status as an indicator variable or as a count variable where many segments have a value of zero. A few transformations were attempted were count variables are concerned. If the count variable had values on almost every segment, the zero value was changed to a very small number that represents zero without actually being written as zero. The variables *parkinglots*, *allaccess*, *parking*, *maccess*, and *residential* were transformed this way. Using the logarithm of the variable also increased the correlation of several of the variables causing some to be eliminated from further model development.

The first attempt at the multiplicative model had a very large coefficient of determination with $R^2 = 0.977$. The large coefficient of determination may be indicating more than that the model is a good fit for the data, but also may be showing that the model is overfit to the data set and not transferable to other data sets. The adjusted coefficient was not as large, but it was still good with $R_a^2 = 0.7241$. Another issue that was found is that of the P-statistic for the overall model. It shows that the overall model is insignificant, with a statistic of 0.3791, implying that there is something incorrect with the model

The variable coefficients for this model also have a P-statistic that shows that none of the variables were significant to the selected level of 0.10. Since none of the variables were significant, for further investigation any variable that was significant to

less than 0.5 was eliminated. This created a basic level to see if the variables were significant in further development and if a multiplicative model developed this way was possible.

In addition to the significance of the model and the variables being a problem, some of the graphical diagnostics also indicated this. The most severe problem was seen in the normal quantile plot, which should show the residuals falling along or near the solid line (See Figure 71). The points in this situation are all well above the line which implies that this model does not do a good job at explaining the variation in the data.

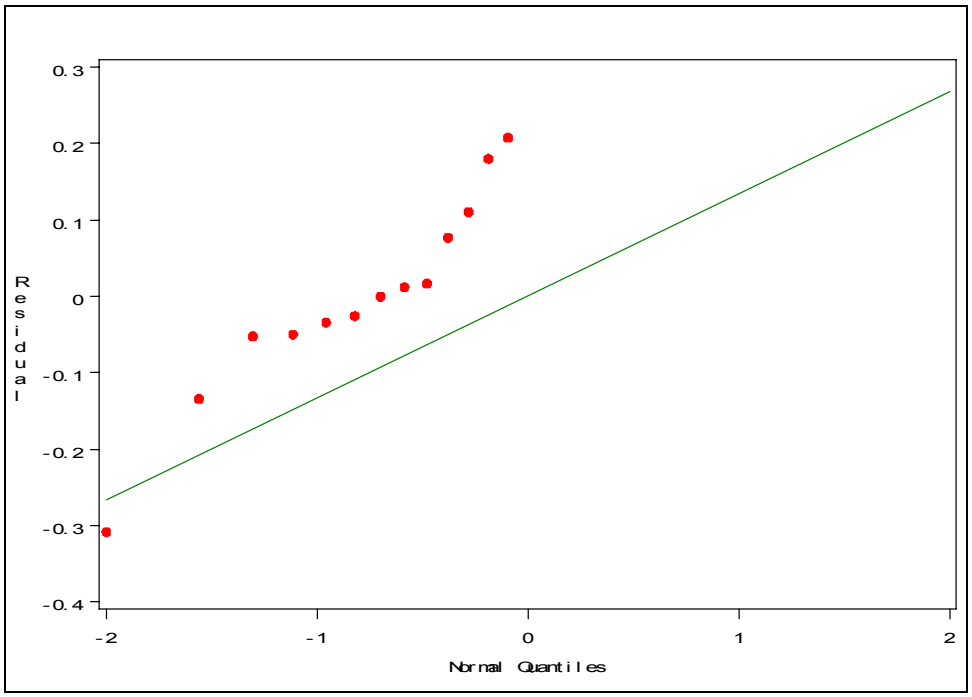


Figure 71: Normal Quantile Plot for First Multiplicative Model

The next step in the multipliable model development looked at the model created from the remaining variables after the five least significant variables were removed from further consideration. The new model has an overall significance that is acceptable as can be seen by the P-statistic of 0.0005. The coefficient of determination is lower than in the previous model, but is still high at 0.6707. In this model, besides the overall model

being significant, some of the individual coefficients are also significant, with *length*, *lighting,* and *pole* being significant to greater than $\alpha = 0.10$.

$$Rate = e^{-23.091} vol^{0.107} length^{-0.968} grade^{-0.0088} crest^{-0.0608} lighting^{7.587} pole^{0.581}$$

There is some concern with some of the coefficients due to their standard errors. Some of the standard errors show that with only one deviation the coefficient could become zero, which causes some concern for the overall model, but this only affects the variables that are not significant in the model to begin with. Unlike the first model, the diagnostics do not give any indication of a violation in model assumption.

The second model, like the first, still had variables included in the final version that were not significant. So despite the overall model working, the insignificant variables were removed in anticipation of the remaining variables keeping their significance and the overall model being significant.

The third model is significant and while the coefficient of determination decreased slightly than from the second model, 0.6672 from 0.6707, the adjusted coefficient of determination increased from 0.5719 to 0.6238, showing that this is the better of the two models. The coefficients and other numbers of interest can be seen in Table 54 below.

Table 54: ANOVA Table for Multiplicative Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	3	10.05655	3.35218	15.37	<0.0001
Error	23	5.01511	0.21805		
Corrected Total	26	15.07167			
Root MSE	0.46696		R-Square	0.6672	
Dependent Mean	2.90341		Adj. R-Sq	0.6238	
Coeff Var	16.08302				

The significance for the individual coefficients also increased slightly with all of the variables, including the intercept, being significant to greater than 0.10. The standard

errors for all of the coefficients are also acceptable in that one deviation can be taken and there is no concern with the parameter estimate possibly becoming zero. This can be seen in Table 55.

Table 55: Parameter Estimates for Multiplicative Model

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	-28.29984	10.22157	-2.77	0.0109
llength	1	-0.95851	0.25768	-3.72	0.0011
llighting	1	7.80913	2.17984	3.58	0.006
lpole	1	0.56736	0.30369	1.87	0.0745

The final diagnostics to check since the model and all variables are significant are the graphs to check model assumptions. The residuals versus the fitted values show that there is a constant variance (See Figure 72). The studentized residuals versus the fitted values shows the same thing with the addition of being able to identify outliers, of which there are none to be concerned about in this model that can skew the model in one direction or the other (See Figure 73).

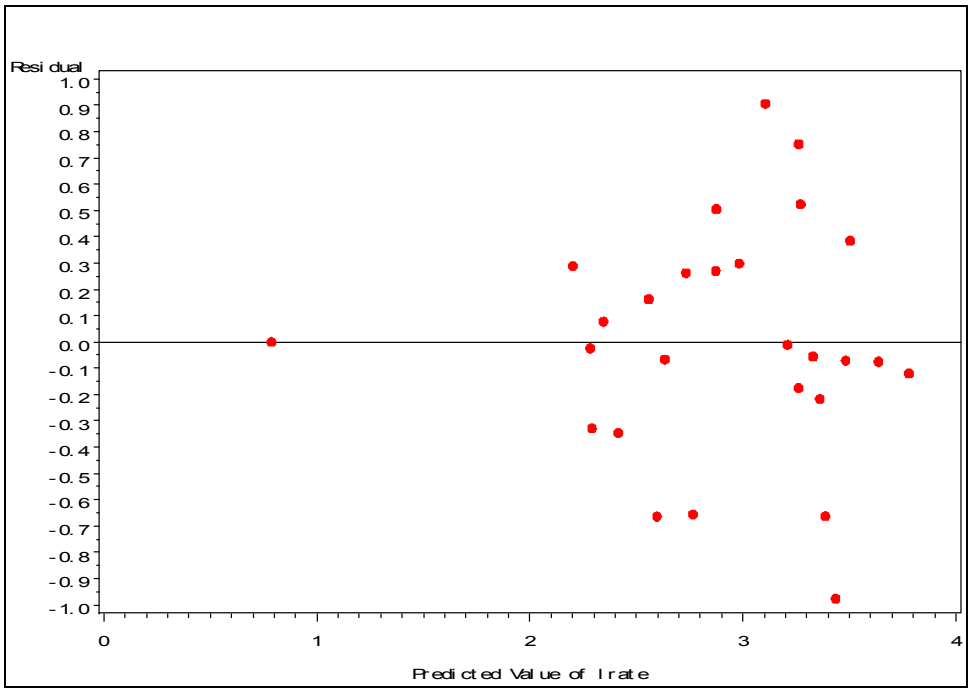


Figure 72: Residuals versus Fitted Values for Multiplicative Model

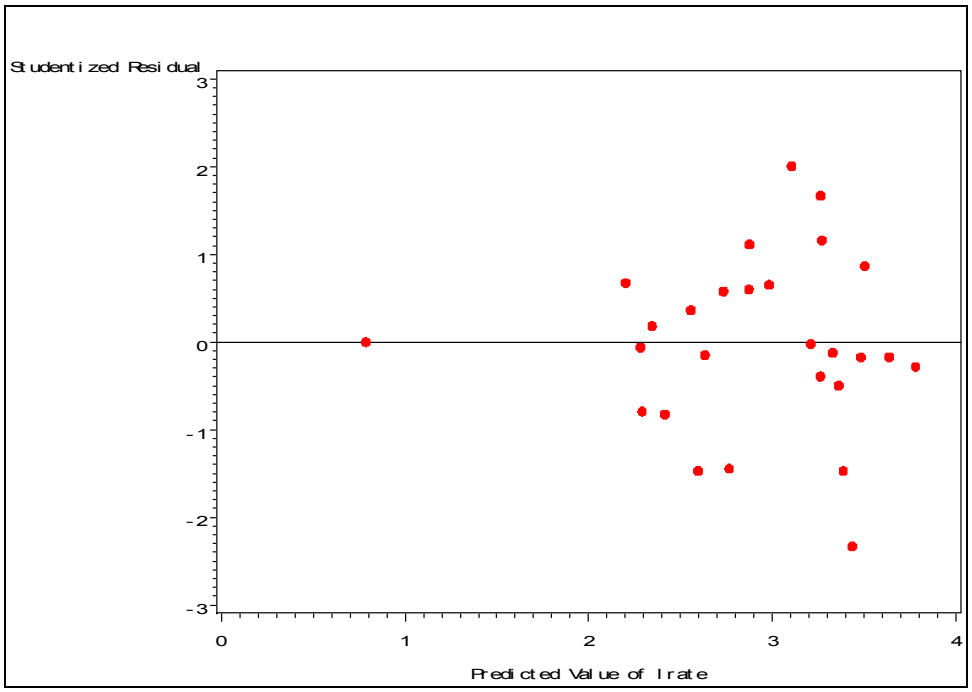


Figure 73: Studentized Residuals versus Fitted Values for Multiplicative Model

The box plot of the residuals in Figure 74 shows that they are highly symmetric with a slight skewness towards positive residuals, which implies that the model will have a tendency to predict a higher accident rate than the actual rate. This is, however, a very minor tendency and not a reason to disregard this model.

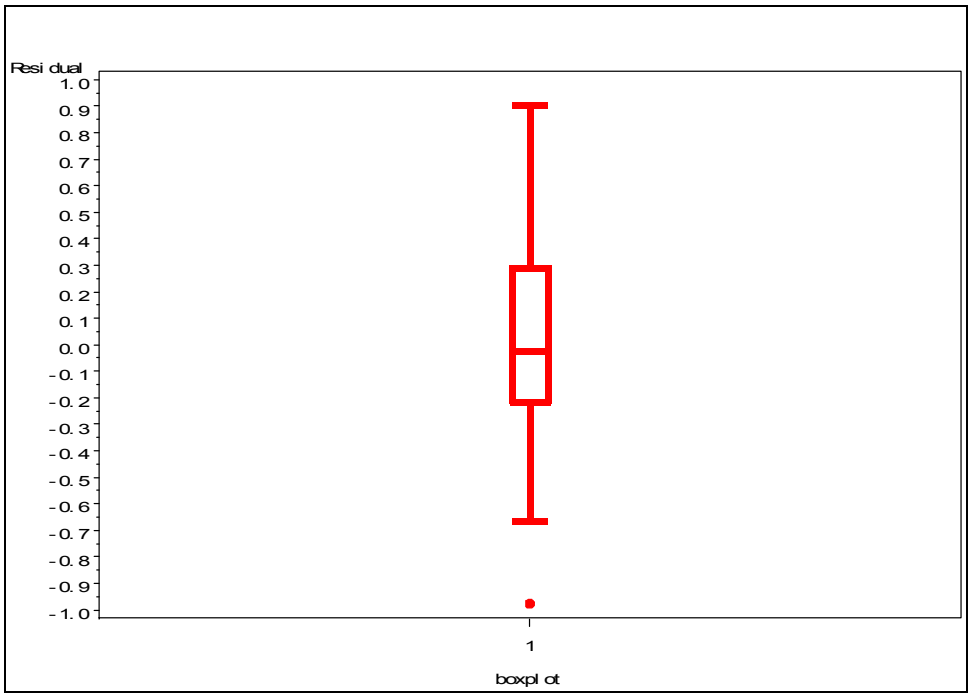


Figure 74: Boxplot of Multiplicative Model

The normal quantile plot, seen in Figure 75, also shows that the model follows the assumptions for a normal distribution with the residuals falling along the line. There is no obvious departure from the normal line in a recognizable pattern that could indicate a model violation.

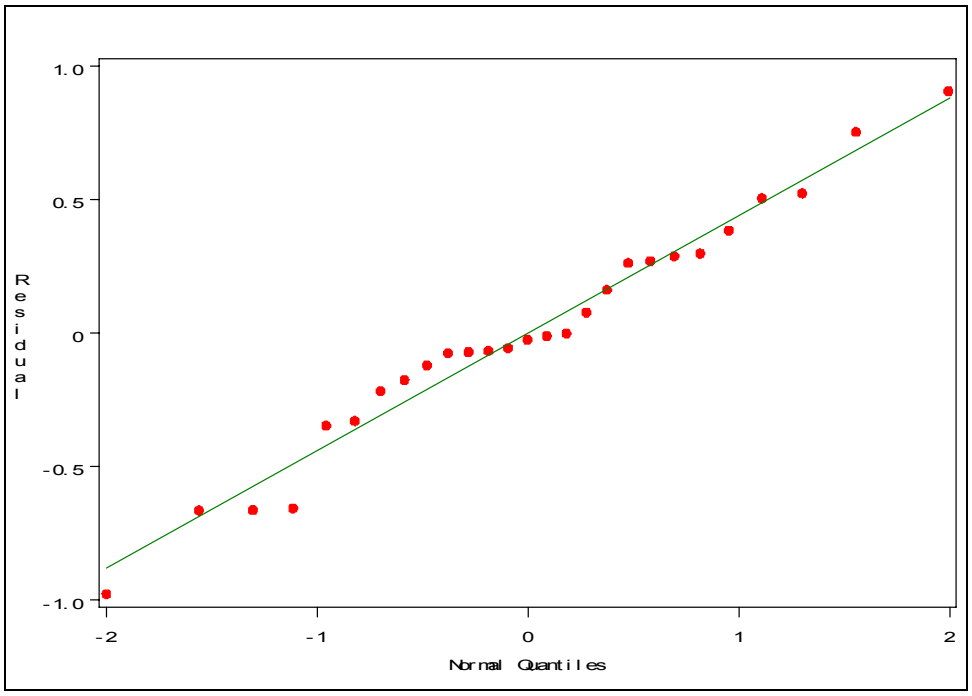


Figure 75: Normal Quantile Plot of Multiplicative Model

There are only very minor deviations from normality that can be seen in the normal probability plot in Figure 76. The dashed line, which represents the model's distribution, almost exactly follows the solid line, which is a normal distribution. This indicates that the model does not violate any of the model assumptions.

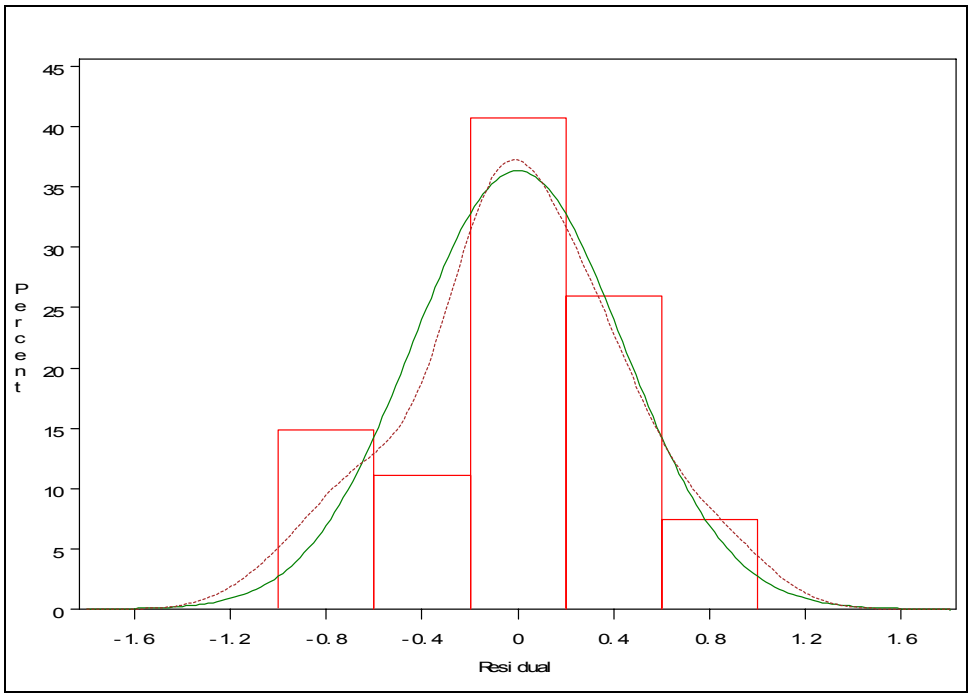


Figure 76: Normal Probability Plot of Multiplicative Model

The parameter estimates for the variable length, in the second and third variations of this model had a negative coefficient near negative one. This is suggestive of a rate. The remaining variables were transformed into densities, to explore whether or not the variable length could be dropped from the model. Despite, the coefficient near negative one, the variable length was never shown to be insignificant even when all the other variables were densities or percentages. This implies that length in this model format remains an important factor towards predicting the crash rates for the total number of accidents.

$Rate = e^{-28.3} length^{-0.959} lighting^{7.809} pole^{0.567}$ is the only model that was developed where the overall model and each of the individual variables passed their significance tests, and while this is the best version of the multiplicative model, only three variables are included in it; *length*, *lighting*, and *pole*. From a modeling standpoint this is fine, but for traffic engineers hoping to tell what part of a road section to improve this is not

completely helpful. The engineers will be able to determine if their road section differs greatly from other similar sections, but with so few variables included in the model there is no clear way to be able to estimate changes in accident rate by improvements. Length can be improved by becoming shorter or longer only by changing signal locations, which is rare in urban settings. Lighting can also be improved only so much until the full segment is lit, but in urban locations most arterial roads are already fully lit. The number of poles can also be changes, but some will be necessary to mark street names and other important driving directions. So while this model does a good job at predicting accident rates, it does not do a good job in helping to make decisions on where to spend the limited roadway improvement/safety dollars.

5.2.6 Injury Accident Model

In the same way that there were three variable groups when looking for the model to predict the total number of accidents, the same three variable groups were used in the process for an injury accident model. This is possible since the same data set is being used and the correlation between variables does not change with a change in dependent variables, allowing the same variables to be eliminated from further consideration. The dependent variable in this model is the accident rate for injury accidents only. This classification includes all types of injuries, including fatalities, and excludes property-damage only accidents. Injury accidents account for approximately one-third of the crashes observed in the study area. Being able to predict the number of injury accidents, or the injury rate is important because the majority of resources for responding to accidents and the care of their victims come from this group. The more injury accidents

prevented the fewer resources needed to be set aside and earmarked toward emergency response and care and could be used for repairing and updating roadway conditions.

5.2.6.1 Variable Group One

Variable group one consisted of the top twenty-four variables under consideration. The possible combinations of variables were sorted by their adjusted coefficients of determination to choose the best possible model that could come from the twenty-four variables. The top model contained seventeen variables with an adjusted coefficient of determination of 0.8508. The coefficient of determination and other values can be seen in Table 56. The same α level of significance is used for the injury accident model as was used for the total accident model, that of 0.01.

Table 56: ANOVA Table of Injury Accident Model Variable Group One Trial One

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	17	924.0355	54.35503	9.72	0.0008
Error	9	50.32513	5.59168		
Corrected Total	26	974.36063			
Root MSE	2.36467		R-Square	0.9484	
Dependent Mean	7.9237		Adj. R-Sq	0.8508	
Coeff Var	29.84303				

The overall model passed the test for significance with a P-statistic of 0.0008. Almost all of the variables in the model also passed their individual significance test with only two failing. The variables that represent the number of minor access points and the percentage of heavy vehicles in the traffic mix did not pass their significance tests. With P-values of 0.2668 and 0.1536, respectively, these two variables needed to be removed from the model by the significance a criteria. The same two variables were the only ones whose 95 percent confidence limits for the parameter estimates included zero which

indicates that there is a chance for the coefficient to be zero and the variable not part of the model. Similarly, the partial coefficient of determination only indicates *maccess* and *heavyveh* for exclusion.

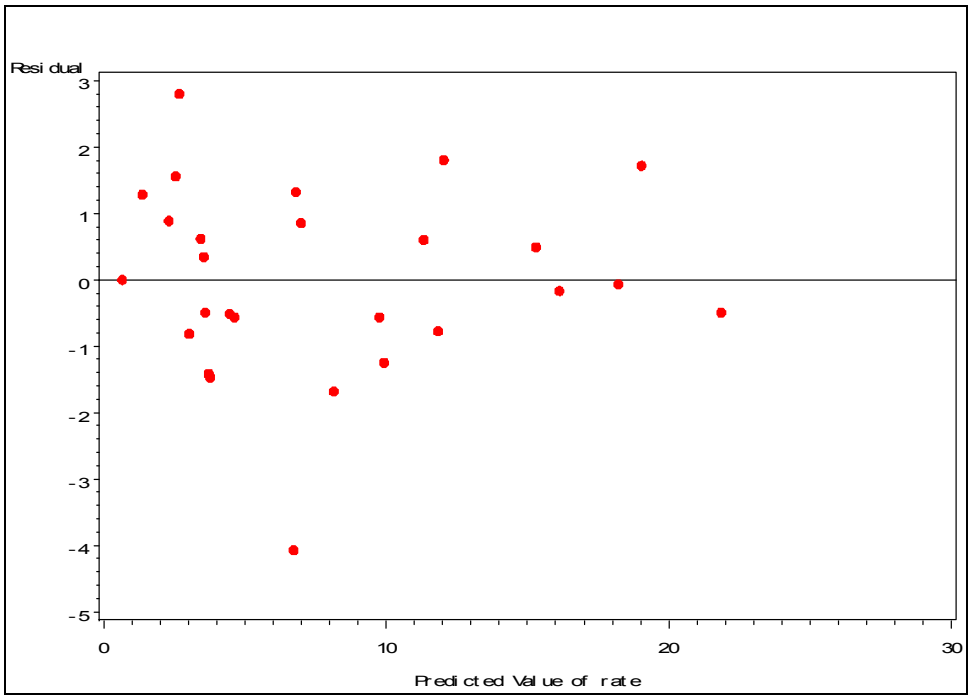


Figure 77: Residuals versus Fitted Values for Injury Accident Rate Variable Group One

The graphical diagnostics for this model indicate that none of the model assumptions are violated. Figure 77 shows the residuals versus the predicted values which indicates that there are not any outlying points and that the variance is approximately constant based on the small data set available.

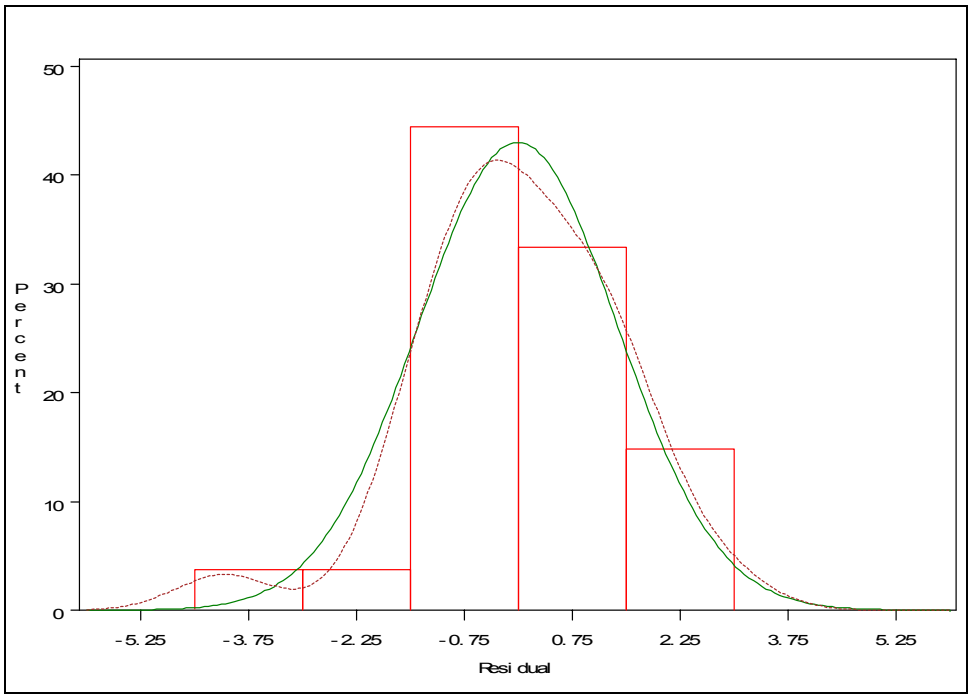


Figure 78: Normal Probability Plot for Injury Accident Rate Variable group One

The normal probability graph indicates how closely the residuals of the model follow a normal distribution. There is a small amount of variation on the left hand side of the graph and on the top as can be seen in Figure 78. Despite the good qualities of this model, there are two variables that are insignificant and further development is needed.

The next step in the model development consisted of a model that had only fifteen variables with maccess and heavyveh being removed from the potential variables. This second model passes the overall significance test with a P-statistic of 0.0002. The coefficient of determination decreased slightly from 0.9484 to 0.9319 with a corresponding minimal decrease in the adjusted coefficient from 0.8508 to 0.839; these and other statistics can be seen in Table 57.

Table 57: ANOVA Table for Variable Group One Final Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	15	907.99025	60.53268	10.03	0.0002
Error	11	66.37038	6.03367		
Corrected Total	26	974.36063			
Root MSE	2.45635		R-Square	0.9319	
Dependent Mean	7.9237		Adj. R-Sq	0.839	
Coeff Var	31.00006				

Surprisingly, at this early stage in the model selection, all of the variables passed their individual significance tests at the specified alpha level of 0.10. The partial coefficient of determination also did not identify any variables for possible elimination. Review of the 95 percent confidence limits, did show one variable whose interval included zero that of *curves*, which indicates the presence of one or more horizontal curves. So there is a possibility that one variable maybe should not be in this model, but only one of the possible identifying traits of that indicates that to be true.

The graphical diagnostics do not indicate any reason for this model to be unacceptable. The residuals versus the predicted values plot indicates that the residuals have a constant variance and are basically symmetric about zero, with perhaps a slight tendency towards the negative side, predicting values that are lower than they really are as can be seen in Figure 79.

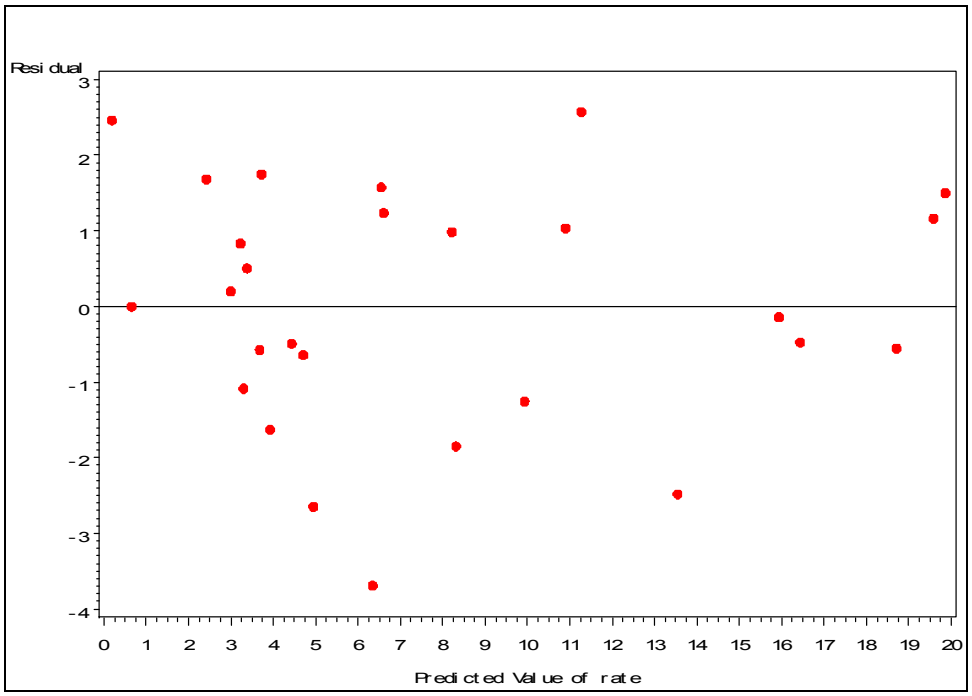


Figure 79: Residuals versus Fitted Values for Variable Group One Final Model

The distribution that the residuals follow almost completely follows that of a normal distribution without the slight extra peak on the left hand side that the previous model had. The residual distribution and a normal distribution can be seen in Figure 80. The normal distribution is the solid line while the dashed line that follows closely is the distribution from the residuals from this data set.

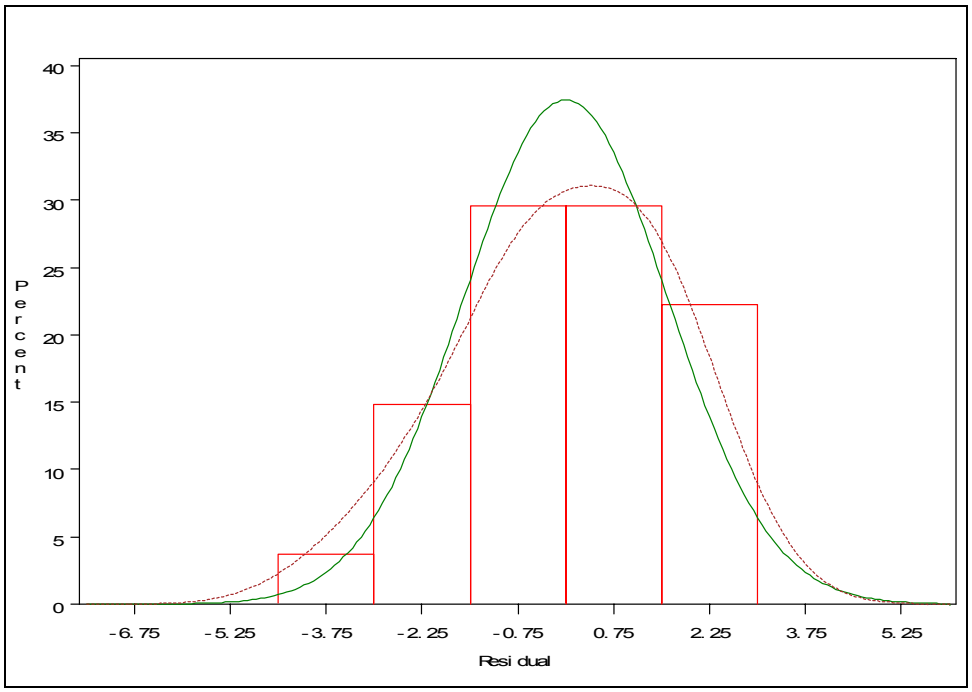


Figure 80: Normal Probability Plot for Variable Group One Final Model

5.2.6.2 Variable Group Two

Variable group two consists of twenty-six possible variables. This includes two more over group one, those of *pole* and *lanelength*. This is the same second group of variables that was used for the development of the prediction models for the total number of accidents that occur on a road segment. The variables were run through a selection process that used the adjusted coefficient of determination to determine the top models. The top model from variable group two consisted of twenty-five variables with a coefficient of determination of 1.0. While this is the maximum allowable value for the coefficient of determination, it is not always a good idea to reach the maximum allowable value. This shows that while the model is a good representation of the given data set, with other data, there will most likely be a problem since the model is over fit to the original database. Event the adjusted coefficient of determination indicates that the model is over fit with a value of 0.9998. Though the coefficients of determination were

very high, almost all of the variables passed their individual significance tests with only one variable failing the test. The variable that represents the percentage of on-street parking was found to not pass the significance test, and only just barely. Parking had a P-statistic of 0.103 and it needed to be smaller than 0.100. So this was a very close call. The overall model was significant with not as a high a P-statistic as would be thought with such a high coefficient of determination. The P-statistic was only 0.0121, but that is enough to call the model significant. The graphical diagnostics hold true to the good quality of the model as expected by the coefficients of determination. The boxplot of the residuals shows that they are symmetrical about zero implying the constant variance of the error residuals. This can be seen in Figure 81.

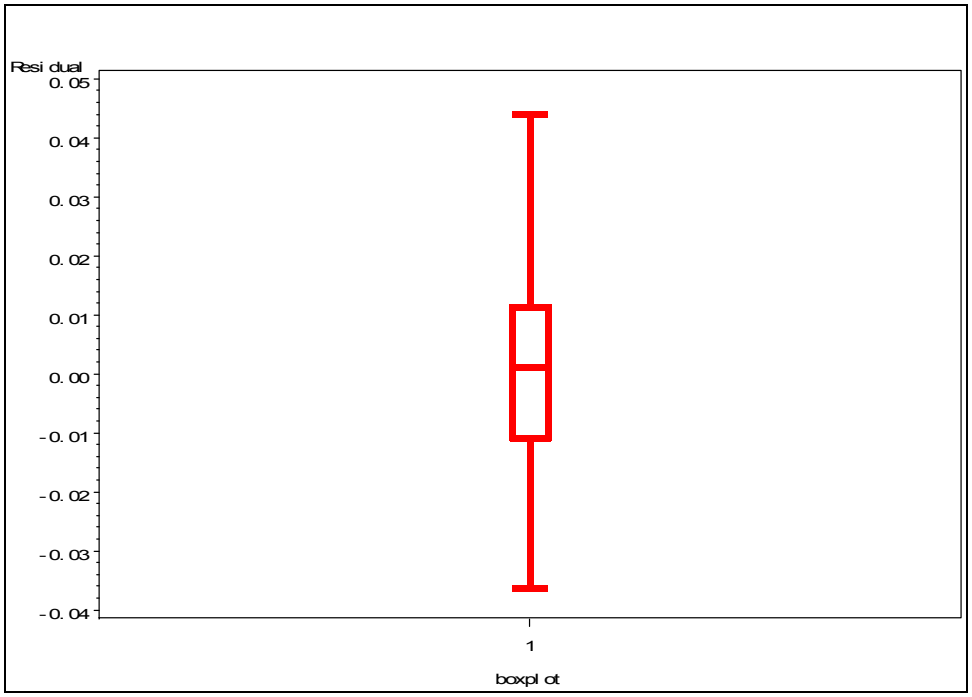


Figure 81: Boxplot for Variable Group Two Preliminary Model

The normal probability plot also indicates the high quality of the model with the distribution created from the residuals closely following that of a normal distribution as can be seen in Figure 82. There are only minor deviations on the left side of the

distribution with the peak of the residual distribution being slightly higher than that of the normal distribution.

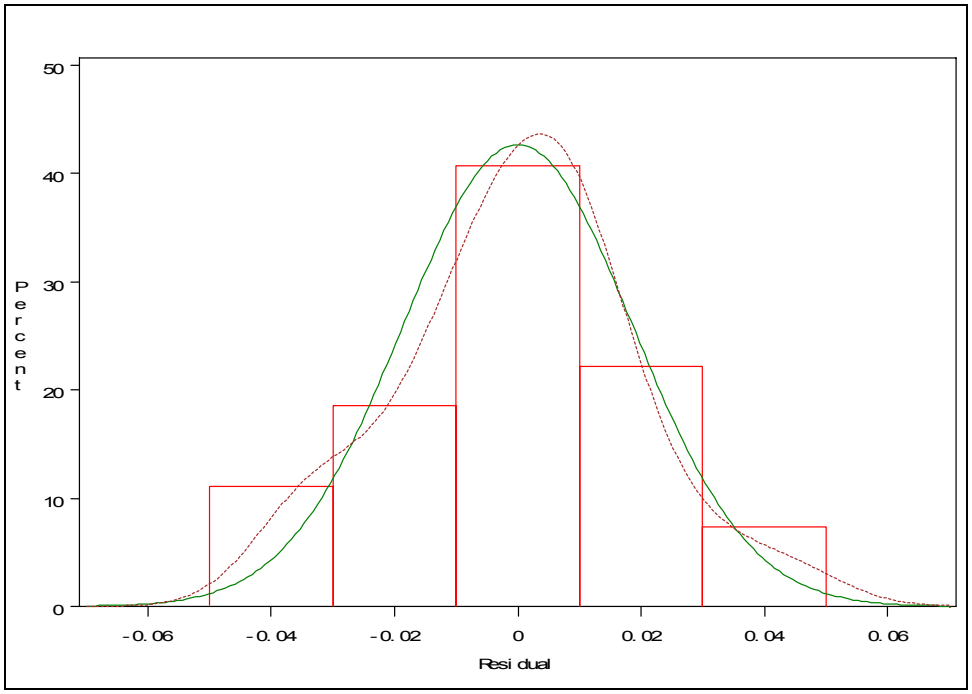


Figure 82: Normal Probability Plot for Variable Group One Preliminary Model

Since one of the variables failed its significance test, it was removed and the model was rerun. There was little change in the coefficients of determination and the adjusted coefficient with a change from 1.0 to 0.9996 and from 0.9998 to 0.9953 respectively. This model, however, passes the overall significance test with a higher statistical value of 0.0043 instead of 0.0121.

In this second draft of the model, all of the remaining twenty-four variables passed their individual significance tests. In this second draft of the model, all of the remaining twenty-four variables passed their individual significance tests. The only variable were there is some concern is that of SD, an indicator variable for problems with stopping sight distance, where the 95 percent confidence limits show that there is a possibility that the coefficient for this variable could be zero. That shows that there is a

small possibility that SD should not be included in the overall model, but since it passed the individual significance test, this variable was left in the model. It has historically been found to be significant in affecting accidents, so there was not a strong concern with leaving the variable in the model.

The graphical diagnostics showed that while the model does not violate any of the model assumptions, such as constant variance, this is not the best possible model available. Figure 83 shows the studentized residuals versus the predicted values which shows the residuals to be evenly distributed about zero and have a constant variance.

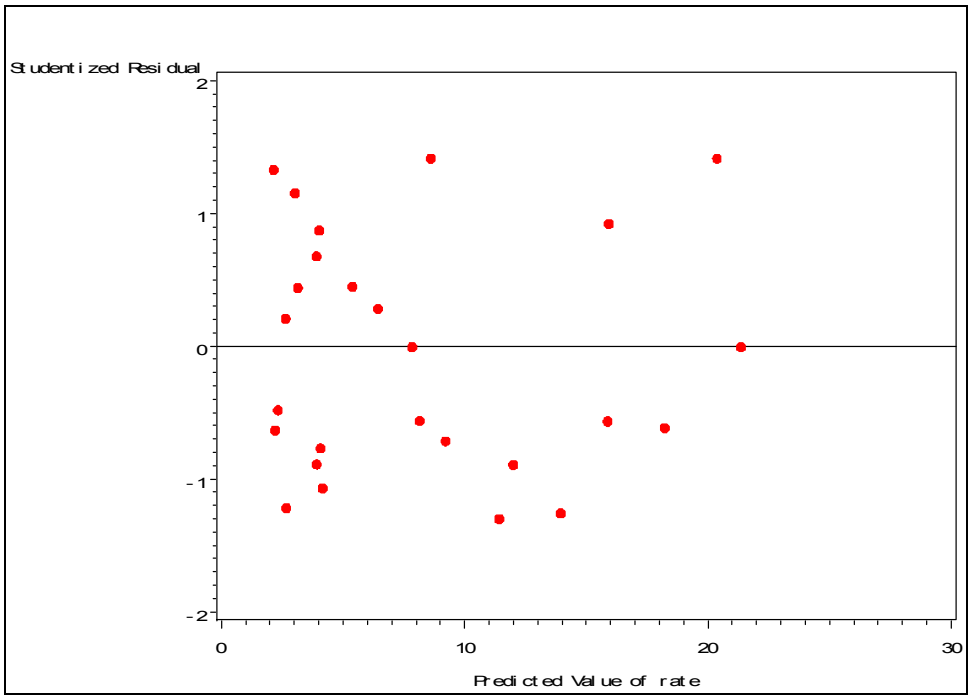


Figure 83: Studentized Residuals versus Predicted Values for Variable Group Two Final Model

The normal quantile plot on the other hand, shows a variation in the data that appears to possibly have a variation that could be describe by some function. The points deviate from normal on the positive or negative side and then abruptly switch with a sharp increase in deviance as can be seen in Figure 84. This implies that there could be a model that follows the normal distribution closer.

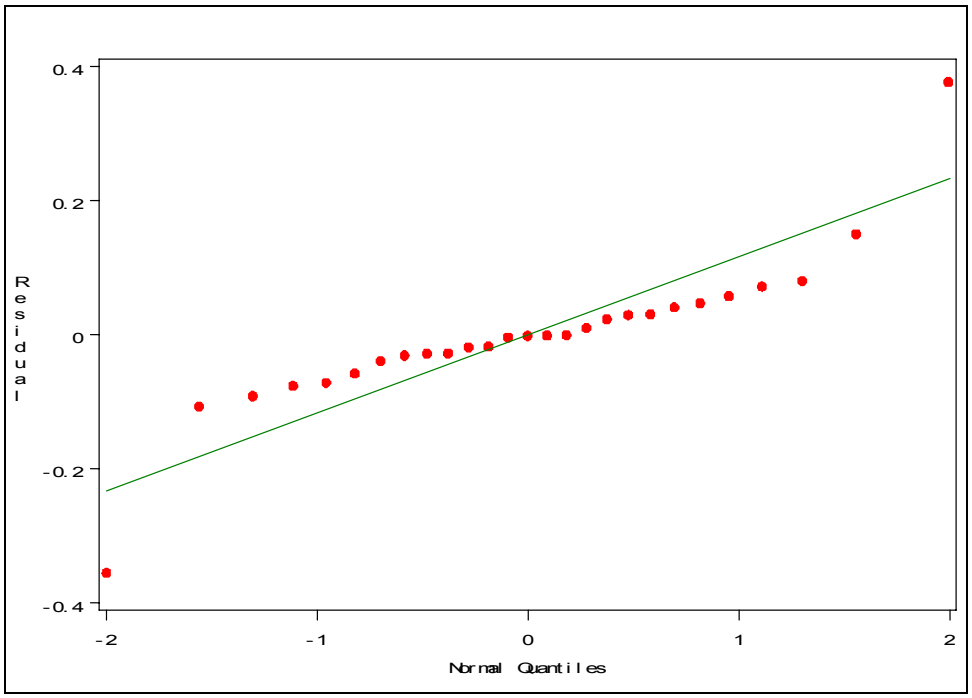


Figure 84: Normal Quantile Plot for Variable Group Two Final Model

The normal probability plot confirms this idea that other models conform to the model assumptions better. The distribution formed from the residuals rises sharply above that of the normal distribution with the peak falling between 30 and 40 percent higher. There is also a deviation in both extreme sides with the distribution formed from the residuals having small peaks on each of the extremities while the normal distribution remains smooth. This can be seen in Figure 85. These graphical diagnostics show that while numerically this model appears to be a close fit to the data and a good representation, there should be a model where the residuals follow the normal distribution closer.

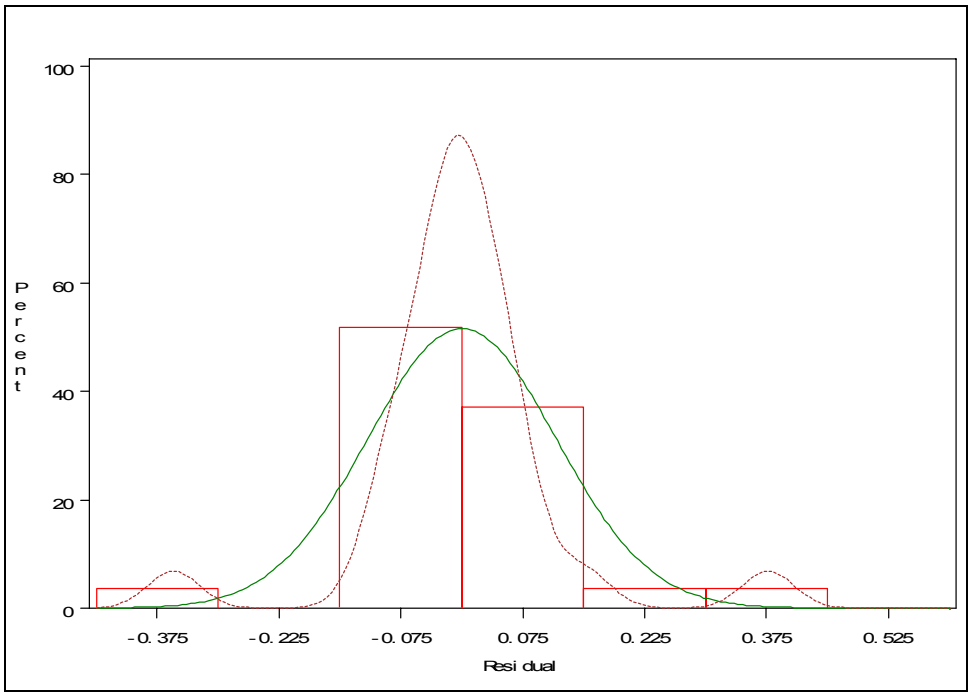


Figure 85: Normal Probability Plot for Variable Group Two Final Model

5.2.6.3 Variable Group Three

Variable group three consist of the variables in group one with the addition of the variable *lanelength* (Refer to Table 47 in section 5.2.4). Using the same methods as the other variable groups, the model selection criteria of the adjusted coefficient of determination was used to choose the top model that could be formed from this group of variables. The first version of this model had the highest adjusted coefficient of determination at 0.9026 and consisted of twenty-one variables. The coefficients of determination can be seen in Table 58.

Table 58: ANOVA Table for Variable Group Three Preliminary Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	21	956.10864	45.52898	12.47	0.0054
Error	5	18.25199	3.65040		
Corrected Total	26	974.36063			
Root MSE	1.91060		R-Square	0.9813	
Dependent Mean	7.92370		Adj. R-Sq	0.9026	
Coeff Var	24.11248				

The model overall passed its significance test with an F-value of 12.47. The individual variables mostly passed their significance test with only the variables *benches* and *curves* not passing. These two variables were only barely insignificant with P-values of 0.1064 and 0.1381 respectively. They were also the only variables were zero appeared in their 95 percent confidence limits for the parameter estimates, which shows that there is a possibility that the variables should not be included in the final model.

The graphical diagnostics support the fact that the model form chosen is the correct one. There was no indication of an inconstant variance and the residuals follow closely along a normal distribution as can be seen in Figure 86.

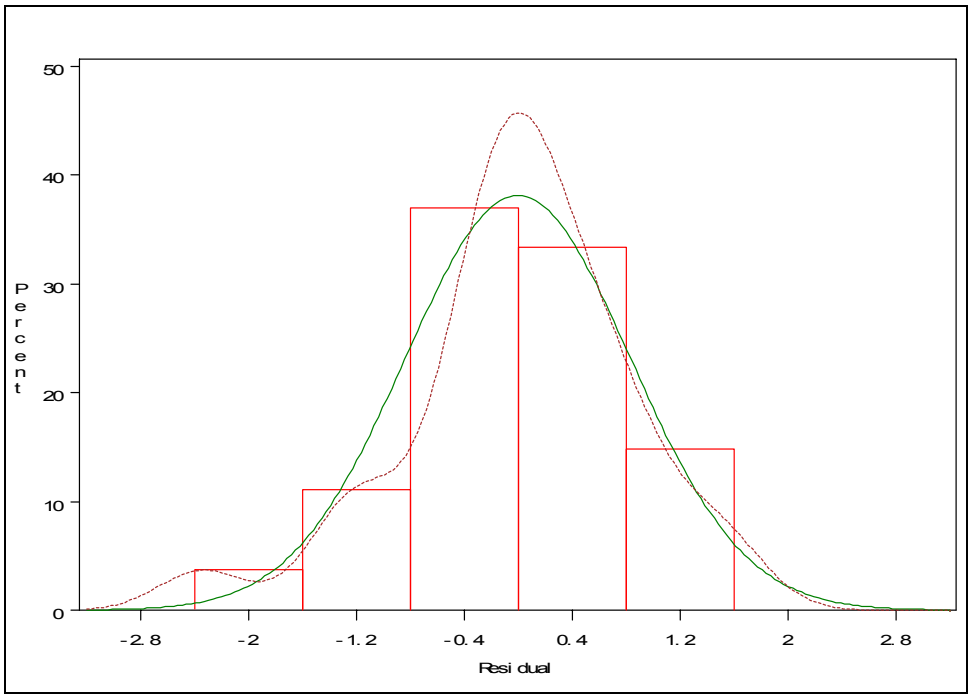


Figure 86: Normal Probability Plot for Variable Group Three Preliminary Model

Since two of the variables were not significant in the model, the model needed to be rerun without those two variables. This second version of the model had nineteen variables and an only slightly lower coefficient of determination at 0.9379 from 0.9813 previously. The adjusted coefficient of determination, however, changed more dramatically at 0.7692 from the previous 0.9026. This is a large change in the coefficient, but the value is still large enough to make exploring this avenue worthwhile.

The model for the second version was found to pass the overall significance test with a P-value of 0.0136. The individual parameter estimates did not fair so well as in the previous model, with four failing to pass their significance tests. The variables *allaccess*, *witha*, *lane* and *markings* were found to be insignificant to this overall model. Due to the variables insignificance the process was repeated again, with the insignificant variables removed from further consideration.

The third version of this model contained fifteen variables. As expected the coefficient of determination and the adjusted coefficient again had lower values, but the model still provides a good prediction value for the injury accident rates as can be seen in Table 59.

Table 59: ANOVA Table for Variable Group Three Final Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	15	850.98591	56.73239	5.06	0.0050
Error	11	123.37472	11.21588		
Corrected Total	26	974.36063			
Root MSE	3.34901		R-Square	0.8734	
Dependent Mean	7.92370		Adj. R-Sq	0.7007	
Coeff Var	42.26574				

The model again passed the overall significance test, but in addition to that all the variables this time passed their individual significance tests. There was no indication of problems with the variables when looking at the partial coefficients of determination. There was a slight indication that one of the variables may not be vital to the model when looking at the 95 percent confidence levels. One variable, *ospole*, the number of overhead sign pole, had a confidence limit that included zero which implies that the variable might not be important to the overall model. As this was the only indication of such a problem, however, the variable was left in the model.

The graphical diagnostics did not indicate that there were any problems with model violations. The plot of the residuals versus the predicted values indicates a constant variance and a symmetric division about zero as seen in Figure 87.

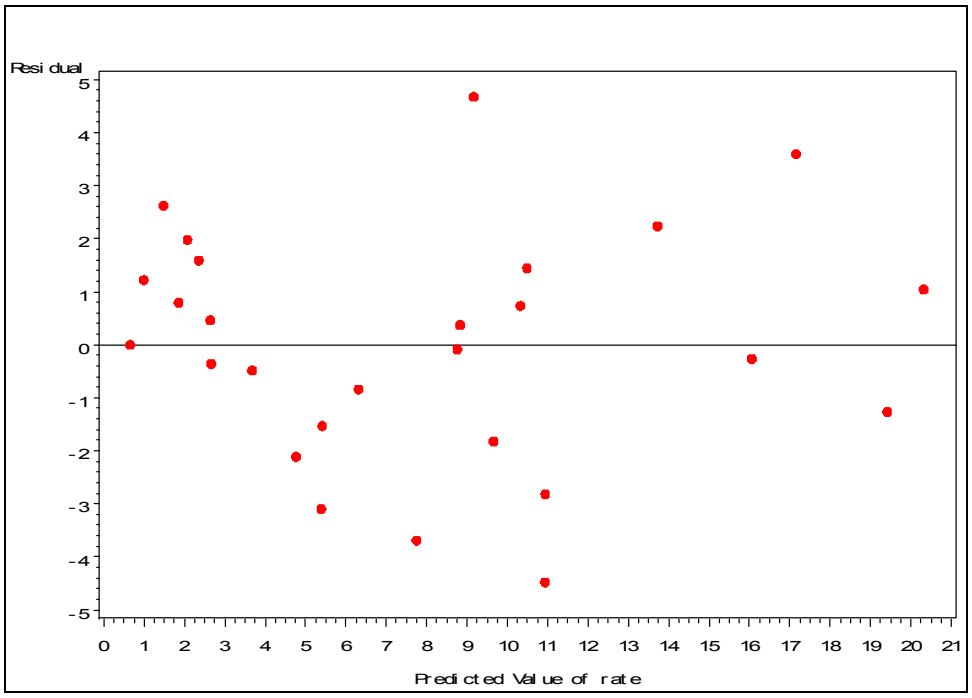


Figure 87: Residuals versus Fitted Values for Variable Group Three Final Model

The normal proability plot also shows that there are no problems with this model's residuals not following a normal distribution. There are only minor variations from the normal as can be seen in Figure 88 where the model's distribution is slightly left of normal and has a slightly lower peak value.

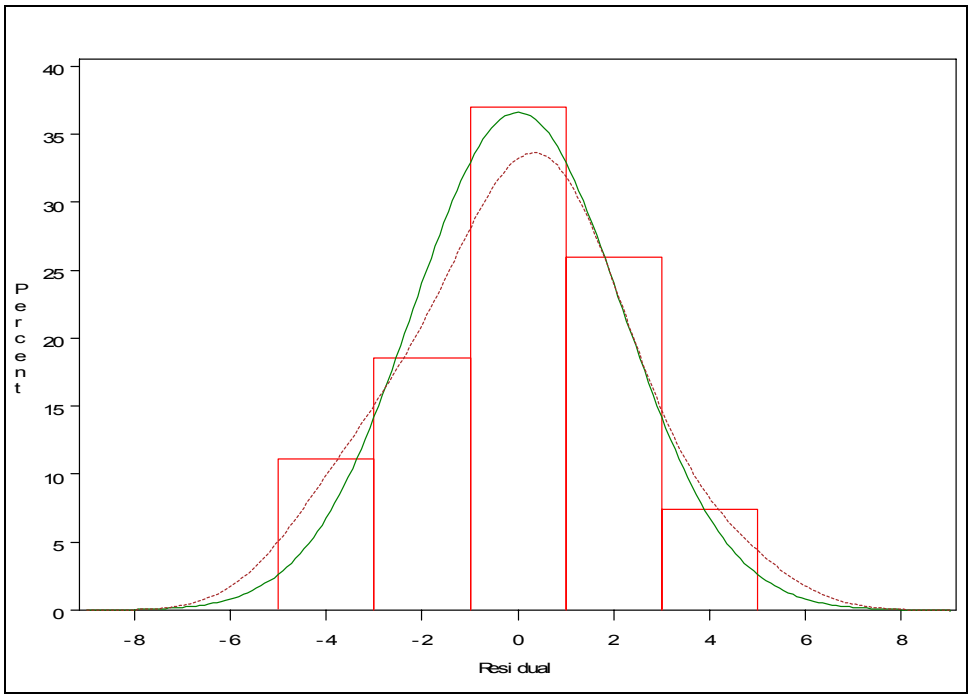


Figure 88: Normal Probability Plot for Variable Group Three Final Model

5.2.6.4 Injury Accident Model Summary

In the search for the best possible model to predict the injury accident rate, three viable contenders were developed. Variable group one and group three yielded models with fifteen variables while variable group two developed into a model with twenty-four variables. Each of these three models had coefficients of determination and adjusted coefficients that would allow them to be used as good models. These coefficients can be seen to compare in Table 60.

Table 60: Comparison of Final Injury Accident Rate Models

Variable Group	# of Variables	R^2	R_a^2
1	15	0.9319	0.839
2	24	0.9996	0.9953
3	15	0.8734	0.7007

Despite having higher coefficients the model developed from variable group two was not selected as the best model. This model appears to be over fit to the database used to develop it, which would make it less useful when applying the model to other data sets.

This model also has a large number of variables which makes is fairly cumbersome to work with. The remaining models from variable groups one and three both have the same number of variables, so that does not separate them. Model one des have both the higher coefficient of determination and adjusted coefficient of determination. Since both are possible models the significance of the models were also compared. Model one had a P-statistic of 0.0002 while Model 3 had a P-statistic of 0.005. The model with the larger significance also had the larger coefficient values and therefore was selected as the best model to predict injury accident rates.

6 Results

The results of this research are three different crash prediction models. One model predicts the total number of accidents on a road segment using an additive model while the second uses a multiplicative or log-linear model. The last model predicts the total number of injury accidents on each road segment. The models predict the total number of crashes meaning the ones that occur on the main (straight segment part) segment and at the major intersection of each segment, which is at the end of the segment with the largest street numbers. This is an important distinction to make since most prediction models are limited by either predicting crashes just at an intersection or just on the segment.

6.1 Final Linear Model

The best model developed fort predicting the total number of accidents on a segment with an additive model consists of six independent variables. The variables are the number of overhead sign poles, the number of parking lot entrances, the percentage of residential land use, an indication of whether or not horizontal curves are present, the percentage of the crest on the road, and the percent of parallel on-street parking allowed on the road segment. This model does a good job at explaining the variation in historical accident data on the segments with a coefficient of determination of 0.7591 and an adjusted coefficient of 0.6869. These coefficients are important in that a coefficient of determination of less than 0.7 is typically considered as the break even point with models with greater coefficients being acceptable for use and models with lower coefficients not being used. The model statistics can be seen in Table 61. The overall model exhibits full significance with an F-value of 10.51 leading to a P-statistics of less than 0.0001. This

indicates that there is only a very small chance that this overall model is not the correct one. The acceptable limit that was set as a model requirement was that this value must be significant to greater than or equal to 90 percent, which the model more than meets.

Table 61: ANOVA Table for the Total Accident Prediction Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	6	4514.59295	752.43216	10.51	<0.0001
Error	20	1432.42057	71.62103		
Corrected Total	26	5947.01352			
Root MSE	8.46292		R-Square	0.7591	
Dependent Mean	23.14741		Adj. R-Sq	0.6869	
Coeff Var	36.56099				

In addition to the overall model being significant, the individual parameters were examined for their significance and to determine what exactly the parameter estimates were saying in the model. The only parameters that did not pass their significance tests were that of the intercept and of the variable *curves* as shown in Table 62. The alpha level for significance was set at 0.10 and both parameter estimates just barely fail their significance tests. The intercept fails by just over one percent with a value 0.1106 and the variable *curves* fails by less than four percent with a value of 0.1359.

Table 62: Parameter Estimates for the Total Accident Prediction Model

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	10.29065	6.16320	1.67	0.1106
Ospole	1	4.92627	1.27494	3.86	0.0010
parkinglots	1	-1.65091	0.29932	-5.52	<0.0001
residential	1	-0.33020	0.05147	-6.42	<0.0001
Curves	1	6.86994	4.42122	1.55	0.1359
Crest	1	3.29592	1.15180	2.86	0.0096
Parking	1	0.12747	0.05631	2.26	0.0349

All the parameter estimates have some flexibility in that based on the standard error of the estimate there is at least one standard error amount of space before there is a question of the parameter estimate becoming zero. The coefficients of partial determination also indicate that all the variables should remain in the model, so that there is some debate that could occur on whether or not curves should be removed. There were two major criteria for allowing a variable to remain in the model, that of the variable's individual significance and that of the coefficient of partial determination. The coefficients of partial determination can be seen in Table 63. Type I SS indicates that that is the value of the coefficient of partial determination if all the previous variables are in the model. The value for residential is 1791.21721 which is the value gained by adding the variable *residential* to a model that already contains the variables of *ospole* and *parkinglots*. Type II SS is the coefficient of partial determination if the variable in question is added to a model already containing the other variables. For instance, the value for *crest* is 586.46737 which is the value gained by adding the variable *crest* to a model that also contains the variables *ospole*, *parkinglots*, *residential*, *curves* and *parking*. The remainder of the table lists the limits within which with a 95 percent confidence it can be stated that the parameter estimate should be located.

Table 63: Parameter Estimate Statistics for the Total Accident Prediction Model

Variable	DF	Type I SS	Type II SS	95% Confidence Limits	
Intercept	1	14467	199.67010	-2.56557	23.14686
Ospole	1	798.66967	1069.28885	2.26679	7.58576
parkinglots	1	565.28903	2178.83710	-2.27527	-1.02655
residential	1	1791.21721	2947.73249	-0.43757	-0.22284
Curves	1	478.60265	172.92665	-2.35257	16.09245
Crest	1	513.75622	586.46737	0.89332	5.69853
Parking	1	367.05818	367.05818	0.01002	0.24492

Since the presence of horizontal curves historically plays a large role in identifying potential accident locations it would be informative if it were left in as a variable in the model. In looking at the 95 percent confidence levels for the parameter estimates, again, the only questionable estimate where the value could be zero is for the one variable that does not reach the full significance that was indicated. The easiest way to notice a problem is when one side of the confidence limit has a negative value and the other side a positive one which happens only with the intercept and the variable *curves*.

Looking closer at the parameter estimates shows that for the most part the signs of the coefficients are as expected or can be explained. The intercept has a positive coefficient, which means that there is a base accident rate for urban arterials. If the coefficient were negative, this would be impossible in reality as there can only be positive accident rates. The coefficient for the variable *residential* also makes sense in much the same way. It is intuitive that residential locations would have lower accident rates than busy commercial areas. The type of traffic in residential areas is mostly restricted to only the people who live or are visiting in the area with the majority of traffic occurring when people are traveling to and from work; otherwise people do not traverse these areas. Commercial areas, on the other hand, have people who can be unfamiliar with the area and large amounts of traffic at most times of the day, leading to a higher possibility for accidents. The negative coefficient for the residential variable demonstrates that where the land use is residential, there is a lowering of the accident rate.

The other parameter estimate that has a negative sign with it is that of the variables *parkinglots*. This states that the more entrances to parking lots the lower the

expected crash rate should be. At first glance this could seem contradictory. Why, with more places for turning vehicles, would the number of parking lots decrease the number of accidents? This can be explained in that the parking lot variable does not really represent just parking lots, but helps to represent the land use and the traffic patterns on the segment. Besides creating places where turning conflicts can occur, parking lots have the affect of removing parked vehicles from the sides of the road and of concentrating pedestrians away from the roadway. Parking lots put many vehicles together in one area and possibly remove some of those vehicles from the street. Parking on the street can cause sight distance problems and create hazards by placing more objects around that can be struck, but also by people entering and exiting their vehicles and entering and exiting their parking spaces. If a driver is not paying attention, a person entering or leaving a parked vehicle can cause a problem with the driver side door opening in the traffic path. The same way a vehicle in the process of parallel parking can potentially cause problems with other inattentive drivers. These problems are removed by having locating the parking vehicles in lots where speed is slower and drivers are aware of the constant parking maneuvers.

In the same way, that parking lots can remove vehicles from the side of the road, the percentage of on-street parallel parking can add to crash rates. The coefficient of the variable *parking*, which represents the percentage of on-street parallel parking that is allow on a road segment, was found to be positive in this model indicating that the more on-street parking is available the higher the crash rates should be expected to be. For similar reasons why the variable *parkinglots* lowered the crash rates, the percent of parking increases them. The presence of vehicles doing parking maneuvers and

pedestrians going to and from their vehicles and nearby buildings can cause situations that a driver is not expecting. While on an arterial, a driver typically expects to be able to continuously move except when at a traffic light. When more pedestrians and parking maneuvers occur on a segment they can startle a driver who is not expecting many of these motions to occur.

The signs of the remaining parameter estimates are what would be intuitively expected. The number of overhead sign poles has a positive coefficient, indicating that the more sign poles the more crashes will occur. This can be for several reasons including the fact that there are more hazards that can be struck by passing vehicles. Overhead signs typically indicate that the entrance to a major arterial is nearby which causes the need for turning movements onto the arterial and also sudden movements of drivers who may have found themselves in the wrong lane to get onto the arterial. Both of these actions can lead to the occurrence of crashes, which implies the positive sign of the parameter estimate.

The variables *crest* and *curves* also have positive values for their parameter estimates. Historically the presence of curves has been an indication of a location where accidents can occur. This has been observed in many studies that have occurred on rural and urban roads and much attention has been given to the proper design of horizontal curvature, so it comes as no surprise that the presence of one or more horizontal curves in this study indicates an increase of accident rates. If drivers are not expecting a change in horizontal alignment or are traveling at speeds that are unsafe for the particular design crashes are more likely to occur. This variable also has the parameter with the largest

value of a coefficient either positive or negative, which implies that the presence of horizontal curvature has a large impact on crashes.

Similarly, the variable crest has a positive coefficient signifying that segments with larger crests will have larger accident rates. This is more likely an indication of the road surface and condition rather than a reflection on the actual crest value because the allowable limits for crests on new roads are rather limited. In New England where problems such as frost heave and freeze-thaw problems are very important, the crest of the road can increase with these problems or with the actual structure of the pavement failing and causing part of the road way to sink. Another environmental problem that develops with large crests, includes that of rain. During heavy rains water can build up in the edge of the crest and cause vehicles to hydroplane and have problems. Variables such as the quality of the pavement and the pavement markings were not found to be significant in this model, but the crest could be representing some of these variables qualities. This is a little difficult to state exactly, due to the small nature of the data set from which this model was built. These parameter estimates all lead to the following model:

$$Rate = 10.29 + 4.92 ospole - 1.65 parkinglots - 0.33 residential + 6.87 curves + 3.30 crest + 0.13 parking$$

Every model needs to ensure that it is not violating any of the model assumptions. Reviewing the graphical analysis of the model mostly covers the model assumptions. The boxplot in Figure 89 shows that the residuals are centered on zero as is expected based on the form of the model. The boxplot also shows where the quarter points of the locations of the residuals fall, this is ideally a symmetric distribution. This plot suggests that this model has a larger variation when it predicts lower than expected rates.

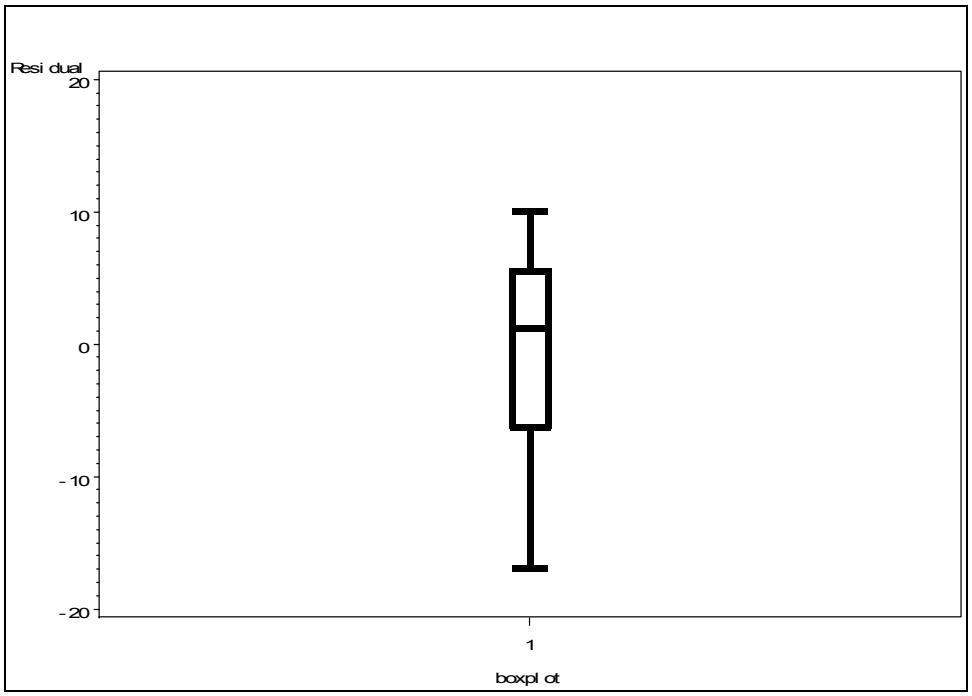


Figure 89: Boxplot of the Total Accident Prediction Model

The graphical diagnostics show that there is no problem perceived with this model violating the linear model assumptions. The plot of the residuals versus the fitted values shows a very constant error variance and an even distribution between positive and negative residuals (See Figure 90).

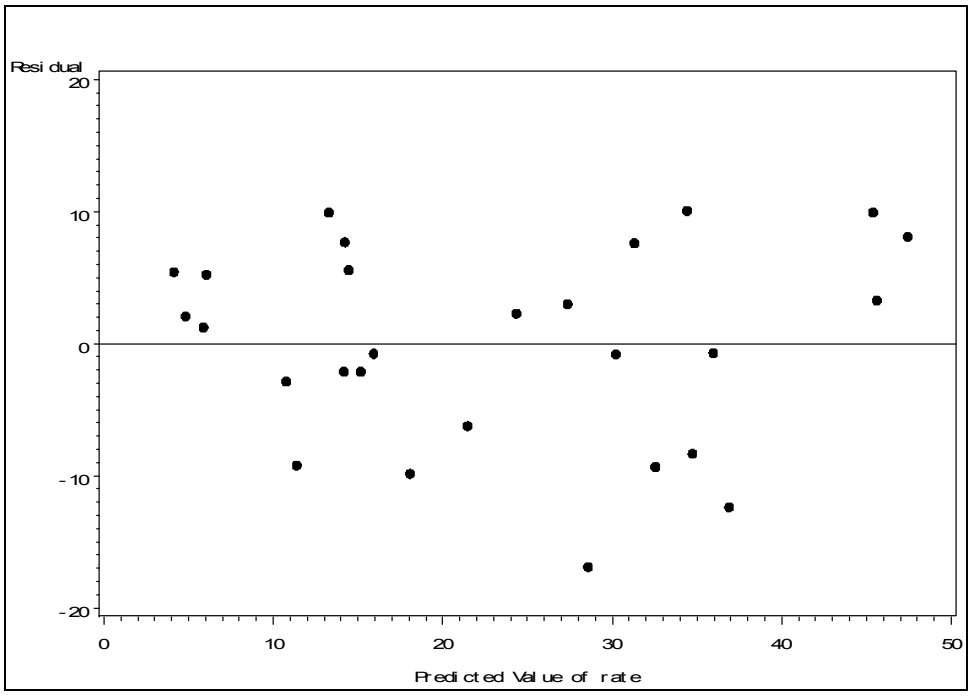


Figure 90: Residuals versus Predicted Values for the Total Accident Prediction Model

There are no points that can be perceived as outliers either. This can more clearly be seen in the studentized residuals versus the predicted values plot in Figure 91. The heuristic for knowing whether to qualify a point as an outlier is if the studentized residual is greater than four. For this model there is not even a point that deserves consideration as an outlier, as the largest studentized residual value that occurred was -2.387.

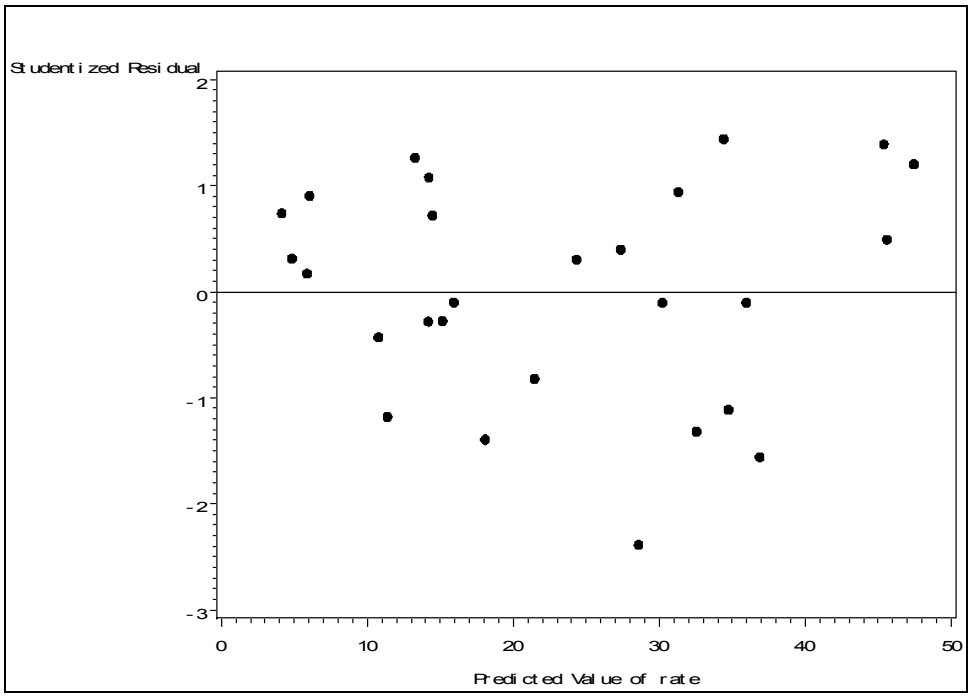


Figure 91: Studentized Residuals versus Predicted Values for the Total Accident Prediction Model

The normal quantile plot in Figure 92 indicates that there is a strong inclination towards normality as the majority of the points closely follow the line that indicates a linear relationship. There are few points that deviate from following the line and are mostly clustered around it. This is an indication that the model assumptions are not violated.

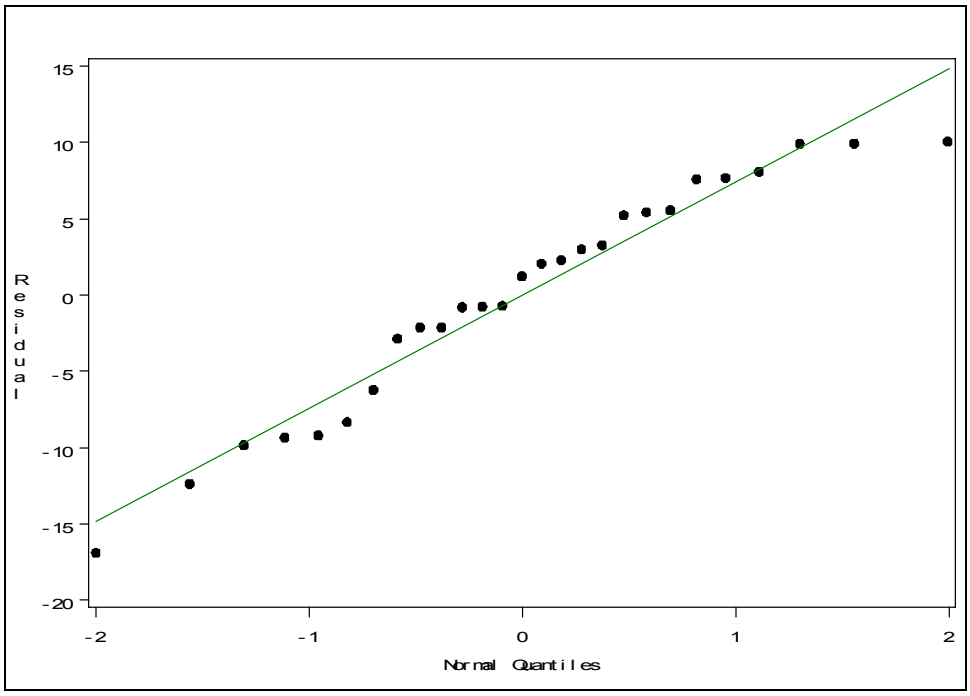


Figure 92: Normal Quantile Plot Values for the Total Accident Prediction Model

Again, only slight departures from normality can be observed in Figure 93 of the normal probability plot. The solid line represents a normal distribution, while the dashed line represents the distribution of the residuals from this model. The distribution for the model has a slightly lower maximum value, and deviates slightly from normal with a small skewness toward the right, but other wise is very similar to the normal distribution.

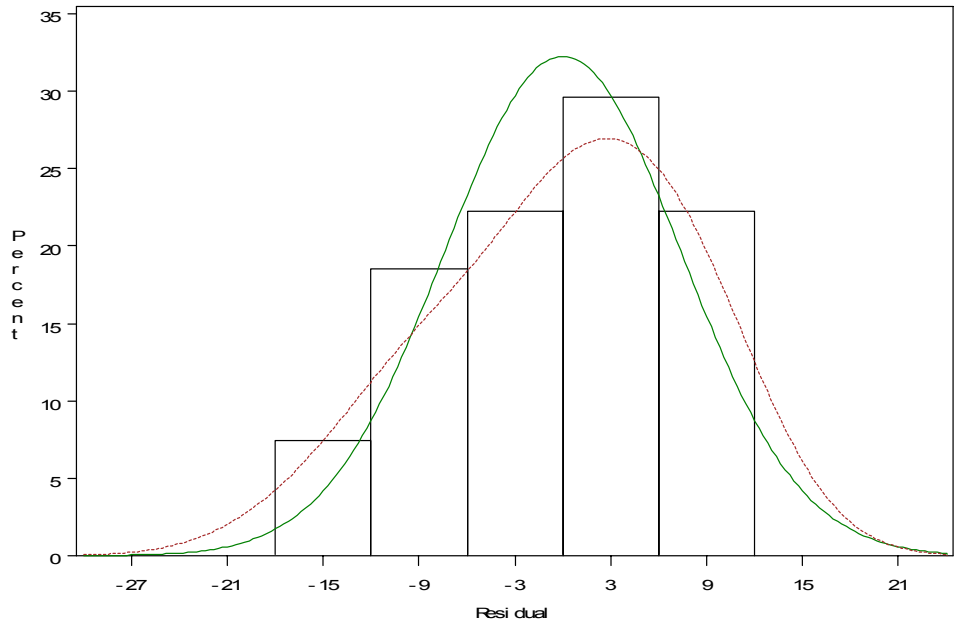


Figure 93: Normal Probability Plot for the Total Accident Prediction Model

This model predicts the rate for the total number of accidents that occur on an arterial road segment. Overall it appears to be a good model to use to predict these crashes and it takes an additive form. The additive form indicates that the variables in question tend to act individually upon the roadway in terms of causing crashes to happen. They do not act together to change crash rates, which will allow each item to be reviewed separately if the segment is about to be repaired or redesigned. This allows each variable to be independently adjusted by traffic engineers and a visible effect to be noticed.

6.2 Final Multiplicative Model

A model that predicted the total number of accidents, but in a multiplicative form, was also developed alongside the previous model. The final model chosen as the best model that can predict the total number of accidents included only three variables: *length*, *lighting* and *pole*. For this the coefficient of determination was 0.6672 and the adjusted

coefficient of determination was 0.6238. These values, while not low, are in a range that is generally not acceptable for an accurate model. The coefficient of determination is lower than that of the linear model at 0.7591. This makes the linear model appear to be the better of the two for predicting the total number of accidents. The coefficients and other statistics can be seen in Table 64 below. Having a lower coefficient of determination and adjusted coefficient does not stop this multiplicative model from passing the overall significance test with a value of less than 0.0001 when anything less than 0.10 would be acceptable.

Table 64: ANOVA Table for Multiplicative Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	3	10.05655	3.35218	15.37	<0.0001
Error	23	5.01511	0.21805		
Corrected Total	26	15.07167			
Root MSE	0.46696		R-Square	0.6672	
Dependent Mean	2.90341		Adj. R-Sq	0.6238	
Coeff Var	16.08302				

The significance for the individual coefficients including the intercept, were found to be significant to greater than 0.10. With pole having the lowest passing statistic at 0.0745. The standard errors for all of the coefficients are also acceptable in that one deviation can be taken for all of the variables and in most cases two standard deviations, eliminating the majority of the concern that the parameter estimates could possibly become zero. This can be seen in Table 65.

Table 65: Parameter Estimates for Multiplicative Model

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	-28.29984	10.22157	-2.77	0.0109
Llength	1	-0.95851	0.25768	-3.72	0.0011
Llighting	1	7.80913	2.17984	3.58	0.006
lpole	1	0.56736	0.30369	1.87	0.0745

The model created has the form of $Rate = e^{-28.3} length^{-0.959} lighting^{7.809} pole^{0.567}$.

The parameter estimate for the variable *length* has a negative coefficient approaching negative one. This is suggestive of a rate. A transformation had been attempted where all the variables were rates or densities in order to try and determine if length should be removed from the model. Despite the estimate near negative one the variable *length* was never shown to be insignificant even when all the other variables were densities or percentages. This implies that length in this model format remains an important factor towards predicting the crash rates for the total number of accidents. The parameters for the other two variables are not suggestive of a rate and so no transformations were tried on them.

The final diagnostics to check, since the model and all variables are significant, are the graphs to check model assumptions. The residuals versus the fitted values show that there is a constant variance (See Figure 94). This can be a little difficult to see due to the way that the majority of the points are all clustered together towards the right hand side of the plot, but the cluster does not show any signs of a systematic departure from normality.

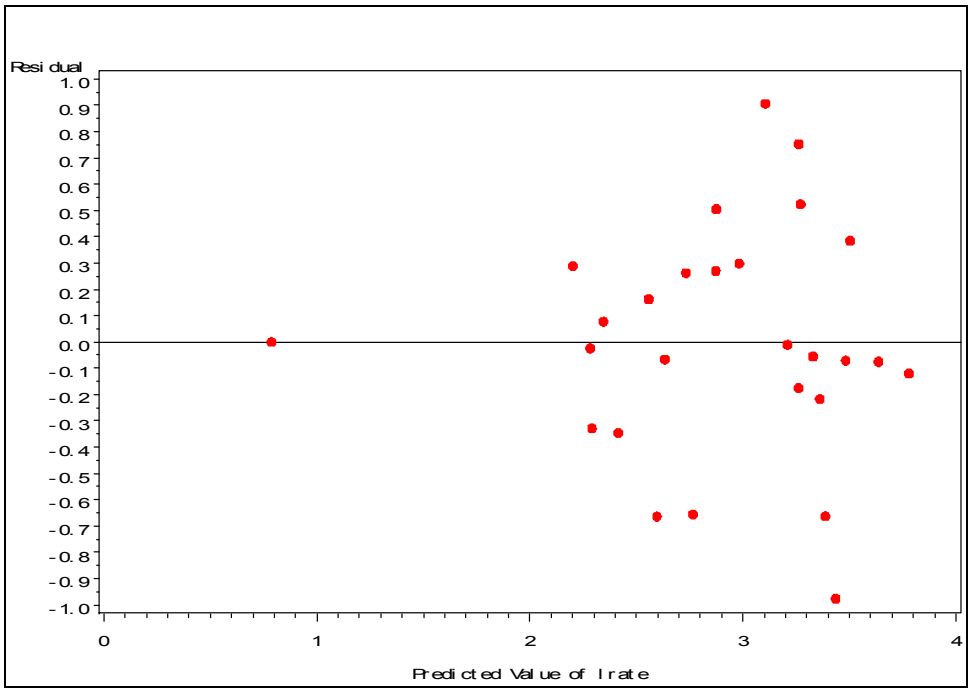


Figure 94: Residuals versus the Predicted Values for the Multiplicative Model

The studentized residuals versus the fitted values show a similar view as the residuals versus the predicted values with the addition of being able to identify outliers. Based on the heuristic of needing to be greater than four before being considered an outlier, none of the points quality or cause concern in this model. Figure 95 shows the studentized residual plot.

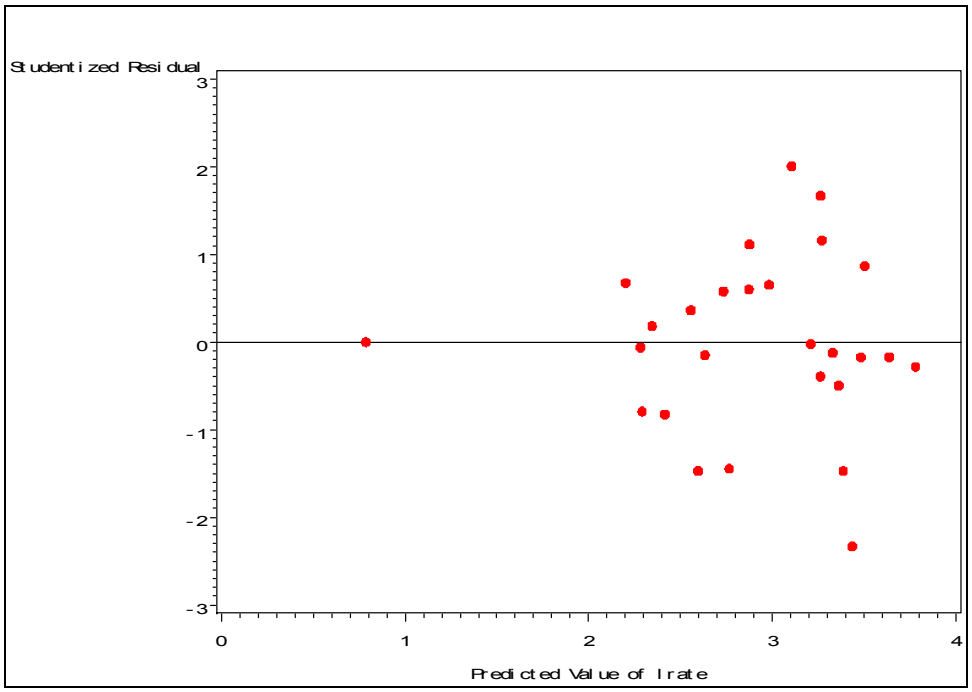


Figure 95: Studentized Residuals versus the Predicted Values for the Multiplicative Model

The box plot of the residuals in Figure 96 shows that the residuals are highly symmetric with a slight skewness on the positive side, which implies that the model will have a tendency to predict with a higher variance when overestimating the accident rate. This is, however, a very minor tendency and not a significant reason to regard this model as suspect.

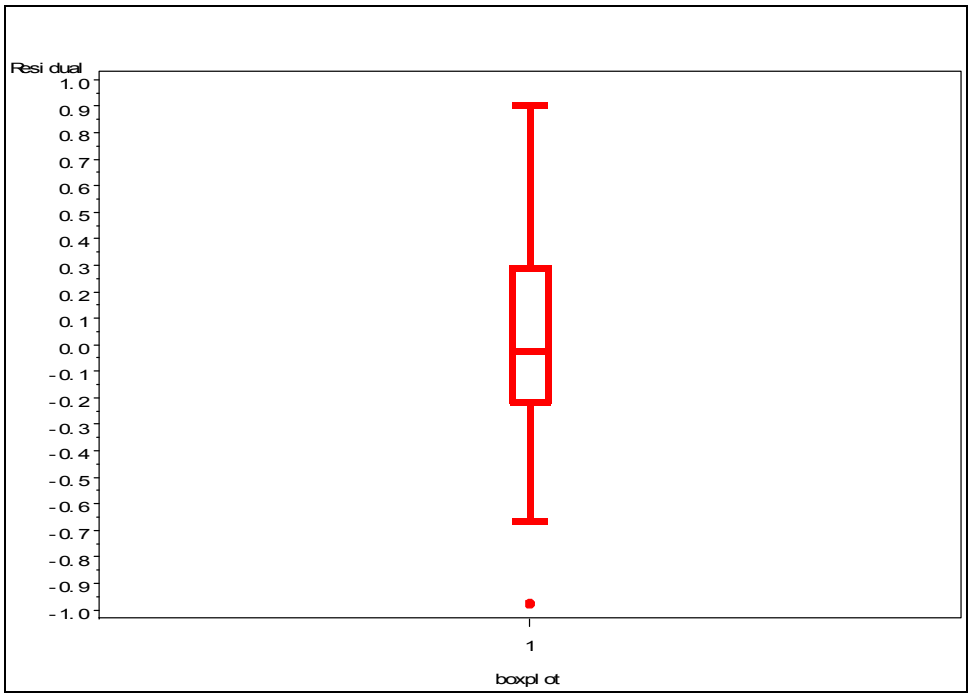


Figure 96: Boxplot for the Multiplicative Model

The normal quantile plot, seen in Figure 97, also shows that the model follows the assumptions for a normal distribution with the residuals falling along the line. There is no obvious departure from the normal line in a recognizable pattern that could indicate a model violation. This is a good indication that the model assumptions are being met.

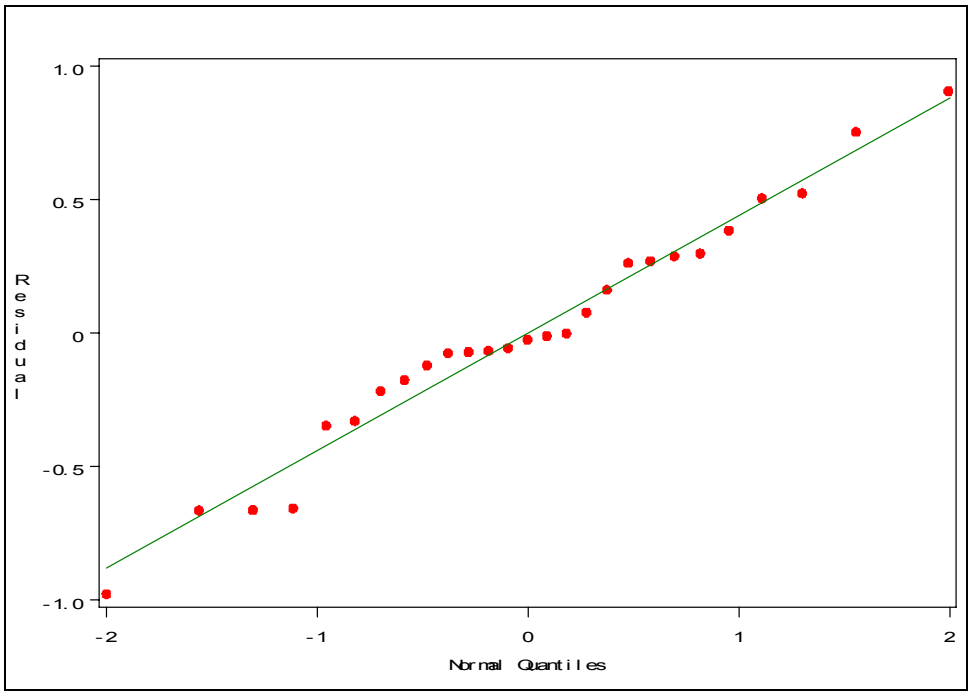


Figure 97: Normal Quantile Plot for the Multiplicative Model

There are only very minor deviations from normality that can be seen in the normal probability plot in Figure 98. The dashed line, which represents the model's distribution, almost exactly follows the solid line, which is a normal distribution. The distribution from the model has a slightly higher peak than does the normal distribution and a small jag on the left side of the distribution. The jag is not duplicated on the right side of the plot where the model's distribution mimics the normal distribution. The graphical diagnostics all indicate that the model does not violate any of the model assumptions and the model form is appropriate for the given dataset.

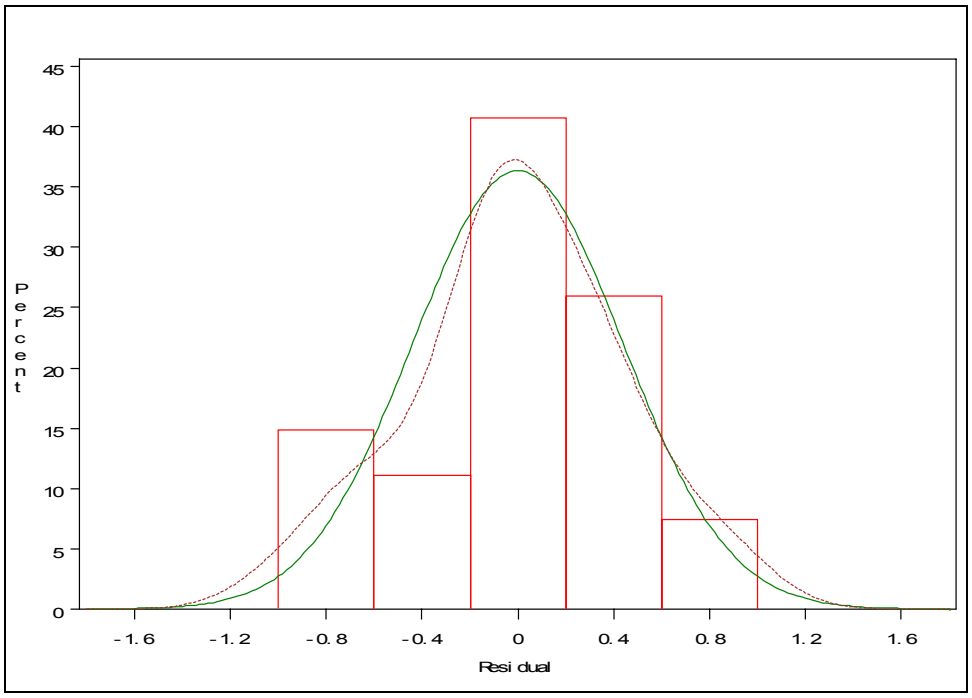


Figure 98: Normal Probability Plot of Multiplicative Model

This model predicts the total number of accidents that occur on an urban roadway segment. It is a fairly good model, but not quite as good as the linear model that predicts the rate of the total number of accidents based on the coefficient of determination. The model form, that of a multiplicative or log-linear model, appears to not be the best choice of functional form for a model in an urban area. This form has been used, but most often in rural areas, where geometric and traffic characteristics greatly effect one another and there combined effects cause the crashes. It appears that a linear, or additive model, is more appropriate in an urban setting where most geometric and traffic characteristics appear to work independently of each other.

6.3 Final Injury Accident Model

The best model developed that predicts the total number of injury accidents only on a segment with an additive model consists of a fifteen independent variable model. The variables include *fence*, *ospole*, *hazards*, *parkinglot*, *vol*, *residential*, *length*, *grade*,

curves, *crest*, *widtha*, *widthsida*, *pavement*, *markings*, and *lighting*. This model does a good job at explaining the variation in historical injury accident data that exists on the segments with a coefficient of determination of 0.9319 and an adjusted coefficient of 0.8390. These coefficients are important in that a coefficient of determination of less than 0.7 is typically considered as the break even point with models with greater coefficients being acceptable for use and models with lower coefficients not being used. These statistics can be seen in Table 66. The overall model exhibits full significance with an F-value of 10.03 leading to a P-statistics of 0.0002. This indicates that there is only a very small chance that this overall model is not the correct one. The acceptable limit that was set as a model requirement was that this value must be significant to greater than or equal to 90 percent, which the model more than meets. In this model the dependent variable is the injury accident rate. The number of injury accidents consists of all types of accidents, including fatalities, because in this data set fatalities were very rare, so they were treated as if they were a very bad injury.

Table 66: ANOVA Table for the Injury Accident Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	15	907.99025	60.53268	10.03	0.0002
Error	11	66.37038	6.03367		
Corrected Total	26	974.36063			
Root MSE	2.45635		R-Square	0.9319	
Dependent Mean	7.92370		Adj. R-Sq	0.8390	
Coeff Var	31.00006				

In addition to the overall model being significant, the individual parameters were examined for their significance and to determine what exactly the parameter estimates were saying in the model. All of the parameters passed their significance tests with an

alpha value of 0.1. Only one variable had a significance value that was greater than even 0.05. This can be seen in Table 67. All the parameter estimates have some flexibility in that based on the standard error of the estimate there is at least two standard error deviations of space before there is a question of any of the parameter estimates becoming zero.

Table 67: Parameter Estimates for the Injury Accident Model

Variable	DF	Parameter Estimate	Standard Error	F Value	Pr>|t|
Intercept	1	-127.70032	26.22728	-4.87	0.0005
Fence	1	2.02433	0.37825	5.35	0.0002
Ospole	1	1.81684	0.44240	4.11	0.0017
hazards	1	0.23576	0.05119	4.61	0.0008
parkinglots	1	-1.58959	0.19219	-8.27	<0.0001
Vol	1	0.00044898	0.00015171	2.96	0.0130
residential	1	-0.13374	0.03427	-3.90	0.0025
Length	1	-0.00999	0.00292	-3.42	0.0058
grades	1	-0.78490	0.26125	-3.00	0.0120
Curves	1	2.90953	1.37192	2.12	0.0575
Crest	1	3.45836	0.55909	6.19	<0.0001
Widtha	1	-1.00710	0.38811	-2.59	0.0249
widthsida	1	2.09122	0.40398	5.18	0.0003
Pavement	1	-16.96193	3.24304	-5.23	0.0003
Markings	1	4.10381	1.08137	3.80	0.0030
lighting	1	1.03749	0.23049	4.50	0.0009

There were two main criteria for allowing a variable to remain in the model, the primary being the variable's individual significance. The coefficients of partial determination can be seen in Table 68 and were also reviewed to see if they indicated that a variable should be removed from the model. There was no specific limit set for the coefficient of partial determination, but if they appeared low then special care was taken in regard to those variables. The table lists the limits within which with a 95 percent confidence it can be stated that the parameter estimate should be located. By reviewing the 95 percent confidence limits, it can be seen whether or not there is a possibility for the

parameter estimate to be zero. As long as the confidence limits have the same sign for the upper and lower limits, there is no concern. One variable only had confidence limits that encompassed both a positive and a negative sign. That variable was *curves*, representing an indication of whether a segment had one or more horizontal curves on it. The confidence interval of -0.11005 to 5.92911 is strongly positive, but there is a small negative range displaying the possibility of the parameter estimate actually being zero and consequently not part of the model. Despite this possibility of the parameter estimate becoming zero, the variable was left in the model for several reasons. There does not appear to be a strong possibility of the estimate becoming zero and also the variable was in the linear model predicting the total number of crashes on a segment. The variable has also played a large role in prediction models for crashes that occur in rural areas, so it was decided to leave it in the model.

Table 68: Parameter Estimate Statistics for the Injury Accident Model

Variable	DF	Type I SS	Type II SS	95% Confidence Limits	
Intercept	1	1695.19717	143.04048	-184.42618	-69.97445
Fence	1	116.16753	172.81621	1.1911	2.85685
Ospole	1	57.39137	101.76015	0.84311	2.79056
hazards	1	114.78938	127.98072	0.12309	0.34843
parkinglots	1	7.18908	412.76577	-2.01259	-1.16659
Vol	1	22.65698	52.84756	0.00011508	0.00078288
residential	1	241.80529	91.88955	-0.20917	-0.05834
Length	1	12.77385	70.37872	-0.01643	-0.00355
grades	1	0.32334	54.46064	-1.35992	-0.20988
Curves	1	24.60478	27.13734	-0.11005	5.92911
Crest	1	24.75725	230.86339	2.22781	4.68892
Widtha	1	6.42181	40.62676	-1.86133	-0.15287
widthsida	1	74.50164	161.68351	1.20207	2.98037
Pavement	1	55.12456	165.05443	-24.09982	-9.82405
Markings	1	27.23838	86.89846	1.72375	6.48388
lighting	1	122.24501	122.24501	0.53017	1.54480

Looking a little closer at the parameter estimates shows that for the most part the signs of the coefficients are as expected or can be explained. The intercept has a negative coefficient, which is not the best possible one. It would be more appropriate if it were positive because there cannot be a negative accident rate in nature. This base rate for injury accidents is negative however due to the fact that the variable *vol*, representing the average daily traffic on each segment, was included in the model. Due to the large volume of the traffic this is somewhat counteracted. The other variable that was included in the model that helps to counteract this large, negative coefficient is that of lighting. The majority of urban streets are fully lit and as lighting has a positive coefficient, it is instrumental in countering the majority of this coefficient.

The coefficients for the variables *fence*, *ospole* and *hazards* are what they would be based on intuition. All three variables represent either a specific roadside hazard or the total number of roadside hazards observed on the segment, with *fence* representing the number of fences or retaining walls observed on the segment, *ospole* representing the number of overhead sign posts and *hazards* representing the total number of roadside hazards observed. These indicated that the more hazards there are on the segment the higher the crash rate is going to be which makes intuitive sense. The more places a driver can run into things, the more likely that will happen.

For the same reasons as were stated in the section on the total number of accidents model above, the parameter estimate of the variable *parkinglots* was negative. The more parking lots on a segment the lower the crash rate becomes. This is mainly due to the fact that the variable is representative of how the traffic is behaving. Removing the slow

traffic and parking maneuvers and confining them to a parking lot, instead of the street, can avoid conflicts.

The variable *vol*, representing the volume or ADT on the segment has a positive coefficient as was expected. The main school of thought behind that is the more traffic on the roadway the more expected accidents. While some researches find that this is not a linear increase but an exponential increase, there is still an upward trend. The parameter estimate is one of the smallest numerically because it is multiplied by the ADT, which is in the tens of thousands for the arterials in the database.

Residential also has the expected coefficient sign of a negative value. This shows that the more residential an area is the less crashes occur because of the differences in mindsets of the drivers. When a driver is in a residential area, he knows that there will be slower traffic more turning vehicles and pedestrians and adjusts his behavior accordingly. There is also a more regular pattern to the traffic, in that the majority of it happens at the beginning and the end of the workday with only scattered times between then. Despite these residential areas occurring on arterials as opposed to residential neighborhoods, there are fewer people who need to access the adjoining land during the day. Commercial areas tend to attract large volumes of traffic throughout the day and do not have a time when people are not going there.

Length is one of the variables where the sign of the parameter estimate at first glance seems contradictory. Intuitively the longer a road segment is the more accidents there should be, but the negative sign implies that the longer the road segment the fewer crashes happen. This is not as counterintuitive as it first seems due to the way that crashes were assigned to segments. The crashes were assigned to a segment by the

location of the incident with crashes occurring on the long stretch of the segment clearly going to that segment, but this model predicts the total number of injury accidents which includes accidents at the major intersection of each segment. Most models focus on either segment or intersection crashes and they are rarely combined in one model, but when traffic engineers are looking at problem locations, they can often include both segment and intersections at the same location when major reconstruction is planned. Due to this inclusion of what would normally be considered intersection accidents, the parameter estimate for the segment length was negative. This means that the longer the segment is the fewer accidents. This is because the short segments have only a small distance before the intersection accidents start taking effect. The longer segments have more space where the intersection does not influence the accidents and intersections have long been agreed to be a location where many crashes happen.

The variables *crest* and *curves* have positive values for their parameter estimates. Historically the presence of curves has been an indication of a location where accidents happen. This has been confirmed by many studies that have looked at rural and urban roads and much attention has been given to the proper design of horizontal curvature, so it comes as no surprise that the presence of one or more horizontal curves in this study indicates an increase of accident rates. If drivers are not expecting a change in horizontal alignment or are traveling at speeds that are unsafe for the particular design, crashes are more likely to occur. Similarly, the variable crest has a positive coefficient signifying that segments with larger crests will have larger accident rates. This is more likely an indication of the road surface and condition rather than a reflection on the actual crest value because the allowable limits for crests on new roads are rather limited. In New

England where problems such as frost heave and freeze-thaw problems are very important, the crest of the road can increase with these problems or with the actual structure of the pavement failing and causing part of the road way to sink. Another environmental problem that can occur is the build up of rain water on the edge of the road when the crest is too large, this can cause vehicles to hydroplane and get into problems.

Continuing to look at the variables that relate to geometric alignment, the variable *grade* has a negative value for its parameter estimate. This appears to mean that the larger the grade becomes the lower the accident rate becomes. This goes against intuitive thought, because it seems that the larger the grade becomes the more crashes should occur. In an urban area, however, there is so much happening that the geometric alignment of the road does not play as important a role as it does on rural arterial roads. There is much more traffic and commotion, that in an urban setting, the steeper the grade becomes on the road, the fewer accidents occur because drivers slow down, so that pedestrians and traffic becomes easier to see and easier to determine the relative distances from these objects.

The variables *widtha* and *widthsida* both relate to the geometric design of the road. *Widthsida* is the average width of the sidewalks on the segment, which is an average of the two sides of the road. This has a positive parameter estimate, which makes intuitive sense. The wider the sidewalk is, the more accidents occur. This is due to similar reasons as that of why the coefficient for the residential parameter is negative. The sidewalks become wider as they are used more and they get used more in areas where there are the most attractions such as shops and parks. It is in these locations where pedestrians can be found in large numbers. The more pedestrians that are around

the more possibilities there are for accidents to occur. This is due to the fact that pedestrian accidents can occur, but by watching to ensure that the pedestrians are safe, drivers may loose sight of the other nearby vehicles or be forced to take actions to protect the pedestrians, such as stopping quickly, that they wouldn't have ordinarily taken. Where the sidewalks are narrower, there are fewer pedestrians and problems are less likely to happen.

On the other side, the coefficient for *widtha* is negative meaning that the wider the traffic lanes are the fewer crashes occur. This is the expected value of the coefficient due to the wider lanes making drivers feel more comfortable with oncoming traffic and putting more distance between the passing vehicles.

The variable *pavement* has a positive value for the parameter estimate. Pavement has two possible values that of zero meaning the pavement is of fair or bad quality and that of one meaning the pavement is of good quality. The sign of the parameter reflects this. The better the pavement is, so if the pavement qualifies as having a good condition, the less crashes occur. This would be the expected condition because when the pavement is in bad shape whether due to patching and cracking, or rutting on the road, there are more problems that could occur. If the cracks are severe or if potholes develop, there is no problem in seeing how crashes can happen. Even if the problems are not so severe, they cause the driver to need to devote more attention to the road surface and remove the driver's attention from the other events that are occurring on the road at the same time, including other drivers.

The parameter estimate for the variable that represents the quality of the pavement markings is positive. At first glance this means that the better the pavement markings are

the more crashes are going to occur. This statement however is not as contradictory as it first may seem. When roads are well marked, drivers are more comfortable with their surroundings and more likely to pay less attention to the task of driving. This parameter does not represent itself as much as it represents more the driver's attitude. If they can clearly see the road and the lane markings and where they should be located, then their attention can wander. If the markings are harder to see, then the drivers pay closer attention in order to determine where they and their vehicle should be located.

The variable *lighting* indicates the percentage of each segment that is lit. The parameter estimate is positive which at first review seem to mean that the more lighting the more accidents occur and conversely the less lighting available the fewer accidents occur. This however is not truly the situation. This variable helps to counteract the majority of the intercept value. Since most urban minor arterials have full lighting, this brings the intercept coefficient closer to zero. So while playing an important role in the model, the value of the coefficient cannot be interpreted in the conventional way.

These parameter estimates all lead to the following model:

$$Rate = -127.7 + 2.02\,fence + 1.82\,ospole + 0.24\,hazards - 1.59\,parkinglots + 0.00045\,Vol - 0.13\,residential - 0.001\,length - 0.78\,grade + 2.91\,curves + 3.46\,crest - 1.01\,widtha + 2.09\,widthsida - 16.96\,pavment + 4.10\,markings + 1.04\,lighting$$

Every model needs to ensure that it is not violating any of the model assumptions. This is mostly done by reviewing the graphical analysis of the model. The boxplot in Figure 99 shows that the residuals are centered on zero as is expected based on the form of the model. The boxplot also shows where the quarter points of the locations of the residuals fall, this is ideally a symmetric distribution. This plot suggests that this model has a larger variation when it predicts lower than the expected rates.

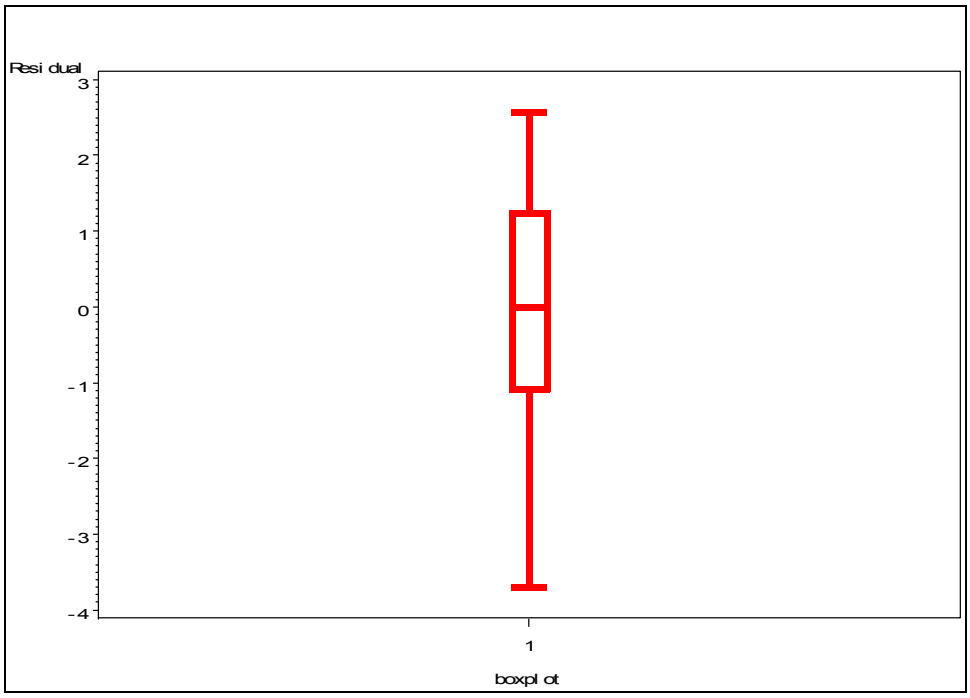


Figure 99: Boxplot of the Injury Accident Model

The graphical diagnostics do not indicate that this model violates any of the model assumptions. The residuals versus the predicted values plot indicates that the residuals have a constant variance and are basically symmetric about zero, as can be seen in Figure 100. The residuals on the positive side can be easily seen to fall under a constant line at approximately 2.75. On the negative side there is one point that falls outside of this range by a small amount with a value of approximately -3.75 but all the other points fall under the -2.75 constant line.

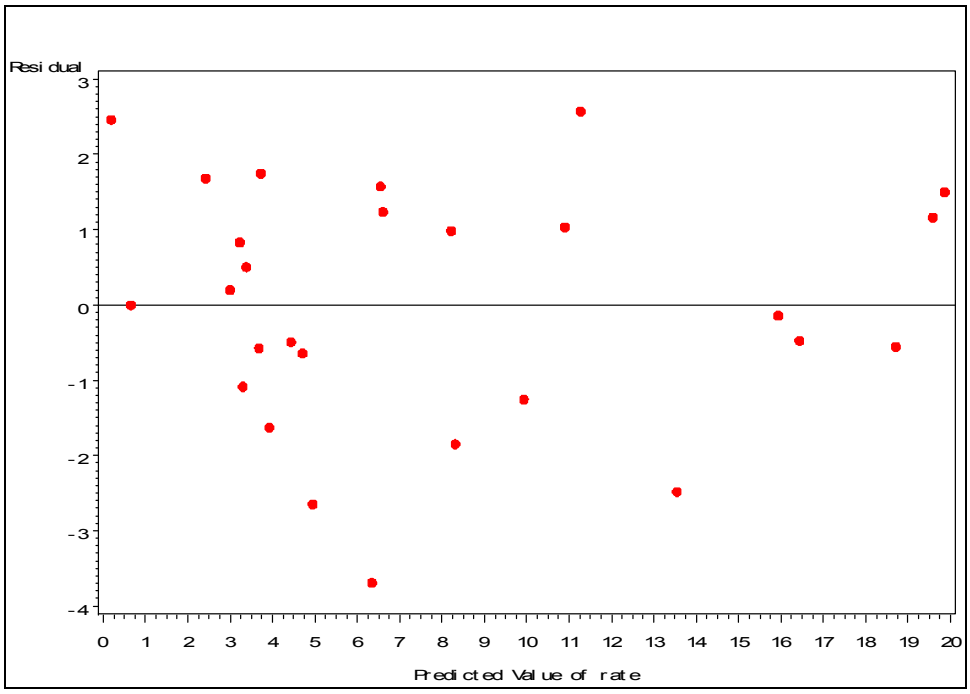


Figure 100: Residuals versus Predicted Values for the Injury Accident Model

There are no points that can be perceived as true outliers despite the one point not exactly behaving in the residual versus predicted values plot. This can more clearly be seen in the studentized residuals versus the predicted values plot in Figure 101. The heuristic for knowing whether to qualify a point as an outlier is if the studentized residual is greater than four. For this model there is not any points that deserve consideration as an outlier as none of the studentized residual values are larger than 2.0. So despite one point not being ideal, there are not any outlying points.

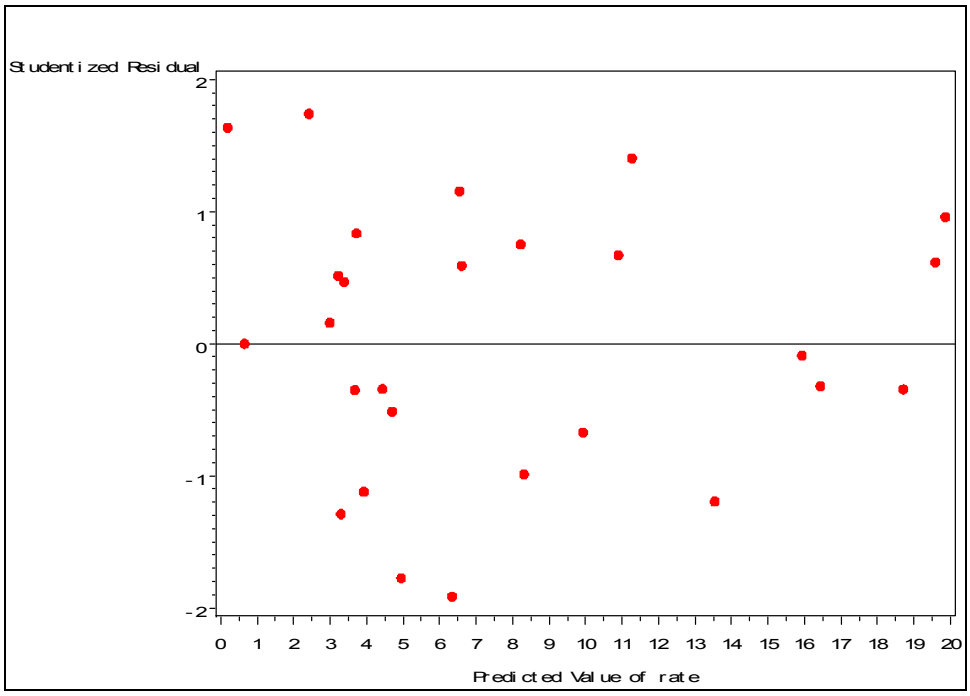


Figure 101: Studentized Residuals versus Predicted Values for the Injury Accident Model

The normal quantile plot in Figure 102 indicates that there is a strong inclination towards normality as the majority of the points closely follow the line that indicates a linear relationship with several even falling on the line. Most of the points cluster around the line with only a few deviating ones. This is an indication that the model assumptions are not violated.

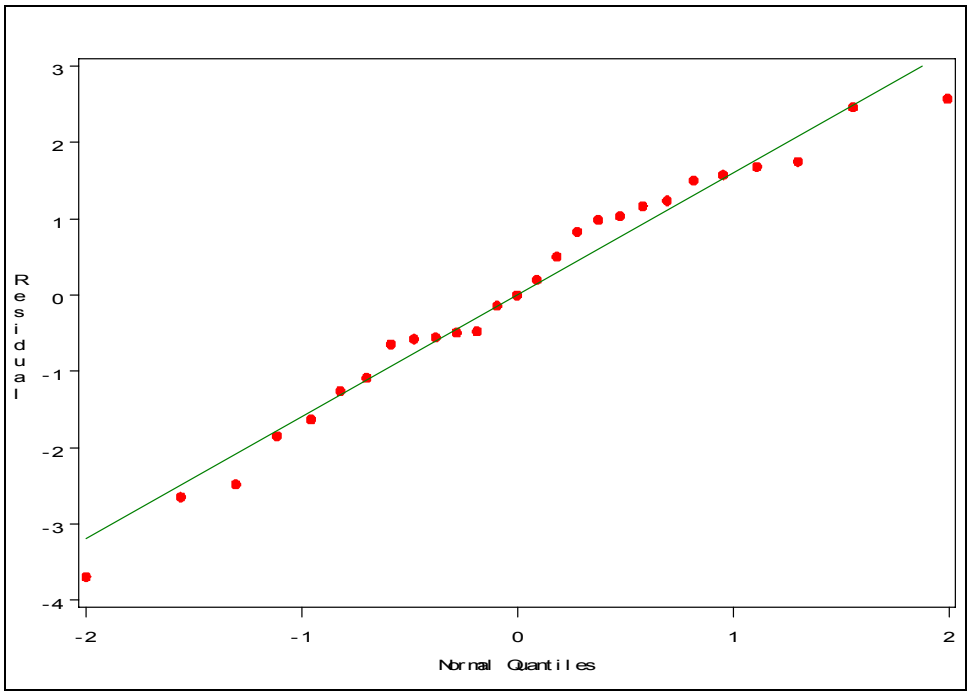


Figure 102: Normal Quantile Plot for the Injury Accident Model

The distribution that the residuals follow almost completely follows that of a normal distribution as can be seen in Figure 103. The peak of the model's distribution is only slightly lower than that of the normal distribution and skewed slightly towards the right. The normal distribution is the solid line while the dashed line that follows closely is the distribution from the residuals from this data set. This indicates that the residuals from the model follow a normal distribution, which is one of the model assumptions.

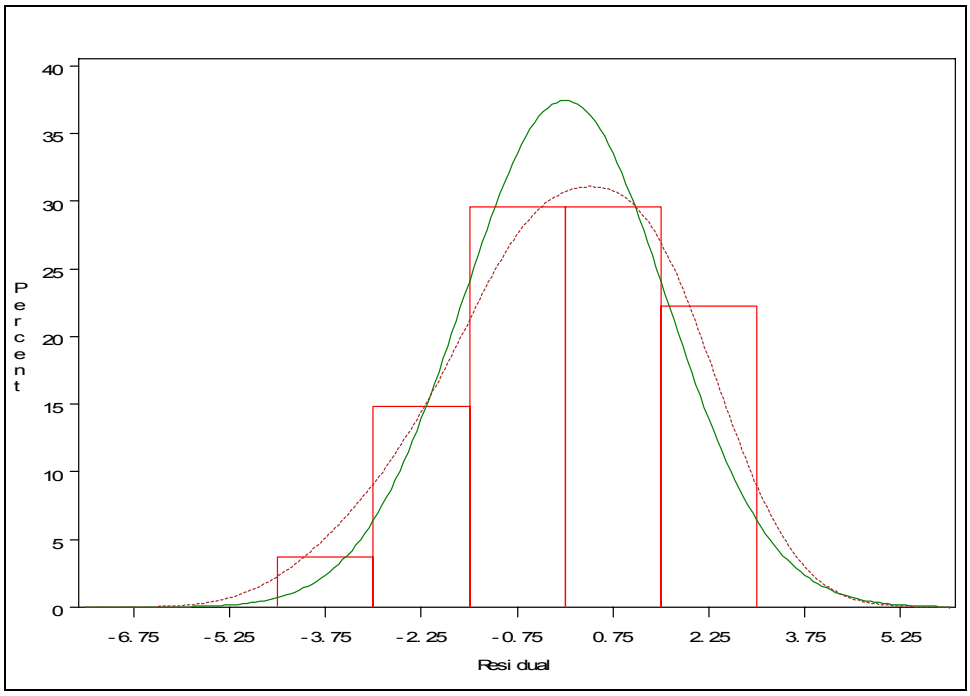


Figure 103: Normal Probability Plot for the Injury Accident Model

This model predicts the rate for the total number of injury accidents that occur on arterial road segments. Overall it appears to be a good model to use to predict these crashes and it takes an additive form. The additive form indicates that the variables in question tend to individual act upon the roadway in terms of causing crashes to happen. They do not act together to change crash rates, which will allow each item to be reviewed separately if the segment is about to get repaired or redesigned. This allows each variable to be independently adjusted by traffic engineers and a visible effect to be noticed. More variables were included in the model that predicts injury accidents than were in the model that predicts the total number of accidents. This is because the total number of accidents is more difficult to predict, since property-damage-only accidents can be caused in many more occasions than are injury accidents. The more exact influence of traffic and geometric characteristics on injury accidents allows for more variables to be included in the final model.

7 Validation

The final step in the modeling process involves validation of the model through independent data by comparing the results from the model with the actual values from a data set that was not used to help create the model. This allows for a review of how well the new data is represented by the model.

For the validation process, two data samples were used. One sample contained what would have been the next segments added to the database had data collection continued. These segments were located on parts of Park Avenue that were not previously sampled. Since these six segments would have been included in the model building database, they fit the exact profile of streets where the model can be appropriately applied. The second data sample consisted of six segments from Shrewsbury Street, which is classified as an urban arterial, though it is not a state primary as were all the other segments. This set of segments was useful in seeing how robust the developed models are and if some further application of the model is appropriate.

7.1 Linear Model Validation

The linear model from a surface review appears to be more robust than the model that predicts injury accident rates. This is due to the fact that only five variables are involved in this model as opposed to the fifteen in the injury accident rate model.

The first data grouping used for validation of the total accident model came from Park Avenue in Worcester. These segments would have been the next to be surveyed if more time had been available for collection of data for the model building. These segments fit the profile of the segments used to develop the model: an urban arterial,

preferably a state primary, with an average volume between ten and fifty thousand vehicles per day.

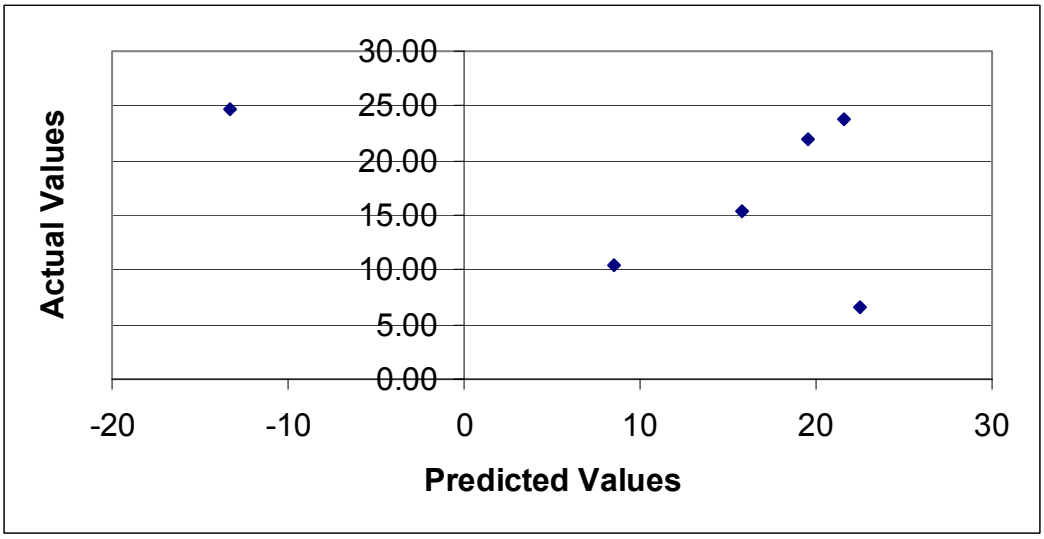


Figure 104: Predicted Values vs. Actual Values for Total Accident Rate Model with Park Avenue Data

When the six Park Avenue segments were entered into the model there was a fairly good result. As can be seen in Figure 104 there was a decent linear trend of the actual values of the total accident rate versus the predicted values from the model for four of the six segments. Two points, however, fall away from the linear trend. One does so due to the model predicting a negative accident rate, which would translate into a zero accident rate occurring on that segment since negative values do not occur. The other outlying point is when the actual accident rate of the segment was very low and the model forecast a much higher one. These both raise different concerns.

The one point where there is a very low actual accident rate may be indicating that this segment, BPP, has an unusually low occurrence of crashes compared with other similar road segments. This is not a bad thing, just a segment with better than average conditions. The accident prediction model, gives what could be considered an average accident rate, based on volume, length, percentage of residential land, number of parking

lots, and several other factors. This allows for segments that are better than 'average' to have low actual rates, while the predicted ones are much higher. Salisbury Street and Sagamore Road bound this segment on Park Ave and the most unusually thing about this road segment is that, while there was some commercial land use, there were no parking lots observed. This is mainly due to the fact that the few businesses were located in converted residential buildings that only had limited space for customer parking with parking provided by driveways and on-street parking. While this is not the most common conditions it is not unheard of and several segments that were used in the model development phase had similar characteristics of combined commercial and residential land use and no observed parking lot entrances.

The second point that leads to some concerns due to its lack of linearity compared with the other points comes from the fact that the model did not predict a positive accident rate. Instead the of the actual crash rate of 24.71 crashes per million vehicle miles, a rate of –13.25 crashes per million vehicle miles was predicted. There does not seem to be a particular reason why this negative rate would be observed. The only unusually characteristic noted on the segment that spans between Chandler Street and May Street, CPP, was a very large number of parking lot entrances, but the number of 27 falls below the maximum of 33 that was used to develop the model.

If the two outlying points are disregarded the amount of error in the predictions is relatively low with the four remaining points all having percent of error of less than twenty percent and two points less than ten percent as can be seen in Table 69.

Table 69: Error Table for Total Accident Rate Model with Park Avenue Data

Segment	Actual Accident Rate	Predicted Rate	% Error
APP	15.39	15.76	2.39
BPP	6.59	22.53	241.65
CPP	24.71	-13.25	153.64
DPP	21.88	19.56	10.62
EPP	23.77	21.60	9.12
FPP	10.49	8.47	19.27

The second data group used to validate the model came from Shrewsbury Street in Worcester. While an urban arterial, this road is not a state primary and throughout its length does not have a large variety in areas such as land use and alignment. The use of these segments will help show how robust the model is in its ability to be applied to more streets than originally designed for.

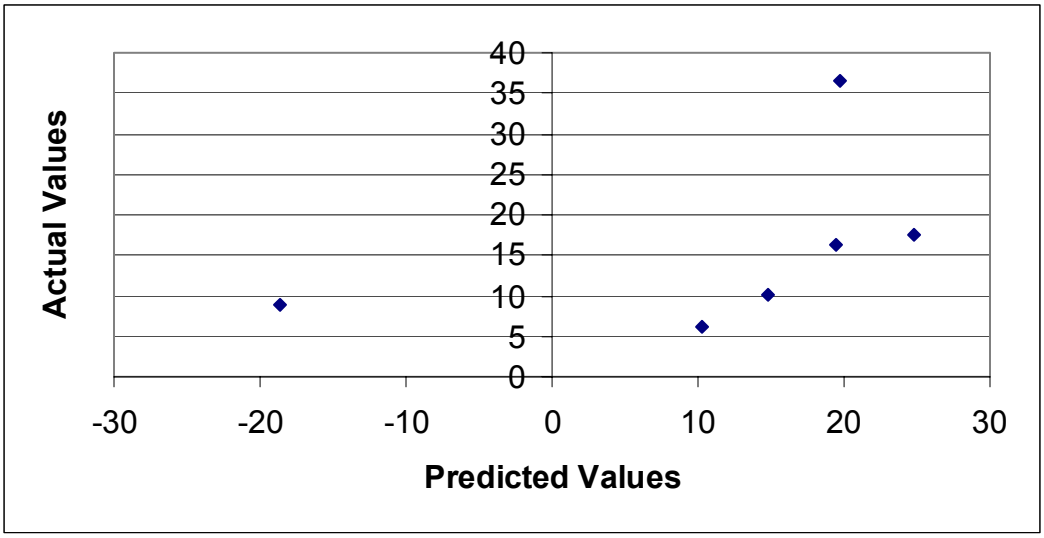


Figure 105: Predicted Values vs. Actual Values for Total Accident Rate Model with Shrewsbury Street Data

The predicted values versus the actual values for the data from Shrewsbury Street can be seen in Figure 105. As with the data from Park Avenue, there are two points that do not follow the linear relationship that is observed with four of the segment points. In terms of linearity Shrewsbury Street appears to perform just as well as Park Avenue does in the model with only two of six points as outliers. Like one of the points in the Park Avenue data, the outlying point on the negative side of the y-axis is due to the

prediction model producing a negative accident rate for the segment of Shrewsbury Street bounded by Adams Street and Fantasia Street (Segment DS). The only thing that appears different in this segment than in the others is again a fairly large number of parking lot entrances at 23 for this segment and while this is less than the maximum number used in the model database, the next highest number of parking lot entrances was in the high teens. On both occasions where large numbers of parking lots were observed, negative crash rates are predicted. This leads to a restriction needing to be placed on the prediction model of segments needing to possess less than a certain number of parking lot entrances. This limit set at sixteen comes from the second highest number of parking lots observed in the database with several segments having parking lot counts in the mid-teens. This sensitivity of the model due to the number of parking lot entrances emphasizes the fact that urban roads especially state primary ones have characteristics that influence crashes that are different than on rural roads where geometry plays the main role.

The second point that appears to be outlying from Figure 105 is similar to the Park Avenue data has a vastly different actual crash rate than would be supposed from the predicted rate. In the Park Ave. data the outlying point had an unusually low crash rate, in this Shrewsbury Street data the opposite is true with the segment displaying a very high crash rate of 36.63 while the predicted rate is 19.75 crashes per million vehicle miles. This segment, FS, is bounded by Belmont Street (Rt. 9) on one side and the entrances to a McDonalds and the Piccadilly Shopping Plaza on the other side. The segment is also relatively short though not so much that it would not fit parameters in the database. The practice of including both link and intersection crashes most likely is the

cause of this deviation between actual and predicted rates. The intersection that is included on this segment is with a state primary route and is in a configuration not of a T-intersection, but of a three-way angled intersection and this combination, despite the traffic lights regulating vehicles is the most probable explanation for the large actual crash rates.

Table 70: Error Table for Total Accident Rate Model with Shrewsbury Street Data

Segment	Actual Accident Rate	Predicted Rate	% Error
AS	17.52	24.84	41.8
BS	16.37	19.48	19.0
CS	10.23	14.78	44.5
DS	8.82	-18.57	310.5
ES	6.14	10.24	66.8
FS	36.63	19.75	46.1

The error observed from the segments on Shrewsbury Street is more than those segments from Park Avenue, but fairly reasonable with the exception of the one segment with a negative accident rate as can be seen in Table 70. Without that segment the error rate is under seventy percent.

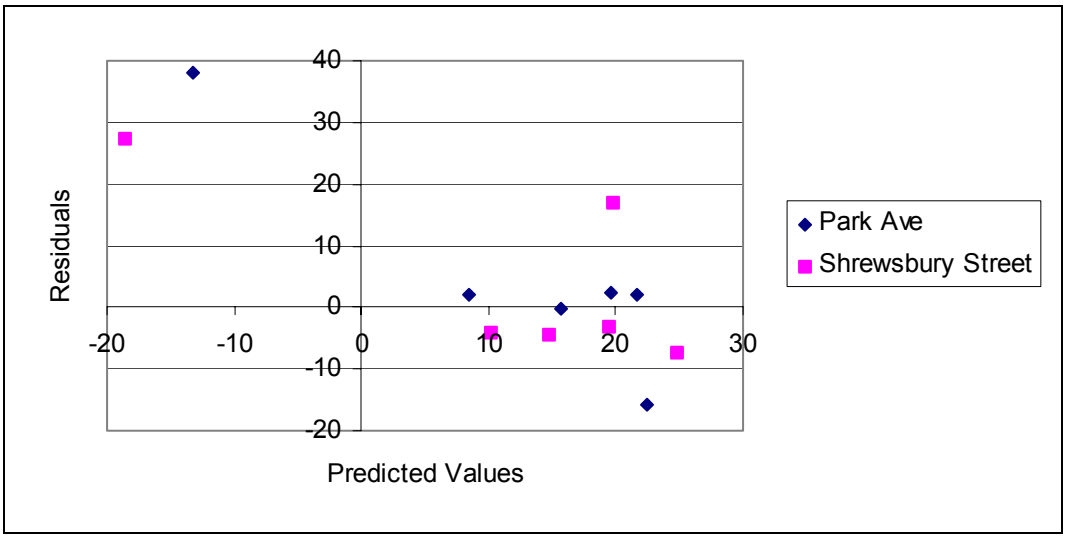


Figure 106: Predicted Values vs. Residuals for Validation of Total Accident Rate Model

The standard graphical diagnostic to check the model assumptions is looking at the plot of the predicted values versus the residuals (See Figure 106). With the exception

of the two points that do not fit the model by having too many parking lot entrances the other points from both Shrewsbury Street and Park Avenue show a constant error variance that follows that of the overall model. The variance for the segments used to develop the model ranged from approximately negative twenty to positive twenty and the validation data follows this trend. Two points even in this range could be considered outlying where the true range would be between negative ten and positive ten. These two points are the ones with either an unusually high or unusually low real crash rate as opposed to what the model predicted.

The linear total accident rate model is fairly robust. Restrictions must be placed on the allowable number of parking lots that can be on a segment in order for it to work properly. There is an indication that predicting the accidents on state primary roads works well, with an error rate at maximum of twenty percent, and the predicting crash rates for urban arterials that are not state primaries has a larger error rate, closer to sixty percent. While not originally designed for general urban arterials this model can be used and if reworked with a larger database, even perform well for these roads.

7.2 Multiplicative Model Validation

The multiplicative model appears to be less robust than the linear model that predicts total accident rates. This is due to the fact that fewer variables are involved in the multiplicative model and the multiplicative model has a lower coefficient of determination, 0.6672.

The same two groups of data were used for validation of the multiplicative model as were used to validate the linear model; Park Avenue and Shrewsbury Street. The first

data grouping used for validation of the total accident model came from Park Avenue in Worcester.

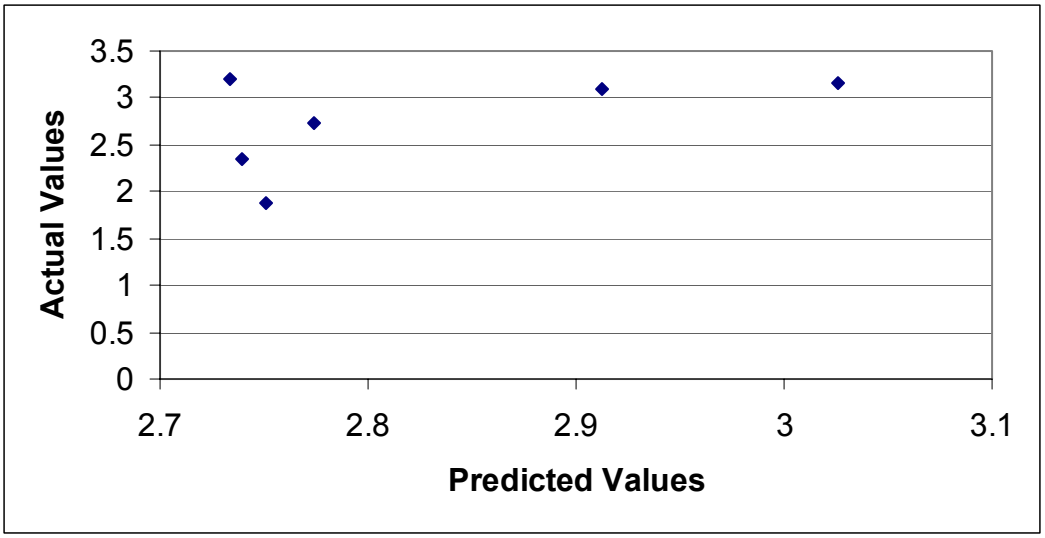


Figure 107: Predicted Values vs. Actual Values for Multiplicative Model with Park Avenue Data

When the six segments were entered into the model there was a fairly good result. As can be seen in Figure 107 there was a decent trend of the actual values of the total accident rate versus the predicted values from the model. Two points, however, fall away from the trend of the remaining points. One of these points is the one that was removed from applying to the total accident rate linear model in the previous section CPP. It is located below the trend line. The second of the two points was also previously discussed due to the segment having a particularly low accident rate and therefore the more average rate developed from the model does not fit segment BPP causing it to be located above the trend of the model. The characteristics observed in the total accident rate linear model remain true with the log-linear model. The same restriction on the database based on the number of parking lot entrances should remain true in spite of the fact that the number of parking lots was not determined to be a significant variable in the multiplicative model.

If the two outlying points are disregarded the amount of error in the predictions is relatively low with the four remaining points all having percent of error of less than twenty percent and three points less than ten percent as can be seen in Table 71. This low error means that the model is doing a good job at predicting values that are near the actual ones.

Table 71: Error Table for Multiplicative Model with Park Avenue Data

Segment	Actual Accident Rate	Predicted Rate	% Error
APP	2.73	2.77	1.5
BPP	1.89	2.75	45.9
CPP	3.21	2.73	14.8
DPP	3.09	2.91	5.6
EPP	3.17	3.03	4.5
FPP	2.35	2.74	16.5

The second data group used to validate the model came from Shrewsbury Street in Worcester. While an urban arterial, this road is not a state primary and throughout its length does not have a large variety in areas such as land use and alignment, but with so few variables in this model, the lack of variety in the data may not have a strong effect on the outcome of the model.

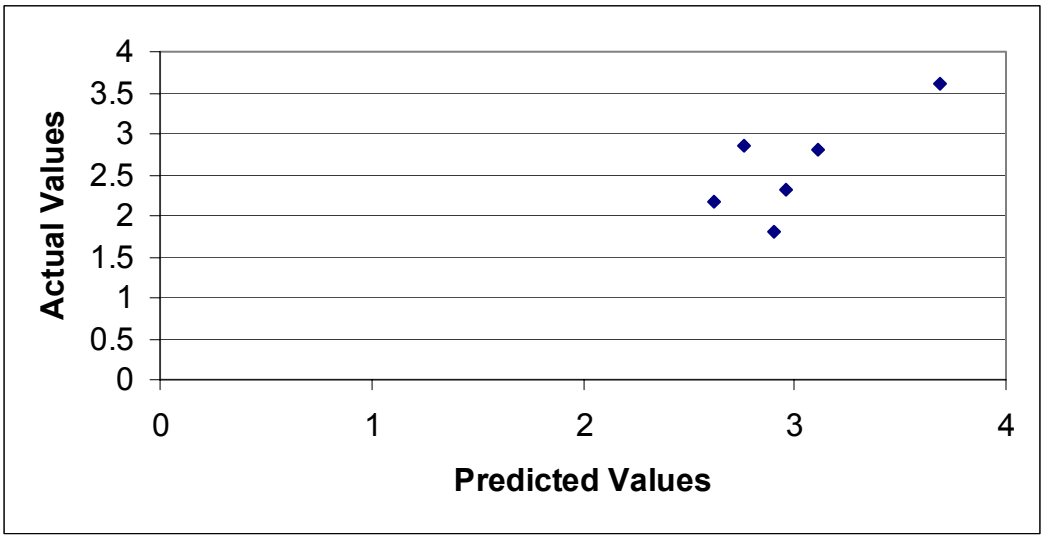


Figure 108: Predicted Values vs. Actual Values for Multiplicative Model with Shrewsbury Street Data

The predicted values versus the actual values for the data from Shrewsbury Street can be seen in Figure 108. As with the data from Park Avenue, there are some points that do not follow the relationship that is observed with the other four segments, however there is not a strong deviation from the noticed trend. These outliers depend on where the trend is assumed to be, but there is not a clear indication of this location. The previous outliers DS and FS are not as apparent in deviating from the remaining points.

Table 72: Error Table for Multiplicative Model with Shrewsbury Street Data

Segment	Actual Accident Rate	Predicted Rate	% Error
AS	2.86	2.76	84.2
BS	2.80	3.11	81.0
CS	2.32	2.96	71.1
DS	2.18	2.62	70.3
ES	1.81	2.90	52.7
FS	3.60	3.69	89.9

The error observed from the segments on Shrewsbury Street is significantly higher than those segments from Park Avenue, but all within the same range of each other as can be seen in Table 72. The jump from error rates of less than twenty percent to error rates around eighty percent show how while the model does work in that it predicts reasonable values for non-state primary roads, it does best with the exact type of roads that it was modeled for. If a larger database was originally collected that included all types of non-access controlled urban arterial roadways then it would probably yield a better match with the data from Shrewsbury Street.

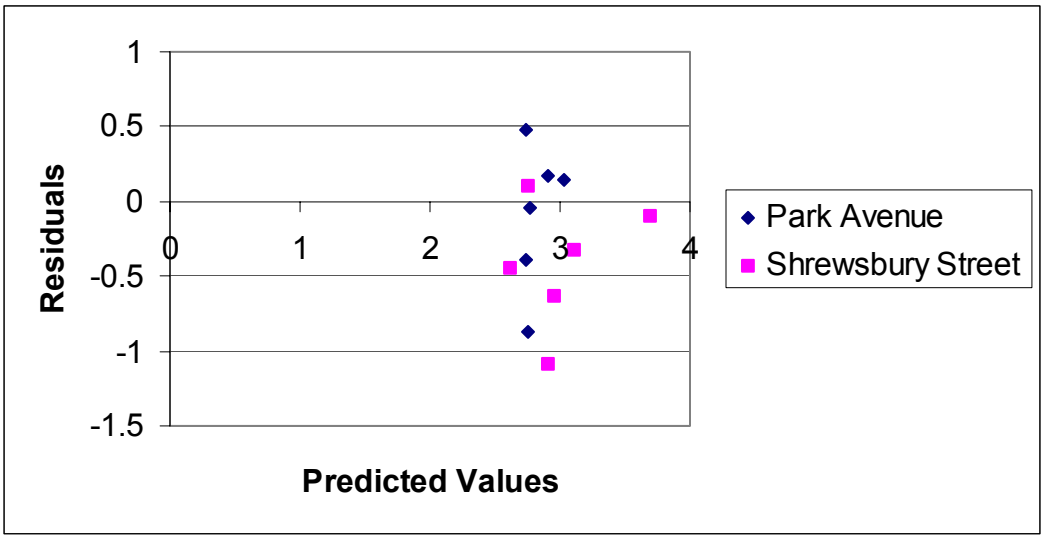


Figure 109: Predicted Values vs. Residuals for Validation of Multiplicative Model

The standard graphical diagnostic to check the model assumptions is looking at the plot of the predicted values versus the residuals (See Figure 109). With the exception of the two points that do not fit the model by having too many parking lot entrances the other points from both Shrewsbury Street and Park Avenue show a constant error variance that follows that of the overall model.

The total accident rate log-linear model is fairly robust. The slightly lower coefficients of determination and the adjusted coefficient have values that are typically not acceptable for working models with 0.6672 and 0.6238 respectively, but that does not prevent the model from giving a general range of what the crash rate on a segment should be near. It was found that restrictions placed on the allowable number of parking lots in other models should also be carried over to this model for it to work properly. There is an indication that predicting the accidents on state primary roads works well, with an error rate at maximum of twenty percent, and the predicting crash rates for urban arterials that are not state primaries has a larger error rate of closer to eighty percent. While this

model works well for the roads it was designed for, extending this exact model to other urban arterial roads is not suggested.

7.3 Injury Accident Model Validation

The total accident rate linear model appears to be more robust than the linear injury accident rate model. Though having the same functional form of a linear model with a normal error distribution, the injury accident model has many more variables, fifteen as opposed to six, which may cause it to be too specific to the model building data set. The use of many more variable shows that more factors are needed when predicting the injury accident rate, but this can be due to the fact that injury crashes compose only approximately one third of all crashes.

The first data grouping used for validation of the injury accident model came from Park Avenue in Worcester. These segments would have been the next to be surveyed if more time had been available for collection of data for the model building. These points fit the parameters of the model, an urban arterial, preferably a state primary, with an average volume between ten and fifty thousand vehicles per day minus the one point that has found to not fit the model parameters by reason of having too many parking lot entrances.

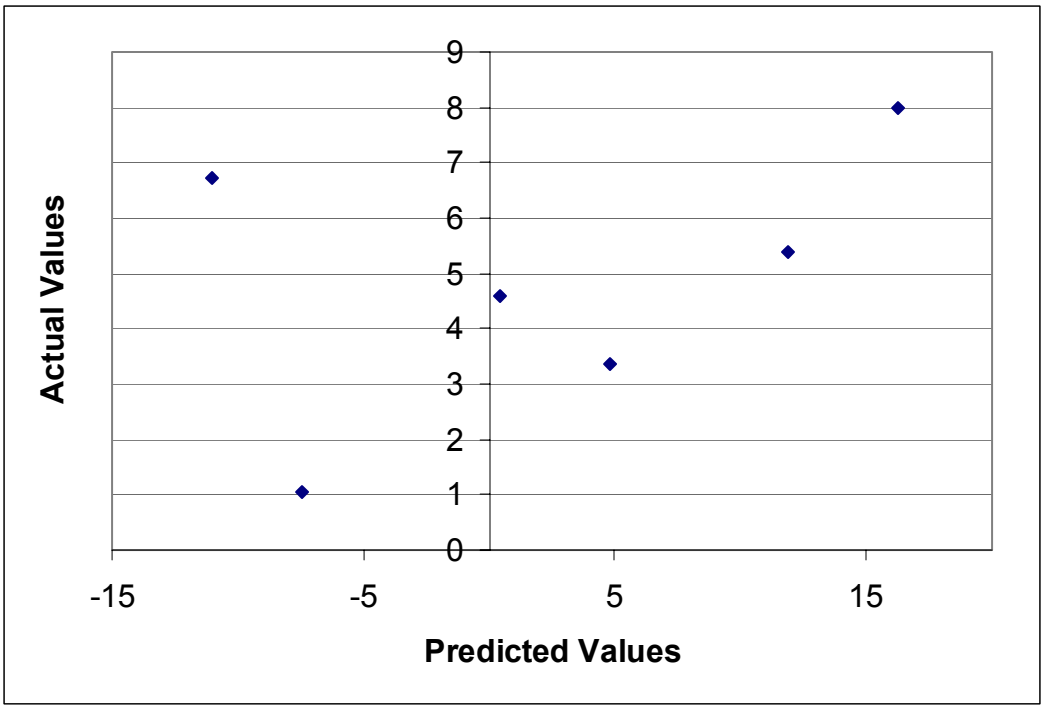


Figure 110: Predicted Values vs. Actual Values for Injury Accident Rate Model with Park Avenue Data

When the six segments were entered into the model there was a fairly good result. As can be seen in Figure 110 there was a fairly linear trend of the actual values of the total accident rate versus the predicted values from the model for four of the six points. This can be more fully seen when segment CPP, that has been removed from eligibility for the model, is no longer in the plot (See Figure 111). Even though one segment has a negative accident rate predicted, it remains along the line defined by the other data points in the plot. The segment removed from the model parameters in the total accident rate model, is again removed based on those same considerations.

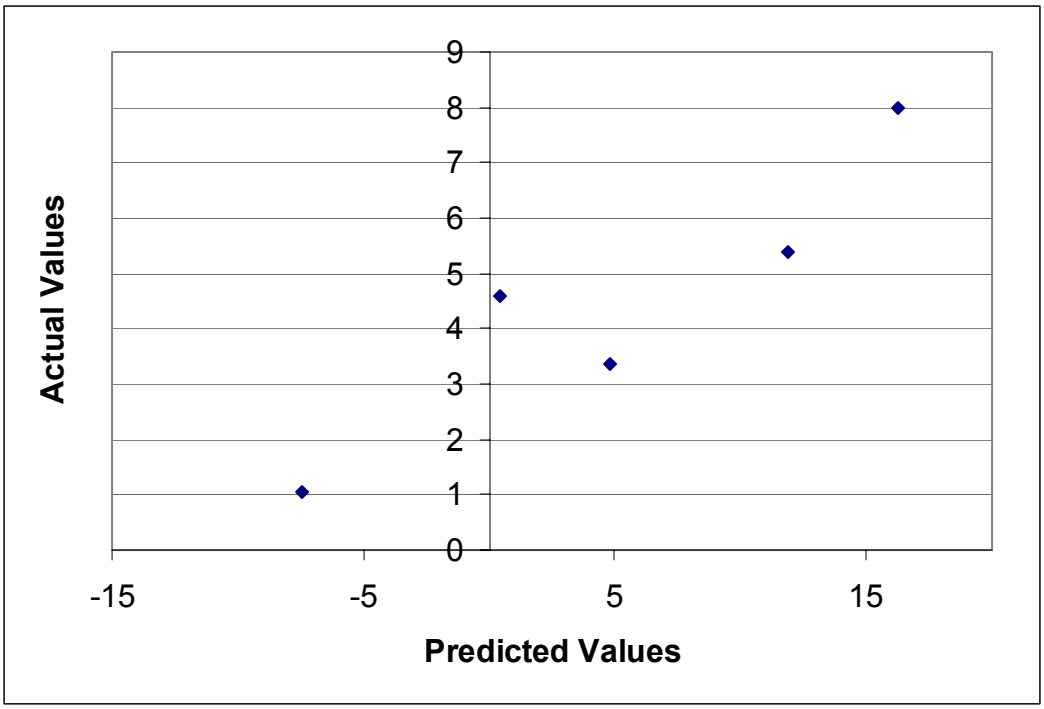


Figure 111: Predicted Values vs. Actual Values for Injury Accident Model with Valid Park Avenue Data

The two points that lead to concern are the same points that brought concern in the total accident rate model. In the injury rate model, both of these points have a prediction of negative crash rates, which should effectively translate into a zero accident rate occurring since negative accidents do not occur. The remaining point that is removed from the range of acceptable predictions does not appear to have any specific area where its characteristics are extreme from those that the model was formed from. The segment does have a relative low actual accident rate, but not by any means the lowest that was used to create the model, so no particular cause can be identified as the reason for the negative injury accident rate.

Even when the extreme points are disregarded the amount of error in the predictions is relatively high. Only two segments had an error less than 100 percent positive or negative and only one segment had a percent error less than fifty percent as

can be seen in Table 73. These large errors show that while the injury accident rate model may have a large coefficient of determination at 0.9319, this does not mean that the model will be robust enough for other data to be well represented and able to be predicted accurately.

Table 73: Error Table for Injury Accident Rate Model with Park Avenue Data

Segment	Actual Accident Rate	Predicted Rate	% Error
APP	4.6	0.45	90.2
BPP	1.06	-7.47	808.4
CPP	6.72	-11.02	264
DPP	5.39	11.90	120.6
EPP	8.00	16.26	103.2
FPP	3.37	4.84	43.6

The second data set used to validate the model came from Shrewsbury Street in Worcester. While an urban arterial, this road is not a state primary and throughout its length does not have a large variety in areas such as land use and alignment.

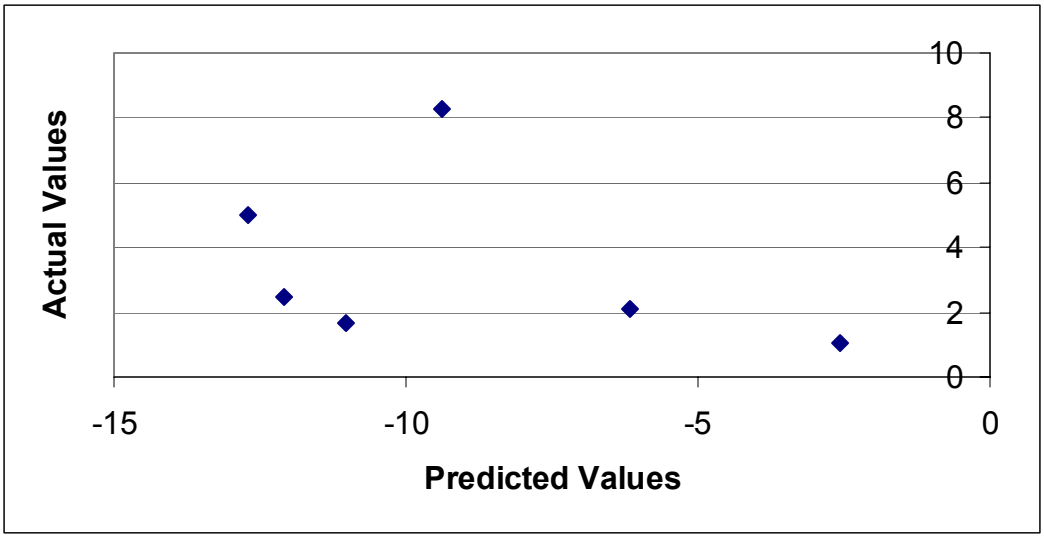


Figure 112: Predicted Values vs. Actual Values for Injury Accident Rate Model with Shrewsbury Street Data

The predicted values versus the actual values for the data from Shrewsbury Street can be seen in Figure 112. As oppose to the data from Park Avenue, none of the points follow the expected linear relationship. With the Shrewsbury Street data, no

positive injury accident rates were predicted, in spite of the fact that injury accidents did occur. The actual injury accident rates are in the same range as those as the Park Avenue data and the same range as those from which the model was built. This lack of any viable accident rates, whether with a large amount of error or not makes this model not applicable to non-state primary roads.

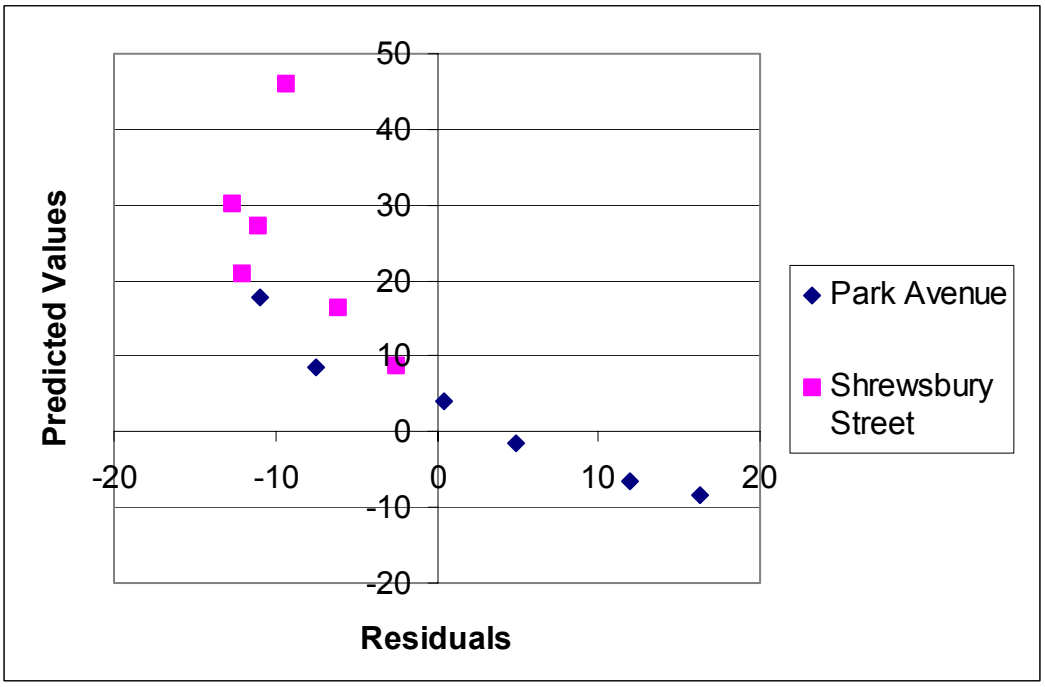


Figure 113: Predicted Values vs. Residuals for Validation of Injury Accident Rate Model

The standard graphical diagnostic to check the model assumptions is looking at the plot of the predicted values versus the residuals (See Figure 113). The data from Park Avenue when the segment that does not fit with the number of parking lots is removed from the data set mostly shows that the error terms follow the normal assumptions. They show that there is a constant variance that falls within that of the model. The Shrewsbury Street data on the other hand does not follow the normal assumptions, and as the model does not appear work for the non-state primary roads, this does not create any surprises. It appears that there is some systematic error in the residuals, but as the residuals did not

exhibit this trait when building the model a transformation is not likely to help at this stage leaving this model to provide very inexact results. By putting the residuals of both validation data sets together it can be easily seen how the Shrewsbury Street data does not work and how the Park Avenue data does work better. The conclusions that can be drawn from this validation process include that the injury accident rate model is not nearly as robust as that of the total accident rate model.

7.4 Summary of Validation

Some important issues have been brought to light during the validation process. One of these is that the model is limited by the number of parking lot entries a segment has. Segments with large number of parking lot entries did not perform well in either the total accident rate model or the injury accident rate model. This sensitivity to the number of parking lot entrances should be further examined in the future.

The total accident rate model was found work well for roads that exactly fit the profile of urban state primary roads with volumes between ten and fifty thousand vehicles per day with error rates of less than twenty percent. With other urban roads the total accident rate model performed adequately but with error rates closer to fifty percent. The total accident rate model can be used with a degree of confidence for state primary roads and with a lesser amount of confidence for other urban roads.

The injury accident rate model was found to be less robust than the total accident rate model. With the data that matched the model specifications (Park Avenue data), the error rates were very high, and when the Shrewsbury Street data was used the model did not perform well at all predicting only negative injury accident rates. While a general

idea can be gained about injury accident rates on state primary roads, this injury accident rate model should not be applied to other urban roads.

The multiplicative model was found to be of median robustness. It works well with error rates under twenty percent for the urban roads it was designed for, but this model's range cannot be extended. When applied to non-state primary roads, the model routinely produced error rates around eighty percent.

8 Conclusions

The study of the causes of vehicle crashes is a complex mixture of vehicle, driver, environment, traffic and road characteristics. These all combine in a myriad of ways that a mathematical model can only attempt to duplicate. The major classifications of rural and urban roads, followed by the classifications of arterial, collector and local roads all have their own patterns and relationships that need to be examined individually and separate from the others. Rural arterials have long been given much attention based on the large number of miles of the roads and the large percentage of crashes that occur on them and many advances have been made in the art of predicting crashes and speed on those roads. But, closer spaced junctions, difference in land use patterns, geometric consideration and traffic patterns along with different layout of link and junctions lend themselves toward a different approach in urban locations than in the longer studied rural ones. The urban environment is similar to the rural one, in that there are geometric and traffic issues that occur, but with the larger populations and numbers of vehicles and pedestrians using the roads, the urban locations become more complex with closely spaced buildings, access points, roadside hazards and people.

In crowded environments the possibilities that exist for crashes to occur are numerically greater leading to more actual crashes with the corresponding damage to property and people. This large number of crashes and limited amount of funds to respond to these incidents and to maintain and improve the roadway network is why the ability to predict where and how many of these incidents will occur is an important skill. A prediction model is also useful in that even if the exact crash rate it predicts is not exact the model does give an idea of what similar road segments should have and allows for

especially hazardous or safe sites to be identified and then examined for the characteristics that are causing the extreme conditions.

The prediction of crashes has many level not the least of which is what should actually be the depended variable, the number of crashes, a crash rate or something else. Historically, crash rates and the total number of crashes have been the choice for dependent variables. Both offer unique challenges as a primary choice. Crash rates are typically normalized by length and volume leading to the question of whether crashes are linearly related to these two items. The other common choice of dependent variable of total number of crashes causes problems in that crashes are discrete and non-negative which causes the normal distribution typically used for the error structure of prediction models to not apply to the dependent variable. The issues relating to the relationships of the variables in crash rates have not been verified repeatedly to be linear or non-linear in nature. Experimenting with the database used in this research no clear relationship between crash rate variables was established as linear or non-linear. The relationship of the number of crashes following a Poisson or negative binomial distributions was found to be equally unclear. This uncertainty lead to no clear trend being identified in the data and the more conventional choice of crash rate chosen as the dependent variable.

Using a dependent variable of crash rate meant that the error structure is normally distributed. The other major choice in modeling that occurs is how the independent variables interact with each other. The forms that were considered as the most likely form for predicting crash rates in urban area were linear relationships and multiplicative relationships. Both have been used to develop models in rural areas but no agreed upon relationship has been found in urban areas. Models were developed to predict the total

crash rate with both a linear and multiplicative form. The linear form was found to have a better fit for the data and to be a more robust model in that both state primary roads and other arterial roads could have crash rates predicted to a better than fifty percent error. The multiplicative model while working well for the state primary roads did not perform well on other urban arterial roads. In addition to the functional form, it is necessary to specify the form of the crash rate. The linear model that predicts the total crash rate has many more independent variables that were found to be significant to predicting the crash rate with fifteen variables as opposed to the six in the total accident rate model.

The models that were developed due to this research help show that the complex nature of crashes in an urban environment need to have a different approach than those in rural areas. The difference in the interaction between variables in the different environments needs to have more exploration since both forms produce workable models and the true model most likely lies in between the two forms. Limitations were also placed on the model due to the small size of the database used to develop the models. With a larger database the relationships between variables should be easier to identify.

9 Reference:

Allison, Paul D. Logistic Regression Using the SAS System: Theory and Application. Cary, NC.: SAS Institute Inc., 1999.

American Association of State Highway and Transportation Officials A Policy on Geometric Design of Highways and Streets. Washington D.C. 2001

Amundes, Astrid H., and Rune Elvik. "Effects on Road Safety of New Urban Arterial Roads." Accident Analysis and Prevention. In Press, Corrected Proof, February 2003.

Bonneson, J. A., and P. T. McCoy, "Effect of Median Treatment on Urban Arterial Safety: An Accident Prediction Model," Highway Research Record 1581, Highway Research Board, Washington, D.C.; (1997).

Botha, J. L, Sullivan, E. C., and X. Zeng, "Level of Service of Two-Lane Rural Highways with Low Design Speeds," Highway Research Record 1457, Highway Research Board, Washington, D.C.; (1994).

Bowman, B.L., and R. L. Vecellio, "Assessment of Current Practice in Selection and Design of Urban Medians to Benefit Pedestrians," Highway Research Record 145, Highway Research Board, Washington, D.C.; (1994).

Bowman, B. L., and R. L. Vecellio, "Effect of Urban and Suburban Median Types on Both Vehicular and Pedestrian Safety," Highway Research Record 1445, Highway Research Board, Washington, D.C.; (1994).

Brown, H. C. and A. P. Tarko, "Effects of Access Control on Safety on Urban Arterial Streets," Highway Research Record 1665, Highway Research Board, Washington, D.C.; (1999).

Central Massachusetts Regional Planning Commission. Daily Traffic Volumes and Peak Period Turning Movement Counts. Worcester, MA contract #30047, January 2001.

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

Davis, G. A., "Estimating Traffic Accident Rates While Accounting for Traffic-Volume Estimation Error," Highway Research Record 1717, Highway Research Board, Washington, D.C.; (2000).

De Leur, P., and T. Sayed, "Development of a Road Safety Risk Index," Highway Research Record 1784, Highway Research Board, Washington, D.C.; (2002).

Devore, Jay L. Probability and Statistics for Engineering and the Sciences. California; Brooks/Cole Publishing Company, 1982.

Donnell, E. T., Ni, Y., Adolini, M., and L. Elefteriadou, "Speed Prediction Models for Trucks on Two-Lane Rural Highways," Highway Research Record 1751, Highway Research Board, Washington, D.C.; (2001).

Easa, S. M., "Design Considerations for Highway Reverse Curves," Highway Research Record 1445, Highway Research Board, Washington, D.C.; (1994).

Elvik, Rune. "The importance of confounding in observational before-and-after studies of road safety measure." Accident Analysis & Prevention. Vol. 34, Great Britain, 2002, pp631-635.

Fitzpatrick, K., Carlson, P., Brewer, M., and M. Wooldridge,"Design Factors that Affect Driver Speed on Suburban Streets," Highway Research Record 1751, Highway Research Board, Washington, D.C.; (2001).

Fitzpatrick, K., Shamburger, C. B., Drammes, R. A. and D. B. Fambro, "Operating speed on Suburban Arterial Curves," Highway Research Record 1579, Highway Research Board, Washington, D.C.; (1997).

Garber, N. J., and A. A Ehrhart, "Effect of Speed, Flow, and Geometric Ccharacteristics on Crash Frequency for Two-Lane Highways," Highway Research Record 1717, Highway Research Board, Washington, D.C.; (2000).

Garber, Nicholas J. and Lester A. Hoel Traffic and Highway Engineering. Revised 2nd. Ed. Boston; PWS Publishing, 1999.

General Laws of Massachusetts <www.state.ma.us/legis/laws/mgl/90-17.htm> (24 November 2003).

Gibreel, G.M., Easa, S.M., Hassan, Y., and I.A. El-Dimeery. "State of the Art of Highway Geometric Design Consistency." Journal of Transportation Engineering. Vol. 125 Issue 4 July/Aug 1999, 305-313.

Greibe, Poul. "Accident Prediction Models for Urban Roads." Accident Analysis and Prevention. Vol 35, Issue 2, March 2003, 273-285.

Hadi, M. A., Aruldhas, J., Chow, L. F., and J. A. Wattleworth, "Estimating Safety Effects of Cross-Section Design for Various Highway Types Using Negative Binomial Regression," Highway Research Record 1500, Highway Research Board, Washington, D.C.; (1995).

Haight, Frank A. Handbook of the Poisson Distribution. New York; John Wiley & Sons, Inc., 1967.

Haselton, C. B., Gibby, A. R., and T. C. Ferrara, "Methodologies Used to Analyze Collision Experience Associated with Speed Limit Changes on Selected California Highways," Highway Research Record 1784, Highway Research Board, Washington, D.C.; (2002).

Hauer Ezra, "Statistical Test of Difference Between Expected Accident Frequencies," Highway Research Record 1542, Highway Research Board, Washington, D.C.; (1996).

Hauer, Ezra, Ng, J.C.N., and J. Lovell, "Estimation of Safety of signalized Intersections," Highway Research Record 1182, Highway Research Board, Washington, D.C.; (1988).

Higle, J. L. and J. M. Witkowski, "Bayesian Identification of Hazardous Locations," Highway Research Record 1185, Highway Research Board, Washington, D.C.; (1988).

Knuiman, M. W., Council, F.M. and D. W. Reinfurt, "Association of Median Width and Highway Accident Rates," Highway Research Record 1401, Highway Research Board, Washington, D.C.; (1993).

Lamm, Ruediger, Basil Psarianos, and Theodor Mailaender. Highway Design and Traffic Safety Engineering Handbook. New York; McGraw-Hill, 1999.

Lau, M. Y. and A. D. May, Jr., "Injury Accident Prediction Models for Signalized Intersections," Highway Research Record 1172, Highway Research Board, Washington, D.C.; (1988).

Lord, D., "Application of Accident prediction Models for computation of Accident Risk on Transportation Networks," Highway Research Record 1784, Highway Research Board, Washington, D.C.; (2002).

Lord, D., and B. N. Persaud, "Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure," Highway Research Record 1717, Highway Research Board, Washington, D.C.; (2000).

Luttinen, R. T., "Uncertainty in Operational Analysis of Two-Lane Highway," Highway Research Record 1802, Highway Research Board, Washington, D.C.; (2002).

Maher, Michael J. and Ian Summersgill. "A Comprehensive Methodology for the Fitting of Predictive Accident Models." Accident Analysis and Prevention. Vol. 28, No. 3, Great Britain, 1996, 281-296.

Mass Highway. <www.state.ma.us/mhd/home.htm> (November 2003).

McFadden, J. and L. Elefteriadou, "Formulation and Validation of Operating Speed-Based design Consistency Models by Bootstrapping," Highway Research Record 1579, Highway Research Board, Washington, D.C.; (1997).

B-McFadden, J., and L. Elefteriadou, "Evaluating Horizontal Alignment Design Consistency of Two-Lane Rural Highways: Development of New Procedure," Highway Research Record 1737, Highway Research Board, Washington, D.C.; (2000).

Miaou, S., Lu, A. and H. S. Lum, "Pitfalls of Using R^2 to Evaluate Goodness of Fit of Accident Prediction Models," Highway Research Record 1542, Highway Research Board, Washington, D.C.; (1996).

A: Miaou, S., Hu, P. S., Wright, T., Rathi, A. K. and S. C. Davis, "Relationship Between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach," Highway Research Record 1376, Highway Research Board, Washington, D.C.; (1992).

Montgomery, Douglas C. and George C. Runger. Applied Statistics and Probability for Engineers. 3rd ed. USA; John Wiley & Sons, Inc., 2003.

Mountain, Linda, Fawaz, B. and D. Jarrett, "Accident Prediction Models for Roads With Minor Junctions," Accident Analysis & Prevention, Vol 28, No. 6, Great Britain, 1996, pp695-707.

Neter, John, M. H. Kutner, C. J. Nachtsheim and William Wasserman. Applied Linear Statistical Models. 4th ed. Boston; WCB McGraw-Hill, 1996.

Pedestrian Safety Roadshow: Facts and Figures. <safety.fhwa.dot.gov/roadshow/walk/facts/mode/ > (January 2004).

Persaud, B., Lord, D., and J. Palmisano, "Calibration and Transterability of Accident Prediction Models for Urban Intersections," Highway Research Record 1784, Highway Research Board, Washington, D.C.; (2002).

Persaud, B. and L. Dzbik, "Accident Prediction Models for Freeways," Highway Research Record 1401, Highway Research Board, Washington, D.C.; (1993).

Petruccelli, Joseph D, B. Nandram and M. Chen. Applied Statistics for Engineers and Scientists. New Jersey; Prentice Hall, 1999.

Poch, Mark and Fred Mannering. "Negative Binomial Analysis of Intersection-Accident frequencies." Journal of Transportation Engineering. March/April 1996, 105-113.

Poe, C. M. and J. M. Mason, Jr., "Analyzing influence of Geometric Design on Operating speeds Along Low-Speed Urban Streets: Mixed-Model Approach," Highway Research Record 1737, Highway Research Board, Washington, D.C.; (2000).

Raub, R. A., "Occurrence of Secondary Crashes on Urban Arterial Roadways," Highway Research Record 1581, Highway Research Board, Washington, D.C.; (1997).

Rice, John A. Mathematical Statistics and Data Analysis. California; Wadsworth & Brooks/Cole Advanced Books & Software, 1988.

Ross, Sheldon. A First Course in Probability. 5th ed. New Jersey; Prentice Hall, 1998.

Saccomanno, F. F. and C. Buyco, "Generalized Loglinear Models of Truck Accident Rates," Highway Research Record 1172, Highway Research Board, Washington, D.C.; (1988).

Saccomanno, F.F., Chong, K.C., and S. A. Nassar, "Geographic Information System Platform for Road Accident Risk Modeling," Highway Research Record 1581, Highway Research Board, Washington, D.C.; (1997).

SAS Institute Inc. User's Guide, Version 8 Cary, www.math.wpi.edu/saspdf/stat/chap29.pdf ,NC: SAS Institute Inc., 1999.

Schurr, K. S., McCoy, P.T., Pesti, G., and R. Huff, "Relationship of Design, Operating, and Posted Speeds on Horizontal Curves of Rural Two-Lane Highways in Nebraska.," Highway Research Record 1796, Highway Research Board, Washington, D.C.; (2002).

Z -Tarris, J. P., Mason, Jr., J. M., and N. D. Antonucii, "Geometric Design of Low-Speed Urban Streets," Highway Research Record 1701, Highway Research Board, Washington, D.C.; (2000).

Tarris, J. P., Poe, C. M., Mason, Jr., J. M. and K. G. Goulias, "Predicting Operating Speeds on Low-Speed Urban Streets: Regression and Panel Analysis Approaches," Highway Research Record 1523, Highway Research Board, Washington, D.C.; (1996).

Wilmink, I.R. and L. H. Immers, "Deriving Incident management Measures Using Incident Probability Models and Simulation," Highway Research Record 1554, Highway Research Board, Washington, D.C.; (1996).

Worcester accident databases 2000, 2001, 2002

A Appendix: Database for Creating Model

This appendix has the datasheets for the arterial segments that were used to create the models in this paper. The summary sheet of that data is also included.

Date	Segment	Route #	volume	3-total	% heavy veh.	3-total-s	3-injury-s	3-PDO-s	3-total-i	3-injury-i	3-PDO-i	2002 total-s
9/27	AC	122, 122A	28046	132		94	36	58	38	13	25	42
9/27	BC	122, 122A	25108	50		39	9	30	11	5	6	12
9/27	CC	122, 122A	25108	60		23	11	12	37	21	16	11
9/27	DC	122, 122A	25108	42		17	7	10	25	13	12	6
9/27	EC	122, 122A	12852	74		44	22	22	30	8	22	13
9/27	FC	122, 122A	12852	145	3	114	39	75	31	17	14	36
9/27	GC	122	12852	110	3	83	31	52	27	9	18	31
9/30	HC	122	11016	26	1.28	18	7	11	8	3	5	8
9/30	IC	122	13223	138	1.28	84	18	66	54	14	40	32
9/30	JC	122	17043	68	0.4	47	6	41	21	4	17	9
9/21	AB	9	34421	157	2.09	113	34	79	44	18	26	52
9/21	BB	9	34421	49	1.11	5	2	3	44	12	32	0
9/21	CB	9	38694	180	1.11	118	33	85	62	23	39	34
9/21	DB	9	23200	169	1.11	161	55	106	8	2	6	67
9/21	EB	9	28726	30	0.41	28	9	19	2	0	2	9
9/21	FB	9	28726	49	0.41	0	0	0	49	10	39	0
9/21	GB	9	46808	100	0.41	7	1	6	93	41	52	3
9/21	HB	9	40698	254	0.55	200	65	135	54	14	40	72
9/16	AH	9	27339	160	3.1	114	30	83	46	16	30	49
9/16	BH	9	27339	58	3.1	8	3	5	50	23	27	3
9/16	CH	9	27339	190	3.1	149	38	111	41	15	26	53
9/15	DH	9	22610	150	3.1	99	22	77	51	17	34	35
10/7	AP	9, 122A	23572	65	1.14	39	13	26	26	8	18	13
10/7	BP	9, 122A	26128	139	1.89	62	16	46	77	22	55	16
9/22	CP	9,12	23332	57	1.89	23	9	14	34	14	20	6
9/22	DP	9,12	26760	146	1.95	87	28	59	59	15	44	33
9/22	EP	9,12	23460	44	1.95	23	4	19	21	6	15	2

Segment	2002 injury-s	2002 PDO-s	2002 total-i	2002 injury-i	2002 PDO-i	2001 total-s	2001 injury-s	2001 PDO-s	2001 total-i
AC	15	27	7	1	6	18	7	11	14
BC	2	10	2	2	0	17	4	13	5
CC	5	6	15	8	7	11	5	6	12
DC	2	4	7	2	5	4	1	3	7
EC	7	6	13	8	5	12	8	4	9
FC	11	25	10	4	6	35	11	24	11
GC	11	20	7	2	5	31	8	23	9
HC	3	5	0	0	0	6	2	4	5
IC	5	27	21	5	16	25	4	21	20
JC	2	7	6	4	2	20	1	19	7
AB	18	34	9	4	5	38	9	29	19
BB	0	0	14	3	11	4	1	3	10
CB	13	21	20	6	14	33	6	27	25
DB	23	44	3	1	2	45	9	36	2
EB	3	6	2	0	2	8	2	6	0
FB	0	0	16	2	14	0	0	0	18
GB	0	3	37	17	20	3	1	2	27
HB	28	44	12	1	11	50	12	38	21
AH	18	30	7	3	4	33	7	26	22
BH	1	2	10	8	2	3	1	2	21
CH	12	41	12	6	6	45	14	31	8
DH	8	27	16	6	10	40	8	32	14
AP	4	9	7	2	5	20	7	13	9
BP	7	9	40	13	27	5	1	4	25
CP	2	4	6	2	4	16	6	10	10
DP	14	19	20	5	15	31	8	23	18
EP	0	2	7	2	5	12	2	10	8

Segment	2001 injury-i	2001 PDO-i	2000 total-s	2000 injury-s	2000 PDO-s	2000 total-i	2000 injury-i	2000 PDO-i	minor access
AC	3	11	34	14	20	17	9	8	2
BC	2	3	10	3	7	4	1	3	3
CC	7	5	1	1	0	10	6	4	0
DC	3	4	7	4	3	11	8	3	0
EC	0	9	19	7	12	8	0	8	1
FC	7	4	43	17	26	10	6	4	5
GC	4	5	21	12	9	11	3	8	11
HC	1	4	4	2	2	3	2	1	6
IC	6	14	27	9	18	13	3	10	8
JC	0	7	18	3	15	8	0	8	0
AB	7	12	23	7	16	16	7	9	2
BB	1	9	1	1	0	20	8	12	0
CB	10	15	51	14	37	17	7	10	2
DB	0	2	49	23	26	3	1	2	10
EB	0	0	11	4	7	0	0	0	11
FB	3	15	0	0	0	15	5	10	0
GB	10	17	1	0	1	29	14	15	1
HB	5	16	78	25	53	21	8	13	8
AH	7	15	32	5	27	17	6	11	0
BH	8	13	2	1	1	19	7	12	0
CH	3	5	51	12	39	21	6	15	8
DH	5	9	24	6	18	21	6	15	13
AP	4	5	6	2	4	10	2	8	3
BP	4	21	41	8	33	12	5	7	2
CP	4	6	1	1	0	18	8	10	0
DP	3	15	23	6	17	21	7	14	2
EP	1	7	9	2	7	6	3	3	0

Segment	driveways	parkinglots	d+p	total	fire hydran	mailboxes	light poles	utility poles	benches	trees	monument
AC	0	3	3	5	4	0	5	0	0	0	0
BC	2	15	17	20	3	4	12	0	0	0	0
CC	2	3	5	5	1	2	4	0	0	0	0
DC	1	4	5	5	0	1	4	0	0	0	0
EC	5	6	11	12	2	0	8	0	0	3	0
FC	5	9	14	19	1	1	8	0	0	4	0
GC	21	33	54	65	8	3	26	1	0	27	0
HC	30	1	31	37	5	0	12	0	0	33	1
IC	66	14	80	88	10	1	24	41	0	84	5
JC	0	16	16	16	1	2	7	2	0	10	0
AB	2	3	5	7	6	1	10	0	1	16	0
BB	0	0	0	0	0	0	2	1	0	0	1
CB	0	1	1	3	2	0	5	0	1		0
DB	18	3	21	31	4	1	17	0	1	17	1
EB	11	1	12	23	3	0	20	0	0	15	1
FB	0	1	1	1	1	0	7	0	0	0	0
GB	0	4	4	5	1	0	4	0	0	0	0
HB	6	11	17	25	4	1	18	0	0	12	0
AH	2	3	5	5	1	0	6	0	0	0	0
BH	3	1	4	4	1	0	4	0	0	0	0
CH	8	12	20	28	3	3	12	0	0	11	1
DH	18	1	19	32	7	2	20	0	3	27	0
AP	4	0	4	7	5	0	10	0	0	15	0
BP	7	13	20	22	2	1	6	0	0	34	0
CP	3	0	3	3	4	0	9	0	0	47	5
DP	3	13	16	18	4	9	8	0	0	6	1
EP	3	15	18	18	4	3	6	0	0	2	0

Segment	fences	buildings	sign poles	overhead s	parking me	rocks	other	total	length	max grade	crest
AC	0	5	10	6	7	0	0	37	464	0.5	2.4
BC	5	14	14	1	25	0	0	78	1163	6.6	3.4
CC	1	5	9	3	0	0	0	25	296	3.2	3.4
DC	1	6	5	1	13	0	0	31	347	3.4	3.4
EC	5	17	12	3	0	0	0	50	943	7	3.4
FC	3	17	26	3	0	0	0	63	983	0.6	6
GC	16	43	57	5	5	0	0	191	3652	4	5.1
HC	1	45	24	0	1	0	0	122	1645	3	5
IC	7	84	80	2	0	0	0	338	5245	2.7	6.5
JC	0	8	21	3	22	0	0	76	728	0.2	5
AB	4	4	24	4	0	0	0	70	896	7.5	3.6
BB	3	0	0	3	0	0	0	10	226	0.2	0.7
CB	2	1	7	4	5	0	0	27	504	4.1	0.3
DB	12	30	29	1	0	0	0	113	2908	9.7	4.8
EB	13	11	34	2	0	0	0	99	2290	8.9	4.7
FB	2	1	14	1	0	0	0	26	539	3	2.4
GB	1	2	10	0	0	0	0	18	470	6.6	3.2
HB	3	9	48	2	0	0	0	97	2306	5.7	4
AH	2	2	6	4	0	0	0	21	508	10.9	6.8
BH	2	3	7	1	9	0	0	27	290	3.9	6.8
CH	6	30	33	2	8	0	1	110	1444	3.9	3
DH	5	31	30	1	5	1	0	132	2106	0.2	6.3
AP	3	4	13	3	0	0	0	53	1865	3.2	4.2
BP	0	10	21	3	0	0	1	78	1280	1.6	4.8
CP	2	0	10	3	0	2	0	82	1485	0.6	3.6
DP	0	13	19	3	20	0	1	84	987	2.3	3.2
EP	0	10	32	6	0	0	0	63	772	2.7	4.1

Segment	terrain type		residential	commercia	industrial	left lanes	right lanes	width	width	width	width
AC	level	3	20	80	0	2	2	0	16.5	12	12
BC	rolling	15	0	100	0	2	2	0	16.5	12	12
CC	rolling	3	0	100	0	2	2	0	16.5	12	12
DC	rolling	4	0	100	0	2	2	0	16.5	12	12
EC	rolling	6	40	55	5	2	2	0	16.5	12	12
FC	level	9	5	95	0	2	2	0	16.5	12	12
GC	rolling	33	25	75	0	1	1	0	0	20	20
HC	level	1	100	0	0	1	1	0	0	20	20
IC	level	14	95	5	0	1	1	0	0	20	20
JC	level	16	0	100	0	1	1	0	0	20	20
AB	rolling	3	0	100	0	3	2	10.6	11	10.6	10
BB	level	0	0	100	0	2	2	0	12	12	11
CB	rolling	1	0	100	0	1	2	0	0	12	11
DB	mountainou	3	80	20	0	1	1	0	0	16	22
EB	mountainou	1	80	20	0	2	2	0	0	11	12
FB	level	1	0	100	0	2	2	0	11.5	11.5	12
GB	rolling	4	0	100	0	3	4	11.5	11	11.5	11
HB	rolling	11	0	100	0	3	3	11.5	11	11.5	11
AH	mountainou	3	0	100	0	2	2	0	12	12	12
BH	rolling	1	0	100	0	2	2	0	12	12	12
CH	rolling	12	10	90	0	1	1	0	0	12.5	21.5
DH	level	1	75	25	0	1	1	0	0	16.5	16.5
AP	rolling	0	100	0	0	2	2	0	16	12	12
BP	level	13	15	85	0	2	2	0	16	12	12
CP	level	0	100	0	0	2	2	0	15	13	13
DP	level	13	0	100	0	2	2	0	15	13	13
EP	level	15	0	100	0	2	2	0	15	13	13

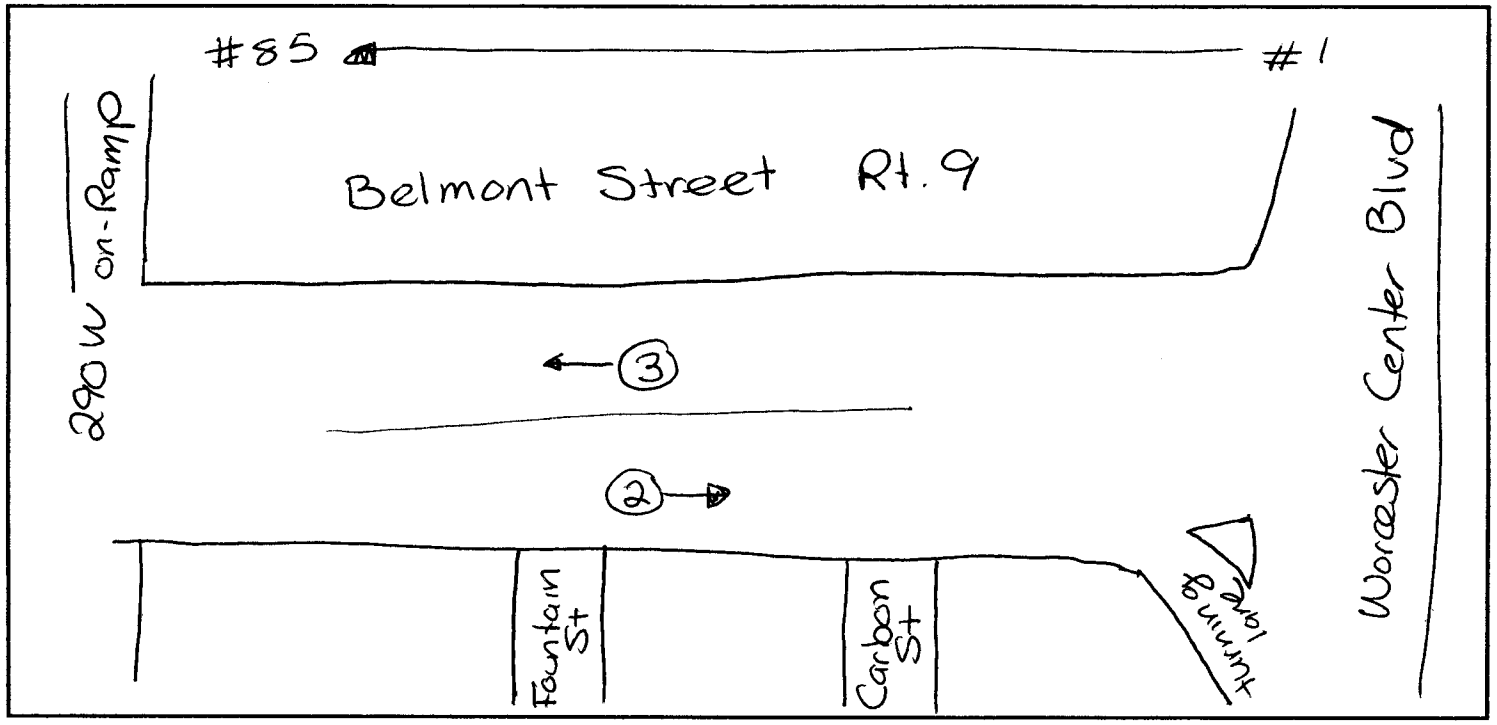
Segment	width	width	width	type should	width	width	sidewalk	width	width	curb	drainage
AC	16.5	0	0	none	0	0	yes	10.5	10.5	both	yes
BC	16.5	0	0	none	0	0	yes	11.5	11	both	yes
CC	16.5	0	0	none	0	0	yes	11	11.5	both	yes
DC	16.5	0	0	none	0	0	yes	11	10	both	yes
EC	16.5	0	0	none	0	0	yes	11.5	10.5	both	yes
FC	16.5	0	0	none	0	0	yes	11	11.5	both	yes
GC	0	0	0	none	0	0	yes	9.5	10.5	both	yes
HC	0	0	0	none	0	0	yes	9.5	10.5	both	yes
IC	0	0	0	none	0	0	yes	10	9.5 (1/2)	both	yes
JC	0	0	0	none	0	0	yes	10	10	both	yes
AB	10	0	0	none	0	0	yes	8	9	both	yes
BB	11	0	0	none	0	0	yes	6	5.5	both	yes
CB	11	0	0	none	0	0	yes	10.5	9	both	yes
DB	0	0	0	none	0	0	yes	8	8.5	both	yes
EB	0	0	0	paved	0	7	yes	6	6	both	yes
FB	12	0	0	none	0	0	yes	6	5	both	yes
GB	11	11	11	none	0	0	yes	5.5	7	both	yes
HB	11	11	0	none	0	0	yes	6	8	both	yes
AH	12	0	0	none	0	0	yes	9	9	both	yes
BH	12	0	0	none	0	0	yes	6	5	both	yes
CH	0	0	0	none	0	0	yes	8	10	both	yes
DH	0	0	0	none	0	0	yes	8	7	both	yes
AP	16	0	0	none	0	0	yes	9.75	12.5	both	yes
BP	16	0	0	none	0	0	yes	9.75	11.5	both	yes
CP	13	0	0	none	0	0	yes	5	5	both	yes
DP	13	0	0	none	0	0	yes	12	11	both	yes
EP	13	0	0	none	0	0	yes	10	12	both	yes

Segment	pavement	comments	markings	comments	% parking	% perpend	% lighting	SD	comment	curvature
AC	fair	minor rutting & cracking	fair	starting to f	20	0	100	no		1
BC	fair	minor rutting, cracking	fair	starting to f	75	0	100	no		1
CC	fair	minor rutting & cracking	good		0	0	100	no		0
DC	fair	minor rutting & cracking	good		100	0	100	no		0
EC	fair	minor rutting & cracking	good		80	0	100	yes	can't see s	1
FC	fair	patching	fair	starting to f	80	0	100	no		1
GC	fair	patching	good		80	0	100	no		3
HC	fair	lots of cracking	fair	starting to f	100	0	100	no		0
IC	good/fair	few cracks	fair	starting to f	80	10	100	no		4
JC	fair	cracking	good/fair	starting to f	100	0	100	no		1
AB	good	a few patches	bad	mostly gon	0	0	100	no		0
BB	fair	rutting patching	fair	starting to f	0	0	100	no		0
CB	fair	rutting, few patches	bad	fading	40	0	100	no		0
DB	fair	cracking, rutting	bad	very faded,	60	0	100	yes	can't see c	2
EB	fair	rutting, some patching	fair	fading	30	0	80	no		0
FB	fair	lots of rutting	fair	fading	0	0	100	no		0
GB	fair	few cracks, deep rutting	good		0	0	100	no		0
HB	fair	rutting	fair	fading	0	0	100	no		0
AH	fair	lots of rutting	bad	most very f	0	0	100	yes	hard to see	0
BH	fair	some rutting	bad	most very f	0	0	100	no		0
CH	fair	some rutting	fair	fading	50	0	100	no		0
DH	good/fair	some rutting	fair	fading	50	0	100	no		0
AP	fair	rutting	fair	starting to f	0	0	100	no		0
BP	fair	minor rutting	fair	starting to f	0	0	100	no		0
CP	fair		fair		0	0	100	no		1
DP	fair	some rutting, cracking	fair	starting to f	80	0	100	no		0
EP	fair	minor cracks	fair	starting to f	0	0	100	no		0

Segment	median type	width
AC	none	0
BC	none	0
CC	none	0
DC	none	0
EC	none	0
FC	none	0
GC	none	0
HC	none	0
IC	none	0
JC	none	0
AB	raised	5.5
BB	none	0
CB	none	0
DB	none	0
EB	none	0
FB	raised	6
GB	raised	6
HB	raised	8
AH	none	0
BH	none	0
CH	none	0
DH	none	0
AP	none	0
BP	none	0
CP	none	0
DP	none	0
EP	none	0

Segment AB

Date: 9-21-03 Weather: Sunny



Posted Speed: 25

Minor Access Points (Road Names)

2

of driveways 11 = 2

# of parking lots 111 = 3

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	1 = 6	1	= 10	1	= 16
	monument	fences	buildings	sign poles	overhead sign
	0	11 = 2 wall 11 = 2	1111 = 4	= 24	1111 = 4
	parking meter	rock			
	0	0			

Section Length: 896 ft

Vertical Grade: 7.5 % Crest on road: 3.6 %

Terrain Type: level (rolling) mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 3 Going Right: 2

width of lanes: 10', 10' 10.5', 11' 10.5'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 8' 9'

curb present: (both) no

drainage present: (yes) no

pavement quality: (good) fair bad
describe: a few patches

pavement marking quality good fair (bad)
describe: mostly gone

parking allowed yes (no) _____ % allowed

road lighting: (present) not 100 %

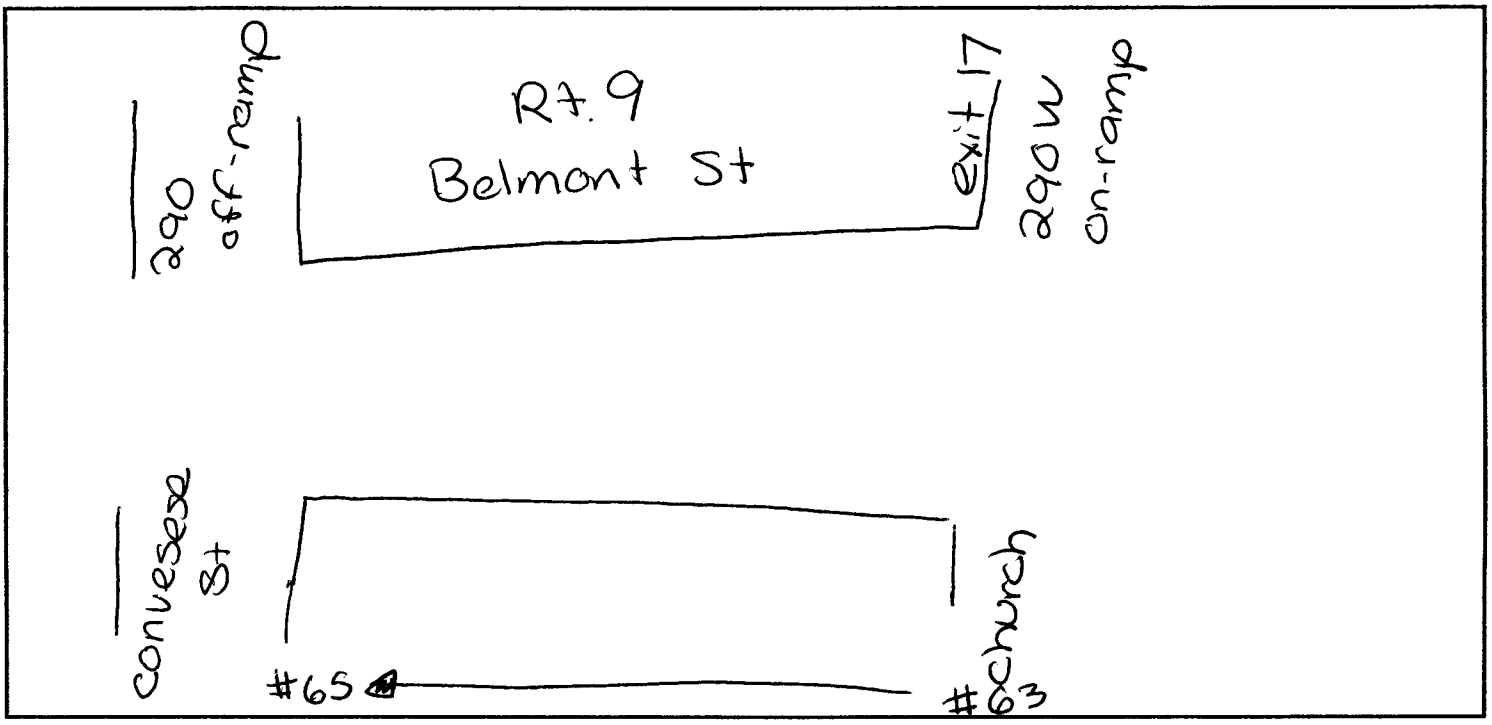
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass (raised paved w/ curb) painted other _____
width 5 ft 2 in

segment BB

Date: 9-21-03 Weather: Sunny



Posted Speed: 25

Minor Access Points (Road Names)

0

of driveways 0

# of parking lots 0

1

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	0	0	11 = 2	0	0
	monument	fences	buildings	sign poles	overhead sign
	1	III = 3	0	0	III = 3
	parking meter	rock			
	0	0			

Section Length: 226 ft

Vertical Grade: 0.2 % Crest on road: 0.7 %

Terrain Type: (level) rolling mountainous

Land Use % residential ____ commerical 100 % industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: 12', 12' 11', 11'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 6' 5.5'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good (fair) bad
describe: rutting, patching

pavement marking quality good (fair) bad
describe: starting to fade

parking allowed yes (no) _____ % allowed

road lighting: (present) not 100 %

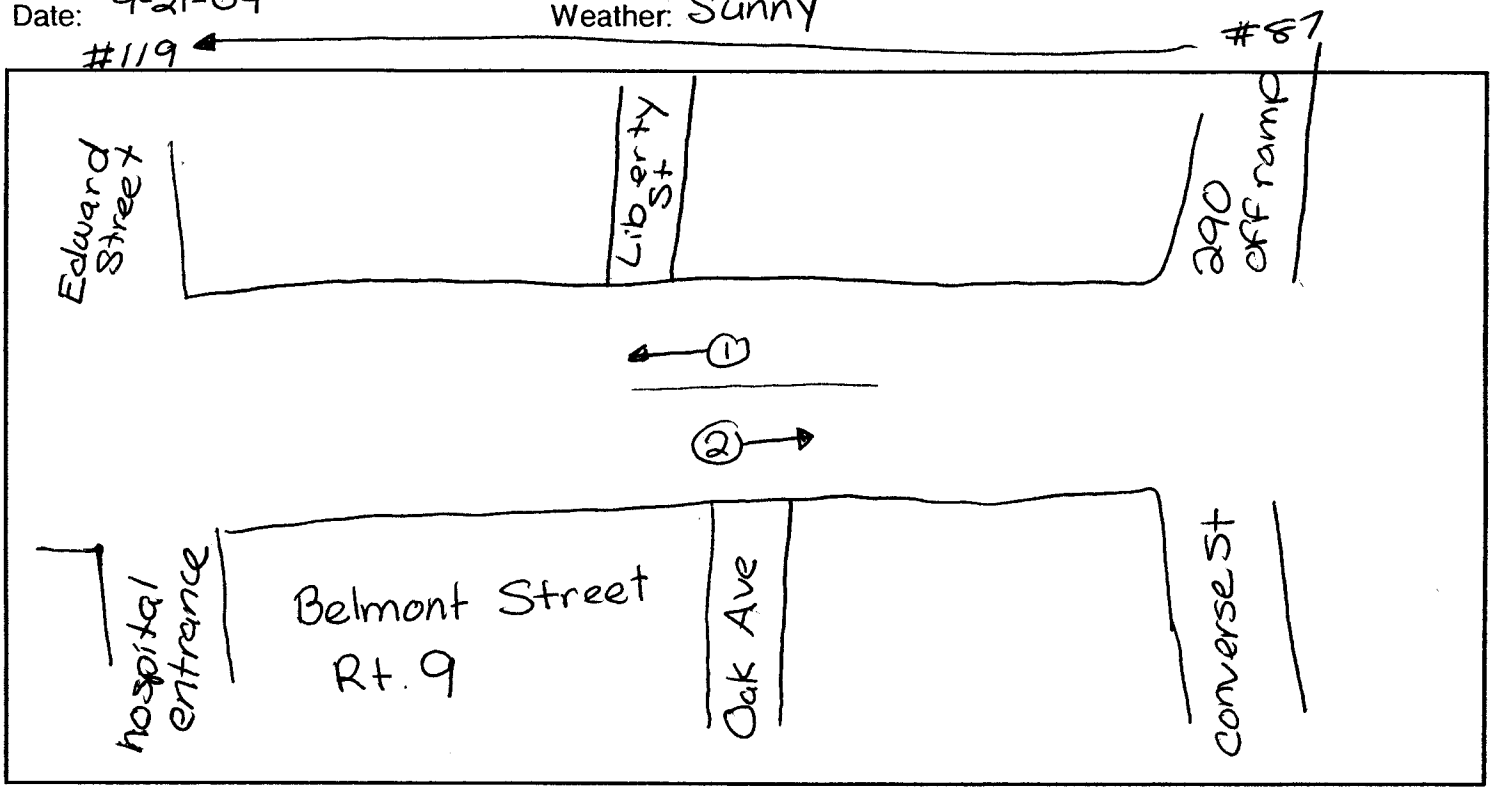
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

segment CB

Date: 9-21-04 Weather: Sunny



Posted Speed: 25

Minor Access Points (Road Names)

2

of driveways 0

# of parking lots II = 2

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	II = 2	0	IIII = 5	1	0
	monument	fences	buildings	sign poles	overhead sign
	0	II = 2	1	III = 7	IIII = 4
	parking meter	rock			
	IIII = 5	0			

Section Length: 503.5 ft

Vertical Grade: 4.1 % Crest on road: 0.3 %

Terrain Type: level (rolling) mountainous

Land Use % residential ____ commerical 100 % industrial ____

lanes: Going Left: 1 Going Right: 2

width of lanes: 11', 11' 12'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 10'5" 10'6"

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good (fair) bad
describe: rutting, few patches

pavement marking quality good fair (bad)
describe: fading

parking allowed (yes) no 40 % allowed

road lighting: (present) not 100 %

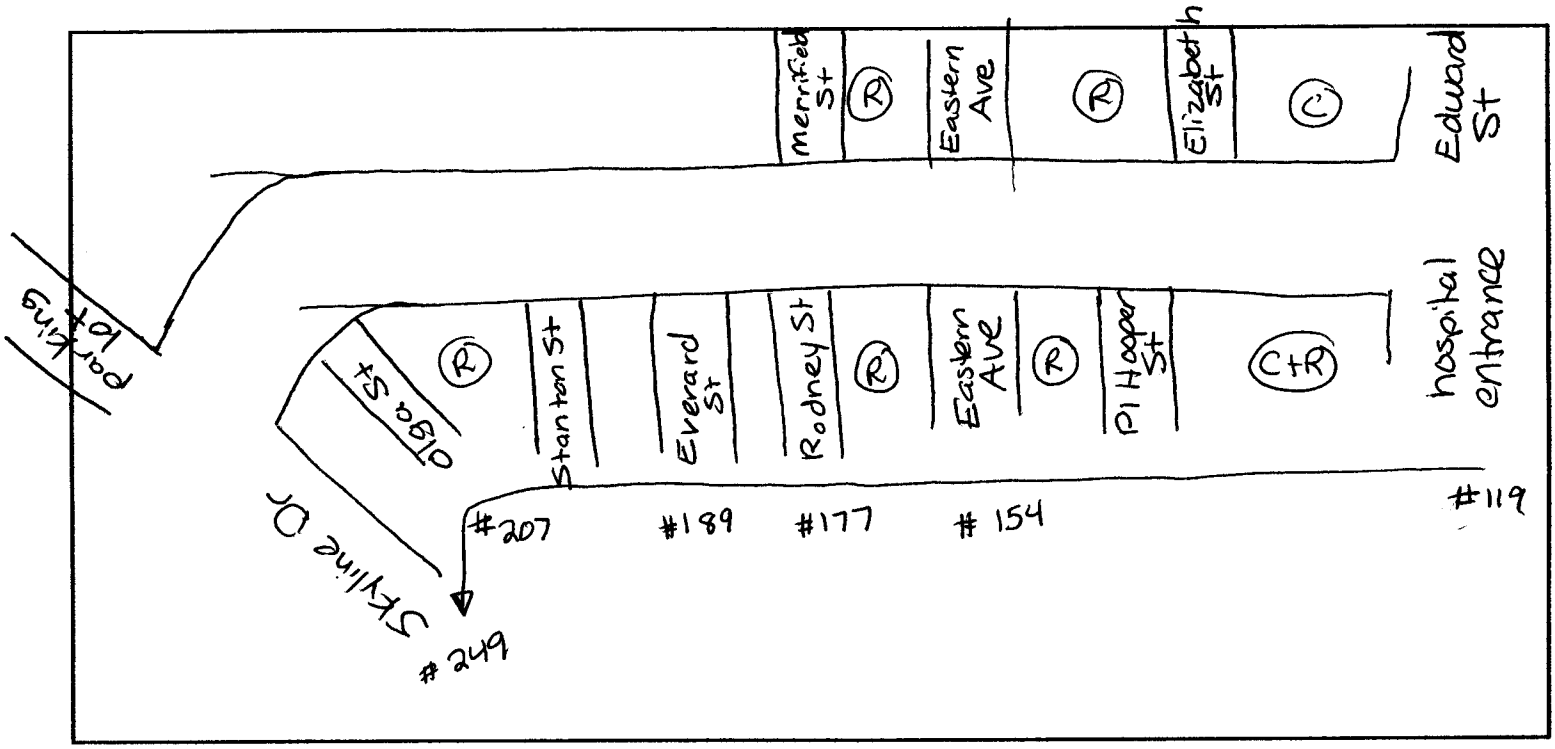
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

Segment DB

Date: 9-21-03 Weather: sunny



Posted Speed: 25

Minor Access Points (Road Names)

11

of driveways 𝍸 𝍸 𝍸 III = 18

# of parking lots III = 3

roadside hazards:	firehydrants 𝍸 = 4	mailboxes 1	utility/light poles 𝍸 II 𝍸 𝍸 = 17	benches 1	trees 𝍸 II 𝍸 𝍸 = 17
	monument 1	fences 𝍸 𝍸 II = 12	buildings 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 = 30	sign poles 𝍸 𝍸 IIII 𝍸 𝍸 𝍸 = 29	overhead sign 1
	parking meter 0	rock 0			

Section Length: 2907 ft 3"

Vertical Grade: 9.7 % Crest on road: 4.8 %

Terrain Type: level (rolling) mountainous

Land Use % residential 80% commerical 20% industrial

lanes: Going Left: 1 Going Right: 1

width of lanes: 16' 22'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 8' 9'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good (fair) bad
describe: cracking, rutting

pavement marking quality good fair (bad)
describe: very faded, can't tell lanes

parking allowed (yes) no 60 % allowed

road lighting: (present) not 100 %

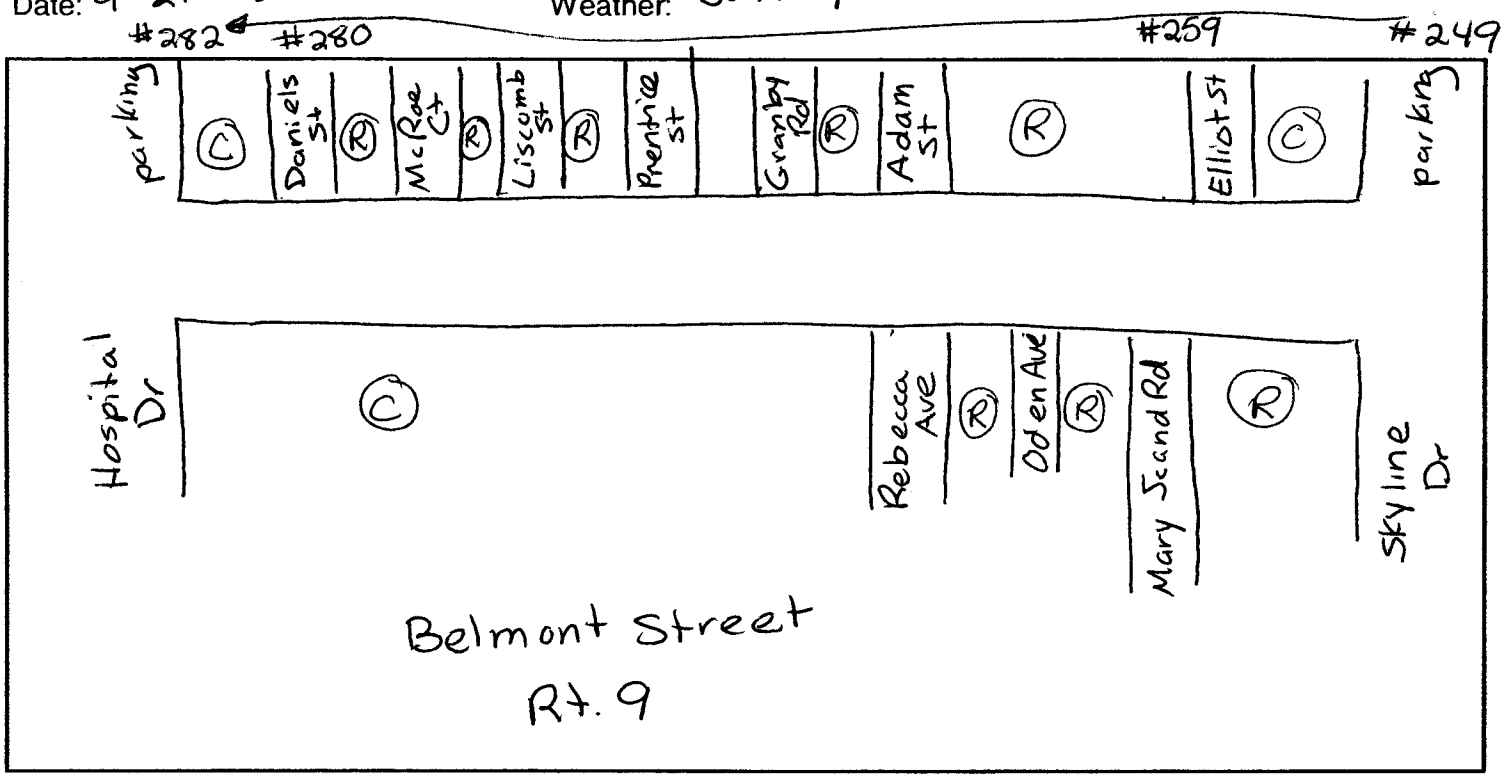
sight distance issues: no (yes)
describe:
curve → ≈200'

horizontal curvature describe: straight 2 (curve)
approximate curve length: 1: 494'
radius: 2: 427'

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

Segment EB

Date: 9-21-03 Weather: sunny



Posted Speed:

Minor Access Points (Road Names)

12

of driveways 𝍸 𝍸 I = 11

# of parking lots 1

roadside hazards:

firehydrants	mailboxes	utility/light poles	benches	trees	
III = 3	0	𝍸 𝍸 = 20		𝍸 = 15	
monument	fences	buildings	sign poles	overhead sign	
1	𝍸 III = 13	𝍸 I = 11	𝍸 𝍸 𝍸 𝍸 = 34	II = 2	
parking meter	rock				
0	0				

Section Length: 2290 ft

Vertical Grade: 8.2 / 8.9 %

Crest on road: 4.7 %

Terrain Type: level rolling (mountainous)

Land Use % residential 80% commerical 20% industrial

lanes: Going Left: 2 Going Right: 2

width of lanes: 12' 12' 11' 11'

type of shoulder: (paved) dirt none

width of shoulder: 6' 6'9"

sidewalk present: (yes) no width: 6' 6'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good (fair) bad
describe: rutting, some patching

pavement marking quality good (fair) bad
describe: fading

parking allowed (yes) no 30 % allowed

road lighting: (present) not 80 %

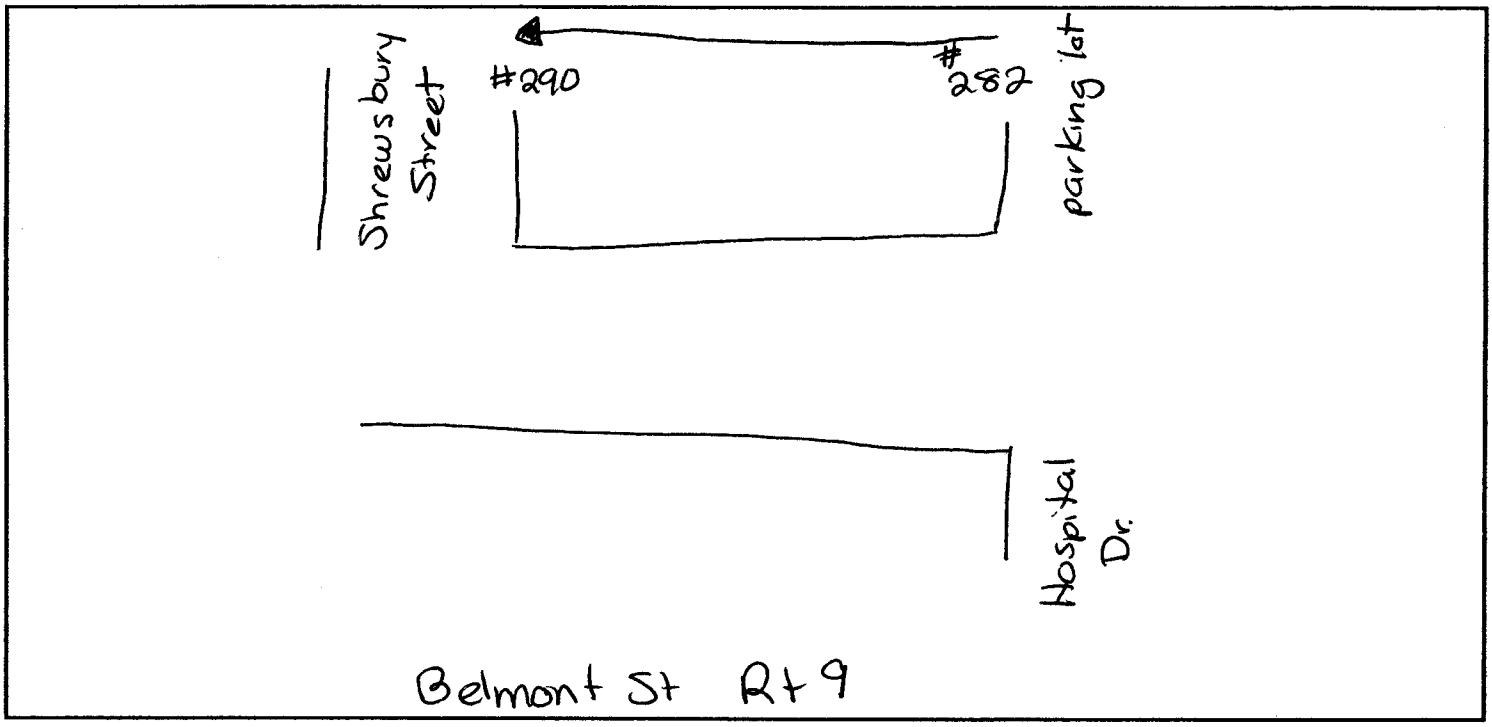
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

segment FB

Date: 9-21-03 Weather: sunny



Posted Speed:
Minor Access Points (Road Names)
0

of driveways 0
# of parking lots 1

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	1	0	𝍸 11 =7	0	0
	monument	fences	buildings	sign poles	overhead sign
	0	11 =2	1	𝍸 𝍸 1111 =14	1
	parking meter	rock			
	0	0			

Section Length: 539 ft

Vertical Grade: 3 % Crest on road: 2.4 %

Terrain Type: (level) rolling mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: 11.5' 11.5' 12' 12'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 6' 5'

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: lots of rutting

pavement marking quality good (fair) bad
describe: fading

parking allowed yes (no) _____ % allowed

road lighting: (present) not 100 %

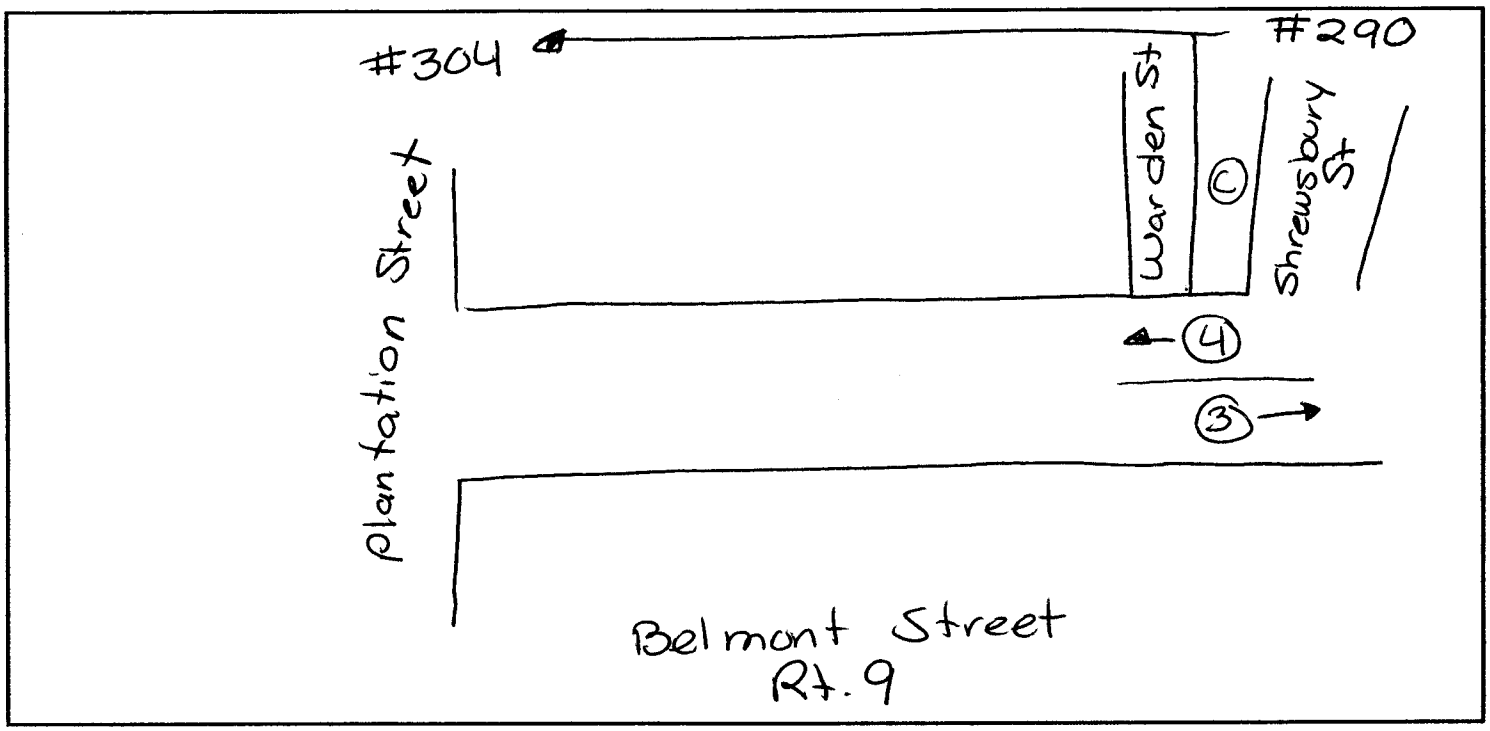
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass (raised paved w/ curb) painted other _____
width _____ ft _____ in
3' 10"
6' ← main section

segment GB

Date: 9-21-03 Weather: Sunny



Posted Speed:
Minor Access Points (Road Names)
1

of driveways 0
# of parking lots |||| =4

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	1	0	|||| =4	0	0

	monument	fences	buildings	sign poles	overhead sign
	0	1	|| =2	##### ##### =10	0

	parking meter	rock
	0	0

Section Length: 470 ft

Vertical Grade: 6.6 % Crest on road: 3.2 %

Terrain Type: level (rolling) mountainous

Land Use % residential _____ commerical 100% industrial _____

lanes: Going Left: 3 Going Right: 4

width of lanes: 11.5', 11, 11.5' 11, 11, 11, 11

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 5'4" 7'

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: few cracks deep rutting

pavement marking quality (good) fair bad
describe:

parking allowed yes (no) _____ % allowed

road lighting: (present) not 100 %

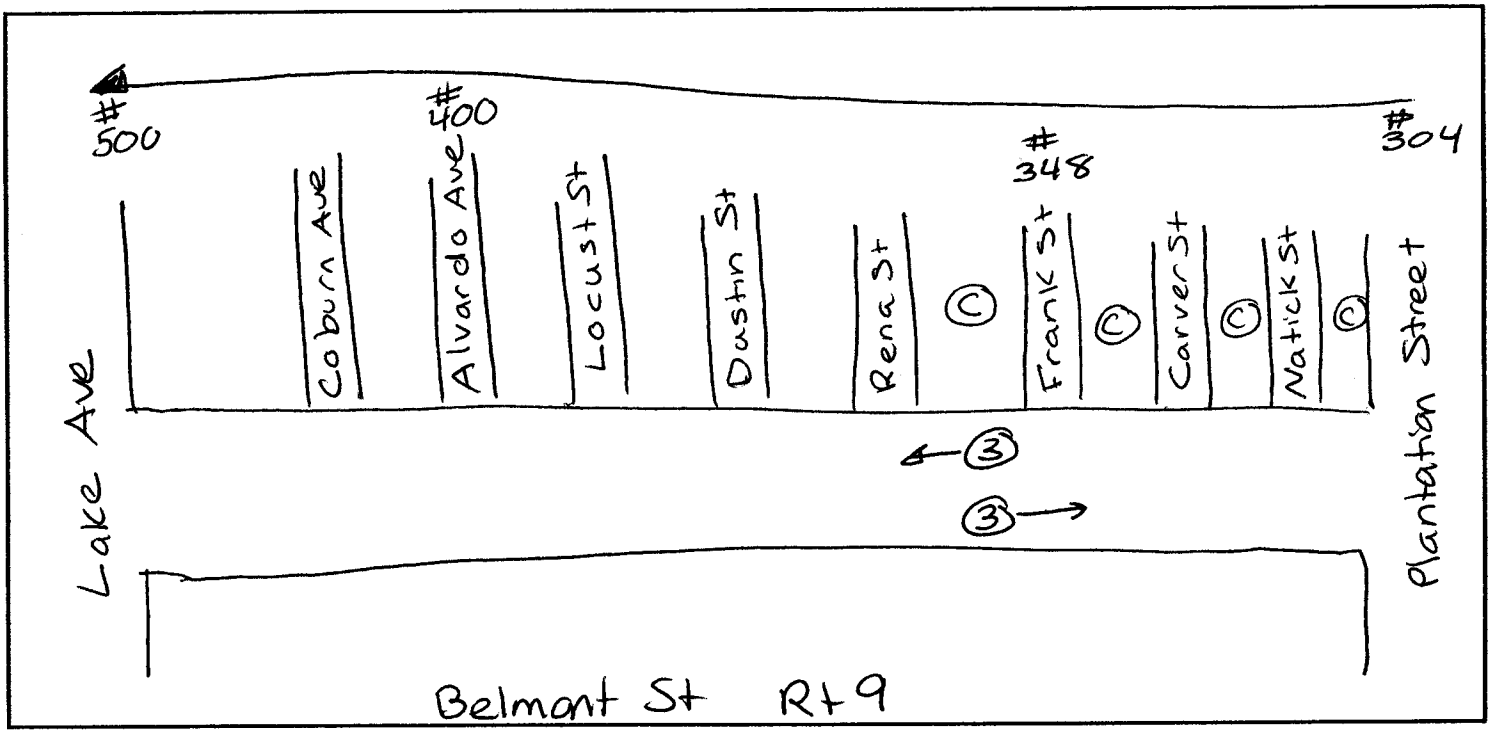
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass (raised paved w/ curb) painted other _____
width 6 ft 0 in

segment HB

Date: 9-21-03 Weather: Sunny



Posted Speed:

Minor Access Points (Road Names)

8

of driveways 𝍸 1 = 6

# of parking lots 𝍸 𝍸 1 = 11

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 4	1	𝍸 𝍸 𝍸 = 18		𝍸 11 = 12
	monument	fences	buildings	sign poles	overhead sign
	0	𝍸 = 3	𝍸 = 9	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 = 48	11 = 2
	parking meter	rock			
	0	0			

Section Length: 2306 ft

Vertical Grade: 5.2, 5.7 %

Crest on road: 4 %

Terrain Type: level (rolling) mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 3 Going Right: 3

width of lanes: 11.5' 11.5' 11' | 11' 11' 11'

type of shoulder: paved | dirt | (none)

width of shoulder: 0' | 0'

sidewalk present: (yes) | no | width: 6' | 8'

curb present: (both) | no

drainage present: (yes) | no

pavement quality: good | (fair) | bad
describe: rutting

pavement marking quality: good | (fair) | bad
describe: fading

parking allowed: yes | (no) | % allowed

road lighting: (present) | not | 100 %

sight distance issues: (no) | yes
describe:

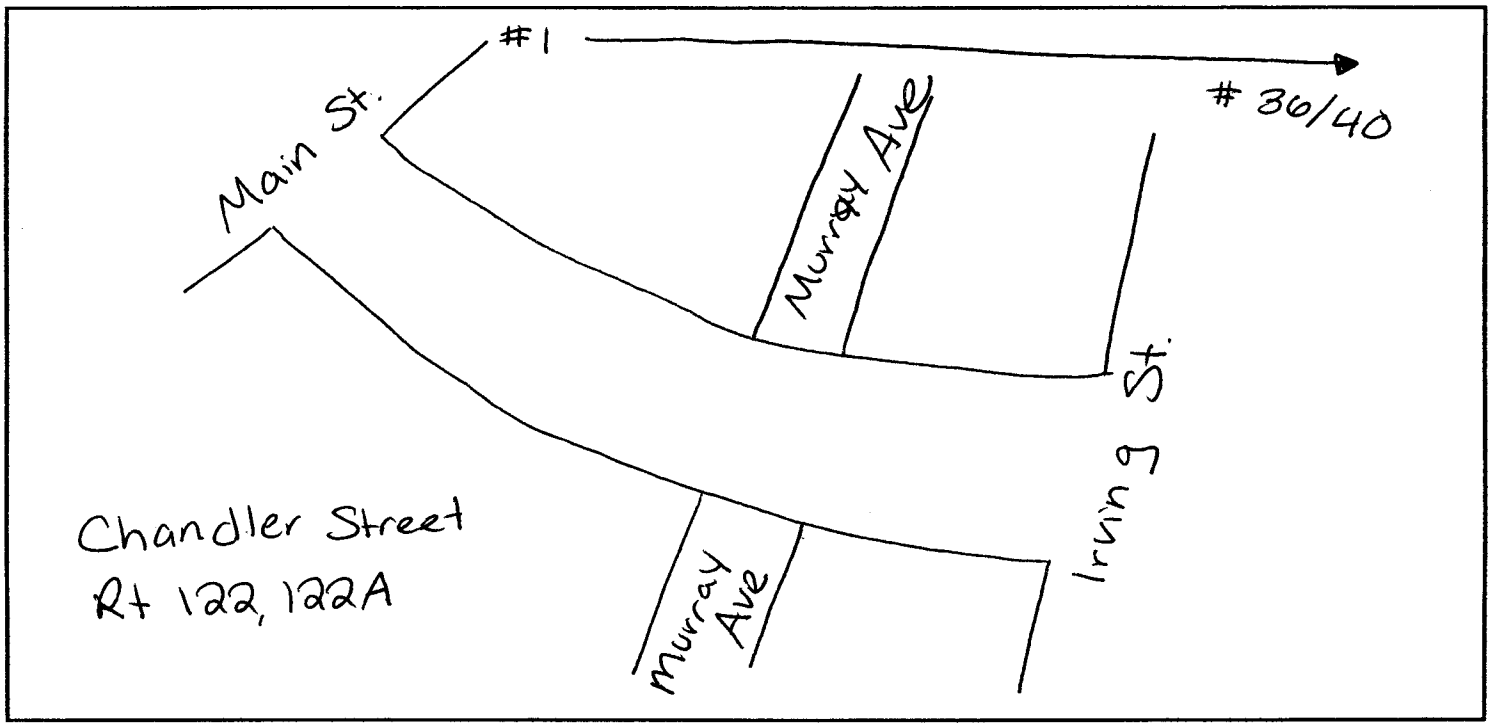
horizontal curvature describe: (straight) | curve
approximate curve length:
radius:

median type: grass | (raised paved w/ curb) | painted | other
width ft in

4' 5"
7' 10" ← main section

segment AC

Date: 9-27-03 Weather: misty



Posted Speed:
Minor Access Points (Road Names)
2

of driveways 0
# of parking lots III = 3

roadside hazards:

firehydrants	mailboxes	utility/light poles	benches	trees
IIII = 4	0	IIII = 5	0	0

monument	fences	buildings	sign poles	overhead sign
0	0	IIII = 5	IIII IIII = 10	IIII I = 6

parking meter	rock
IIII II = 7	0

Section Length: 464 ft

Vertical Grade: 0.5 % Crest on road: 2.4 %

Terrain Type: (level) rolling mountainous

Land Use % residential 20% commerical 80% industrial

lanes: Going Left: 2 Going Right: 2

width of lanes: 16.5' 12' | 12' 16.5'

type of shoulder: paved | dirt | (none)

width of shoulder: 0' | 0'

sidewalk present: (yes) | no | width: 10.5' | 10.5'

curb present: (both) | no

drainage present: (yes) | no

pavement quality: good | (fair) | bad
describe: minor rutting + cracking

pavement marking quality: good | (fair) | bad
describe: starting to fade

parking allowed: (yes) | no | 20 % allowed

road lighting: (present) | not | 100 %

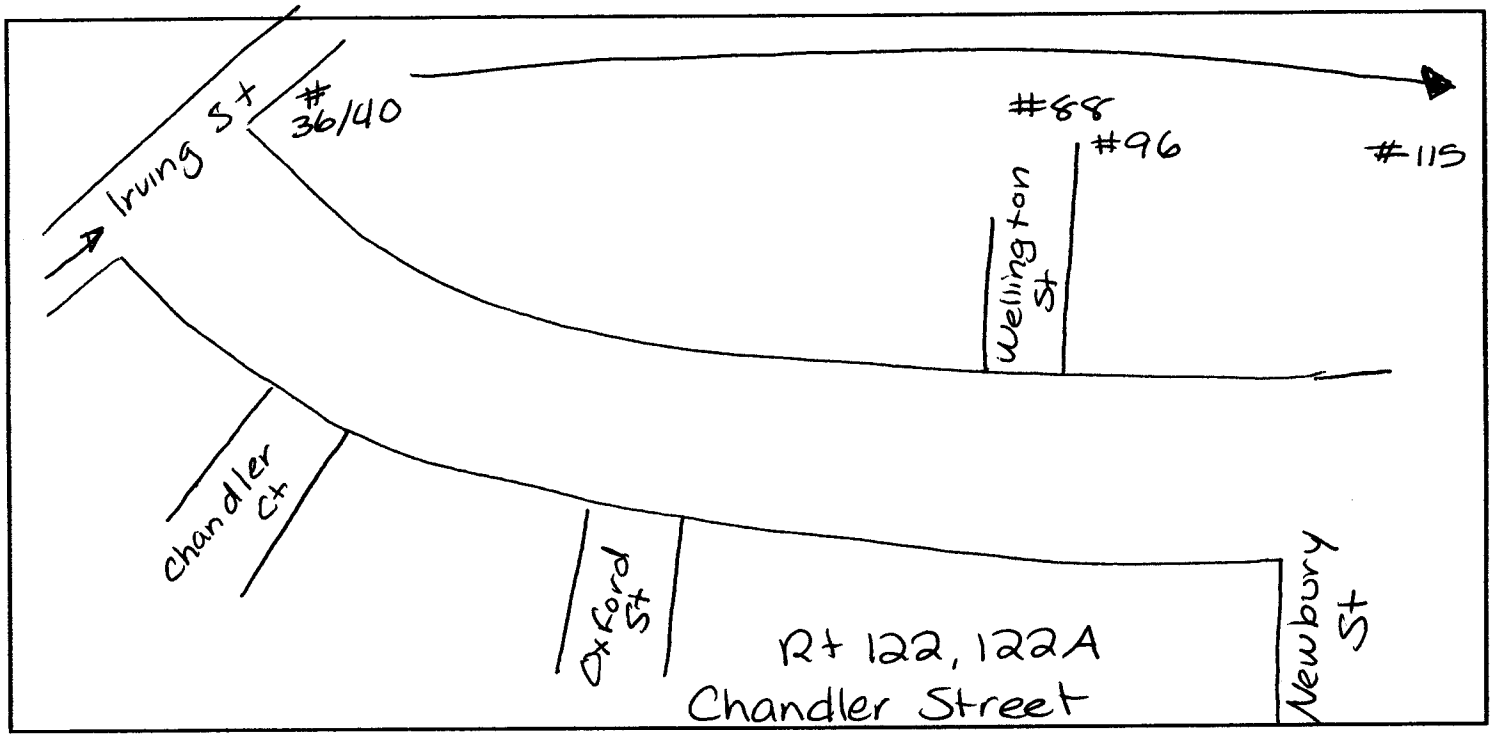
sight distance issues: (no) | yes
describe:

horizontal curvature describe: straight | (curve)
approximate curve length:
radius: 464'

median type: grass | paved w/ curb | painted | other (none)
width ____ ft ____ in

segment 13C

Date: 9-27-03 Weather: misty



Posted Speed:

Minor Access Points (Road Names)

3

of driveways || = 2

# of parking lots 𝍸 𝍸 𝍸 = 15

roadside hazards:

firehydrants	mailboxes	utility/light poles	benches	trees
||| = 3	|||| = 4	𝍸 || = 12	0	0

monument	fences	buildings	sign poles	overhead sign
0	𝍸 = 5	𝍸 𝍸 |||| = 14	𝍸 𝍸 |||| = 14	1

parking meter	rock
𝍸 𝍸 𝍸 𝍸 𝍸 = 25	0

Section Length: 1163 ft

Vertical Grade: 3.0, 3.9, 6.0 % Crest on road: 3.4 %

Terrain Type: level (rolling) mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: 16.5, 12' | 12', 16.5'

type of shoulder: paved | dirt | (none)

width of shoulder: _____ | _____

sidewalk present: (yes) no width: 11.5' 11'1"

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good (fair) bad
describe: minor rutting, cracking

pavement marking quality good (fair) bad
describe: starting to fade

parking allowed (yes) no 75 % allowed

road lighting: (present) not 100 %

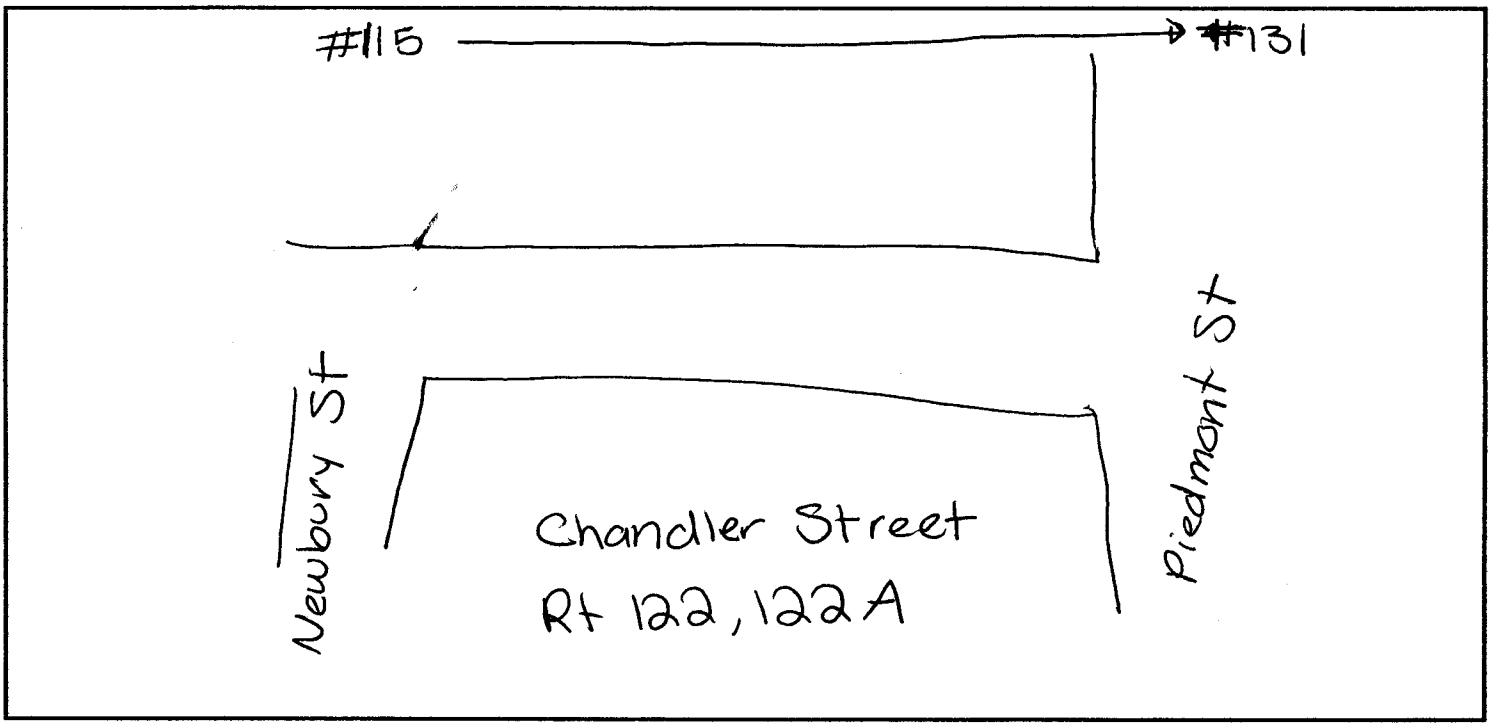
sight distance issues: (no) yes
describe:

horizontal curvature describe: straight (curve)
approximate curve length:
radius: 880 ft

median type: grass | paved w/ curb | painted | other (none)
width _____ ft _____ in

Segment CC

Date: 9-27-03 Weather: misty



Posted Speed:
Minor Access Points (Road Names)
0

of driveways || = 2
# of parking lots ||| = 3

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	1	||=2	|||| = 4	0	0
	monument	fences	buildings	sign poles	overhead sign
	0	1	𝍸 = 5	𝍸 𝍸 = 9	||| = 3
	parking meter	rock			
	0	0			

Section Length: 296 ft

Vertical Grade: 3.2 % Crest on road: 3.4 %

Terrain Type: level (rolling) mountainous

Land Use % residential ____ commerical 100 % industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: 16.5, 12 | 12, 16.5

type of shoulder: paved | dirt | (none)

width of shoulder: 0' | 0'

sidewalk present: (yes) | no | width: 11' | 11.5'

curb present: (both) | no

drainage present: (yes) | no

pavement quality: good | (fair) | bad
describe: minor rutting + cracking

pavement marking quality (good) | fair | bad
describe:

parking allowed: yes | (no) | ____ % allowed

road lighting: (present) | not | 100 %

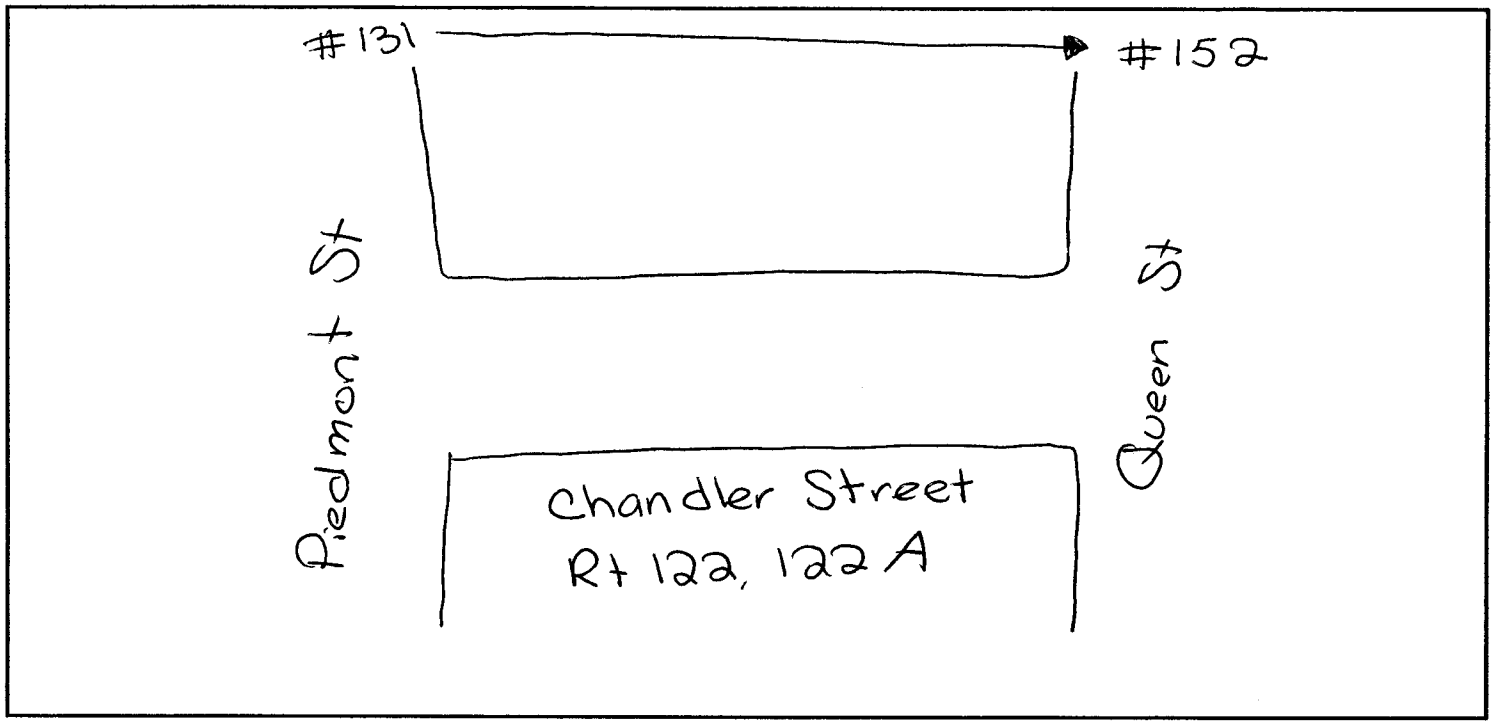
sight distance issues: (no) | yes
describe:

horizontal curvature describe: (straight) | curve
approximate curve length:
radius:

median type: grass | paved w/ curb | painted | other (none)
width ____ ft ____ in

Segment DC

Date: 9-27-03 Weather: misty



Posted Speed:
Minor Access Points (Road Names)
0

of driveways 1
# of parking lots 1111 =4

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	0	1	1111 =4	0	0
	monument	fences	buildings	sign poles	overhead sign
	0	1	HHT 1 =6	HHT =5	1
	parking meter	rock			
	HHT HHT 111 =13	0			

Section Length: 346 ft 10"

Vertical Grade: 3.4 % Crest on road: 3.4 %

Terrain Type: level (rolling) mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: 16.5' 12' 12' 16.5'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 11' 10'

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: minor cracks + ruts

pavement marking quality (good) fair bad
describe:

parking allowed (yes) no 100 % allowed

road lighting: (present) not 100 %

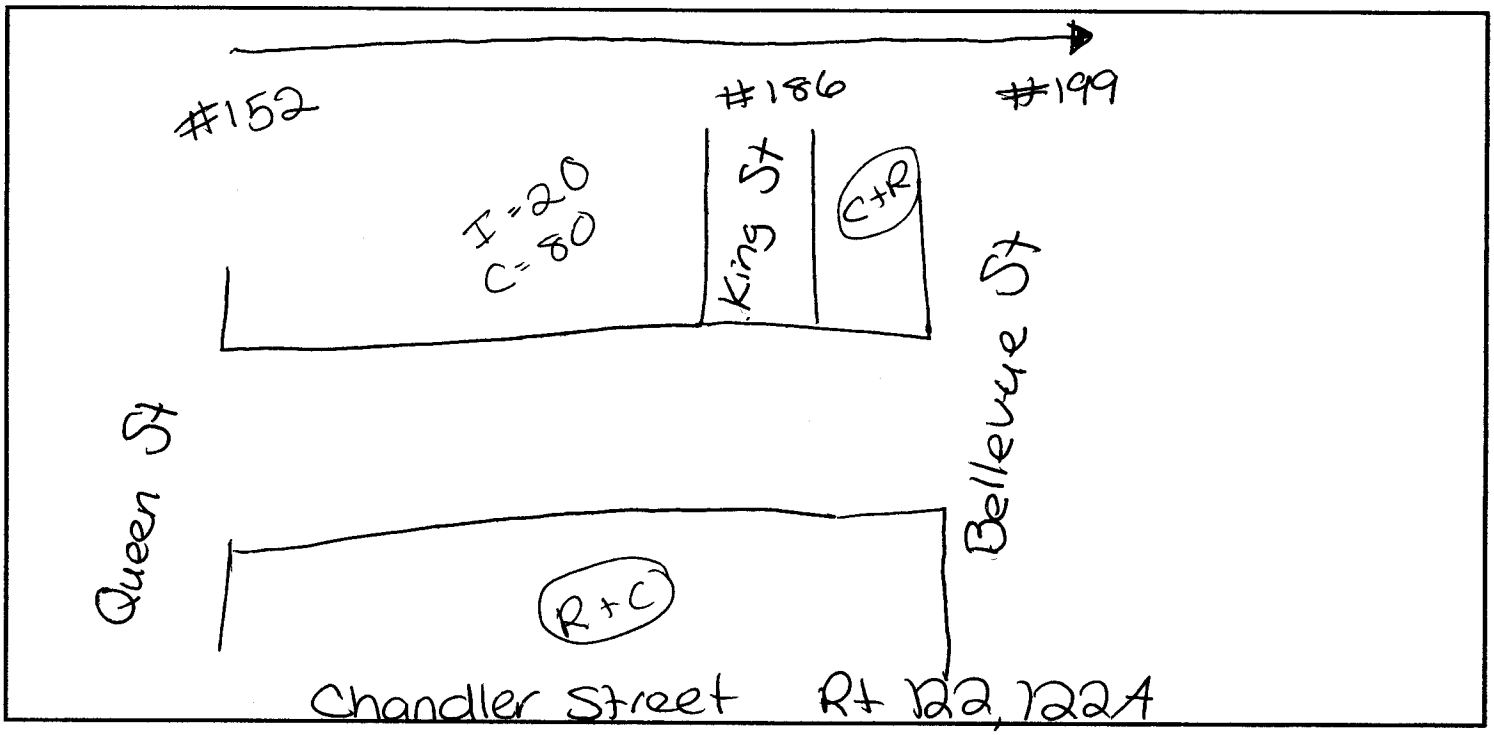
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width ft in

Segment EC

Date: 9-27-03 Weather: misty



Posted Speed:
Minor Access Points (Road Names)
1

of driveways 𝍸 = 5
# of parking lots 𝍸 1 = 6

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	II = 2	0	𝍸 III = 8	0	III = 3
	monument	fences	buildings	sign poles	overhead sign
	0	𝍸 = 5	𝍸 𝍸 = 17	𝍸 II = 12	III = 3
	parking meter	rock			
	0	0			

Section Length: 942.5 ft

Vertical Grade: 7.0 % Crest on road: 3.4 %

Terrain Type: level (rolling) mountainous

Land Use % residential 40% commerical 55% industrial 5%

lanes: Going Left: 2 Going Right: 2

width of lanes: 16.5', 12' | 12', 16.5'

type of shoulder: paved | dirt | (none)

width of shoulder: 0' | 0'

sidewalk present: (yes) | no | width: 11' 4" | 10.5'

curb present: (both) | no

drainage present: (yes) | no

pavement quality: good | (fair) | bad
describe: minor rutting, minor cracks

pavement marking quality (good) | fair | bad
describe:

parking allowed (yes) | no | 80 % allowed

road lighting: (present) | not | 100 %

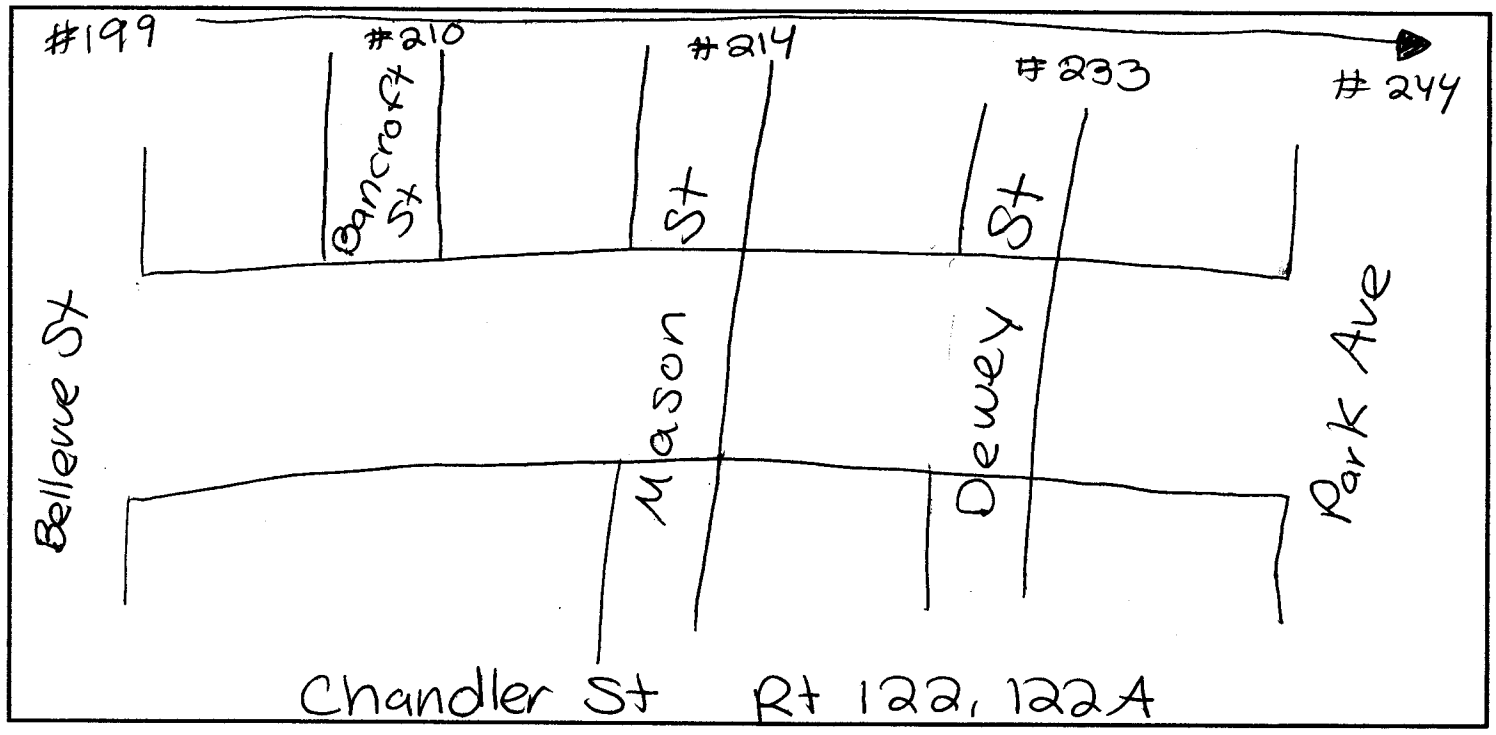
sight distance issues: no | (yes)
describe: can't see signal - sign to tell you signal color

horizontal curvature describe: straight | (curve)
approximate curve length:
radius: ~353

median type: grass | paved w/ curb | painted | other (none)
width ____ ft ____ in

Segment FC

Date: 9-27-03 Weather: misty



Posted Speed: 30mph

Minor Access Points (Road Names)

5

of driveways ||||- = 5

# of parking lots ||||- |||| = 9

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	1	1	||||- ||| = 8	0	|||| = 4

	monument	fences	buildings	sign poles	overhead sign
	0	||| = 3	||||- ||||- ||| = 17	||||- ||||- ||||-	= 26	||| = 3

parking meter	rock
0	0

Section Length: 982 ft 8"

Vertical Grade: 0.6 %

Crest on road: 6.0 %

Terrain Type: (level) rolling mountainous

Land Use % residential 5% commerical 95% industrial _____

lanes: Going Left: 2 Going Right: 2

width of lanes: 16.5', 12' 12', 16.5'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 10' 11" 11' 6"

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: patching

pavement marking quality good (fair) bad
describe: starting to fade

parking allowed (yes) no 80 % allowed

road lighting: (present) not 100 %

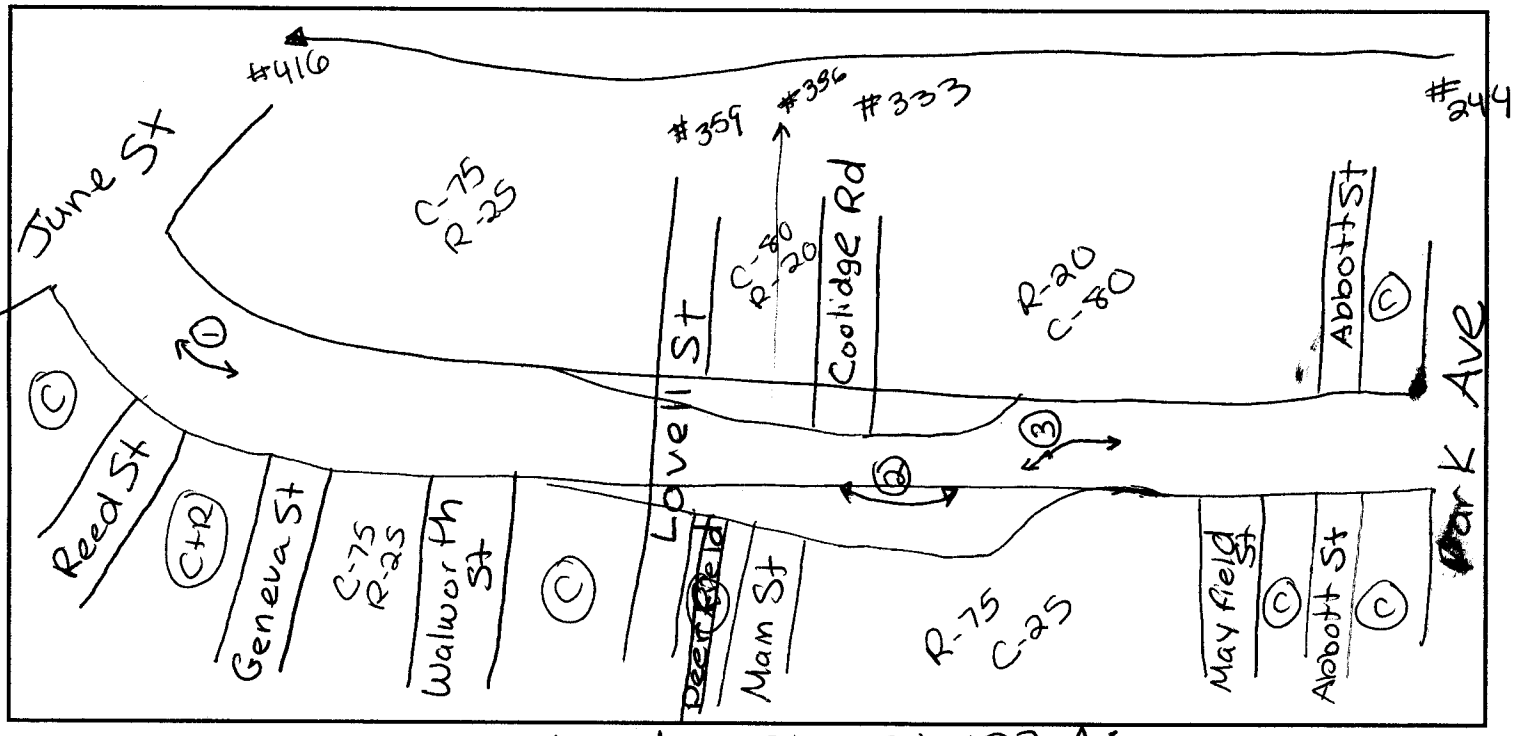
sight distance issues: (no) yes
describe:

horizontal curvature describe: straight (curve)
approximate curve length:
radius: ≈ 200'

median type: grass paved w/ curb painted other (none)
width ft in

Segment GC

Date: 9-27-03 Weather: misty



Posted Speed:

Minor Access Points (Road Names)

11

of driveways 𝍸 𝍸 𝍸 𝍸 / = 21

# of parking lots 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 III = 33

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 III = 8	III = 3	𝍸 𝍸 𝍸 𝍸 𝍸 = 26	0	𝍸 𝍸 II = 27
	monument	fences	buildings	sign poles	overhead sign
	0	𝍸 𝍸 𝍸 = 16	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 III = 43	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 II = 57	𝍸 = 5
	parking meter	rock			
	𝍸 = 5	0			

Section Length: 3651 ft 2"

Vertical Grade: 4.0 % Crest on road: 5.1 %

Terrain Type: level (rolling) mountainous

Land Use % residential 25% commerical 75% industrial

lanes: Going Left: 1 Going Right: 1

width of lanes: 20' 20'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 9'4" 10'4"

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: patching

pavement marking quality (good) fair bad
describe:

parking allowed (yes) no 80 % allowed

road lighting: (present) not 100 %

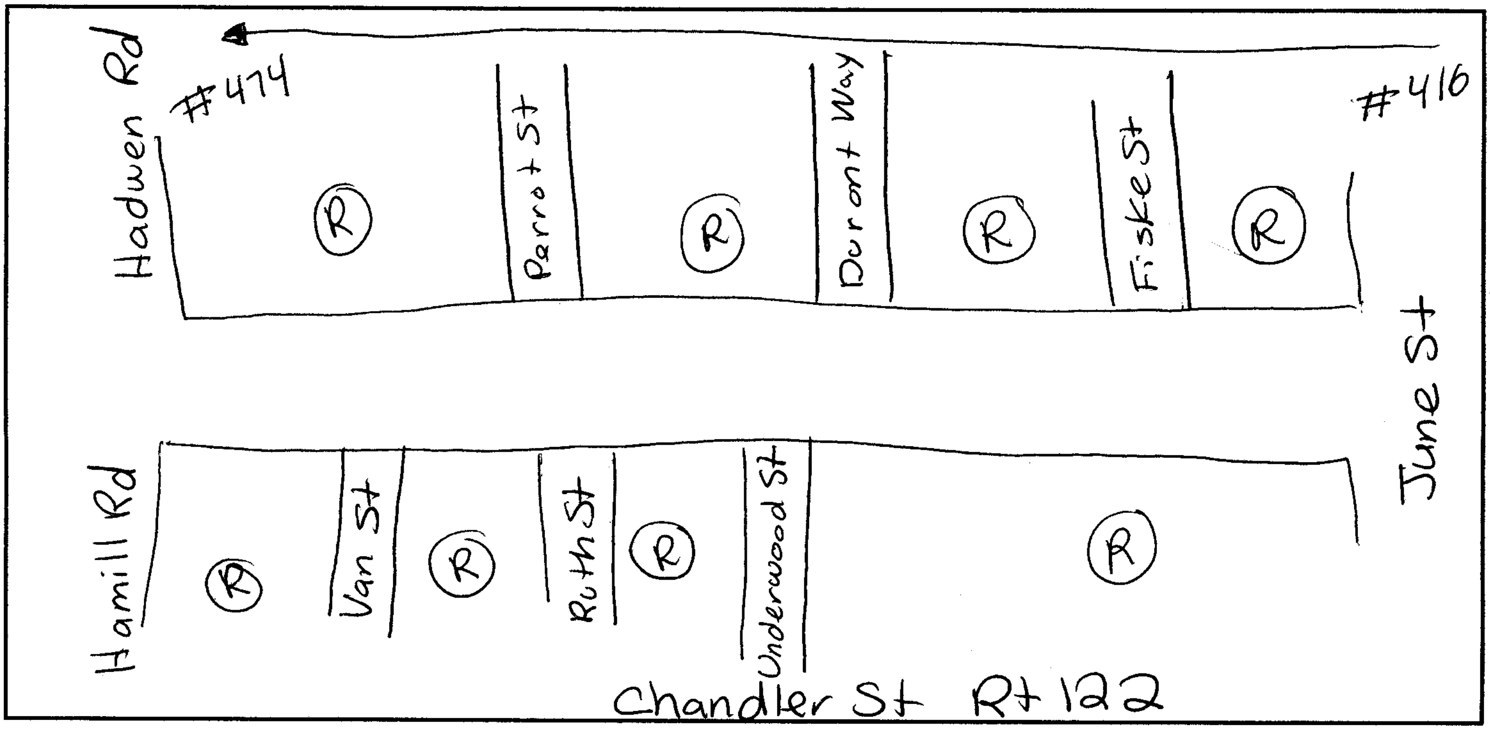
sight distance issues: (no) yes
describe:

horizontal curvature describe: straight 3 (curve)
approximate curve length: 1: 347' 2: 277' 3: 363'
radius:

median type: grass paved w/ curb painted other (none)
width ft in

Segment HC

Date: 9-30-03 Weather: 60's sunny



Posted Speed:

Minor Access Points (Road Names)

\# 6

\# of driveways ~~HHH HHH HHH HHH HHH HHH~~ = 30

\# of parking lots 1

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	HHH = 5	0	HHH HHH II = 12	0	HHH HHH HHH HHH HHH HHH III = 33
	monument	fences	buildings	sign poles	overhead sign
	1	1	HHH HHH HHH HHH HHH HHH HHH HHH HHH = 45	HHH HHH HHH HHH IIII = 24	0
	parking meter	rock			
	1	0			

Section Length: 1645 ft 1"

Vertical Grade: 2.3, 3.0 % Crest on road: 5.0, 4.2 %

Terrain Type: level (rolling) mountainous

Land Use % residential 100% commerical ______ industrial ______

\# lanes: Going Left: 1 Going Right: 1

width of lanes: 20' 20'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 9'4" 10'4"

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) (bad)
describe: lots of cracking

pavement marking quality good (fair) bad
describe: starting to fade

parking allowed (yes) no 100 % allowed

road lighting: (present) not 100%

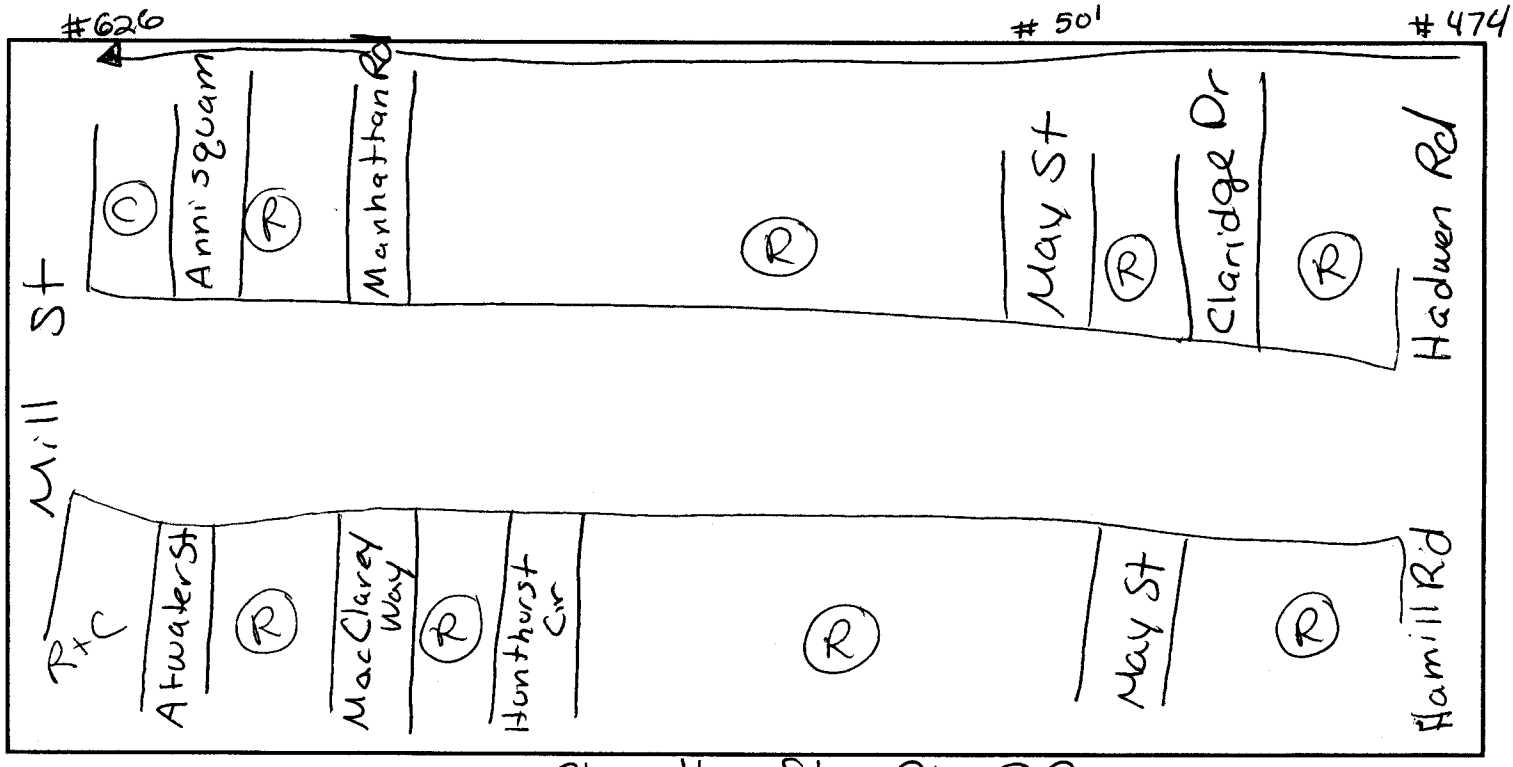
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width ft in

Segment 1C

Date: 9-30-03 Weather: 60's Sunny



Posted Speed:

Minor Access Points (Road Names)

8

of driveways 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 I = 66

# of parking lots 𝍸 𝍸 IIII = 14

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 = 10	1	𝍸 𝍸 IIII = 24	0	IIII = 84 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸

	monument	fences	buildings	sign poles	overhead sign
	𝍸 = 5	II 𝍸 = 7	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 IIII = 84	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 = 80	II = 2
	parking meter	rock			
	0	0			

utility pole 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 I = 41

Section Length: 5245 ft

Vertical Grade: 1.1, 2.7 % Crest on road: 6.5, 0.7 %

Terrain Type: (level) rolling mountainous

Land Use % residential 95% commerical 5% industrial

lanes: Going Left: 1 Going Right: 1

width of lanes: 20' 20'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: all 10' 1" ½ 9' 4"

curb present: (both) no

drainage present: (yes) no

pavement quality: (good) (fair) bad
describe: few cracks

pavement marking quality good (fair) bad
describe: starting to fade

parking allowed (yes) no 80 % allowed 10% ⊥

road lighting: (present) not 100 %

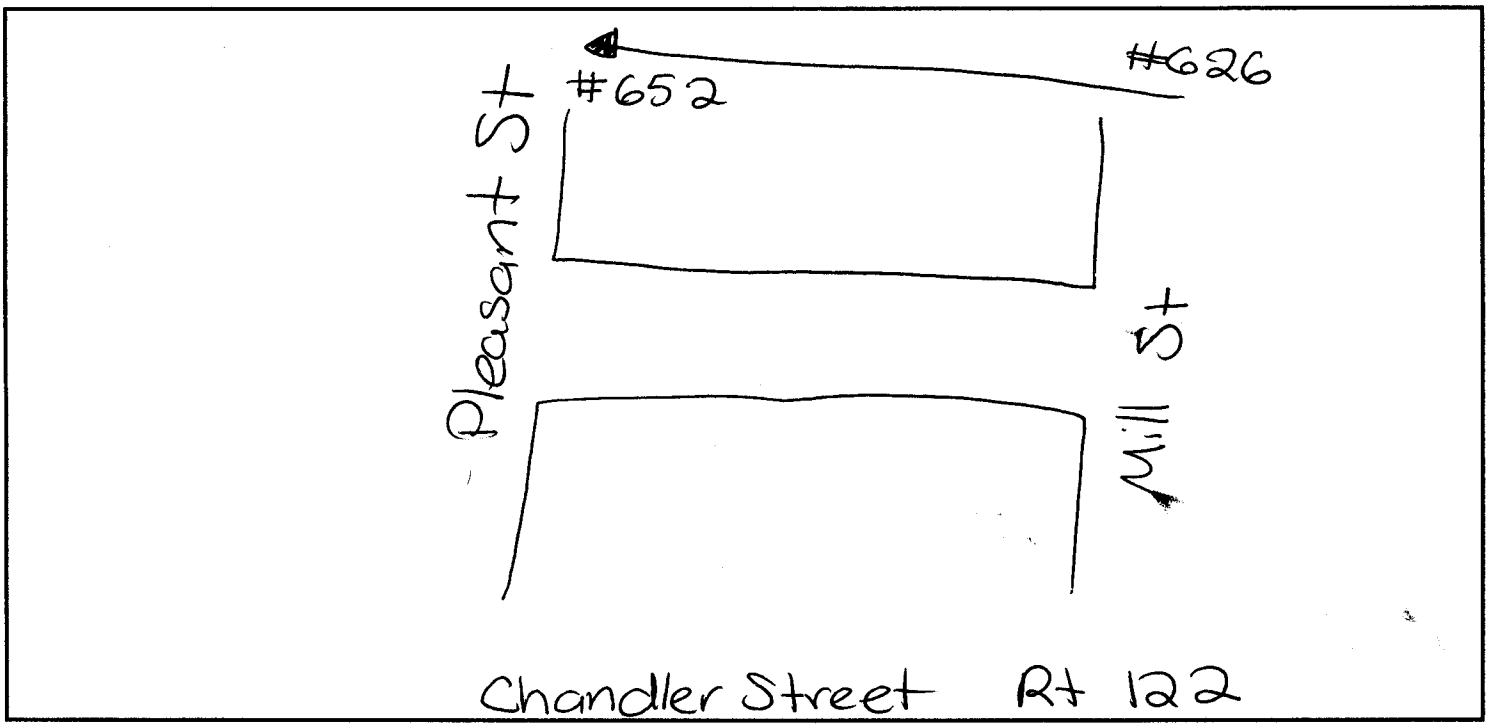
sight distance issues: (no) yes
describe:

horizontal curvature describe: straight 4 (curve)
approximate curve length: 1: 541' 3: 757'
radius: 2: 236' 4: 704'

median type: grass paved w/ curb painted other (none)
width ft in

Segment JC

Date: 9-30-03 Weather: 60's sunny



Posted Speed:

Minor Access Points (Road Names)

0

of driveways 0

# of parking lots 𝍸 𝍸 𝍸 | = 16

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	1	𝍸 = 2	𝍸 𝍸 = 7	0	𝍸 𝍸 = 10
	monument	fences	buildings	sign poles	overhead sign
	0	0	𝍸 𝍸 = 8	𝍸 𝍸 𝍸 𝍸 = 21	𝍸 = 3
	parking meter	rock	utility pole		
	𝍸 𝍸 𝍸 𝍸 𝍸 = 22	0	𝍸 = 2		

Section Length: 727 ft 7"

Vertical Grade: 0.2 % Crest on road: 5.0 %

Terrain Type: (level) rolling mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 1 Going Right: 1

width of lanes: 20' 20'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 9'10" 10'

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: cracking

pavement marking quality (good) (fair) bad
describe: starting to fade

parking allowed (yes) no 100 % allowed

road lighting: (present) not 100 %

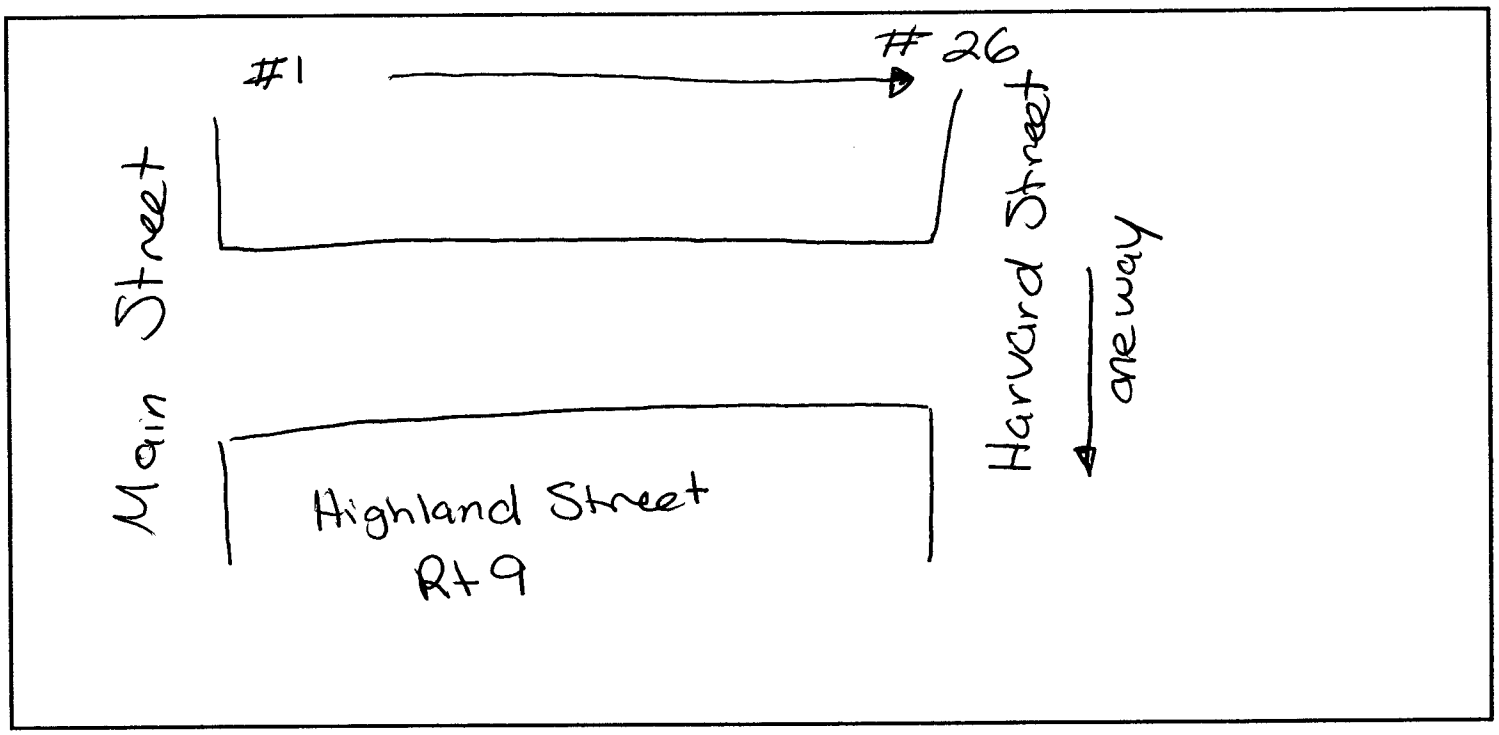
sight distance issues: (no) yes
describe:

horizontal curvature describe: straight | (curve)
approximate curve length:
radius: ≈ 143'

median type: grass paved w/ curb painted other (none)
width ft in

Segment AH

Date: 9-16-03 Weather: Sunny



Posted Speed:
Minor Access Points (Road Names)
0

of driveways // 2
# of parking lots /// 3

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	1	0	HHT / =6	0	0
	monument	fences	buildings	sign poles	overhead sign
	0	// =2	// =2	HHT / =6	//// =4
	parking meter	rock			
	0	0			

Section Length: 508 ft

Vertical Grade: 10.9 % Crest on road: 6.8 %

Terrain Type: level (rolling) mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: 12' 12' 12' 12'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 8'9" 8'6"

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: lots of rutting

pavement marking quality good fair (bad)
describe: most very faded, some all gone

parking allowed yes (no) _____ % allowed

road lighting: (present) not 100 %

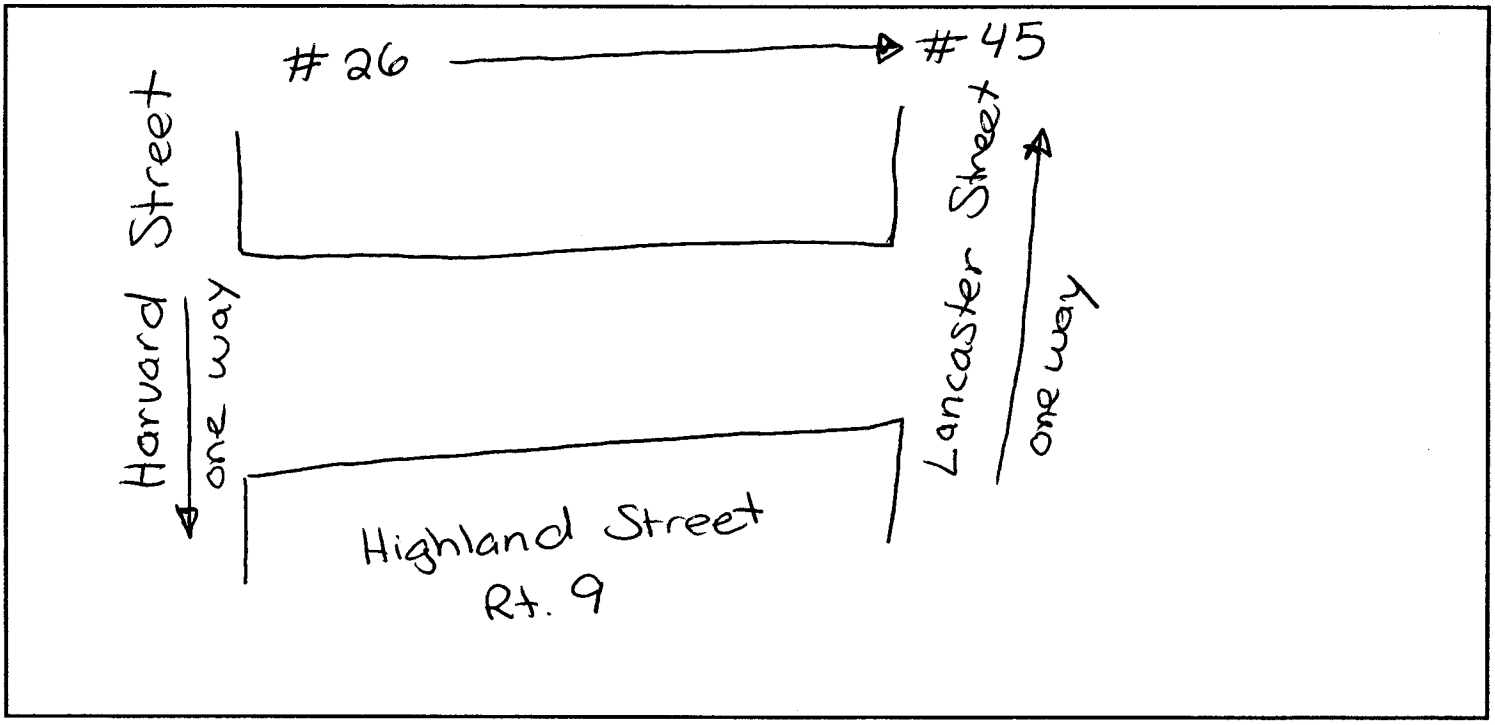
sight distance issues: no (yes)
describe: going up hill hard to see traffic light

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

Segment BH

Date: 9-16-03 Weather: Sunny



Posted Speed:
Minor Access Points (Road Names)
0

of driveways |||| =3
# of parking lots 1

roadside hazards:	firehydrants	mailboxes	utility (light) poles	benches	trees
	1	0	|||| =4	0	0

	monument	fences	buildings	sign poles	overhead sign
	0	|| =2	||| =3	|||| || =7	1

	parking meter	rock
	|||| |||| =9	0

Section Length: 289 ft 9"

Vertical Grade: 0.2, 3.9 % Crest on road: 6.8 %

Terrain Type: level (rolling) mountainous

Land Use % residential ____ commerical 100 % industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: 12', 12' 12', 12'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 6' 5'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good (fair) bad
describe: some rutting

pavement marking quality good fair (bad)
describe: very faded, can't see most of them

parking allowed yes (no) _____ % allowed

road lighting: (present) not 100 %

sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

Segment CH

Date: 9-16-03 Weather: sunny + cloudy

Posted Speed:

Minor Access Points (Road Names)

𝍸 III = 8

of driveways 8

# of parking lots 12

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	III = 3	III = 3	𝍸 II = 12		𝍸 I = 11
	monument	fences	buildings	sign poles	overhead sign
	I	𝍸 I = 6	𝍸 𝍸 𝍸 = 30	𝍸 𝍸 𝍸 III = 33	II = 2
	parking meter	rock	electrical box		
	𝍸 III = 8		I		

Section Length: 1443 ft 2"

Vertical Grade: 3.9 % Crest on road: 3.0 %

Terrain Type: level (rolling) mountainous

Land Use % residential 10% commerical 90% industrial

lanes: Going Left: 1 Going Right: 1

width of lanes: 12.5' 21.5'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 8' 9'9"

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: some rutting

pavement marking quality good (fair) bad
describe: Fading

parking allowed (yes) no 50 % allowed

road lighting: (present) not 100 %

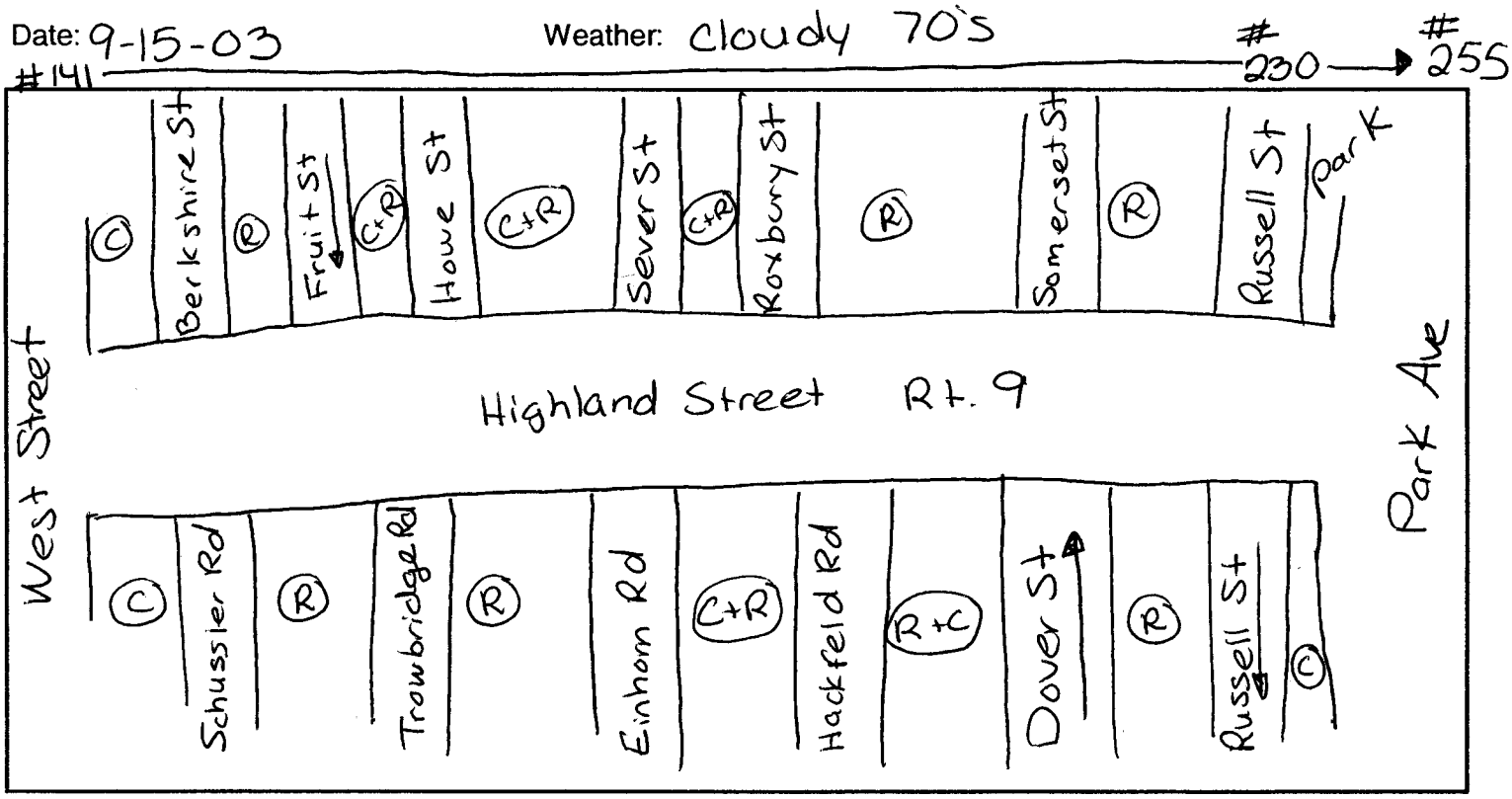
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width ft in

Segment DH

Date: 9-15-03 Weather: Cloudy 70's



Posted Speed:

Minor Access Points (Road Names)

13

of driveways 𝍸 𝍸 𝍸 111 = 18

# of parking lots 1

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 11 = 7	11 = 2	𝍸 𝍸 = 20	111 = 3	𝍸 𝍸 11 𝍸 𝍸 = 27
	monument	fences	buildings	sign poles	overhead sign
		𝍸 = 5	𝍸 𝍸 𝍸 𝍸 𝍸 1 = 31	𝍸 𝍸 𝍸 𝍸 = 30	1
	parking meter	rock			
	𝍸 = 5	1			

Section Length: 2105 ft 5"

Vertical Grade: 0.2 %

Crest on road: 5.8, -6.3 %

Terrain Type: level rolling mountainous

Land Use % residential 75% commerical 25% industrial

lanes: Going Left: 1 Going Right: 1

width of lanes: 16.5' 16.5'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 8' 7'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good (fair) bad
describe: some rutting

pavement marking quality good (fair) bad
describe: fading

parking allowed (yes) no 50 % allowed

road lighting: (present) not 100 %

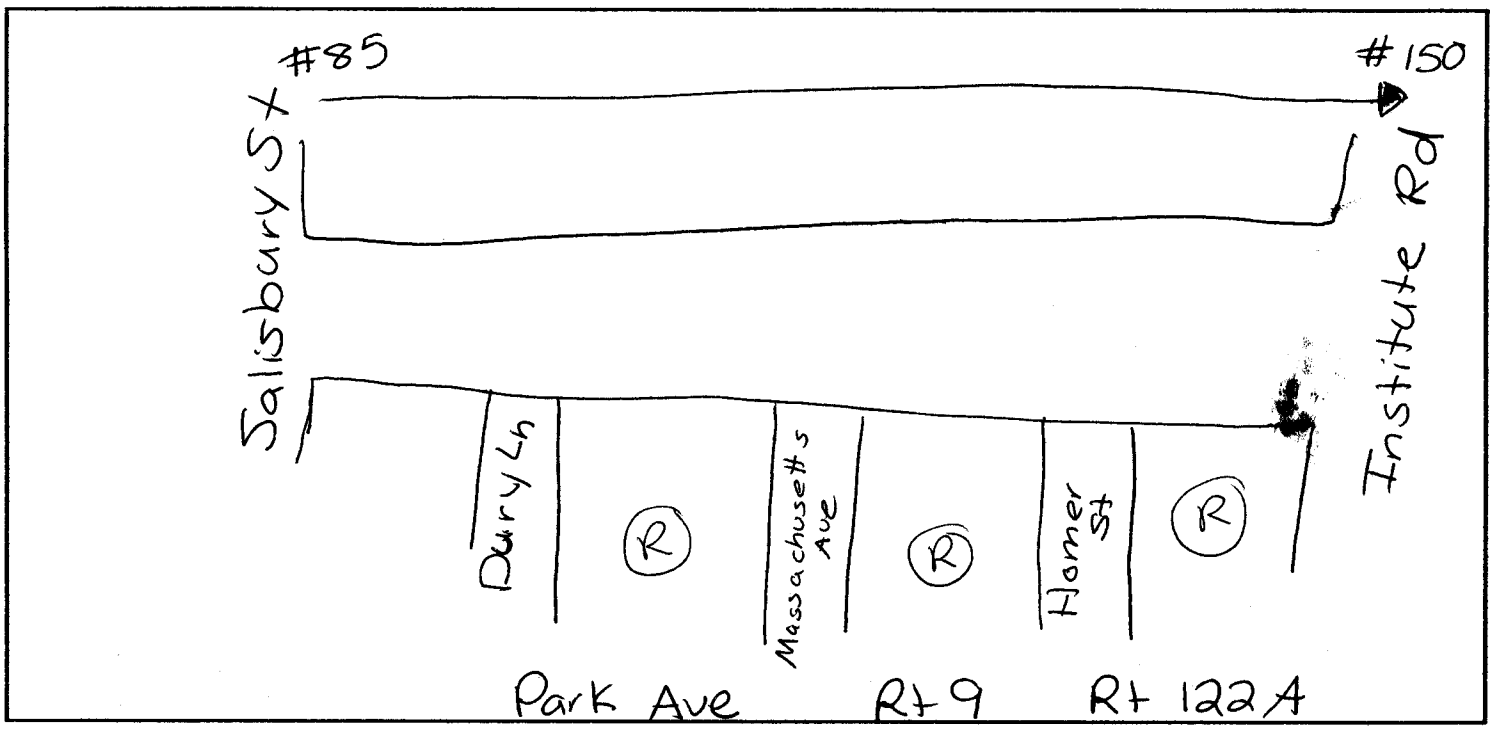
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

Segment AP

Date: 10-7-03 Weather: Sunny



Posted Speed: 35

Minor Access Points (Road Names)

3

of driveways |||| = 4

# of parking lots 0

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 = 5	0	𝍸 = 10	0	𝍸 𝍸 = 15
	monument	fences	buildings	sign poles	overhead sign
	0	||| = 3	|||| = 4	𝍸 ||| = 13	||| = 3
	parking meter	rock			
	0	0			

Section Length: 1865 ft

Vertical Grade: 3.0, 3.2 % Crest on road: 4.2 %

Terrain Type: level (rolling) mountainous

Land Use % residential 100 % commerical _____ industrial _____

lanes: Going Left: 2 Going Right: 2

width of lanes: 16', 12' | 12', 16'

type of shoulder: paved | dirt | (none)

width of shoulder: 0' | 0'

sidewalk present: (yes) | no | width: 9'8" | 12'5"

curb present: (both) | no

drainage present: (yes) | no

pavement quality: good | (fair) | bad
describe: rutting

pavement marking quality: good | (fair) | bad
describe: starting to fade

parking allowed: yes | (no) | _____ % allowed

road lighting: (present) | not | 100 %

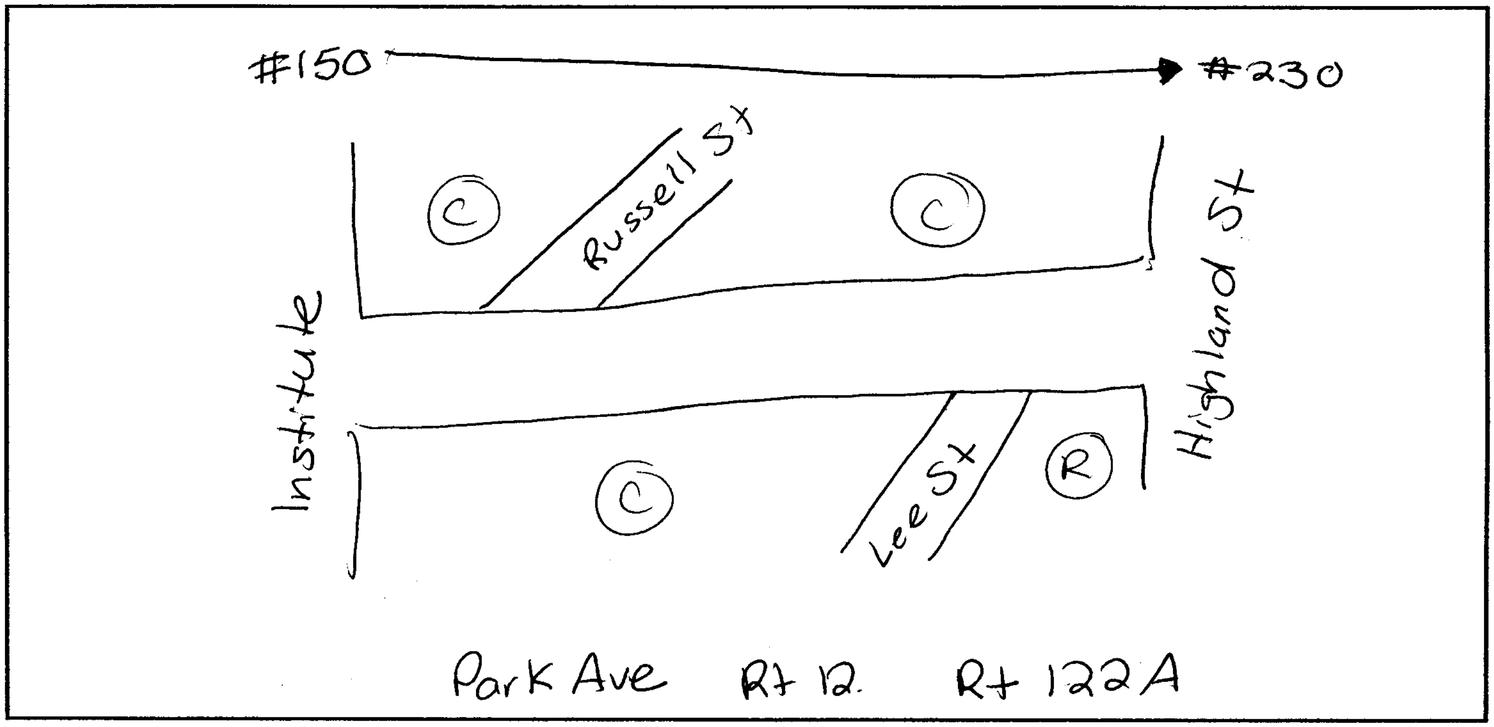
sight distance issues: (no) | yes
describe:

horizontal curvature describe: (straight) | curve
approximate curve length:
radius:

median type: grass | paved w/ curb | painted | other (none)
width _____ ft | _____ in

segment 13P

Date: 10-7-03 Weather: Sunny



Posted Speed: 35

Minor Access Points (Road Names)

2

of driveways ~~||||~~ || =7

# of parking lots ~~||||~~ ~~||||~~ ||| =13

roadside hazards:

firehydrants	mailboxes	utility (light) poles	benches	trees
|| =2	1	~~||||~~ | =6	0	~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ |||| =34
monument	**fences**	**buildings**	**sign poles**	**overhead sign**
0	0	~~||||~~ ~~||||~~ =10	~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ | =21	||| =3
parking meter	**rock**	**electrical box**		
0	0	1		

Section Length: 1280 ft

Vertical Grade: 1.6 % Crest on road: 4.8 %

Terrain Type: (level) rolling mountainous

Land Use % residential 15% commerical 85% industrial ______

lanes: Going Left: 2 Going Right: 2

width of lanes: 16' 12' | 12' 16'

type of shoulder: paved | dirt | (none)

width of shoulder: 0' | 0'

sidewalk present: (yes) | no | width: 9'8" | 11'4"

curb present: (both) | no

drainage present: (yes) | no

pavement quality: good | (fair) | bad
describe: minor rutting

pavement marking quality: good | (fair) | bad
describe: starting to fade

parking allowed: yes | (no) | ____ % allowed

road lighting: (present) | not | 100 %

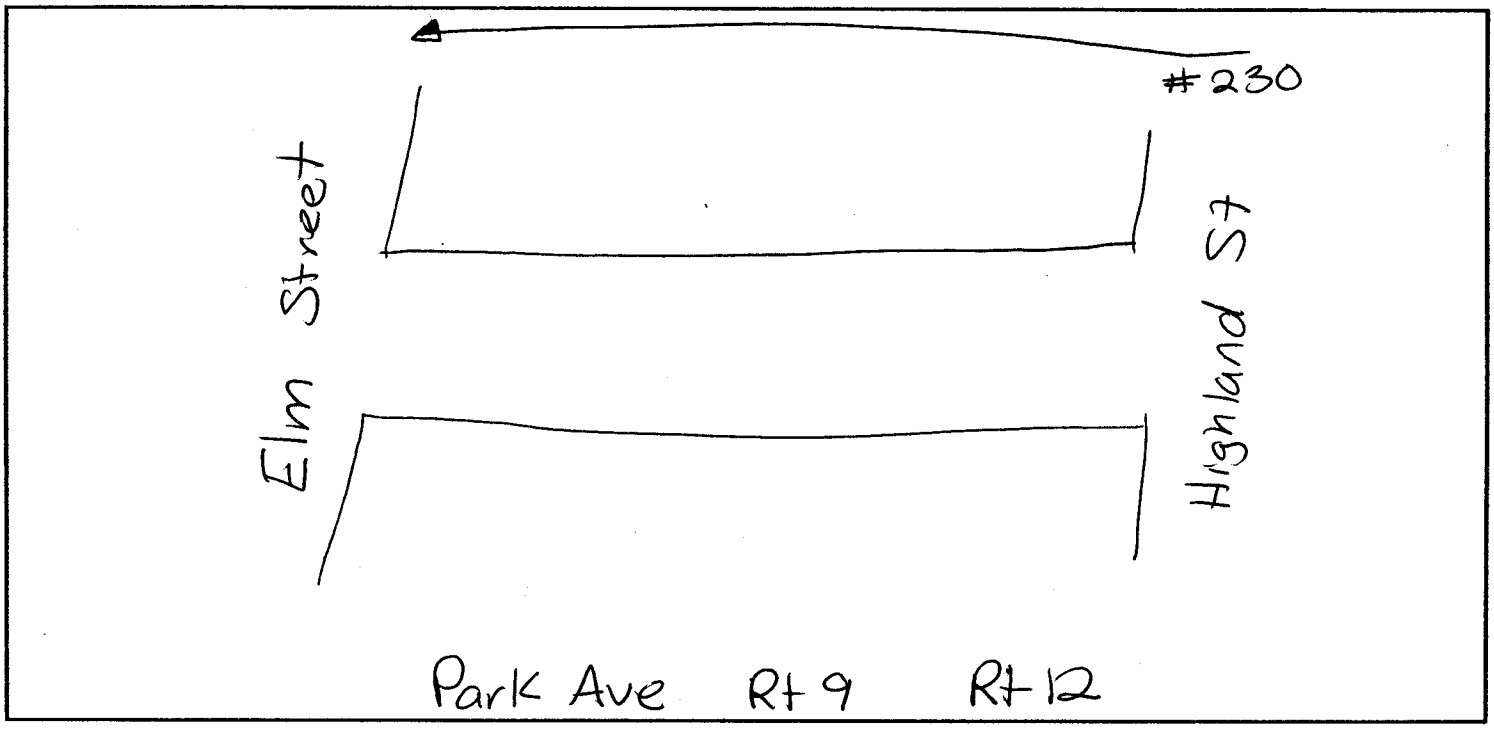
sight distance issues: (no) | yes
describe:

horizontal curvature describe: (straight) | curve
approximate curve length:
radius:

median type: grass | paved w/ curb | painted | other (none)
width ____ ft ____ in

segment CP

Date: 9-22-03 Weather: Sunny mid 60's



Posted Speed: 30
Minor Access Points (Road Names)
0

of driveways III = 3
# of parking lots 0

roadside hazards:

firehydrants IIII = 4	mailboxes 0	utility/light poles 𝍸 IIII = 9	benches 0	trees 𝍸 𝍸 𝍸 𝍸 II = 47	overhead
monument 𝍸 = 5	fences II = 2	buildings 0	sign poles 𝍸 = 10	sign III = 3	
parking meter 0	rock II = 2				

Section Length: 1484 ft 5"

Vertical Grade: 0.6 %

Crest on road: 3.6 %

Terrain Type: (level) rolling mountainous

Land Use % residential 100% commerical _____ industrial _____

lanes: Going Left: 2 Going Right: 2

width of lanes: 15' 13' 13' 13'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 3' 3'

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe:

pavement marking quality good (fair) bad
describe:

parking allowed yes (no) _____ % allowed

road lighting: (present) not 100 %

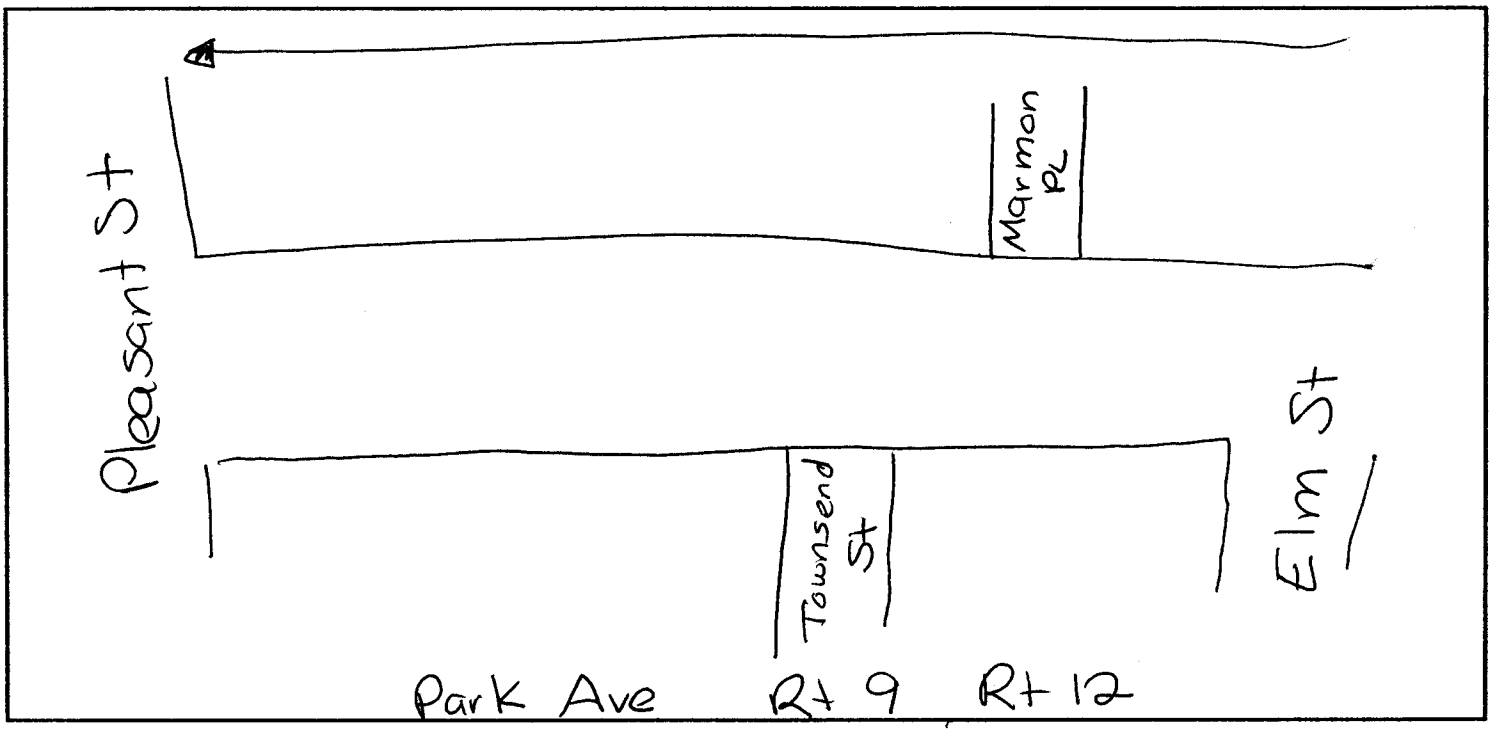
sight distance issues: (no) yes
describe:

horizontal curvature describe: straight (curve)
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

segment DP

Date: 9-22-03 Weather: Sunny



Posted Speed:
Minor Access Points (Road Names)
2

of driveways |||| = 3
# of parking lots ||||-||||-||| = 13

roadside hazards:

firehydrants |||| = 4	mailboxes |||| = 9	utility/light poles ||||-||| = 8	benches 0	trees ||||-| = 6
monument 1	fences 0	buildings ||||-||||-||| = 13	sign poles ||||-||||-|||| = 19	overhead sign ||| = 3
parking meter ||||-||||-||||-|||| = 20	rock 0	utility 1		

Section Length: 487 ft

Vertical Grade: 2.3 % Crest on road: 3.2 %

Terrain Type: (level) rolling mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: 15', 13' 13', 15'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 11' 10" 11'

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: some rutting, craclang

pavement marking quality good (fair) bad
describe: starting to fade

parking allowed (yes) no 80 % allowed

road lighting: (present) not 100 %

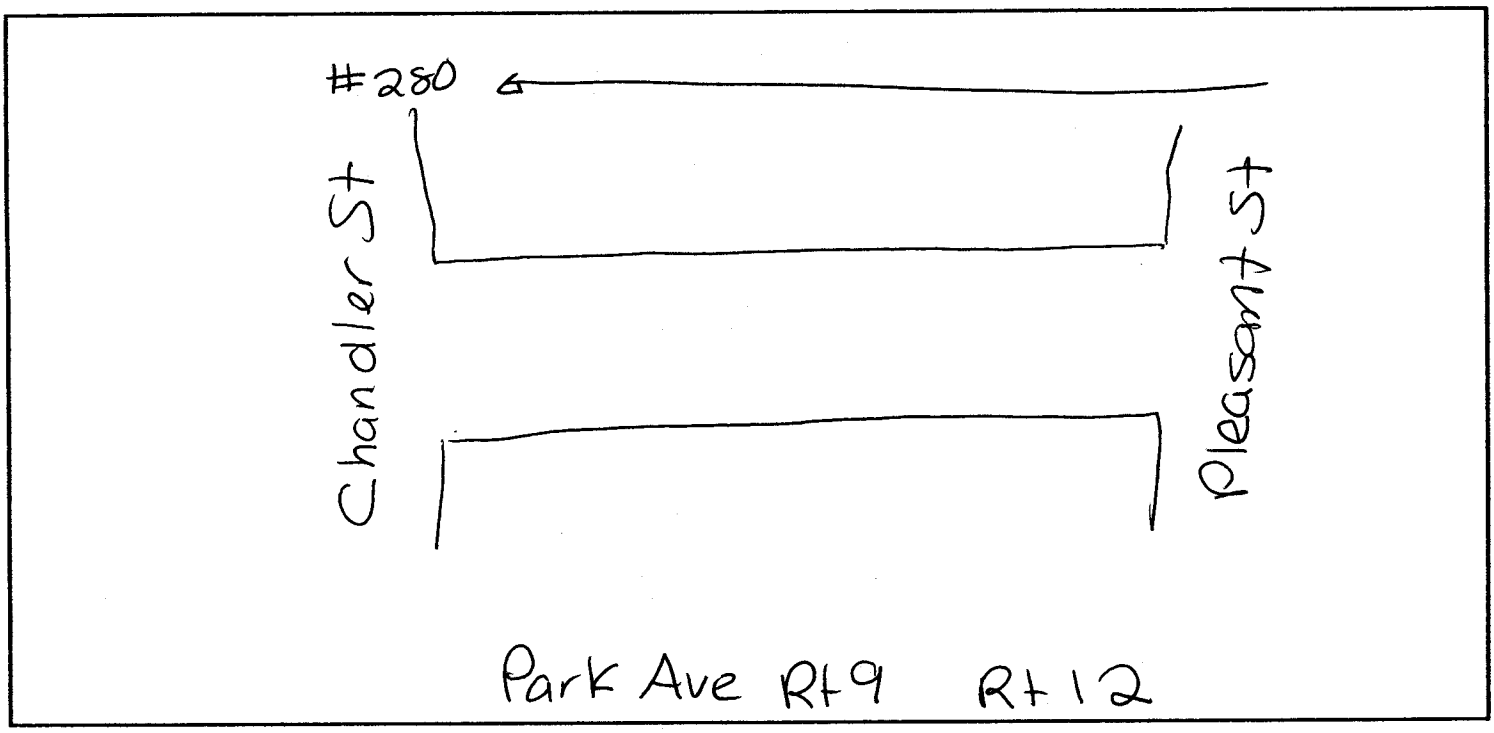
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width ft in

segment EP

Date: 9-22-03 Weather: Sunny



Posted Speed:
Minor Access Points (Road Names)
0

of driveways III = 3
# of parking lots 𝍸 𝍸 𝍸 = 15

roadside hazards:

firehydrants	mailboxes	utility/light poles	benches	trees
IIII = 4	III = 3	𝍸 I = 6	0	II = 2

monument	fences	buildings	sign poles	overhead sign
0	0	𝍸 𝍸 = 10	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 II = 32	𝍸 I = 6

parking meter	rock
0	0

Section Length: 771 ft 10"

Vertical Grade: 2.7 % Crest on road: 4.1 %

Terrain Type: (level) rolling mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: 15', 13' 13', 13'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 10' 12'

curb present: (both) no

drainage present: (yes) no

pavement quality: (good) fair bad

describe: minor cracks

pavement marking quality good (fair) bad

describe: starting to fade

parking allowed yes (no) _____ % allowed

road lighting: (present) not 100 %

sight distance issues: (no) yes

describe:

horizontal curvature describe: (straight) curve

approximate curve length:

radius:

median type: grass paved w/ curb painted other (none)

width _____ ft _____ in

B Appendix: Databases for Validation Data

This appendix has the datasheets for the arterial segments that were used to validate the models in this paper. This includes the data from both Park Avenue and Shrewsbury Street. The summary sheets of that data are also included. The data from Park Avenue is first followed by that of Shrewsbury Street starting on page B-16.

Segment	Date	speed	volume	3-total	3-injury	3-total-s	3-injury-s	3-PDO-s	3-total-i	3-injury-i
APP	4-15	35	24163	87	26	68	19	49	19	7
BPP	4-15	35	24163	50	8	27	4	23	23	4
CPP	4-16	30	25990	228	62	160	41	119	68	21
DPP	4-16	30	25990	142	35	128	31	97	14	4
EPP	4-16	30	26780	101	34	56	23	33	45	11
FPP	4-16	30	26780	84	27	72	23	49	12	4

Segment	3-PDO-i	2002 total-	2002 injur	2002 PDO	2002 total-	2002 injury	2002 PDO-i	2001 total-s	2001 injury	2001 PDO-s
APP	12	27	9	18	6	3	3	9	3	6
BPP	19	7	2	5	3	1	2	12	1	11
CPP	47	50	14	36	20	7	13	57	11	46
DPP	10	51	14	37	8	3	5	39	8	31
EPP	34	23	10	13	17	4	13	12	6	6
FPP	8	36	14	22	3	2	1	18	3	15

Segment	2001 total-	2001 injur	2001 PDO	2000 total-	2000 injur	2000 PDO-s	2000 total-i	2000 injury-i	2000 PDO-i	minor access
APP	8	1	7	32	7	25	5	3	2	1
BPP	3	0	3	8	1	7	17	3	14	2
CPP	23	7	16	53	16	37	25	7	18	3
DPP	4	1	3	38	9	29	2	0	2	4
EPP	18	4	14	21	7	14	10	3	7	3
FPP	4	0	4	18	6	12	5	2	3	0

Segment	driveways	parkinglot	d+p	total	fire hydra	mailboxes	light poles	utility poles	benches	trees
APP	3	14	17	18	3	8	7	2	1	11
BPP	14	0	14	16	6	0	12	4	0	30
CPP	0	27	27	30	8	1	14	1	0	4
DPP	3	12	15	19	4	2	10	0	0	8
EPP	1	8	9	12	1	0	6	0	0	8
FPP	5	14	19	19	5	2	11	0	0	5

Segment	monumen	fences	buildings	sign poles	overhead	parking me	rocks	other	total	pole
APP	1	1	11	15	2	0	0	1	63	26
BPP	3	2	21	22	3	0	0	2	105	41
CPP	0	4	26	33	1	0	0	0	92	49
DPP	1	3	19	25	2	0	0	1	75	37
EPP	0	2	12	14	1	0	0	2	46	21
FPP	0	5	21	26	0	0	0	1	76	37

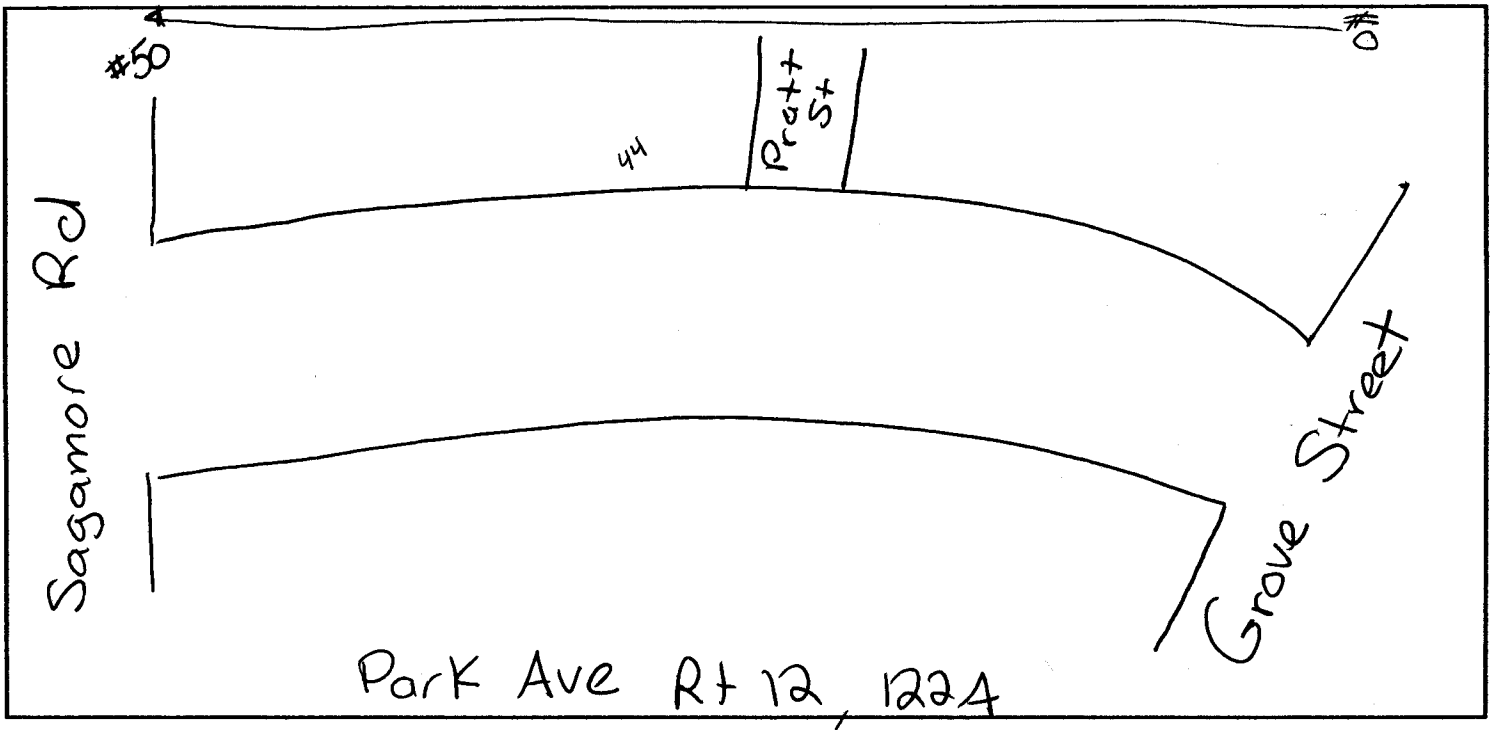
Segment	length	max grade	crest	terrain typ	residentia	commercia	industrial	left lanes	right lanes	width
APP	1128	1	4.6	level	10	90	0	2	2	16
BPP	1513	3.2	4	rolling	80	20	0	2	2	14
CPP	1712	2.5	5	level	5	95	0	2	2	16
DPP	1204	0.9	4.4	level	5	95	0	2	2	16
EPP	765	2	5.4	level	10	90	0	2	2	16
FPP	1442	3.7	5.8	rolling	5	95	0	2	2	20

Segment	width	width	width	average w	type shou	sidewalk	average wid	curb	drainage	pavement
APP	12	12	16	14	none	yes	9.5	both	yes	fair
BPP	13	13	14	13.5	none	yes	4.5	both	yes	good
CPP	12	12	16	14	none	yes	10	both	yes	fair
DPP	12	12	16	14	none	yes	11.5	both	yes	fair
EPP	12	12	16	14	none	yes	10.5	both	yes	fair
FPP	12	12	20	16	none	yes	10	both	yes	fair

Segment	markings	% parking	% lighting	SD	curve	curves	median type	width
APP	fair	0	100	no	1	1	none	0
BPP	fair	30	100	no	1	1	none	0
CPP	good	10	100	no	0	0	none	0
DPP	fair	50	100	no	0	0	none	0
EPP	good	40	100	no	0	0	none	0
FPP	fair	30	100	no	0	0	none	0

Segment APP

Date: 4-15-04 Weather: Sunny 50's



Posted Speed: 35

Minor Access Points (Road Names)

1

of driveways III = 3

# of parking lots 𝍸 𝍸 III = 14

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	III = 3	𝍸 III = 8	𝍸 II = 7	1	𝍸 𝍸 I = 11
	monument	fences	buildings	sign poles	overhead sign
	1	1	𝍸 𝍸 I = 11	𝍸 𝍸 = 15	II = 2
	parking meter	rock	utility poles	electrical	
	0	0	II = 2	I = 1	

Section Length: 1128 ft

Vertical Grade: 1.0 % Crest on road: 4.6 %

Terrain Type: (level) rolling mountainous

Land Use % residential 10% commerical 90% industrial 0%

lanes: Going Left: 2 Going Right: 2

width of lanes: 16' 12' 12' 16'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 10' 9'

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: minor rutting

pavement marking quality good (fair) bad
describe: starting to fade

parking allowed yes (no) 0 % allowed

road lighting: (present) not 100 %

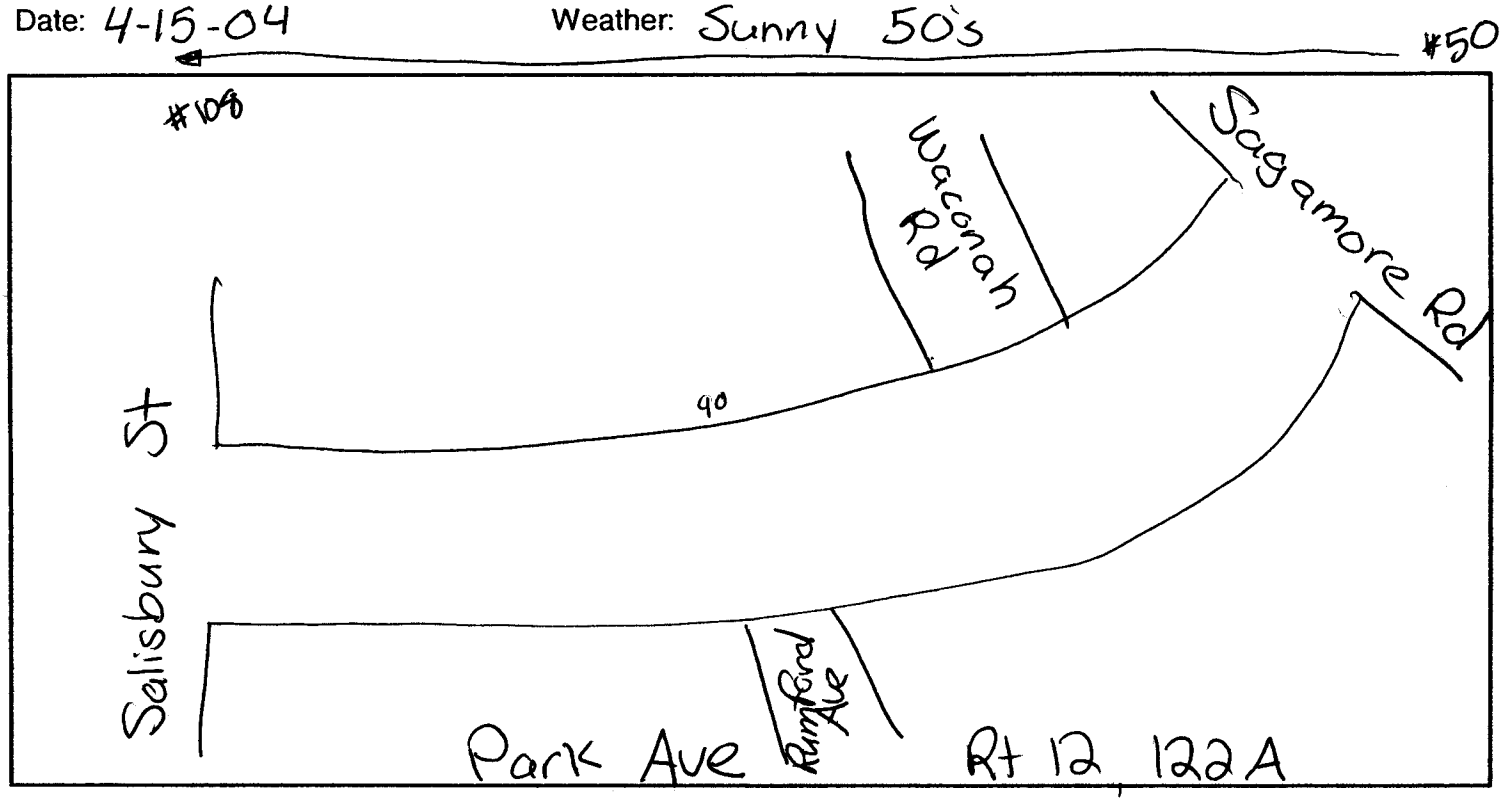
sight distance issues: (no) yes
describe:

horizontal curvature describe: straight 1 (curve)
approximate curve length:
radius: ≈250

median type: grass paved w/ curb painted other (none)
width ft in

Segment B PP

Date: 4-15-04 Weather: Sunny 50's



Posted Speed: 35

Minor Access Points (Road Names)

11 = 2

of driveways 𝍸 𝍸 1111 = 14

# of parking lots 0

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 1 = 6	0	𝍸 𝍸 11 = 12	0	𝍸 𝍸 𝍸 = 30
	monument	fences	buildings	sign poles	overhead sign
	111 = 3	11 = 2	𝍸 𝍸 = 20	𝍸 𝍸 11 = 22	111 = 3
	parking meter	rock	utility pole	electrical	
	0	0	1111 = 4	11	

Section Length: 1513 ft

Vertical Grade: 3.2 % Crest on road: 4.0 %

Terrain Type: level (rolling) mountainous

Land Use % residential 80% commerical 20% industrial 0%

lanes: Going Left: 2 Going Right: 2

width of lanes: 14' 13' 13' 14'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 4.5' 4.5'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: (good) fair bad
describe:

pavement marking quality good (fair) bad
describe: starting to fade

parking allowed (yes) no 30 % allowed

road lighting: (present) not 100 %

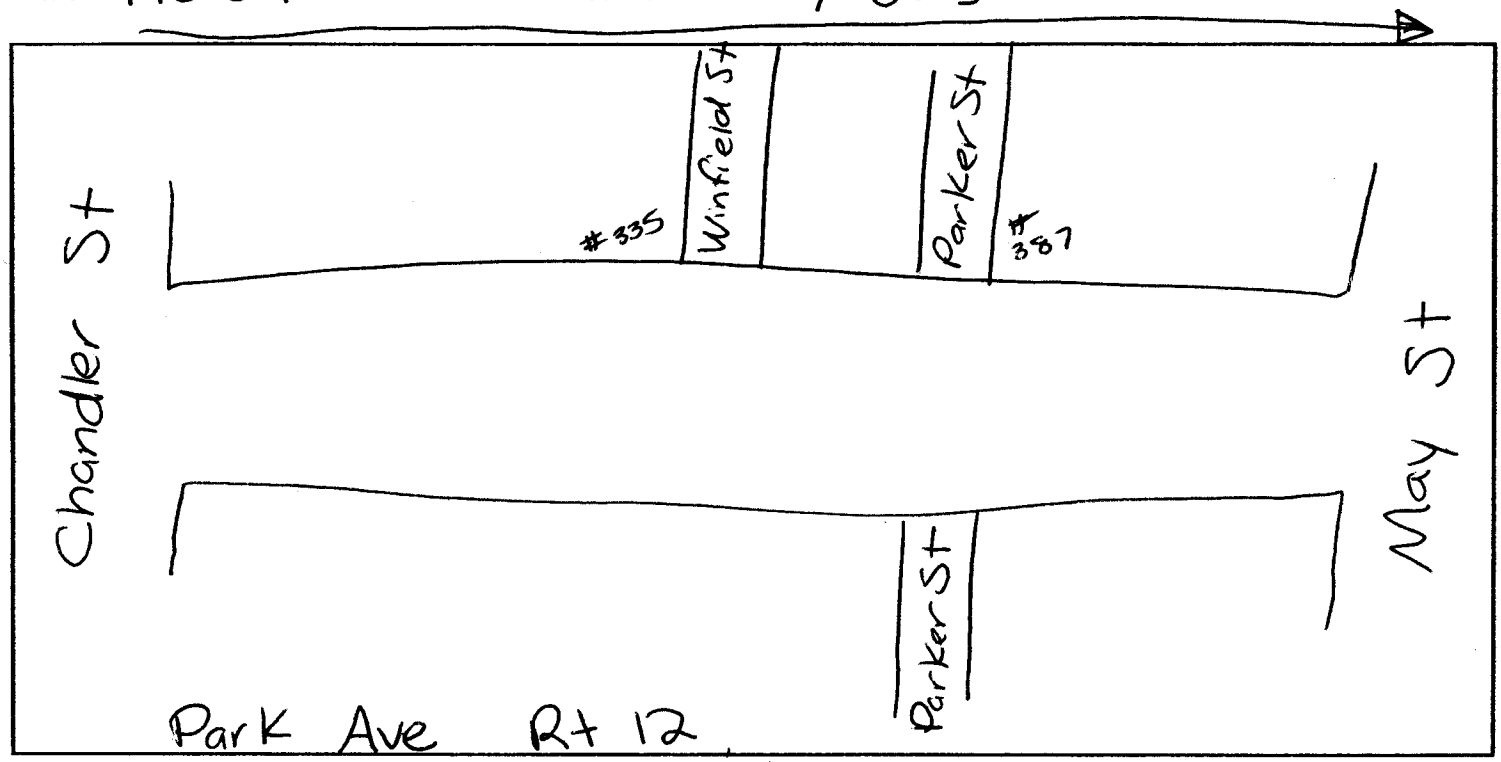
sight distance issues: (no) yes
describe:

horizontal curvature describe: straight 1 (curve)
approximate curve length:
radius: ≈500

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

Segment CPP

Date: 4-16-04 Weather: Sunny 50's



Posted Speed: 30

Minor Access Points (Road Names)

||| = 3

of driveways

# of parking lots 𝍸 𝍸 𝍸 𝍸 𝍸 || = 27

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 = 8	1	𝍸 𝍸 |||| = 14	0	|||| = 4
	monument	fences	buildings	sign poles	overhead sign
	0	|||| = 4	𝍸 𝍸 𝍸 𝍸 = 26	𝍸 𝍸 𝍸 𝍸 = 33	1
	parking meter	rock	utility pole		
	0	0	1		

Section Length: 1712 ft

Vertical Grade: 2.5 % Crest on road: 5.0 %

Terrain Type: (level) rolling mountainous

Land Use % residential 5% commerical 95% industrial 0%

lanes: Going Left: 2 Going Right: 2

width of lanes: 16' 12' 12' 16'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 10' 10'

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: cracking, rutting

pavement marking quality (good) fair bad
describe:

parking allowed (yes) no 10 % allowed

road lighting: (present) not 100 %

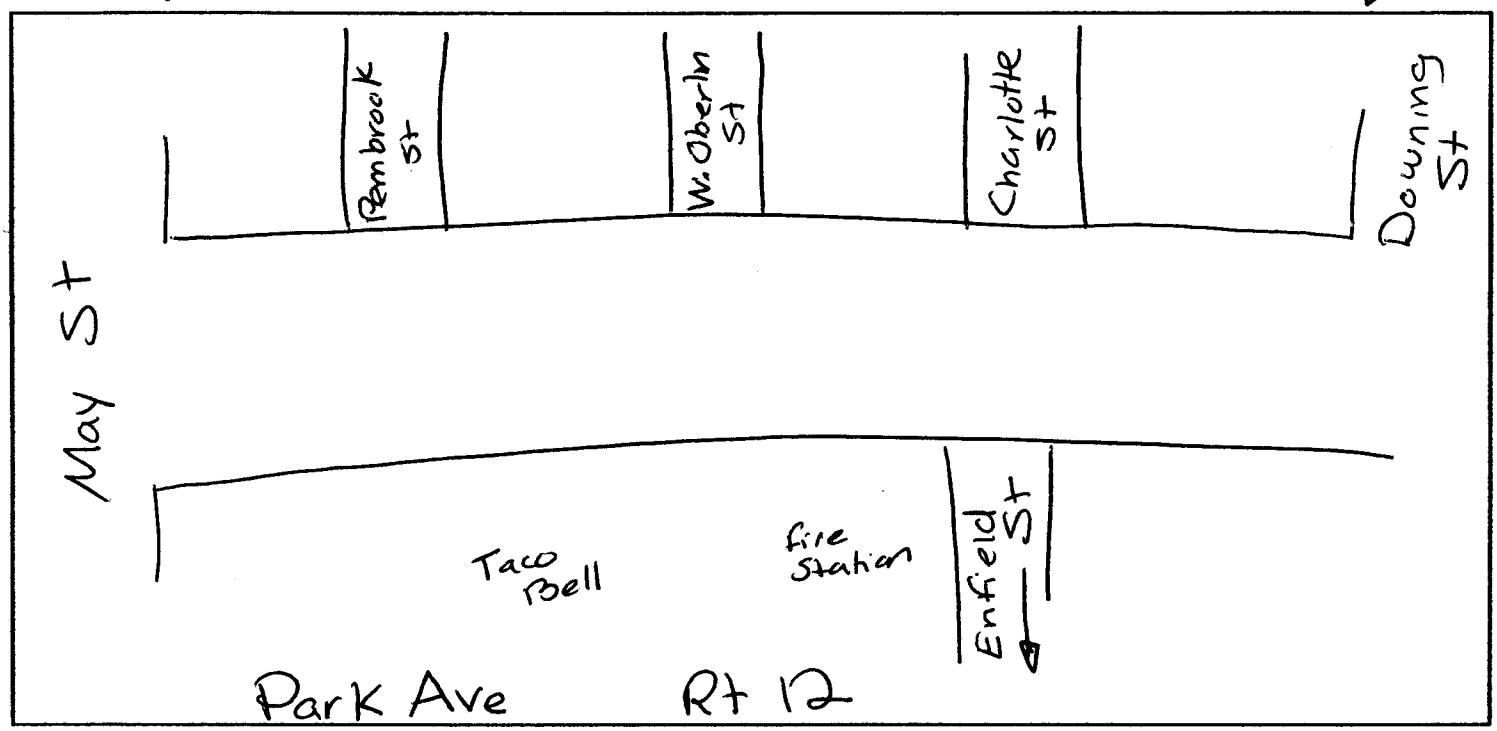
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width ft in

segment DPP

Date: 4-16-04 Weather: Sunny 50's



Posted Speed: 30

Minor Access Points (Road Names)

|||| = 4

of driveways||| = 3

# of parking lots 𝍸 𝍸 || = 12

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	|||| = 4	|| = 2	𝍸	0	𝍸 ||| = 8
	monument	fences	buildings	sign poles	overhead sign
	1	||| = 3	𝍸 𝍸 |||| = 19	𝍸 𝍸 𝍸 = 25	|| = 2
	parking meter	rock	electrical		
	0	0	1		

Section Length: 1204 ft

Vertical Grade: 0.9 % Crest on road: 4.4 %

Terrain Type: (level) rolling mountainous

Land Use % residential 5% commerical 95% industrial 0%

lanes: Going Left: 2 Going Right: 2

width of lanes: 16' 12' 12' 16'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 13' 10'

curb present: (both) no

drainage present: (yes) no

pavement quality: good (fair) bad
describe: ruthing, cracking

pavement marking quality good (fair) bad
describe: starting to fade

parking allowed (yes) no 50 % allowed

road lighting: (present) not 100 %

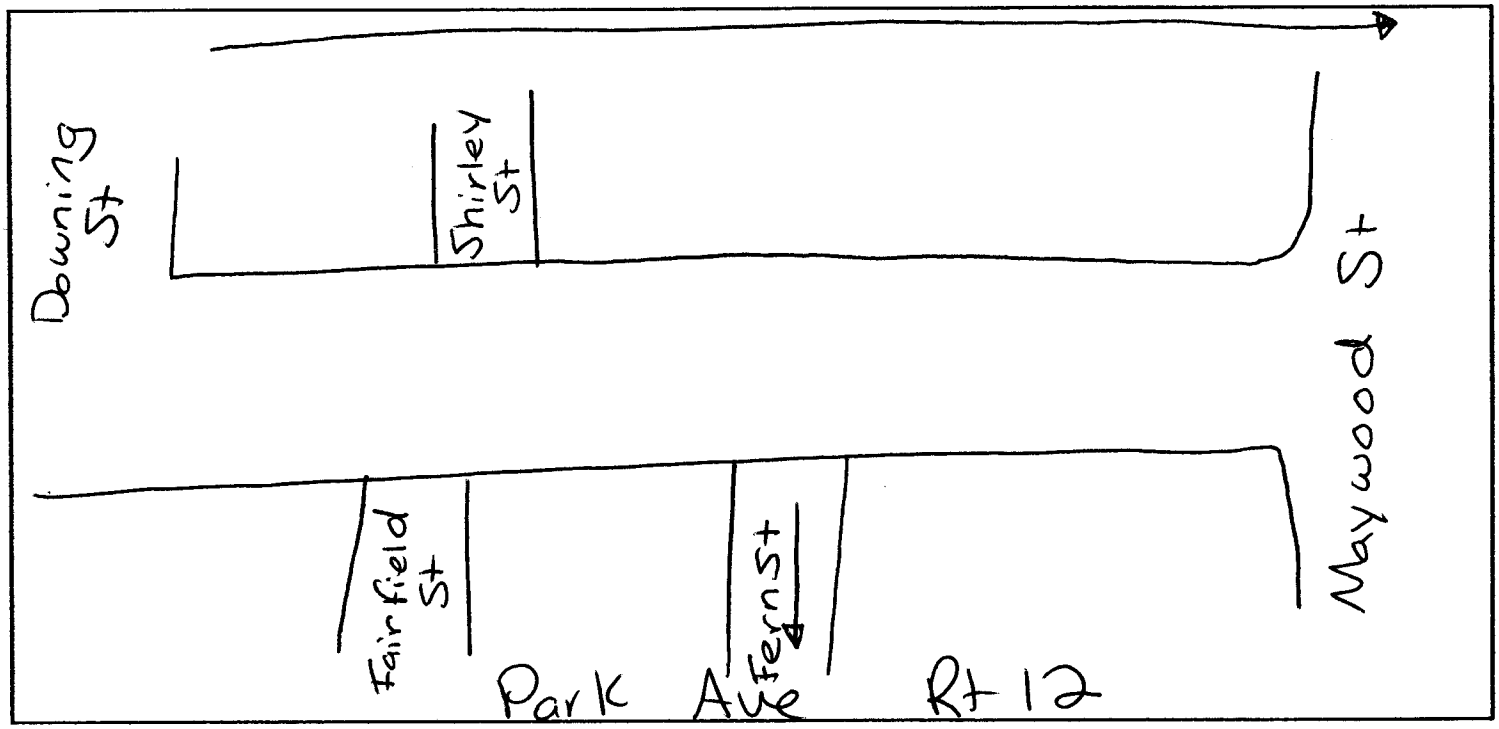
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width ft in

Segment EPP

Date: 4-16-04 Weather: Sunny 50's



Posted Speed: 30

Minor Access Points (Road Names)

III = 3

of driveways I

# of parking lots ~~IIII~~ III = 8

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	I	0	~~IIII~~ I = 6		III ~~IIII~~ = 8
	monument	fences	buildings	sign poles	overhead sign
	0	II = 2	~~IIII~~ II = 12	~~IIII~~ IIII = 14	I
	parking meter	rock	electrical		
	0	0	II		

Section Length: 765 ft

Vertical Grade: 2.0 % Crest on road: 5.4 %

Terrain Type: (level) rolling mountainous

Land Use % residential 10% commerical 90% industrial 0%

lanes: Going Left: 2 Going Right: 2

width of lanes: 16' 12' 12' 16'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 10.5' 10.5'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good (fair) bad
describe: rutting, cracking

pavement marking quality (good) fair bad
describe:

parking allowed (yes) no 40 % allowed

road lighting: (present) not 100 %

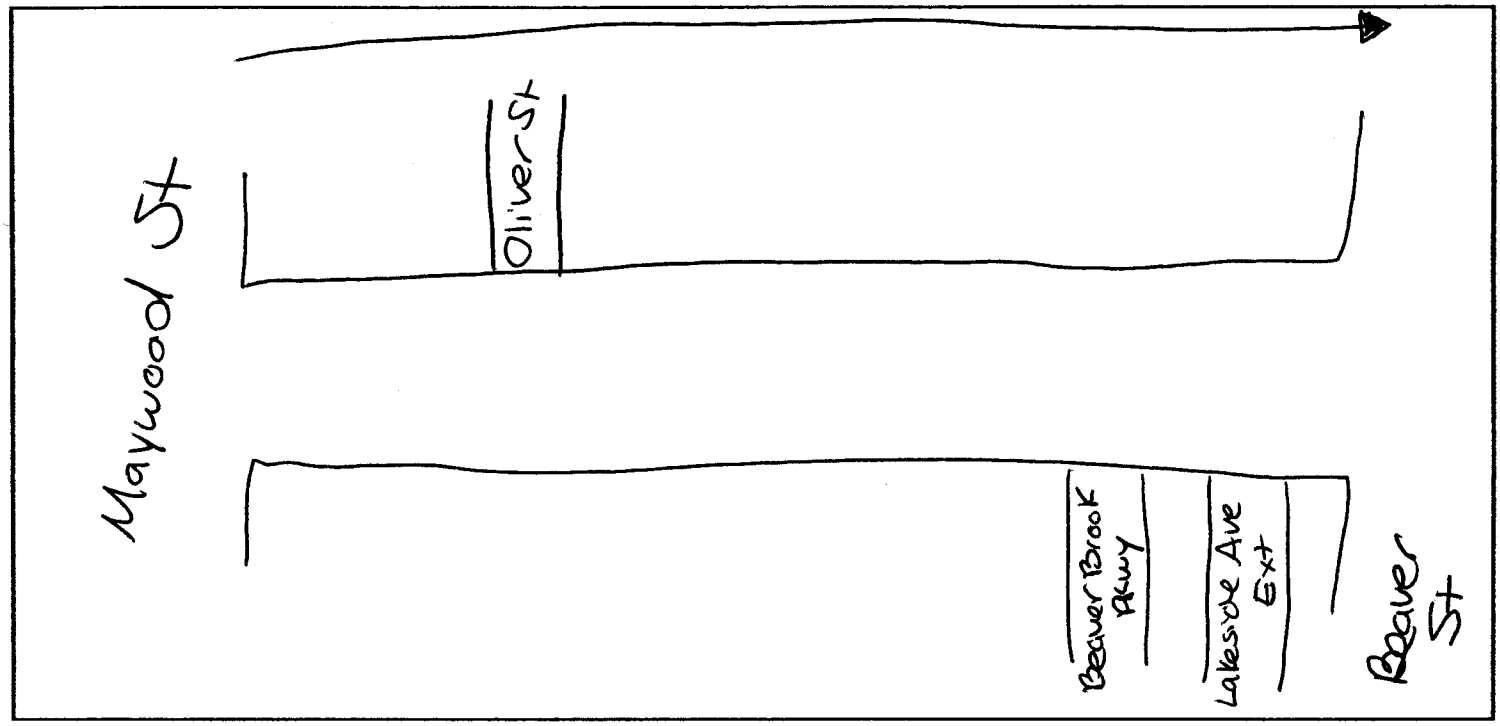
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

Segment FPP

Date: 4-16-04 Weather: Sunny 50's



Posted Speed: 30

Minor Access Points (Road Names)

III = 3

of driveways 𝍸 = 5

# of parking lots 𝍸 𝍸 IIII = 14

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 = 5	II = 2	𝍸 I = 11	0	𝍸 = 5
	monument	fences	buildings	sign poles	overhead sign
	0	𝍸 = 5	𝍸 𝍸 I = 21	𝍸 𝍸 I = 26	0
	parking meter	rock	electrical		
	0	0	1		

Section Length: 1442 ft

Vertical Grade: 3.7 % Crest on road: 5.8 %

Terrain Type: level rolling mountainous

Land Use % residential 5% commerical 95% industrial 0%

lanes: Going Left: 2 Going Right: 2

width of lanes: 20' 12' 12' 20'

type of shoulder: paved dirt (none)

width of shoulder: 0' 0'

sidewalk present: (yes) no width: 10' 10'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good (fair) bad
describe: patching

pavement marking quality good (fair) bad
describe: fading

parking allowed (yes) no 30 % allowed

road lighting: (present) not 100 %

sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

Segment	Date	speed	volume	3-total	3-injury	3-total-s	3-injury-s	3-PDO-s	3-total-i	3-injury-i	3-PDO-i
AS	9-23	30	16830	70	20	46	15	31	24	5	19
BS	9-23	30	16830	30	3	12	1	11	18	2	16
CS	9-24	30	21726	44	9	36	8	28	8	1	7
DS	10-3	30	21726	93	26	79	19	60	14	7	7
ES	10-3	30	19074	46	8	42	8	34	4	0	4
FS	10-3	30	19074	71	16	17	5	12	54	11	43

Segment	2002 total-s	2002 injury-s	2002 PDO-s	2002 total-i	2002 injury-i	2002 PDO-i	2001 total-s	2001 injury-s	2001 PDO-s	2001 total-i	2001 injury-i
AS	9	4	5	7	2	5	23	7	16	10	2
BS	5	1	4	5	0	5	4	0	4	4	1
CS	22	5	17	2	1	1	8	1	7	4	0
DS	31	8	23	7	3	4	27	4	23	2	1
ES	18	4	14	0	0	0	13	1	12	0	0
FS	5	2	3	13	2	11	7	2	5	21	4

Segment	2001 PDO-i	2000 total-s	2000 injury-s	2000 PDO-s	2000 total-i	2000 injury-i	2000 PDO-i	minor access	driveways	parkinglots	d+p
AS	8	14	4	10	7	1	6	3	2	5	7
BS	3	3	0	3	9	1	8	1	0	6	6
CS	4	6	2	4	2	0	2	5	1	6	7
DS	1	21	7	14	5	3	2	10	8	23	31
ES	0	11	3	8	4	0	4	8	5	12	17
FS	17	5	1	4	20	5	15	1	0	4	4

Segment	total	fire hydrants	mailboxes	light poles	utility poles	benches	trees	monuments	fences	buildings	sign poles
AS	10	3	1	12	0	2	4	1	1	8	14
BS	7	1	2	6	0	0	3	2	2	10	7
CS	12	3	4	12	0	4	5	5	3	11	15
DS	41	12	11	23	0	1	34	3	9	33	45
ES	25	8	3	29	0	1	9	2	4	27	49
FS	5	1	1	10	0	0	3	0	2	5	22

Segment	overhead signs	parking meters	rocks	other	total	pole	length	max grade	crest	terrain type	residential
AS	0	0	0	1	47	26	1145	1.7	2.9	level	0
BS	0	0	0	0	33	13	525	1.9	2.7	level	0
CS	0	0	0	0	62	27	955	2.6	0.5	level	0
DS	0	0	0	0	171	68	2340	2	0.9	level	20
ES	0	0	0	1	133	78	1895	4.2	2.9	rolling	0
FS	0	0	0	1	45	32	490	4.3	4.1	rolling	0

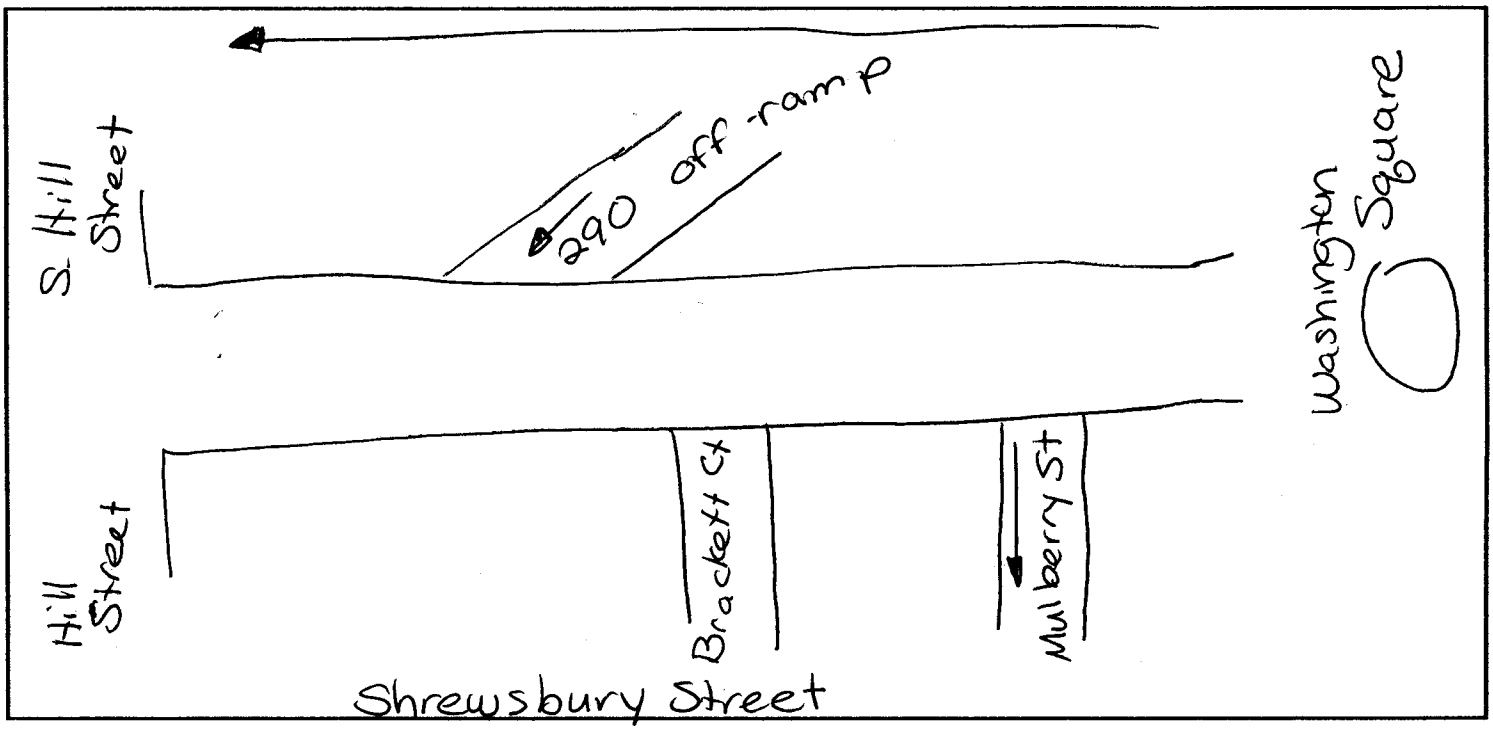
Segment	commercial	industrial	left lanes	right lanes	width	width	width	width	average width	type shoulder	sidewalk
AS	100	0	2	2	12	12	12	12	12	paved	yes
BS	100	0	2	2	12	12	12	12	12	paved	yes
CS	100	0	2	2	12	12	12	12	12	paved	yes
DS	80	0	2	2	12	12	12	12	12	paved	yes
ES	100	0	2	2	12	12	12	12	12	paved	yes
FS	100	0	2	2	12	12	12	12	12	paved	yes

Segment	average width	curb	drainage	pavement	markings	% parking	% lighting	SD	curve	curves	median type
AS	5	both	yes	bad	bad	50	100	0	1	1	none
BS	6	both	yes	fair	bad	80	100	0	0	0	curb&grass
CS	9	both	yes	bad	bad	100	100	0	0	0	curb&grass
DS	8	both	yes	bad	bad	100	100	0	0	0	curb&grass
ES	8	both	yes	bad	bad	80	100	0	0	0	curb&grass
FS	8	both	yes	good	fair	20	100	0	0	0	curb&grass

Segment	width
AS	0
BS	8
CS	0
DS	0
ES	0
FS	0

segment AS

4-7-04
Date: 9-23-03 Weather: Sunny



Posted Speed: 30
Minor Access Points (Road Names)
3

of driveways 𝍸 = 2
# of parking lots 𝍸 = 5

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 = 3	1	𝍸 𝍸 = 12	𝍸 = 2	𝍸 = 4
	monument	fences	buildings	sign poles	overhead sign
	1	1	𝍸 = 8	𝍸 = 14	0
	parking meter	rock	electrical box		
	0	0	1		

Section Length: 1145 ft

Vertical Grade: 1.7 % Crest on road: 2.9 %

Terrain Type: (level) rolling mountainous

Land Use % residential ____ commerical 100 % industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: _____ _____

type of shoulder: (paved) dirt none

width of shoulder: _____ _____

sidewalk present: (yes) no width: 4'8" _____

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good fair (bad)
describe: cracked, rutting

pavement marking quality good fair (bad)
describe: fading

parking allowed (yes) no 50 % allowed

road lighting: (present) not 100 %

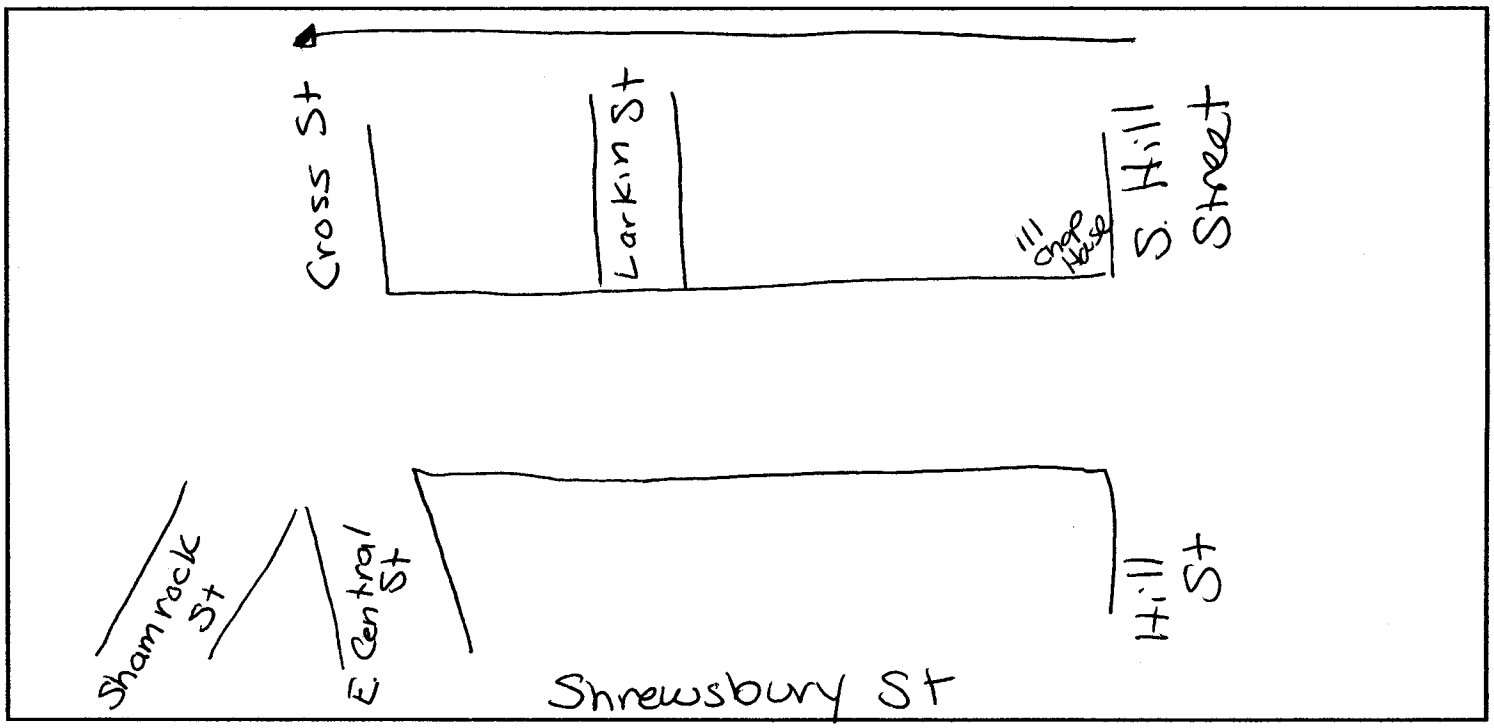
sight distance issues: (no) yes
describe:

horizontal curvature describe: straight (curve)
approximate curve length:
radius: ≈300

median type: grass paved w/ curb painted other (none)
width _____ ft _____ in

segment B5

Date: 4-7-04 / 9-23-03 Weather: Sunny



Posted Speed: 30

Minor Access Points (Road Names)

1

of driveways 0

# of parking lots 𝍸 1 = 6

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	1	𝍸 =2	𝍸 1 =6	0	𝍸 =3
	monument	fences	buildings	sign poles	overhead sign
	𝍸 =2	𝍸 =2	𝍸 𝍸 =10	𝍸 𝍸 =7	0
	parking meter	rock			
	0	0			

Section Length: 525 ft

Vertical Grade: 1.9 % Crest on road: 2.7 %

Terrain Type: (level) rolling mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: _____ _____

type of shoulder: (paved) dirt none

width of shoulder: _____ _____

sidewalk present: (yes) no width: 6' _____

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good (fair) bad
describe:

pavement marking quality good fair (bad)
describe:

parking allowed (yes) no 80 % allowed

road lighting: (present) not 100 %

sight distance issues: (no) yes
describe:

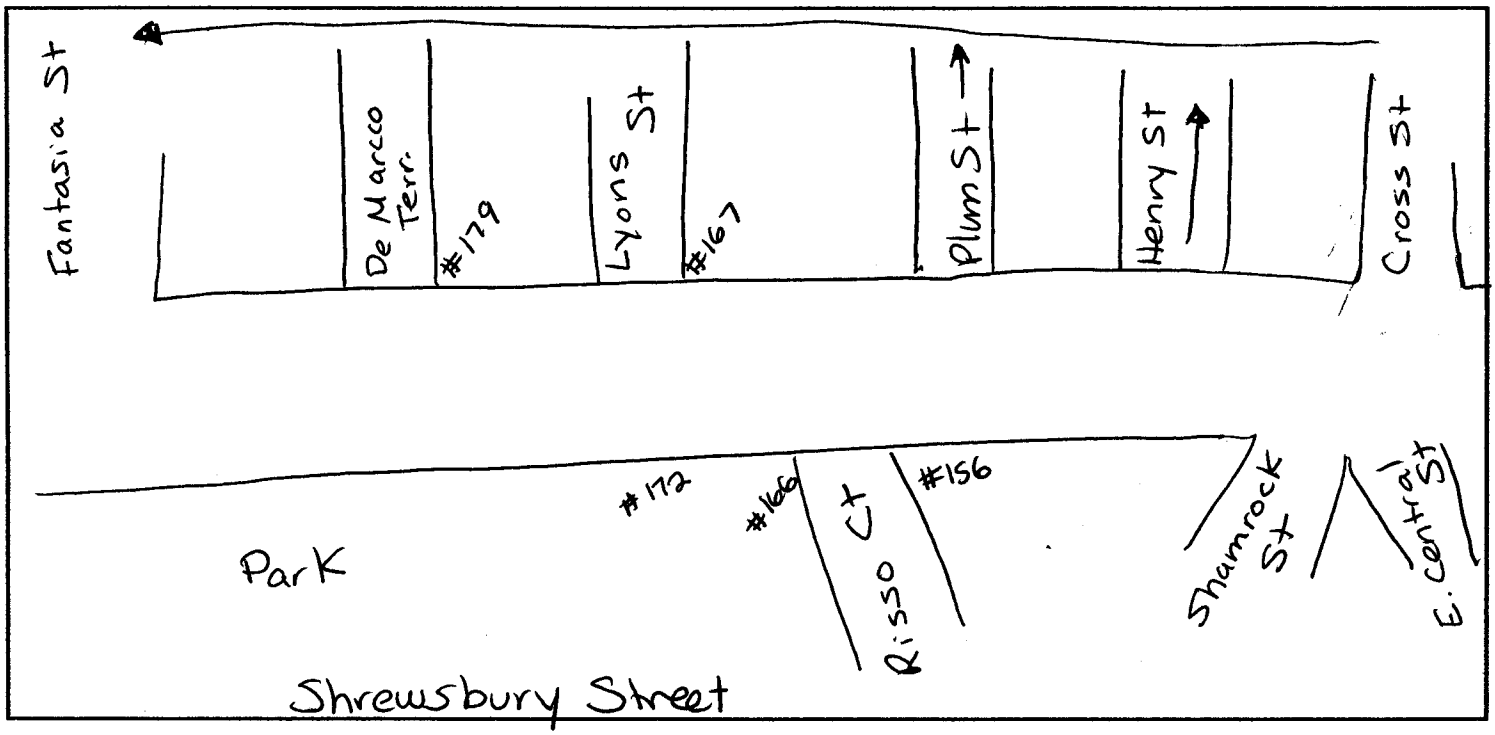
horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (grass + curb)
width 8 ft _____ in

Segment CS

4-7-04

Date: 9-24-03 Weather: Sunny



Posted Speed: 30

Minor Access Points (Road Names)

5

of driveways 1

# of parking lots 𝍸 1 = 6

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 = 3	𝍸 = 4	𝍸 𝍸 = 12	𝍸 = 4	𝍸 = 5
	monument	fences	buildings	sign poles	overhead sign
	𝍸 = 5	𝍸 = 3	𝍸 1 = 11	𝍸 𝍸 = 15	0
	parking meter	rock			
	0	0			

Section Length: 955 ft

Vertical Grade: 2.6 % Crest on road: 0.5 %

Terrain Type: (level) rolling mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: _____ _____

type of shoulder: (paved) dirt none

width of shoulder: _____ _____

sidewalk present: (yes) no width: 9' _____

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good fair (bad)
describe:

pavement marking quality good fair (bad)
describe:

parking allowed (yes) no 100 % allowed

road lighting: (present) not 100 %

sight distance issues: (no) yes
describe:

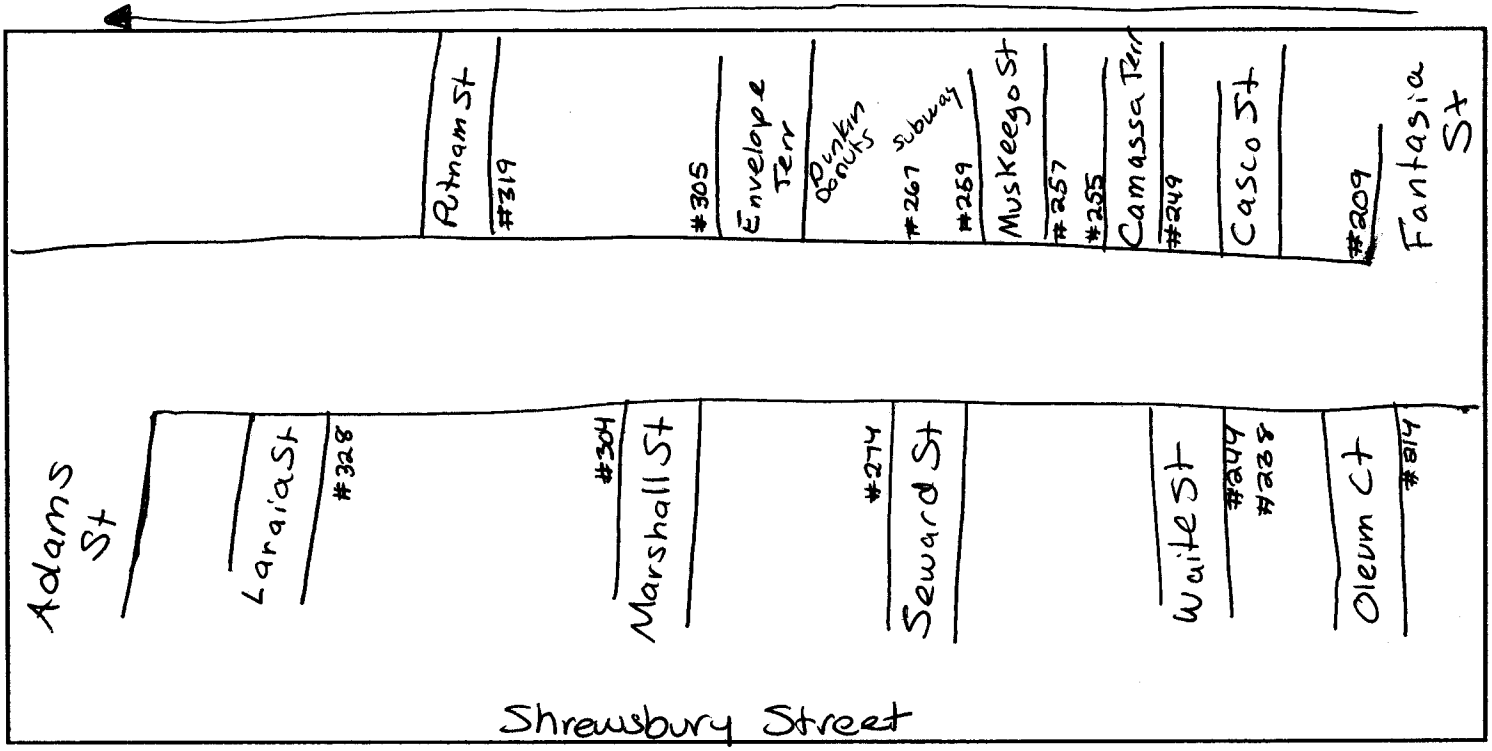
horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (grass + curb)
width _____ ft _____ in

segment D.S

4-7-04

Date: 10-3-03 Weather: sunny



Posted Speed: 30

Minor Access Points (Road Names)

10

of driveways 𝍸 III = 8

# of parking lots 𝍸 𝍸 𝍸 𝍸 III = 23

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 II =12	𝍸 I =11	𝍸 𝍸 III =23	1	𝍸 𝍸 𝍸 𝍸 IIII =34
	monument	fences	buildings	sign poles	overhead sign
	III =3	IIII 𝍸 =9	𝍸 𝍸 𝍸 III =33	𝍸 𝍸 𝍸 𝍸 =45	0
	parking meter	rock			
	0	0			

Section Length: 2340 ft

Vertical Grade: 20 % Crest on road: 0.9 %

Terrain Type: (level) rolling mountainous

Land Use % residential 20% commerical 80% industrial _____

lanes: Going Left: 2 Going Right: 2

width of lanes: _____ _____

type of shoulder: (paved) dirt none

width of shoulder: _____ _____

sidewalk present: (yes) no width: 8' 8'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good fair (bad)
describe:

pavement marking quality good fair (bad)
describe:

parking allowed (yes) no 100 % allowed

road lighting: (present) not 100 %

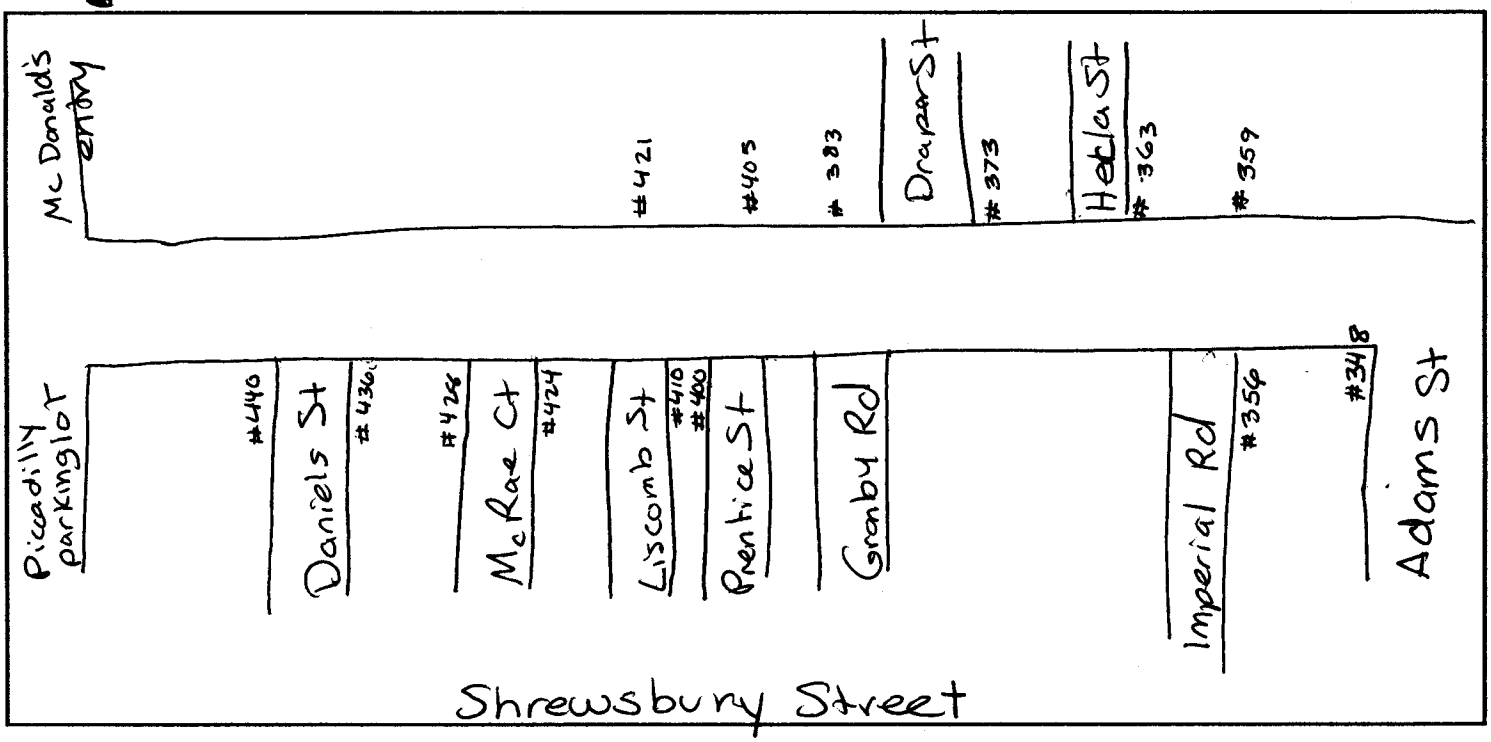
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (grass + curb)
width _____ ft _____ in

Segment ES

4-7-04
Date: 10-3-03 Weather: Sunny



Posted Speed: 30
Minor Access Points (Road Names)
8

of driveways 𝍸 = 5
# of parking lots 𝍸 𝍸 || = 12

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	𝍸 ||| = 8	||| = 3	𝍸 𝍸 |||| = 29	1	𝍸 = 9
	monument	fences	buildings	sign poles	overhead sign
	|| = 2	|||| = 4	𝍸 𝍸 || = 27	𝍸 𝍸 𝍸 |||| = 49	0
	parking meter	rock	electric box		
	0	0	1		

Section Length: 1895 ft

Vertical Grade: 4.2 % Crest on road: 2.9 %

Terrain Type: level (rolling) mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: _____ _____

type of shoulder: (paved) dirt none

width of shoulder: _____ _____

sidewalk present: (yes) no width: 8' 8'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: good fair (bad)
describe: cracks, patching, rutting

pavement marking quality good fair (bad)
describe: faded - mostly gone

parking allowed (yes) no 80 % allowed

road lighting: (present) not 100 %

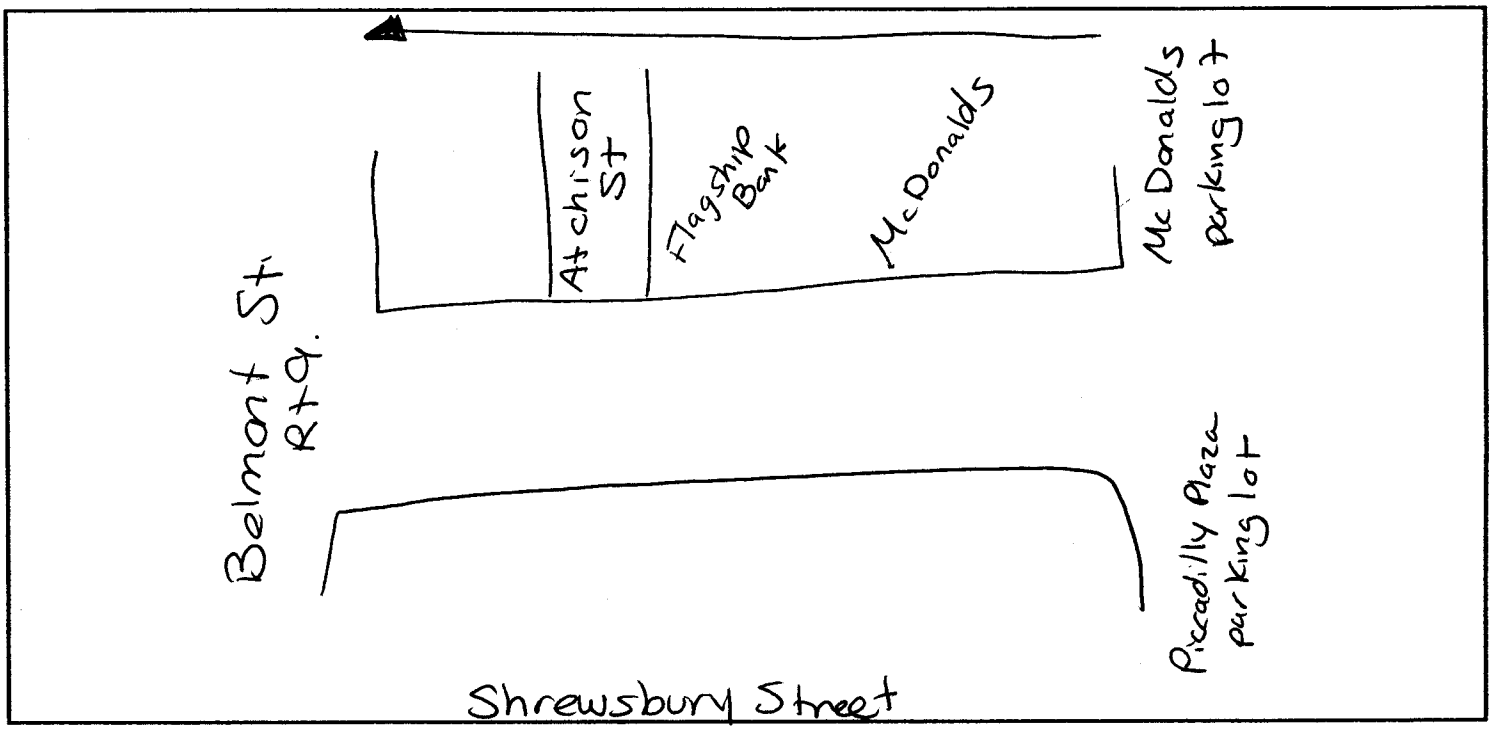
sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (grass + curb)
width _____ ft _____ in

segment FS

4-7-04
Date: 10-3-03 Weather: sunny



Posted Speed: 30
Minor Access Points (Road Names)
1

of driveways 0
# of parking lots IIII = 4

roadside hazards:	firehydrants	mailboxes	utility/light poles	benches	trees
	1	1	𝍸 = 10	0	III = 3
	monument	fences	buildings	sign poles	overhead sign
	0	II = 2	𝍸 = 5	𝍸 𝍸 II = 22	0
	parking meter	rock	electrical box		
	0	0	1		

Section Length: 490 ft

Vertical Grade: 4.3 % Crest on road: 4.1 %

Terrain Type: level (rolling) mountainous

Land Use % residential ____ commerical 100% industrial ____

lanes: Going Left: 2 Going Right: 2

width of lanes: _____ _____

type of shoulder: (paved) dirt none

width of shoulder: _____ _____

sidewalk present: (yes) no width: 8' 8'

curb present: (both) no _____

drainage present: (yes) no

pavement quality: (good) fair bad
describe:

pavement marking quality good (fair) bad
describe: fading

parking allowed (yes) no 20 % allowed

road lighting: (present) not 100 %

sight distance issues: (no) yes
describe:

horizontal curvature describe: (straight) curve
approximate curve length:
radius:

median type: grass paved w/ curb painted other (grass + curb)
width _____ ft _____ in

C Appendix: SAS Code and Output

This appendix has SAS code that was used to create the three final models for this paper. The main output from that code is also included. The code and output for the linear total accident rate model is first starting on page C-2 followed by the multiplicative model starting on page C-10. The injury accident rate model's code and output is the last part of this appendix starting on page C-21.

SAS Code and Output

```
options ls=70 ps=65 pageno=1 nocenter;
```

Linear: Total Accident Rate Model and graphical diagnostics Code

```
proc reg data=thesis;
     model rate=ospole parkinglots residential curves crest
parking/p r clb clm cli ss1 ss2 pcorr2;
     output out=thesis1 p=pred r=resid student=student;
     run;
     */studentized residuals vs predicted*/;
symbol i=none v=dot c=red width=1;
proc gplot data=thesis1;
     plot student*pred/vref=0;
run;
          */boxplot of residuals*/;
symbol i=boxt c=red width=5;
proc gplot data=thesis1;
     plot resid*boxplot;
run;
     */residual vs predicted */;
symbol i=none v=dot c=red width=1;
proc gplot data=thesis1;
     plot resid*pred/vref=0;
run;
     */normal quantile plot*/;
symbol v=dot w=1 i=none c=red;
proc univariate data=thesis1;
     var resid;
     qqplot resid/normal (L=1 mu=est sigma=est);
     run;
     */normal probability plot*/;
proc univariate data=thesis1 plot;
var resid;
histogram resid/normal kernal (L=2);
run;
```

SAS Output
Linear: Total Accident Rate Model

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	4515.67464	752.61244	10.52	<.0001
Error	20	1431.33888	71.56694		
Corrected Total	26	5947.01352			

Root MSE	8.45972	R-Square	0.7593
Dependent Mean	23.14741	Adj R-Sq	0.6871
Coeff Var	36.54718		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.37975	6.16058	1.68	0.1076
ospole	1	4.92593	1.27469	3.86	0.0010
parkinglots	1	-1.65140	0.29927	-5.52	<.0001
residential	1	-0.32989	0.05145	-6.41	<.0001
curves	1	6.88388	4.41977	1.56	0.1350
crest	1	3.28343	1.15060	2.85	0.0098
parking	1	0.12701	0.05631	2.26	0.0355

Parameter Estimates

Variable	DF	Type I SS	Type II SS	Corr	Squared Partial Type II
Intercept	1	14467	203.16183		.
ospole	1	798.66967	1068.76438		0.42749
parkinglots	1	570.80662	2179.17560		0.60356
residential	1	1790.74362	2942.29802		0.67273
curves	1	479.64386	173.61223		0.10817
crest	1	511.73310	582.80331		0.28936
parking	1	364.07778	364.07778		0.20278

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	-2.47101	23.23050
ospole	1	2.26698	7.58488

parkinglots	1	-2.27566	-1.02713
residential	1	-0.43721	-0.22257
curves	1	-2.33559	16.10336
crest	1	0.88333	5.68354
parking	1	0.00955	0.24447

Output Statistics

Obs	Dep Var rate	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	48.9100	45.6876	5.2262	34.7860	56.5893
2	8.2600	18.0749	4.6679	8.3377	27.8120
3	38.9300	31.3467	2.5308	26.0677	36.6258
4	23.2400	32.5316	4.6389	22.8550	42.2082
5	29.4400	30.2618	3.5020	22.9567	37.5669
6	55.3400	45.3906	4.5306	35.9401	54.8412
7	11.3000	6.0561	6.1864	-6.8486	18.9608
8	6.9200	4.8574	5.2732	-6.1424	15.8572
9	9.5900	4.1594	4.2138	-4.6305	12.9493
10	26.4300	34.7040	3.9436	26.4778	42.9303
11	24.5500	36.9293	2.9476	30.7808	43.0779
12	30.3700	27.2705	3.9494	19.0322	35.5089
13	44.5000	34.4644	4.7857	24.4817	44.4472
14	12.0800	14.2252	3.8996	6.0908	22.3596
15	2.2000	11.4314	3.2849	4.5793	18.2836
16	15.2600	21.5142	3.7228	13.7485	29.2800
17	21.9200	14.2608	4.5560	4.7572	23.7645
18	13.0500	15.1797	3.4919	7.8957	22.4636
19	55.5500	47.4363	5.1358	36.7233	58.1494
20	35.2800	35.9613	4.9762	25.5813	46.3414
21	23.2100	13.3166	3.1793	6.6848	19.9485
22	15.1900	15.9486	3.8230	7.9740	23.9232
23	7.1300	5.8066	4.4594	-3.4956	15.1087
24	20.0400	14.5142	3.4725	7.2706	21.7577
25	7.9300	10.7204	5.2396	-0.2092	21.6500
26	26.6500	24.3240	3.8704	16.2504	32.3976
27	11.7100	28.6061	4.6398	18.9276	38.2847

Output Statistics

Obs	95% CL Predict		Residual	Std Error Residual	Student Residual
1	24.9451	66.4301	3.2224	6.652	0.484
2	-2.0799	38.2297	-9.8149	7.055	-1.391

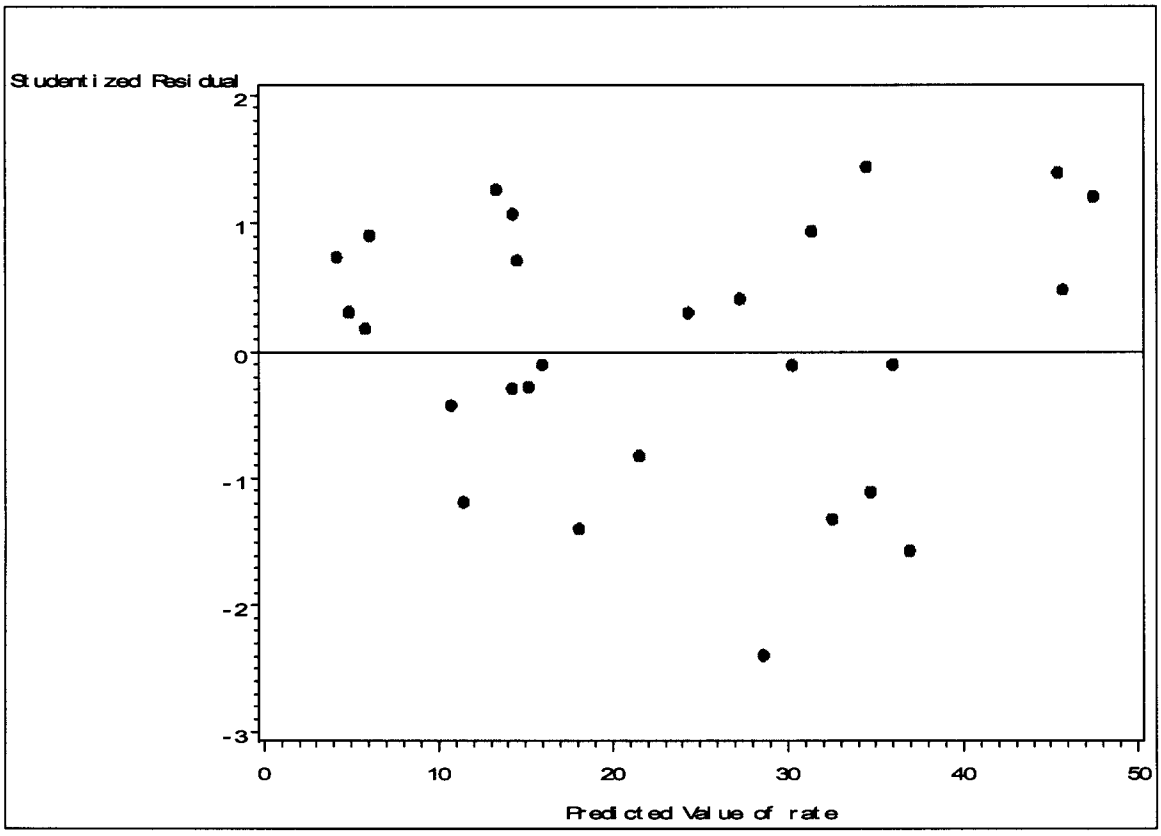
3	12.9273	49.7661	7.5833	8.072	0.939
4	12.4059	52.6573	-9.2916	7.074	-1.313
5	11.1628	49.3607	-0.8218	7.701	-0.107
6	25.3727	65.4086	9.9494	7.144	1.393
7	-15.8057	27.9178	5.2439	5.770	0.909
8	-15.9369	25.6516	2.0626	6.615	0.312
9	-15.5553	23.8740	5.4306	7.336	0.740
10	15.2342	54.1739	-8.2740	7.484	-1.106
11	18.2422	55.6165	-12.3793	7.930	-1.561
12	7.7955	46.7455	3.0995	7.481	0.414
13	14.1898	54.7391	10.0356	6.976	1.439
14	-5.2061	33.6564	-2.1452	7.507	-0.286
15	-7.4989	30.3618	-9.2314	7.796	-1.184
16	2.2344	40.7941	-6.2542	7.597	-0.823
17	-5.7822	34.3039	7.6592	7.128	1.075
18	-3.9112	34.2705	-2.1297	7.705	-0.276
19	26.7923	68.0803	8.1137	6.722	1.207
20	15.4882	56.4345	-0.6813	6.841	-0.0996
21	-5.5351	32.1683	9.8934	7.840	1.262
22	-3.4163	35.3135	-0.7586	7.547	-0.101
23	-14.1417	25.7549	1.3234	7.189	0.184
24	-4.5613	33.5897	5.5258	7.714	0.716
25	-10.0368	31.4776	-2.7904	6.642	-0.420
26	4.9181	43.7299	2.3260	7.522	0.309
27	8.4796	48.7327	-16.8961	7.074	-2.389

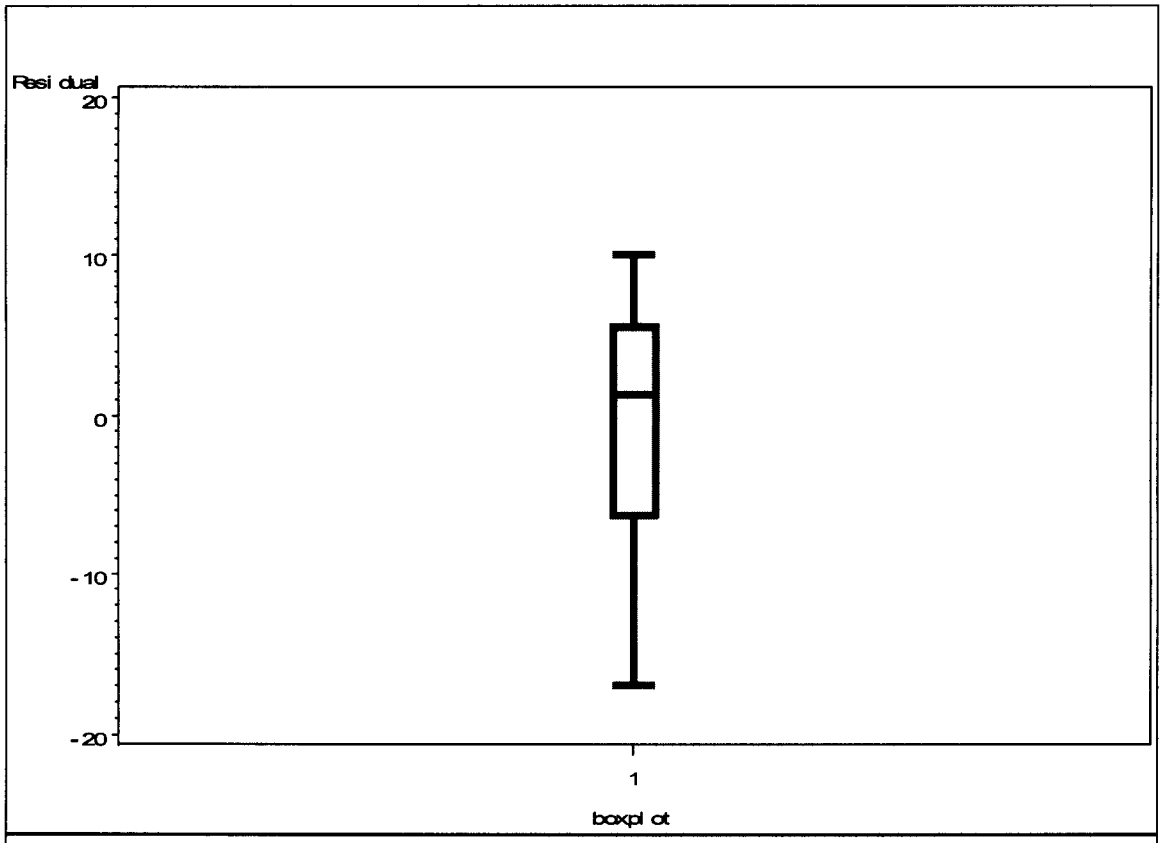
Output Statistics

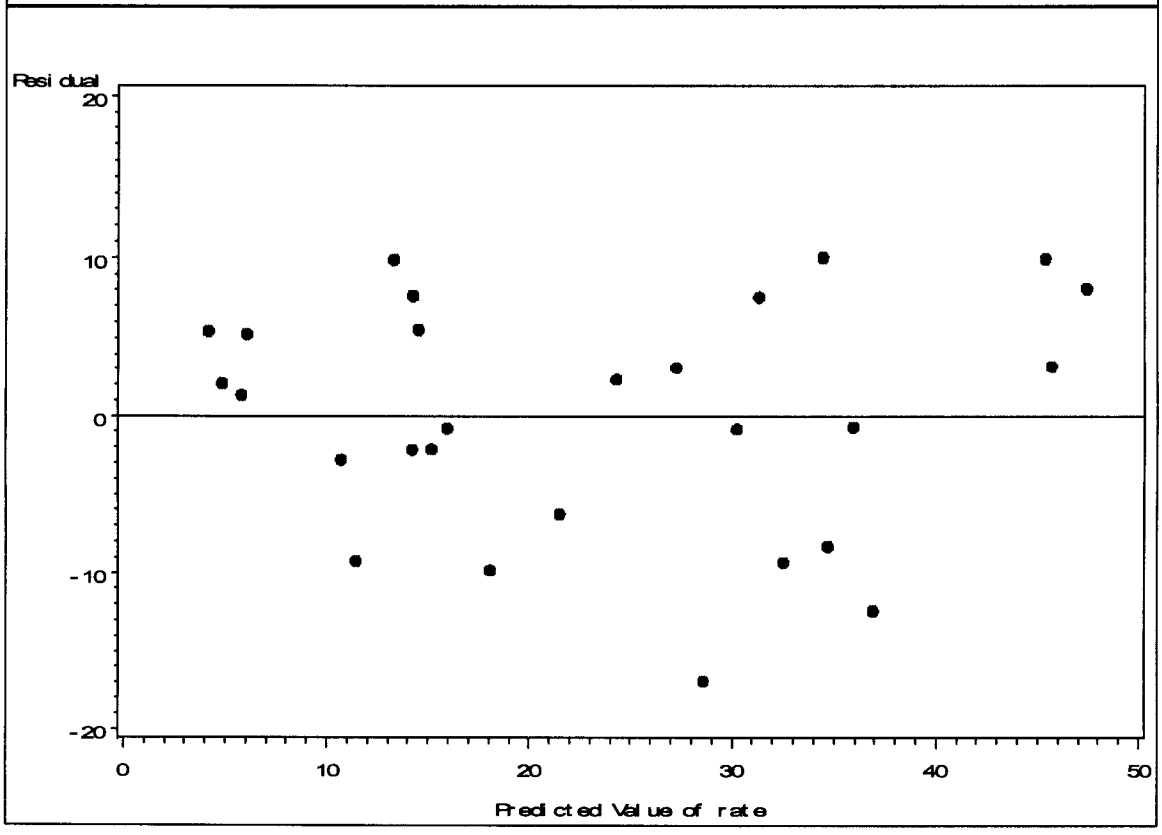
Obs	-2-1 0 1 2	Cook's D
1	| | |	0.021
2	| ** | |	0.121
3	| | * |	0.012
4	| ** | |	0.106
5	| | |	0.000
6	| | ** |	0.111
7	| | * |	0.136
8	| | |	0.009
9	| | * |	0.026
10	| ** | |	0.048
11	| *** | |	0.048
12	| | |	0.007
13	| | ** |	0.139
14	| | |	0.003
15	| ** | |	0.036
16	| * | |	0.023

17		**	0.067
18			0.002
19		**	0.121
20			0.001
21		**	0.037
22			0.000
23			0.002
24		*	0.015
25			0.016
26			0.004
27	****		0.351

Sum of Residuals	0
Sum of Squared Residuals	1431.33888
Predicted Residual SS (PRESS)	2671.35680







The UNIVARIATE Procedure

Variable: resid (Residual)

Moments

N	27	Sum Weights	27
Mean	0	Sum Observations	0
Std Deviation	7.41966949	Variance	55.0514954
Skewness	-0.525266	Kurtosis	-0.5223208
Uncorrected SS	1431.33888	Corrected SS	1431.33888
Coeff Variation	.	Std Error Mean	1.42791606

Basic Statistical Measures

Location		Variability	
Mean	0.000000	Std Deviation	7.41967
Median	1.323432	Variance	55.05150
Mode	.	Range	26.93169
		Interquartile Range	11.78004

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	0.5	Pr >= |M|	1.0000
Signed Rank	S	11	Pr >= |S|	0.7972

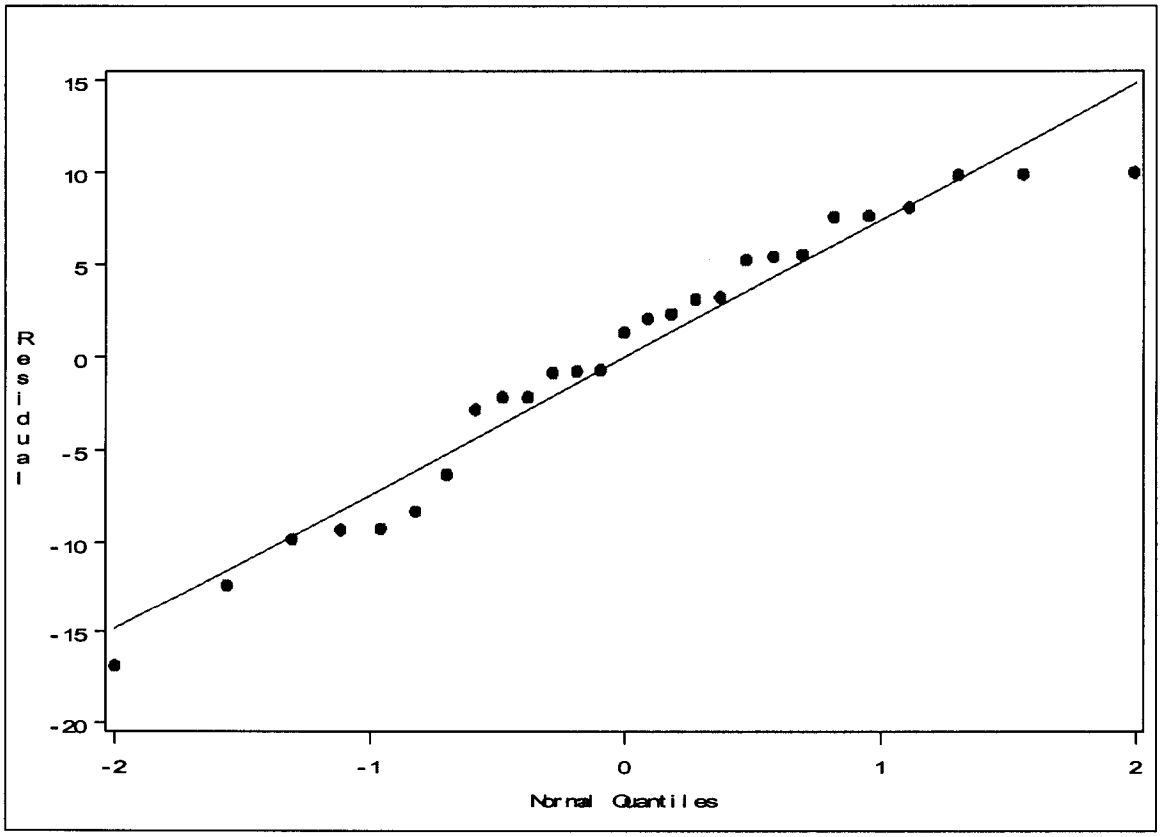
Quantiles (Definition 5)

Quantile	Estimate
100% Max	10.03556
99%	10.03556
95%	9.94936
90%	9.89336
75% Q3	5.52581
50% Median	1.32343
25% Q1	-6.25423
10%	-9.81488

5%	-12.37934
1%	-16.89613
0% Min	-16.89613

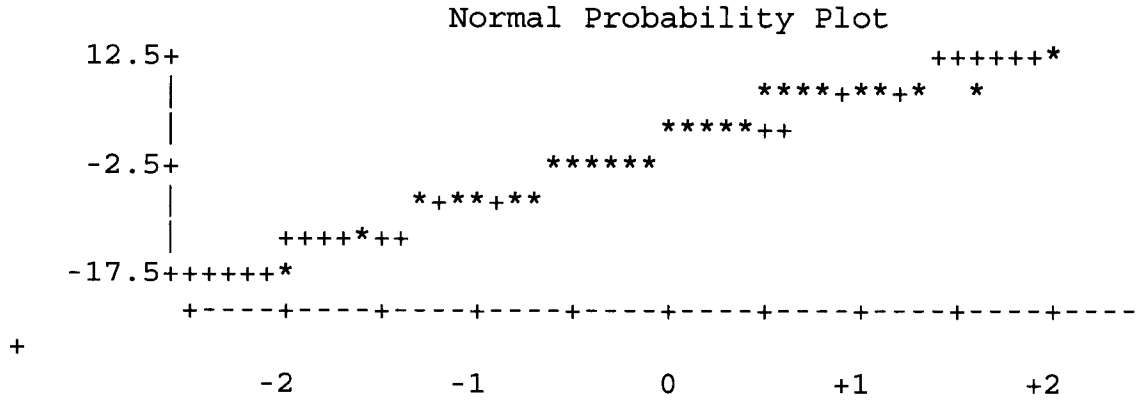
Extreme Observations

------Lowest------		------Highest-----	
Value	Obs	Value	Obs
-16.89613	27	7.65915	17
-12.37934	11	8.11367	19
-9.81488	2	9.89336	21
-9.29160	4	9.94936	6
-9.23143	15	10.03556	13



The UNIVARIATE Procedure
Variable: resid (Residual)

```
   Stem Leaf                     #  Boxplot
      1 000                      3      |
      0 556888                   6  +-----+
      0 12233                    5  *--+--*
     -0 322111                   6  |     |
     -0 9986                     4  +-----+
     -1 20                       2      |
     -1 7                        1      |
        ----+----+----+----+
    Multiply Stem.Leaf by 10**+1
```



The UNIVARIATE Procedure
Fitted Distribution for resid

Parameters for Normal Distribution

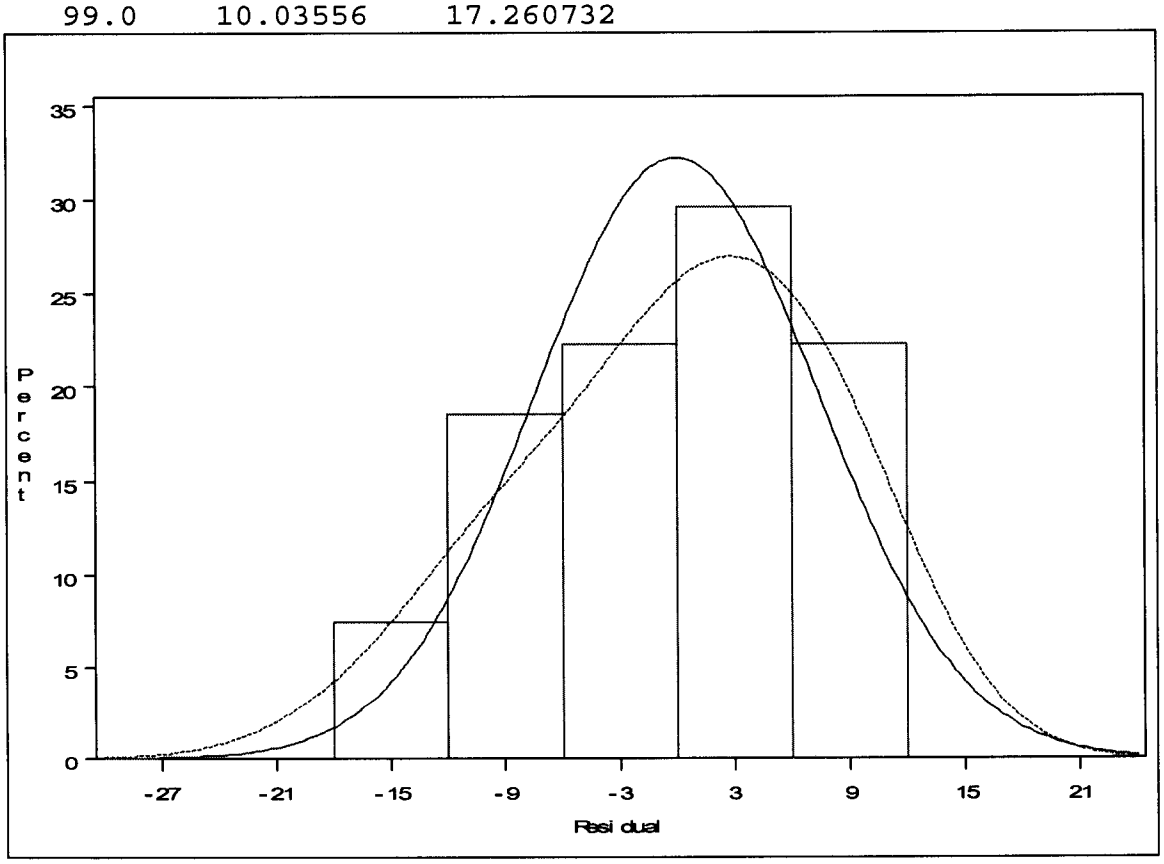
Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	7.419669

Goodness-of-Fit Tests for Normal Distribution

Test	---Statistic----		-----p Value-----	
Kolmogorov-Smirnov	D	0.09416938	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.06418911	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.43319375	Pr > A-Sq	>0.250

Quantiles for Normal Distribution

Percent	Quantile Observed	Quantile Estimated
1.0	-16.89613	-17.260732
5.0	-12.37934	-12.204270
10.0	-9.81488	-9.508689
25.0	-6.25423	-5.004491
50.0	1.32343	-0.000000
75.0	5.52581	5.004491
90.0	9.89336	9.508689
95.0	9.94936	12.204270
99.0	10.03556	17.260732



Multiplicative: Total Accident Rate Model and graphical diagnostics Code

```
proc reg data=thesis;
     model lrate=llength llighting lpole/p r;
     output out=thesis2 p=pred r=resid student=student;
     run;
*/studentized residuals vs predicted*/;
symbol i=none v=dot c=red width=1;
proc gplot data=thesis2;
     plot student*pred/vref=0;
run;
*/boxplot of residuals*/;
symbol i=boxt c=red width=5;
proc gplot data=thesis2;
     plot resid*boxplot;
run;
     */residual vs predicted */;
symbol i=none v=dot c=red width=1;
proc gplot data=thesis2;
     plot resid*pred/vref=0;
run;
     */normal quantile plot*/;
symbol v=dot w=1 i=none c=red;
proc univariate data=thesis2;
     var resid;
     qqplot resid/normal (L=1 mu=est sigma=est);
     run;
     */normal probability plot*/;
proc univariate data=thesis2 plot;
var resid;
histogram resid/normal kernal (L=2);
run;
```

SAS Output

Multiplicative: Total Accident Rate Model

The REG Procedure
Model: MODEL1
Dependent Variable: lrate

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	10.05655	3.35218	15.37	<.0001
Error	23	5.01511	0.21805		
Corrected Total	26	15.07167			

Root MSE	0.46696	R-Square	0.6672
Dependent Mean	2.90341	Adj R-Sq	0.6238
Coeff Var	16.08302		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-28.29984	10.22157	-2.77	0.0109
llength	1	-0.95851	0.25768	-3.72	0.0011
llighting	1	7.80913	2.17984	3.58	0.0016
lpole	1	0.56736	0.30369	1.87	0.0745

The REG Procedure
Model: MODEL1
Dependent Variable: lrate

Output Statistics

Obs	Dep Var lrate	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual
1	3.8900	3.5047	0.1447	0.3853	0.444	0.868
2	2.1114	2.7666	0.1114	-0.6551	0.453	-1.445
3	3.6618	3.7813	0.1888	0.1195	0.427	-0.280
4	3.1459	3.3623	0.1677	-0.2164	0.436	-0.497
5	3.3824	2.8766	0.1091	0.5058	0.454	1.114
6	4.0135	3.1065	0.1187	0.9070	0.452	2.008

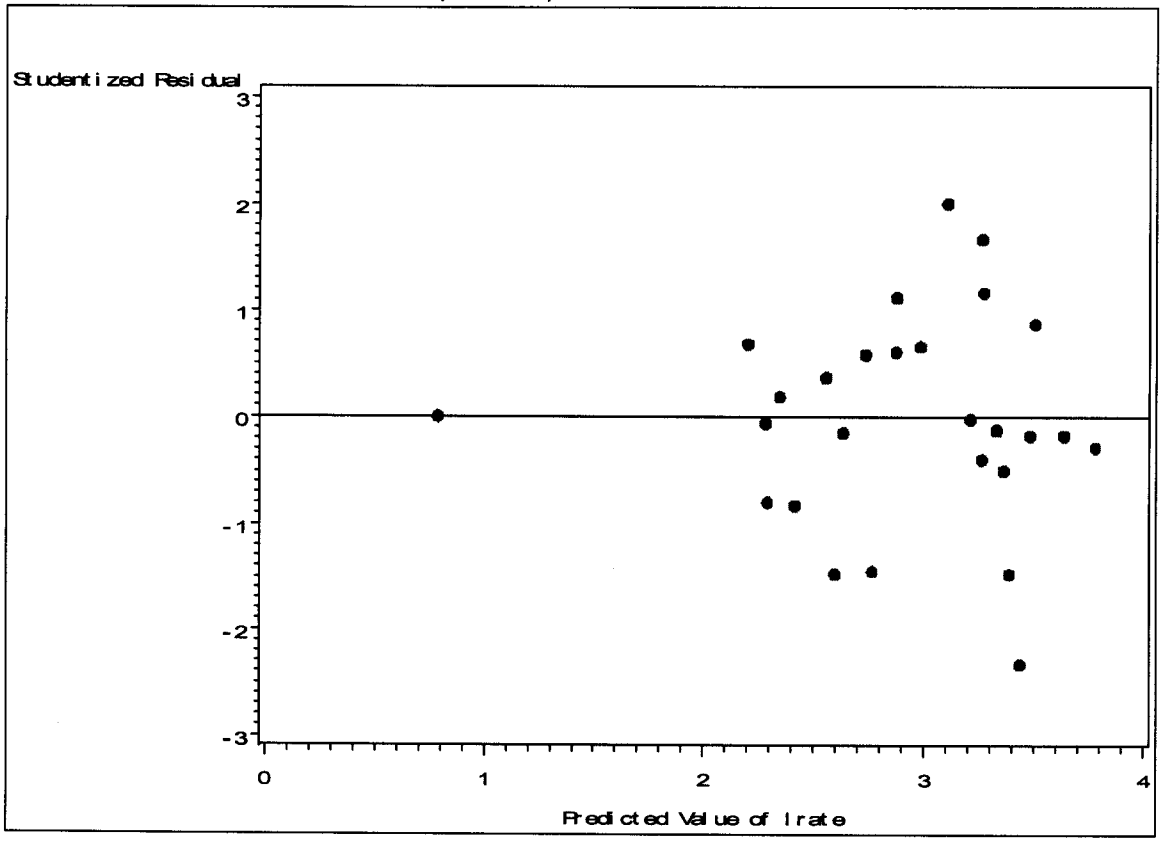
7	2.4248	2.3465	0.1832	0.0783	0.429	0.182
8	1.9344	2.5974	0.1226	-0.6630	0.451	-1.471
9	2.2607	2.2842	0.2410	-0.0235	0.400	-0.0588
10	3.2745	3.3294	0.1463	-0.0549	0.443	-0.124
11	3.2007	3.2104	0.1395	-0.009735	0.446	-0.0218
12	3.4135	3.4835	0.2324	-0.0700	0.405	-0.173
13	3.7955	3.2712	0.1191	0.5243	0.452	1.161
14	2.4916	2.2026	0.1877	0.2889	0.428	0.676
15	0.7885	0.7885	0.4670	-1.89E-11	1.62E-6	-16E-7
16	2.7252	3.3875	0.1254	-0.6623	0.450	-1.472
17	3.0874	3.2624	0.1312	-0.1750	0.448	-0.390
18	2.5688	2.6345	0.1506	-0.0657	0.442	-0.149
19	4.0173	3.2636	0.1190	0.7537	0.452	1.669
20	3.5633	3.6377	0.1639	-0.0744	0.437	-0.170
21	3.1446	2.8736	0.1192	0.2710	0.451	0.600
22	2.7206	2.5582	0.1292	0.1624	0.449	0.362
23	1.9643	2.2925	0.2159	-0.3282	0.414	-0.793
24	2.9977	2.7345	0.1095	0.2633	0.454	0.580
25	2.0707	2.4161	0.2064	-0.3454	0.419	-0.825
26	3.2828	2.9836	0.0923	0.2992	0.458	0.654
27	2.4604	3.4364	0.2068	-0.9760	0.419	-2.331

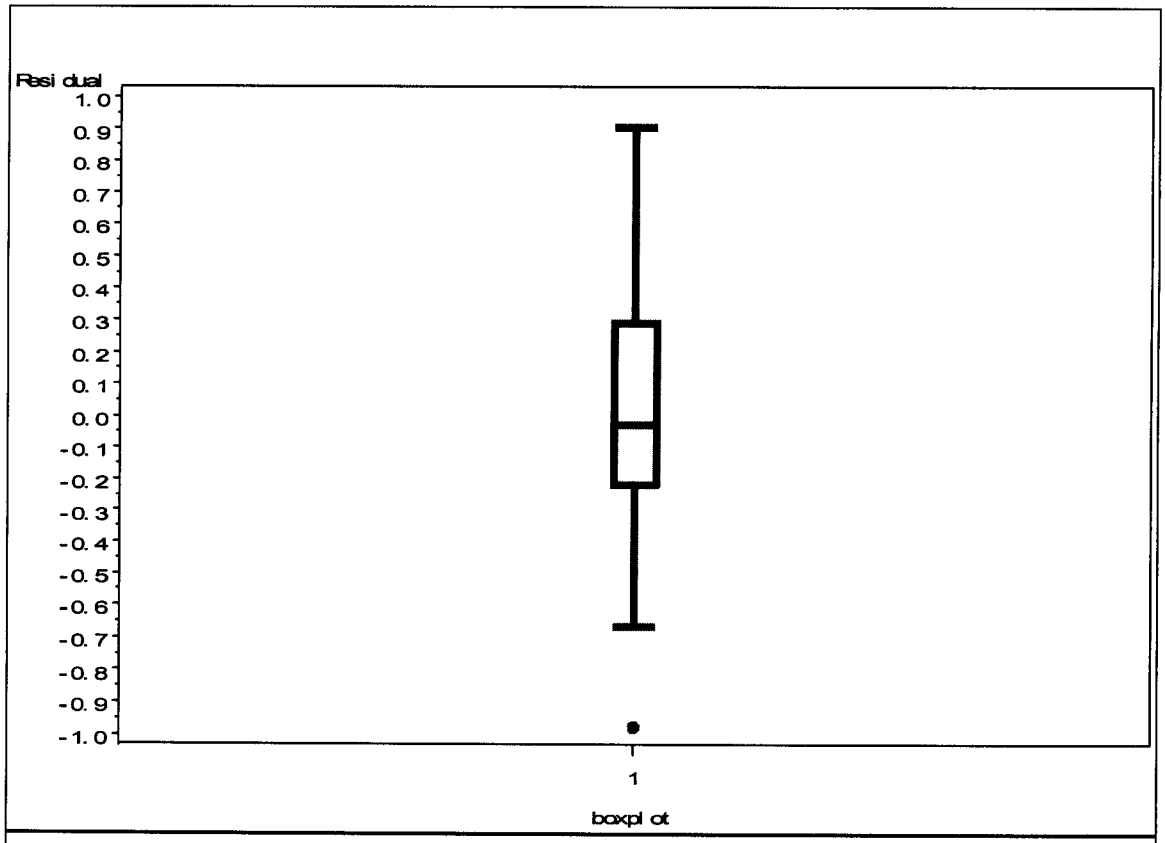
Output Statistics

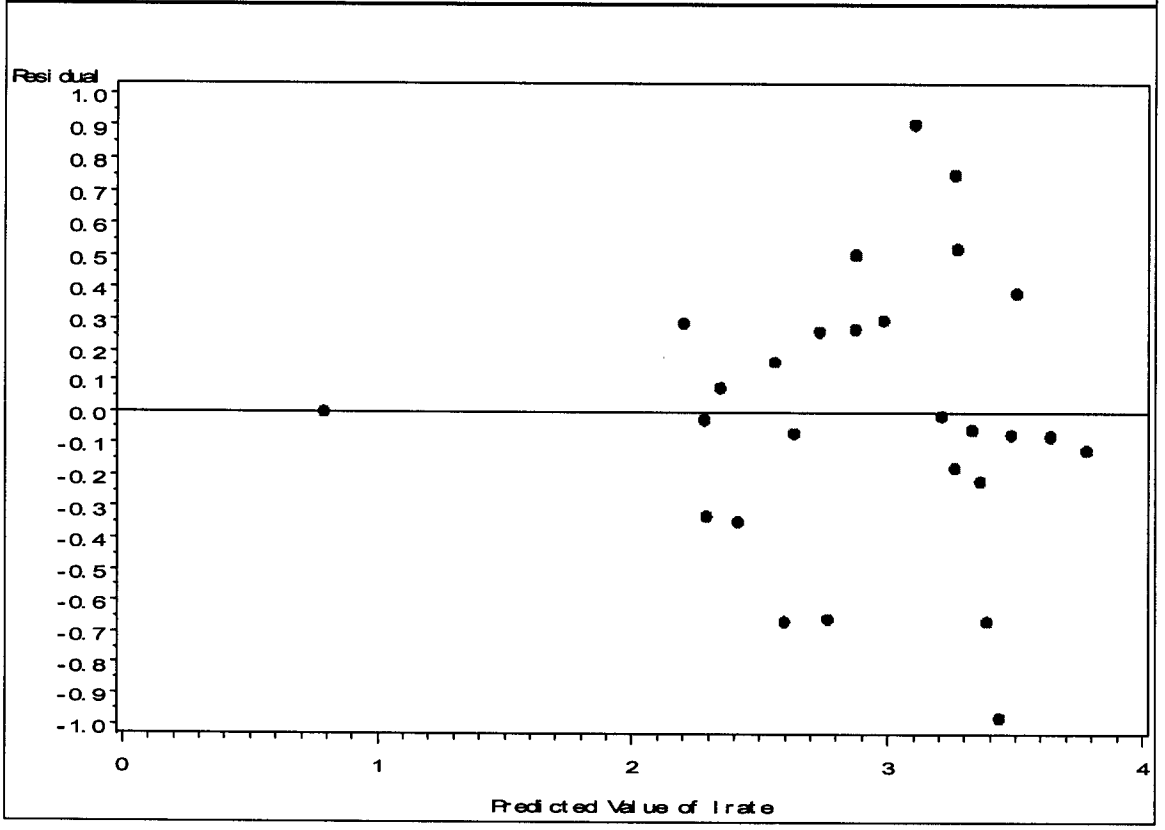
Obs	-2-1 0 1 2	Cook's D
1	| * |	0.020
2	| ** |	0.032
3	| |	0.004
4	| |	0.009
5	| ** |	0.018
6	| **** |	0.070
7	| |	0.002
8	| ** |	0.040
9	| |	0.000
10	| |	0.000
11	| |	0.000
12	| |	0.002
13	| ** |	0.023
14	| * |	0.022
15	| |	2.795
16	| ** |	0.042
17	| |	0.003
18	| |	0.001
19	| *** |	0.048
20	| |	0.001
21	| * |	0.006

22				0.003
23		*		0.043
24			*	0.005
25		*		0.041
26			*	0.004
27	****			0.331

Sum of Residuals	-5.0919E-13
Sum of Squared Residuals	5.01511
Predicted Residual SS (PRESS)	8.70301







The UNIVARIATE Procedure

Variable: resid (Residual)

Moments

N	27	Sum Weights	27
Mean	0	Sum Observations	0
Std Deviation	0.43919123	Variance	0.19288893
Skewness	-0.1416285	Kurtosis	0.12987266
Uncorrected SS	5.01511227	Corrected SS	5.01511227
Coeff Variation	.	Std Error Mean	0.08452239

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	0.43919
Median	-0.02350	Variance	0.19289
Mode	.	Range	1.88296
		Interquartile Range	0.50534

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-2.5	Pr >= |M|	0.4421
Signed Rank	S	-2	Pr >= |S|	0.9628

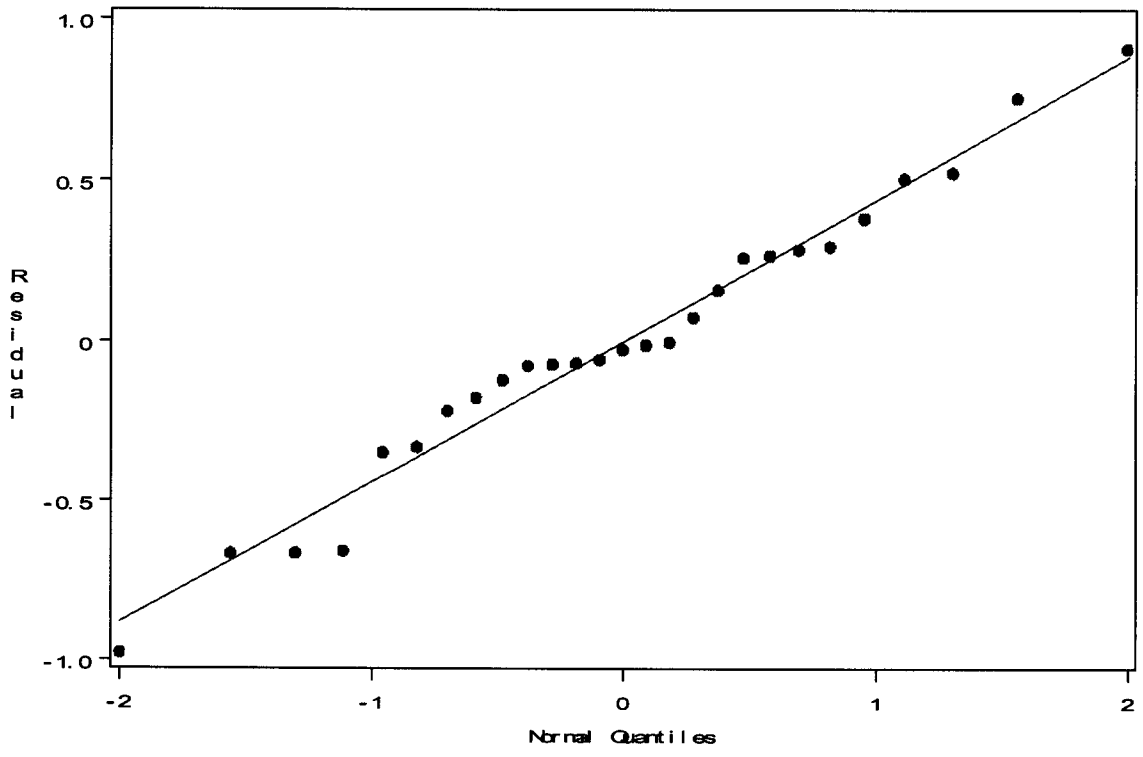
Quantiles (Definition 5)

Quantile	Estimate
100% Max	0.9070042
99%	0.9070042
95%	0.7536889
90%	0.5243173
75% Q3	0.2889438
50% Median	-0.0235034

25% Q1	-0.2164000
10%	-0.6622623
5%	-0.6630052
1%	-0.9759572
0% Min	-0.9759572

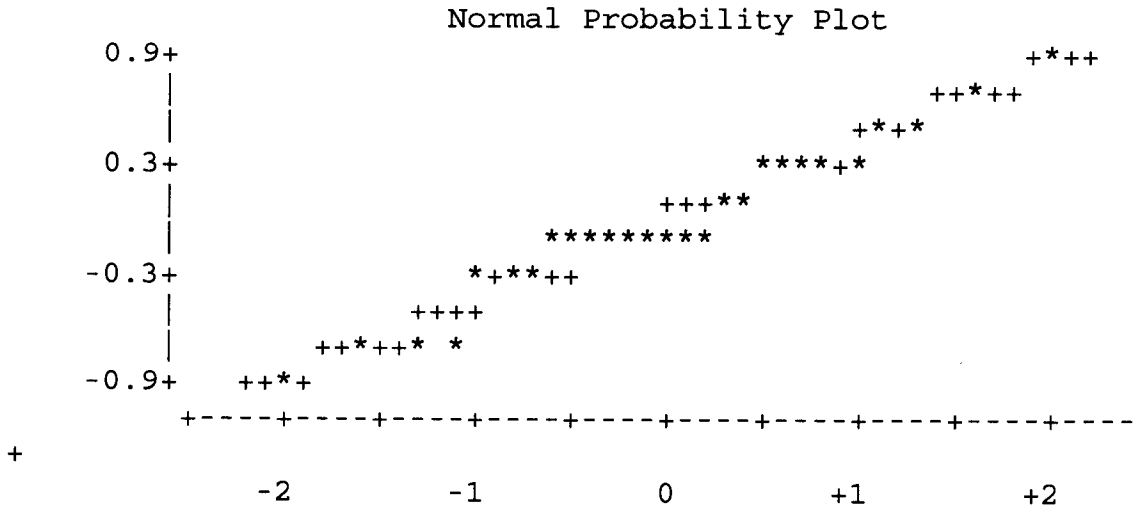
Extreme Observations

------Lowest------		------Highest-----	
Value	Obs	Value	Obs
-0.975957	27	0.385263	1
-0.663005	8	0.505782	5
-0.662262	16	0.524317	13
-0.655128	2	0.753689	19
-0.345435	25	0.907004	6



The UNIVARIATE Procedure
Variable: resid (Residual)

```
Stem Leaf                     #  Boxplot
   8 1                        1      |
   6 5                        1      |
   4 12                       2      |
   2 67909                    5  +-----+
   0 86                       2  |  +  |
  -0 727775210                9  *-----*
  -2 532                      3  +-----+
  -4                                |
  -6 666                      3     |
  -8 8                        1     0
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-1
```



The UNIVARIATE Procedure
Fitted Distribution for resid

Parameters for Normal Distribution

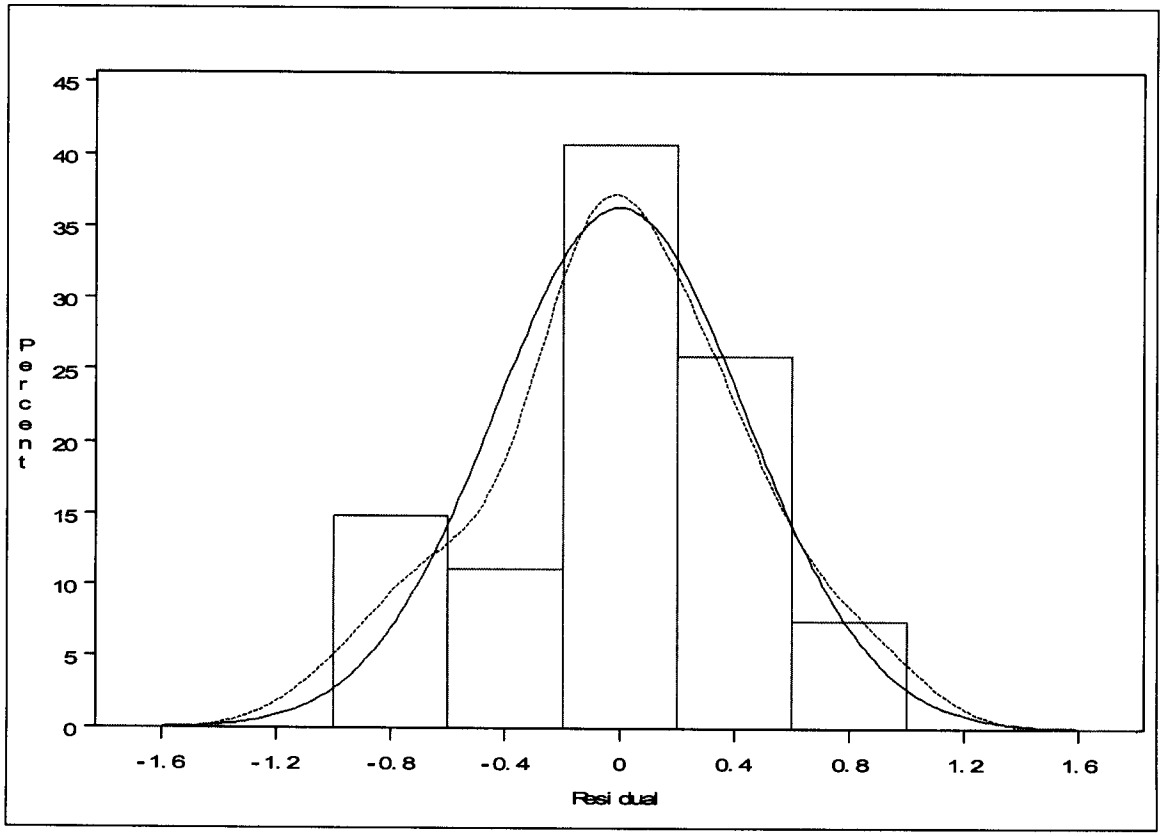
Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	0.439191

Goodness-of-Fit Tests for Normal Distribution

Test		---Statistic----		-----p Value-----
Kolmogorov-Smirnov	D	0.09940601	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.05035555	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.29758669	Pr > A-Sq	>0.250

Quantiles for Normal Distribution

	------Quantile------	
Percent	Observed	Estimated
1.0	-0.97596	-1.021712
5.0	-0.66301	-0.722405
10.0	-0.66226	-0.562846
25.0	-0.21640	-0.296230
50.0	-0.02350	-0.000000
75.0	0.28894	0.296230
90.0	0.52432	0.562846
95.0	0.75369	0.722405
99.0	0.90700	1.021712



Linear: Injury Accident Rate and graphical diagnostics Code

```
proc reg data=thesis;
     model rate=fence ospole hazards parkinglots vol
residential length grades curves crest widtha widthsida
pavement markings lighting/p r clb clm cli ss1 ss2 pcorr2;
     output out=thesis3 p=pred r=resid student=student;
     run;
     */studentized residuals vs predicted*/;
symbol i=none v=dot c=red width=1;
proc gplot data=thesis3;
     plot student*pred/vref=0;
run;
          */boxplot of residuals*/;
symbol i=boxt c=red width=5;
proc gplot data=thesis3;
     plot resid*boxplot;
run;
     */residual vs predicted */;
symbol i=none v=dot c=red width=1;
proc gplot data=thesis3;
     plot resid*pred/vref=0;
run;
     */normal quantile plot*/;
symbol v=dot w=1 i=none c=red;
proc univariate data=thesis3;
     var resid;
     qqplot resid/normal (L=1 mu=est sigma=est);
     run;
     */normal probability plot*/;
proc univariate data=thesis3 plot;
var resid;
histogram resid/normal kernal (L=2);
run;
```

SAS Output
Linear: Injury Accident Rate Model

The REG Procedure
Model: MODEL1
Dependent Variable: rate

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	5169.87253	344.65817	4.88	0.0058
Error	11	777.14098	70.64918		
Corrected Total	26	5947.01352			

Root MSE	8.40531	R-Square	0.8693
Dependent Mean	23.14741	Adj R-Sq	0.6911
Coeff Var	36.31209		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-193.35194	89.81139	-2.15	0.0544
fence	1	2.93044	1.29497	2.26	0.0449
ospole	1	4.86890	1.51462	3.21	0.0082
hazards	1	0.38116	0.17511	2.18	0.0522
parkinglots	1	-3.15499	0.65683	-4.80	0.0006
vol	1	0.00075693	0.00051916	1.46	0.1728
residential	1	-0.38469	0.11714	-3.28	0.0073

Parameter Estimates

Variable	DF	Type I SS	Type II SS	Corr	Squared Partial Type II
Intercept	1	14467	327.44729		.
fence	1	933.87327	361.78892		0.31766
ospole	1	715.58091	730.06986		0.48438
hazards	1	534.69451	334.72221		0.30105
parkinglots	1	18.34893	1630.01234		0.67715
vol	1	55.58497	150.17803		0.16195
residential	1	1686.75489	761.90869		0.49505

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	-391.02547	4.32159
fence	1	0.08024	5.78064
ospole	1	1.53526	8.20255
hazards	1	-0.00426	0.76658
parkinglots	1	-4.60067	-1.70930
vol	1	-0.00038575	0.00190
residential	1	-0.64251	-0.12686

The REG Procedure
Model: MODEL1
Dependent Variable: rate

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
length	1	-0.01504	0.01001	-1.50	0.1612
grade	1	-1.32956	0.89444	-1.49	0.1652
curves	1	7.75536	4.69458	1.65	0.1268
crest	1	6.26575	1.91241	3.28	0.0074
widtha	1	-0.50217	1.32845	-0.38	0.7126
widthsida	1	3.34846	1.38181	2.42	0.0338
pavement	1	-28.58297	11.10120	-2.57	0.0258
markings	1	3.61247	3.70068	0.98	0.3500
lighting	1	1.51331	0.78949	1.92	0.0816

Parameter Estimates

Variable	DF	Type I SS	Type II SS	Corr	Squared Partial Type II
length	1	58.39700	159.43970		0.17024
grade	1	0.68897	156.10644		0.16727
curves	1	177.70870	192.80457		0.19878
crest	1	175.99482	758.38432		0.49389
widtha	1	149.71812	10.09520		0.01282
widthsida	1	173.25068	414.85496		0.34803
pavement	1	226.86876	468.36239		0.37604
markings	1	2.83018	67.32135		0.07972
lighting	1	259.57783	259.57783		0.25038

Parameter Estimates

Variable	DF	95% Confidence Limits	
length	1	-0.03707	0.00699
grade	1	-3.29821	0.63909
curves	1	-2.57734	18.08805
crest	1	2.05655	10.47494
widtha	1	-3.42607	2.42173
widthsida	1	0.30710	6.38981
pavement	1	-53.01655	-4.14940
markings	1	-4.53267	11.75760
lighting	1	-0.22435	3.25096

The REG Procedure
Model: MODEL1
Dependent Variable: rate

Output Statistics

Obs	Dep Var rate	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	48.9100	52.1308	6.3534	38.1471	66.1146
2	8.2600	15.7924	6.6857	1.0772	30.5075
3	38.9300	44.8829	5.3699	33.0639	56.7019
4	23.2400	30.7329	4.4769	20.8792	40.5865
5	29.4400	24.1787	6.5355	9.7941	38.5632
6	55.3400	48.6952	6.4822	34.4280	62.9624
7	11.3000	8.1453	7.7265	-8.8606	25.1512
8	6.9200	3.7802	6.6439	-10.8429	18.4033
9	9.5900	10.8410	7.8919	-6.5289	28.2109
10	26.4300	22.4717	7.5614	5.8290	39.1143
11	24.5500	26.1746	6.9933	10.7825	41.5667
12	30.3700	28.7010	5.4258	16.7589	40.6431
13	44.5000	35.5084	5.6089	23.1632	47.8537
14	12.0800	17.6899	7.2281	1.7811	33.5988
15	2.2000	2.2000	8.4053	-16.3000	20.7000
16	15.2600	16.2404	6.2462	2.4927	29.9881
17	21.9200	17.3075	7.1061	1.6672	32.9479
18	13.0500	13.7902	6.3466	-0.1786	27.7589
19	55.5500	47.0950	6.7091	32.3283	61.8617
20	35.2800	41.9510	6.3961	27.8733	56.0286
21	23.2100	23.3214	5.4480	11.3304	35.3123
22	15.1900	12.3144	6.8274	-2.7126	27.3414
23	7.1300	8.4843	6.7752	-6.4278	23.3964

24	20.0400	16.1659	4.4007	6.4800	25.8519
25	7.9300	9.3351	7.2157	-6.5466	25.2167
26	26.6500	21.2945	4.4001	11.6099	30.9791
27	11.7100	25.7555	5.1974	14.3161	37.1948

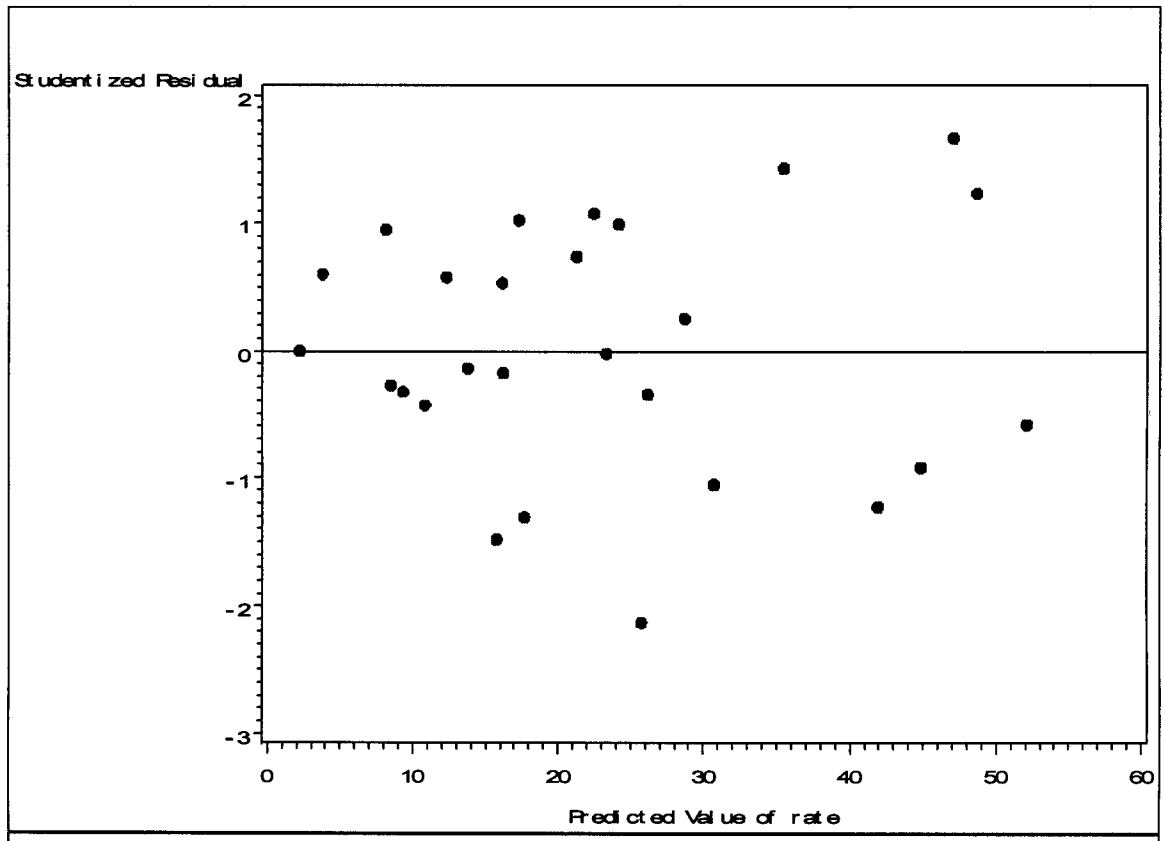
Output Statistics

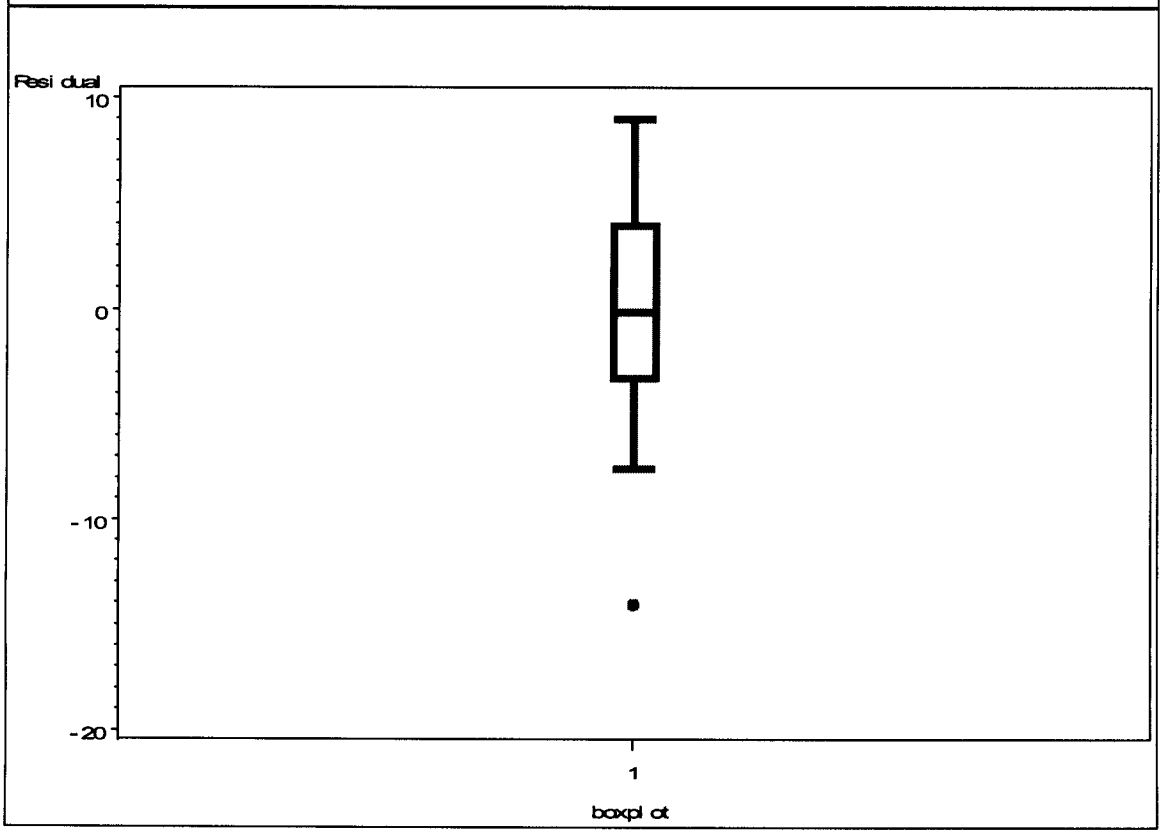
Obs	95% CL Predict		Residual	Std Error Residual	Student Residual
1	28.9404	75.3212	-3.2208	5.503	-0.585
2	-7.8463	39.4310	-7.5324	5.094	-1.479
3	22.9299	66.8360	-5.9529	6.466	-0.921
4	9.7723	51.6934	-7.4929	7.114	-1.053
5	0.7444	47.6129	5.2613	5.285	0.995
6	25.3328	72.0576	6.6448	5.351	1.242
7	-16.9834	33.2740	3.1547	3.309	0.953
8	-19.8012	27.3616	3.1398	5.149	0.610
9	-14.5354	36.2174	-1.2510	2.893	-0.432
10	-2.4126	47.3559	3.9583	3.671	1.078
11	2.1088	50.2405	-1.6246	4.663	-0.348
12	6.6814	50.7206	1.6690	6.419	0.260
13	13.2677	57.7492	8.9916	6.260	1.436
14	-6.7096	42.0895	-5.6099	4.290	-1.308
15	-23.9629	28.3629	-2.51E-12	1.85E-6	-136E-8
16	-6.8085	39.2892	-0.9804	5.624	-0.174
17	-6.9179	41.5329	4.6125	4.489	1.027
18	-9.3912	36.9715	-0.7402	5.511	-0.134
19	23.4242	70.7657	8.4550	5.063	1.670
20	18.7038	65.1981	-6.6710	5.453	-1.223
21	1.2753	45.3675	-0.1114	6.401	-0.0174
22	-11.5196	36.1484	2.8756	4.903	0.587
23	-15.2774	32.2460	-1.3543	4.975	-0.272
24	-4.7163	37.0481	3.8741	7.161	0.541
25	-15.0468	33.7169	-1.4051	4.311	-0.326
26	0.4130	42.1760	5.3555	7.162	0.748
27	4.0044	47.5065	-14.0455	6.606	-2.126

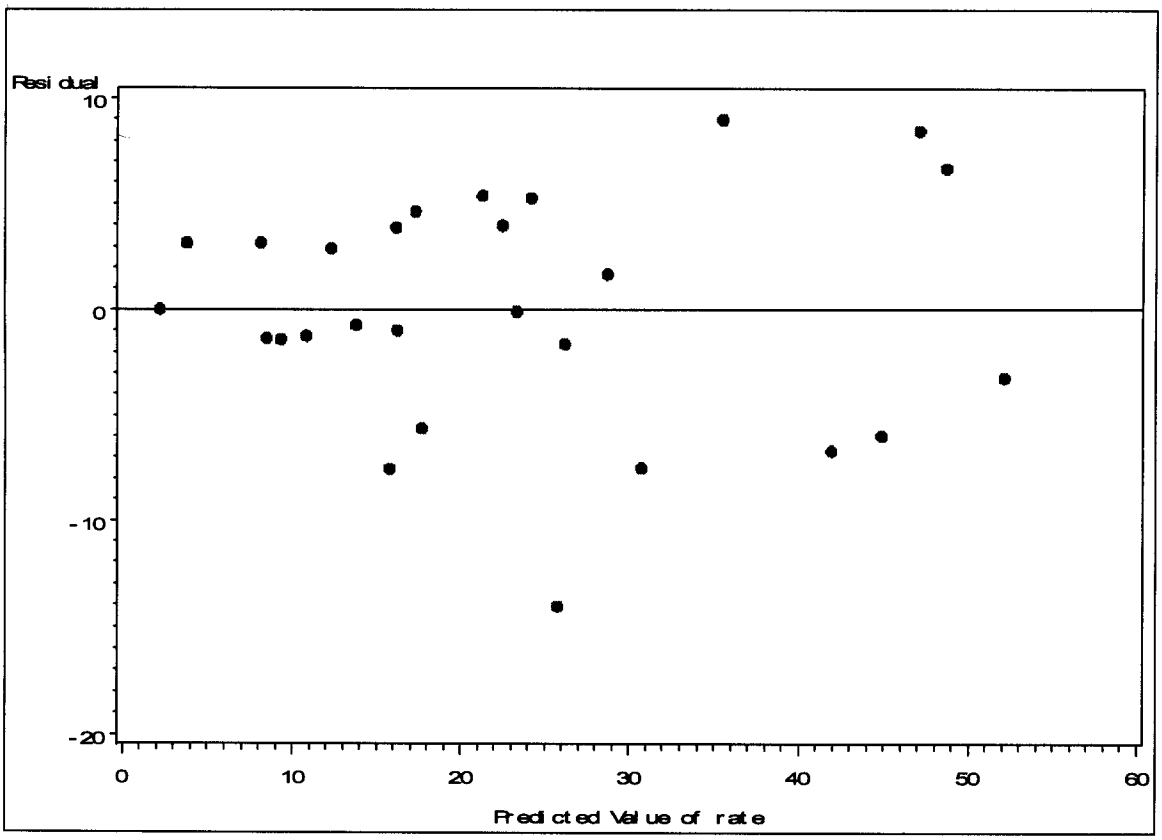
Output Statistics

Obs	-2-1 0 1 2	Cook's D
1	| * |	0.029
2	| ** |	0.235
3	| * |	0.037
4	| ** |	0.027
5	| * |	0.095
6	| ** |	0.141
7	| * |	0.310
8	| * |	0.039
9	| |	0.087
10	| ** |	0.308
11	| |	0.017
12	| |	0.003
13	| ** |	0.104
14	| ** |	0.303
15	| |	2.393
16	| |	0.002
17	| ** |	0.165
18	| |	0.001
19	| *** |	0.306
20	| ** |	0.129
21	| |	0.000
22	| * |	0.042
23	| |	0.009
24	| * |	0.007
25	| |	0.019
26	| * |	0.013
27	| **** |	0.175

Sum of Residuals	0
Sum of Squared Residuals	777.14098
Predicted Residual SS (PRESS)	7406.37292


Studentized Residual
2
1
0
-1
-2
-3
0
10
20
30
40
50
60
Predicted Value of rate


Residual
10
0
-10
-20
1
boxplot



The UNIVARIATE Procedure
Variable: resid (Residual)

Moments

N	27	Sum Weights	27
Mean	0	Sum Observations	0
Std Deviation	5.46717823	Variance	29.8900378
Skewness	-0.5379714	Kurtosis	0.20121435
Uncorrected SS	777.140984	Corrected SS	777.140984
Coeff Variation	.	Std Error Mean	1.05215894

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	5.46718
Median	-0.11137	Variance	29.89004
Mode	.	Range	23.03702
		Interquartile Range	7.17915

Tests for Location: Mu0=0

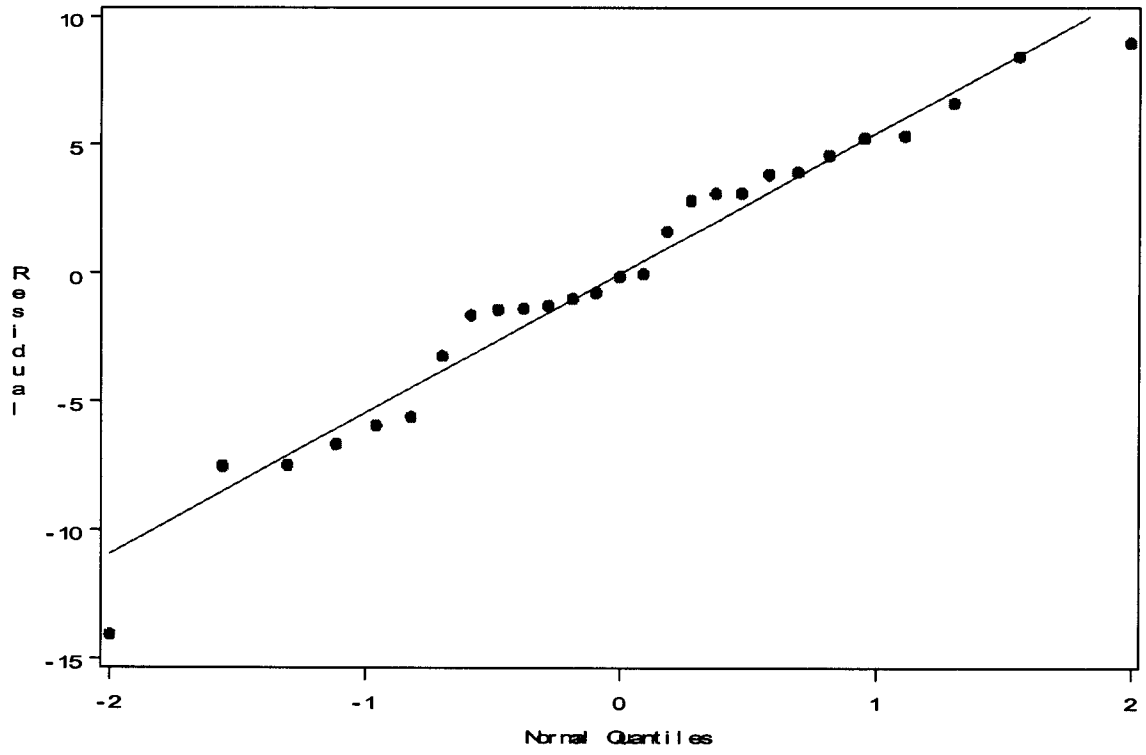
Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-1.5	Pr >= |M|	0.7011
Signed Rank	S	5	Pr >= |S|	0.9070

Quantiles (Definition 5)

Quantile	Estimate
100% Max	8.991556
99%	8.991556
95%	8.455016
90%	6.644799
75% Q3	3.958335
50% Median	-0.111366
25% Q1	-3.220813
10%	-7.492856
5%	-7.532351
1%	-14.045460
0% Min	-14.045460

Extreme Observations

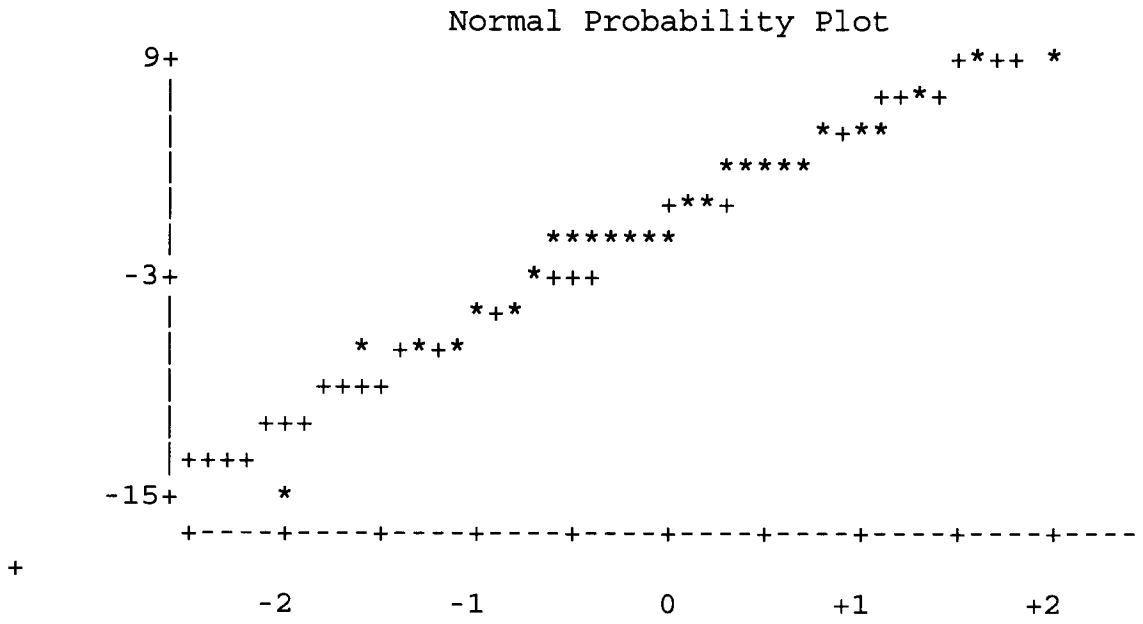
------Lowest------		-----Highest-----	
Value	Obs	Value	Obs
-14.04546	27	5.26134	5
-7.53235	2	5.35550	26
-7.49286	4	6.64480	6
-6.67096	20	8.45502	19
-5.95292	3	8.99156	13



The UNIVARIATE Procedure
Variable: resid (Residual)

Stem	Leaf	#	Boxplot
8	50	2	
6	6	1	
4	0634	4	+-----+
2	9129	4	| |
0	7	1	| + |
-0	64430710	8	*-----*
-2	2	1	+-----+
-4	6	1	
-6	5570	4	
-8			
-10			
-12			
-14	0	1	0

----+----+----+----+



The UNIVARIATE Procedure
Fitted Distribution for resid

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	5.467178

Goodness-of-Fit Tests for Normal Distribution

Test		---Statistic----		-----p Value-----
Kolmogorov-Smirnov	D	0.12391328	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.05215304	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.32431521	Pr > A-Sq	>0.250

Quantiles for Normal Distribution

Percent	Quantile Observed	Quantile Estimated
1.0	-14.04546	-12.718558
5.0	-7.53235	-8.992708
10.0	-7.49286	-7.006471
25.0	-3.22081	-3.687556
50.0	-0.11137	-0.000000
75.0	3.95833	3.687556
90.0	6.64480	7.006471
95.0	8.45502	8.992708
99.0	8.99156	12.718558