LINEAR REGRESSION CRASH PREDICTION MODELS: ISSUES AND PROPOSED SOLUTIONS

FINAL REPORT

PennDOT/MAUTC Agreement Contract No. 510401
VT-2008-02
DTRS99-G-0003

Prepared for

Virginia Transportation Research Council

By

H. Rakha, M. Arafeh, A. G. Abdel-Salam, F. Guo and A. M. Flintsch

Virginia Tech Transportation Institute

May 2010

This work was sponsored by the Virginia Department of Transportation and the U.S. Department of Transportation, Federal Highway Administration.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
LINEAR REGRESSION CRASH PREDICTION MODELS: ISSUES AND PROPOSED SOLUTIONS	May 2010
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
H. Rakha, M. Arafeh, A. G. Abdel-Salam, F. Guo, and A. M. Flintsch	VT-2008-02

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Virginia Tech Transportation Institute 3500 Transportation Research Plaza Blacksburg, VA 24061	**11. Contract or Grant No.** DTRS99-G-003

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Virginia Department of Transportation ****** U.S. Department of Transportation Research and Innovative Technology Administration UTC Program, RDT-30 1200 New Jersey Ave., SE Washington, DC 20590	Final Report
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

The paper develops a linear regression model approach that can be applied to crash data to predict vehicle crashes. The proposed approach involves novice data aggregation to satisfy linear regression assumptions; namely error structure normality and homoscedasticity. The proposed approach is tested and validated using data from 186 access road sections in the state of Virginia. The approach is demonstrated to produce crash predictions consistent with traditional negative binomial and zero inflated negative binomial general linear models. It should be noted however that further testing of the approach on other crash datasets is required to further validate the approach.

17. Key Words	18. Distribution Statement
Linear regression model, novice data aggregation, negative binomial, zero inflated negative binomial general linear models.	No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	21	

Form DOT F 1700.7 (8-72) **Technical Report Documentation Page Reproduction of completed page authorized**

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Linear regression crash prediction models: issues and proposed solutions

H. Rakha
Charles Via, Jr. Department of Civil and Environmental Engineering, Blacksburg, Virginia, USA

M. Arafeh, A.G. Abdel-salam, F. Guo and A.M. Flintsch
Virginia Tech Transportation Institute, Blacksburg, Virginia,USA

ABSTRACT: The paper develops a linear regression model approach that can be applied to crash data to predict vehicle crashes. The proposed approach involves novice data aggregation to satisfy linear regression assumptions; namely error structure normality and homoscedasticity. The proposed approach is tested and validated using data from 186 access road sections in the state of Virginia. The approach is demonstrated to produce crash predictions consistent with traditional negative binomial and zero inflated negative binomial general linear models. It should be noted however that further testing of the approach on other crash datasets is required to further validate the approach.

INTRODUCTION

The current state-of-the-art for developing Crash Prediction Models (CPMs) is to adopt General Linear Models (GLMs) considering either a Poisson or a negative binomial error structure (Lord et al. 2004; Lord et al. 2005; Sawalha et al. 2006). Recently, researchers have also proposed the use of zero inflated negative binomial regression models in order to address the high propensity of zero crashes within typical crash data (Shankar et al. 1997; Shankar et al. 2003). The use of Linear Regression Models (LRMs) is not utilized because crash data typically do not satisfy the assumptions of such models, namely: normal error structure and constant error variance.

The objectives of the research presented in this paper are two-fold. First, the paper demonstrates how through the use of data manipulation it is possible to satisfy the assumptions of LRMs and thus develop robust LRMs. Second, the paper compares the LRM approach to the traditional GLM approach considering a negative binomial error structure to demonstrate the adequacy of the proposed approach. The objectives of the paper are achieved by applying the models to crash, traffic, and roadway geometric data obtained from 186 freeway access roads in the state of Virginia in the U.S.

In terms of the paper layout, initially a brief background of CPMs is presented. Subsequently, the unique characteristics of the crash, traffic, and roadway geometry data that are utilized to validate the proposed approach are described. Next, the two modeling approaches are described and applied to the access road data. Finally, the study conclusions are presented.

BACKGROUND

An earlier publication (Lord et al. 2004) indicated that “there has been considerable research conducted over the last 20 years focused on predicting motor vehicle crashes on transportation facilities. The range of statistical models commonly applied includes binomial, Poisson, Poisson-gamma (or Negative Binomial), Zero-Inflated Poisson and Negative Binomial Models (ZIP and ZINB), and Multinomial probability models. Given the range of possible modeling

approaches and the host of assumptions with each modeling approach, making an intelligent choice for modeling motor vehicle crash data is difficult at best." The authors further indicate that "in recent years, some researchers have applied "zero-inflated" or "zero altered" probability models, which assume that a dual-state process is responsible for generating the crash data." The authors indicated that "these models have been applied to capture the 'excess' zeroes that commonly arise in crash data—and generally have provided improved fit to data compared to Poisson and Negative Binomial (NB) regression models."

Lord et al. (Lord et al. 2004) conducted a simulation experiment to demonstrate how crash data may give rise to "excess" zeroes. They demonstrated that under certain (fairly common) circumstances excess zeroes are observed—and that these circumstances arise from low exposure and/or inappropriate selection of time/space scales and not an underlying dual state process. They concluded that a careful selection of the time/space scales for analysis, including an improved set of explanatory variables and/or unobserved heterogeneity effects in count regression models, or applying small area statistical methods (observations with low exposure) represent the most defensible modeling approaches for datasets with a preponderance of zeros. We partially agree with these conclusions, however modelers my not have much choice in their time/space scale selection given the limitation of traffic and crash data.

In this paper we present an alternative approach that combines novice data aggregation with LRMs to address the challenges of crash data that were described earlier.

DATA DESCRIPTION

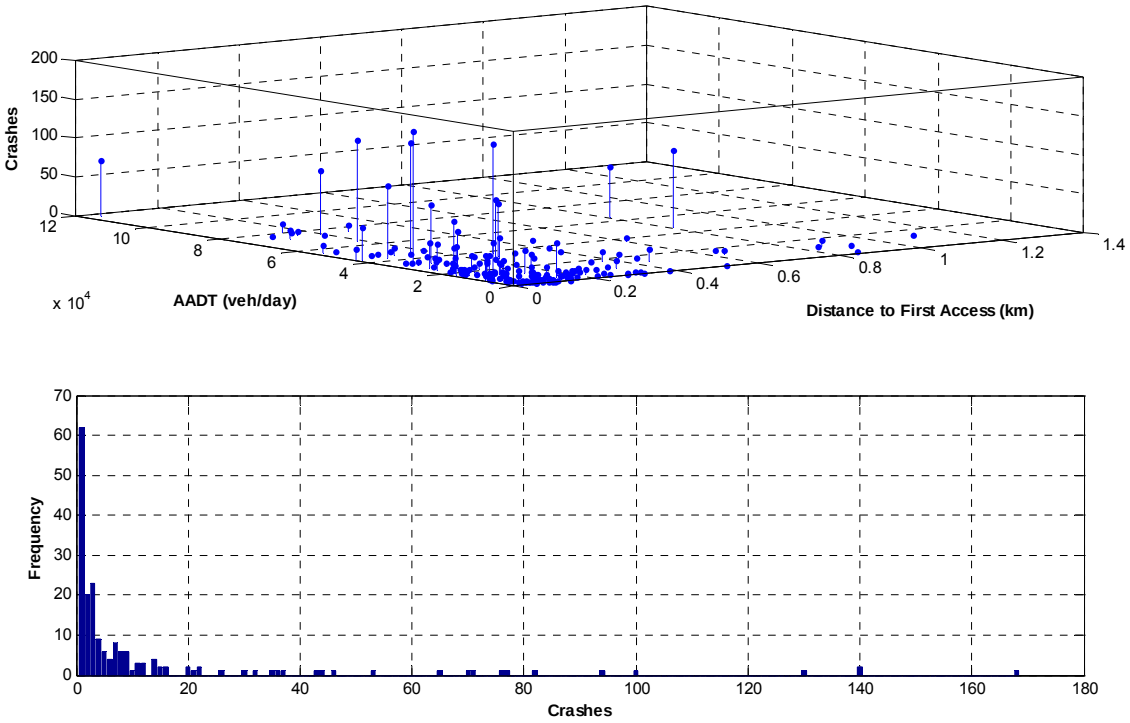
As was mentioned earlier, the two approaches for developing CPMs are demonstrated using a database of crash, traffic, and geometric data obtained from 186 randomly selected arterial access roads connected to freeway ramps. The sections that were considered included the following:

1. Areas designated as rural (79 observations) and others designated as urban (107 observations),
2. Sections with (76 observations) and without acceleration lanes (110 observations),
3. Arterial facilities with a median (121 observations) versus without (65 observations),
4. With 2, 4, or 6 lanes (57, 95, and 34 observations, respectively),
5. First intersection either signalized, stop-sign controlled, or no control on the arterial, and
6. Sections with a left turn bay (78 observations) versus without (108 observations).

The length of the access roads varied from 3 to 1110 m with an average length of 169 m while the distance to the first intersection varied from 6 to 2285 m with an average value of 298 m. The Average Annual Daily Traffic (AADT) varied from 112 to 117,314 with an average AADT of 19,456 veh/day. The number of crashes over 5 years varied from 0 to 169 crashes in a section with an average number of crashes of 12 over the 186 study sections.

The crash data were extracted from the Highway Traffic Records Information System Crash Database (HTRIS) for the years 2001 through 2005. The traffic data were obtained from the Road inventory VDOT database. The data represents the AADT of the section with a few exceptions were the traffic data represent a 24 hour count.

After fusing the crash, traffic, and geometric data it was possible to plot the data, as illustrated in **Error! Reference source not found.**. Specifically, the figure demonstrates a general increase



in the number of crashes as the facility Average Annual Daily Traffic (AADT) increases. The figure also illustrates a high cluster of data at the short access road distances with minimum observations for access roads in excess of 400 m. Similarly, AADTs in excess of 8,000 veh/day are a rare occurrence. The figure does illustrate a number of sections with high AADTs and short access roads with a small number of crashes. Conversely, observations with high crashes are also observed for low AADTs and long access roads.

Figure 1. Data Distribution.

The crash frequency distribution demonstrates that the number of crashes ranges from a minimum of 0 to a maximum of 169 crashes over 5 years, as illustrated in **Error! Reference source not found.**. The higher frequency of zero crashes is typical of crash data when the exposure rate is low as was described in the literature (Lord et al. 2004). The higher propensity of zero observations has lead some researchers to apply zero-inflated negative binomial models to the modeling of crashes considering two underlying processes (Shankar et al. 1997; Shankar et al. 2003). The frequency distribution is consistent with a negative binomial distribution with a consistent decrease in the frequency as the number of crashes increases.

A statistical analysis of the data demonstrated that the type of road (rural versus urban), number of lanes, the availability of a median, the type of signal control at the nearest intersection, and availability of an acceleration lane were not statistically significant. The details of these statistical tests are beyond the scope of this paper, but are provided elsewhere (Medina Flintsch et al. 2008).

MODEL DEVELOPMENT

This section describes the two approaches that were tested in the paper for developing crash prediction models. The first approach is the common approach that is reported in the literature, which is based on the use of Poisson, Negative Binomial (NB), Zero Inflated Poisson (ZIP), and Zero Inflated Negative Binomial (ZINB) regression models. An alternative approach that is developed in this paper is the use of LRMs. Initially, a discussion of the NB and ZINB

regression approach is presented followed by a discussion of the proposed LRM approach. Both approaches are then compared using crash data from 186 access roads in the state of Virginia in the U.S.

Prior to describing the various models, the model structure is discussed. Specifically, the study considers a crash rate that is formulated as

$$CR = \frac{C}{L_2 V^p} \cdot \frac{10^6}{(365 \times 5)^p} = \exp(\beta_0 + \beta_1 L_1), \quad [1]$$

where *CR* is the crash rate (million vehicle crashes per vehicle kilometer of exposure over a 5 year period), *C* is the total number of crashes over the study section of length L_2 in the 5-year analysis period (crashes), L_2 is the length of the section which is the distance between the freeway off-ramp and the first intersection (km), L_1 is the distance between the freeway off-ramp and the first access road (may equal L_2 if the first access road is an intersection) (km), V is the section AADT (veh/day), B_0 and B_1 are the model constants.

The model of Equation [1] can then be manipulated to produce a linear model of the form

$$C = \frac{\left(365 \times 5\right)^p}{10^6} \cdot \exp\left(\beta_0 + \beta_1 L_1 + \ln(L_2) + p\ln(V)\right) + E \quad . \quad [2]$$

Here *E* is a random error term that accounts for the error that is not captured in the model. The advantage of this model is that (a) it is linear in structure after applying a logarithmic transformation; (b) it ensures that the crashes are positive (greater than or equal to zero); and (c) it produces zero crashes when the exposure is set to zero (i.e. when L_2 or V is zero).

Poisson or Negative Binomial Model Approach

The Poisson distribution is more frequently applied to models with count data. The probability mass function of a Poisson crash random variable (λ is the mean crashes per unit time) is given by

$$P(C) = \frac{\lambda^C}{C!} e^{-C};\ C = 0,1,... \quad [3]$$

One feature of the Poisson random variable (*C*) is the identity of the mean and variance. The NB model overcomes this limitation by considering the λ parameter to be a random variable that is distributed following a Poisson distribution. Consequently, the negative binomial distribution has the same support (*C*=0,1,2, …) as the Poisson distribution but allows for greater variability within the data (variance can be greater than the mean). Consequently, a good alternative model for count data that exhibit excess variation compared to a Poisson model is the negative binomial model.

The structure of the GLM model that computes the expected number of crashes (E(*C*) can be written as

$$\mathrm{E}\left(C\right) = \left(\frac{\left(365 \times 5\right)^p}{10^6}\right) L_2 \cdot \exp\left(\beta_0 + \beta_1 L_1 + p\ln(V)\right). \quad [4]$$

Two sets of Poisson and two NB models were developed, one a standard model (NB and Poisson) and one a zero inflated model (ZINB and ZIP), as summarized in Table 1. Unfortunately, the Poisson regression model suffered from over-dispersion as indicated by the value of the deviance divided by the degrees of freedom which was much greater than a value of 1.0 (was 184.7). Consequently, a Modified Poisson Regression was also applied using an over-dispersion parameter. The model produces the same parameter values; however, the deviance is reduced to 1.0 and thus is valid from a statistical standpoint. In the case of the NB model the value of the deviance divided by the degrees of freedom was close to a value of 1.0 (1.2063) and

thus demonstrating the adequacy of the negative binomial error structure. It should be noted that the model predictions for the zero-inflated models are computed my multiplying the results of Equation [4] by $1-\exp(\theta)\big/\big(1+\exp(\theta)\big)$ as

$$\mathrm{E}\big(C\big)=\left(\frac{\big(365\times 5\big)^{p}}{10^{6}}\right)L_2\cdot\exp\big(\beta_0+\beta_1 L_1+p\ln(V)\big)\times\left(1-\frac{\exp(\theta)}{1+\exp(\theta)}\right). \qquad [5]$$

Table 1: Summary Results of Regression Models

Parameter	LRM	NB	Poisson	ZINB	ZIP
B_0	4.27	4.76	3.42	4.76	3.83
B_1	-6.88	-3.64	-6.90	-3.64	-2.00
p	0.86	0.81	0.92	0.81	0.84
θ	0.00	0.00	0.00	-16.00	-1.69
SSE	365608	389589	437926	389589	254307
SSE (%)	0%	7%	20%	7%	-30%
Slope	0.472	0.512	0.549	0.512	0.339
Error (%)	2.12	1.95	1.82	1.95	2.95

The results of Table 1 demonstrate that the model parameters a practically identical for both the NB and ZINB models except for the θ parameter. Given that the θ parameter is much less than zero the predictions of the NB and ZINB are very similar. In the case of the Poisson models (Poisson and ZIP) the model parameter values are significantly different, as demonstrated in Table 1.

Linear Regression Modeling Approach

In this section we consider a linear regression approach for the development of a model. If we consider the number of crashes per unit distance as our dependent variable the model of Equation [2] can be cast as

$$\frac{\mathrm{E}(C)}{L_2}=\mathrm{E}(C')=\frac{\big(365\times 5\big)^{p}}{10^{6}}\cdot\exp\big(\beta_0+\beta_1 L_1+p\ln(V)\big). \qquad [6]$$

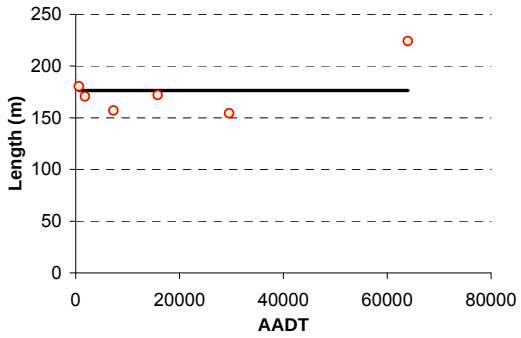
Equation [6] is a linear model with two independent variables: V and L_1. It should be noted that an analysis of crashes per unit distance ensures that the data are normalized across the different section lengths. The development of a LRM using the least squares approach requires that the data follow a normal distribution. A statistical analysis of the data revealed that there was insufficient evidence to conclude that the data were normal. Furthermore, the dispersion parameter, which measures the amount of variation in the data, was significantly greater than 1.0 indicating that a negative binomial model would be appropriate for the data.

Here we present an approach for normalizing the data in order to apply a least squared LRM to the data. The approach involves sorting the data based on one of the independent variables and then aggregating the data using a variable bin size to ensure that the second independent variable remains constant across the various bins. Data transformations can then be applied to the data to ensure normality and homoscedasticity (equal variance). Once the parameters of the first independent variable are computed, the data are sorted on the second independent variable. The data are then aggregated in order to ensure normality and homoscedasticity and then linear models are fit to the data to compute the variable coefficient. The approach is demonstrated using the access road crash data in the following sub-sections.

Selecting Exposure Measures

The typical exposure measure for crashes is million vehicle-miles or million vehicle-kilometers of travel. However, researchers have argued that the exponent of the volume variable (V) in the exposure measure is not necessarily equal to 1.0 [7, 8]. Consequently, the first step in the analysis was to compute the exponent of V (denoted as p).

Given that the data vary as a function of two variables L_1 and V, it was important to normalize one of the variables while analyzing the second variable. In order to estimate the volume exponent, the data were sorted based on their AADT values and aggregated using variable bin sizes to ensure that the L_1 variable remained constant across the various bins, as illustrated in Figure 1. The figure demonstrates that by performing a linear regression of L_1 against V that the slope of the line is insignificant ($p > 0.05$) and thus there is insufficient evidence to conclude that the L_1 variable varies across the aggregated data.



Regression Statistics	
R^2	0.509
R^2 (Adj)	0.386
Obs.	6

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Reg.	1	17608.38	17608.38	4.14	0.11
Residual	4	17003.46	4250.86		
Total	5	34611.84			

	Coeff.	*SE*	*t Stat*	*P-value*	*Low 95%*	*Up 95%*
Intercept	530.039	35.820	14.80	0.00	430.59	629.49
AADT	0.002	0.001	2.04	0.11	0.00	0.01

Figure 1. Effect of Access Section Length within AADT Binning.

In estimating crash rates it is important that the measure of exposure ensures that the data are normalized. In doing so a multiplicative crash adjustment factor (F_i) for each bin i was computed as

$$F_i = \frac{\min_i \left[\max_j \left(\frac{C_{ij}}{L_{ij}} \right) \right]}{\max_j \left(\frac{C_{ij}}{L_{ij}} \right)}, \qquad [7]$$

where C_{ij} is the number of crashes for section j in bin i and L_{ij} is the length of section j in bin i. The F_i correction factor ensures that the maximum number of crashes remains constant (equal to the minimum number of crashes) across the various bins, which is by definition what an exposure measure is. The correction factor is also equal to

$$F_i = \alpha V_i^{\beta}, \qquad [8]$$

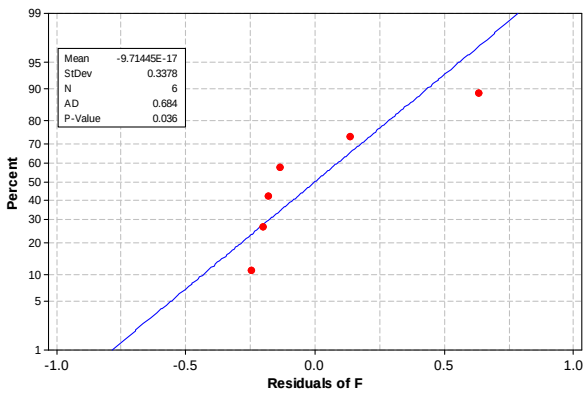
where V_i is the mean AADT volume across all observations j in bin i and α and β are model coefficients. By solving Equation [7] and [8] simultaneously we derive

$$CR = F_i \cdot \max_j \left(\frac{C_{ij}}{L_{ij}} \right) = \alpha V_i^{\beta} \cdot \max_j \left(\frac{C_{ij}}{L_{ij}} \right) = \alpha \cdot \frac{C_i}{L_i V_i^{-\beta}} = \frac{\text{Crashes}}{\text{Length} \cdot \text{AADT}^p}. \qquad [9]$$

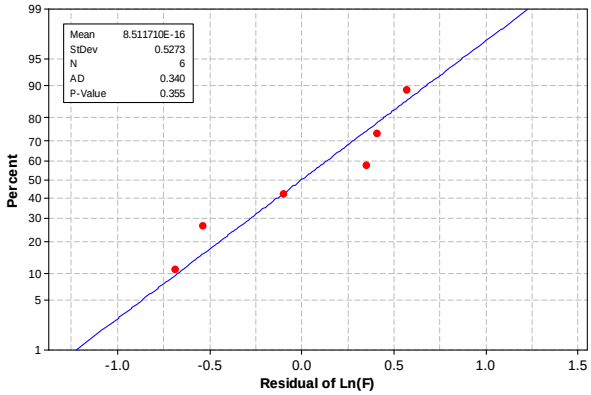
Consequently, β is equivalent to $-p$ and can be solved for by fitting a regression line to the logarithmic transformation of Equation [8] as

$$\ln(F_i) = \ln(\alpha) + \beta \ln(V_i). \qquad [10]$$

After applying a least squared fit to the data, the model residual errors were tested for normality. As illustrated in Figure 2, although in the case of the original non-transformed data the residual error did not pass the normality test, the log-transformed residual errors did pass the test ($p = 0.355$).



(a) Normality Test on F_i

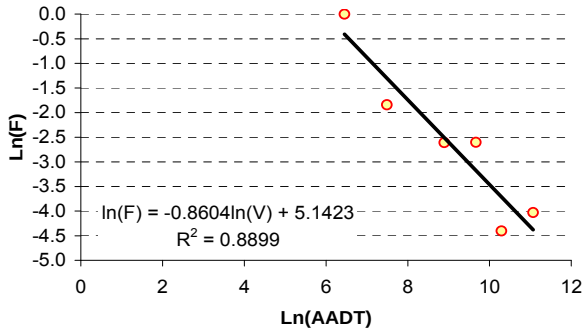


(b) Normality Test on Log-transformed F_i

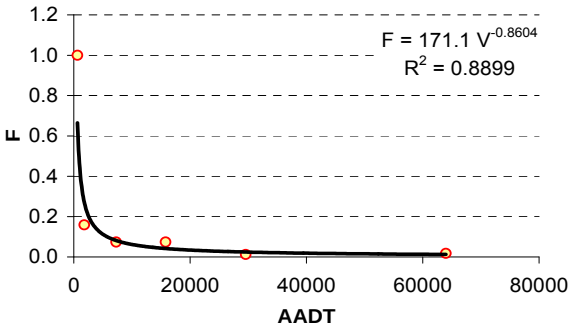
Figure 2. Test of Normality of AADT Adjustment Factors.

A least squares LRM was then fit to the log-transformed data producing an R^2 of 0.89, as illustrated in Figure 3. The model was statistically significant (p << 0.005) and both the intercept and slope coefficients were significant (p = 0.02 and 0.00, respectively). Consequently, the exponent of the AADT for utilization in the exposure measure is 0.86, which is very similar to what was derived from the negative binomial fit to the data (p = 0.81).

(a) Observed and Fitted F after Log-transformation of Data



(b) Observed and Fitted F



Regression Statistics	
R^2	0.890
R^2 (Adj)	0.862
Obs.	6

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Reg.	1	11.24	11.24	32.33	0.00
Residual	4	1.39	0.35		
Total	5	12.63			

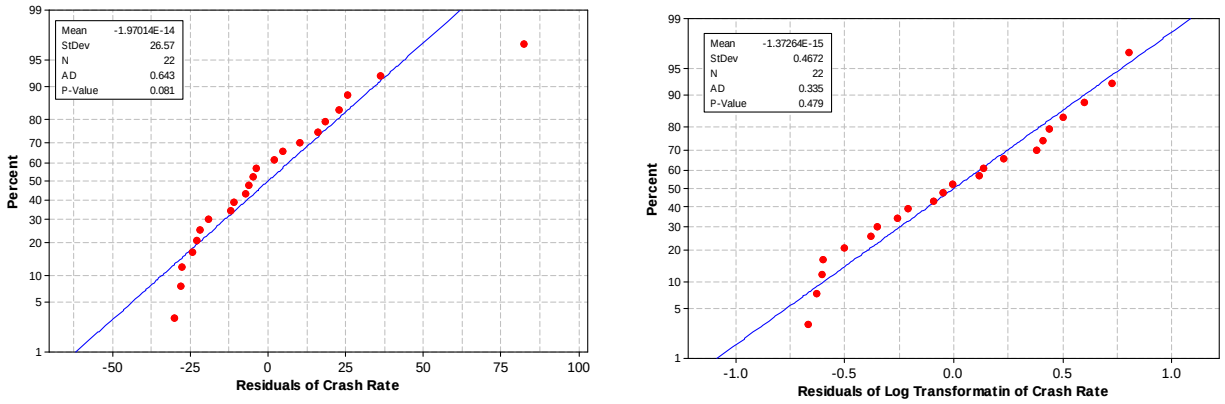
	Coeff.	*SE*	*t Stat*	*P-value*	*Low 95%*	*Up 95%*
Intercept	5.1423	1.38	3.73	0.02	1.31	8.97
ln(AADT)	-0.8604	0.15	-5.69	0.00	-1.28	-0.44

Figure 3. Computation of Exposure Measure.

Once the exponent of the AADT was estimated, the crash rate was computed for each of the 186 study sections. A linear regression model in a single independent variable of the form

$$\ln(CR) = \beta_0 + \beta_1 L_1 \qquad [11]$$

was fit to the data. In order to satisfy data normality, the data were sorted based on L_1 and aggregated into equally sized bins of 8 observations. It should be noted that the typical approach to binning is to use equal intervals for binning as opposed to equally sized bins. This unique data aggregation approach is equivalent to considering a longer analysis period (in this case considering an analysis period of 8×5=40 years). The data aggregation increases the level of exposure and thus reduces the number of zero crash observations (in this case zero observations are removed), given that it is highly unlikely to have no crashes over a 40-year period. For each bin the average section length (L_1) and crash rate (CR) was computed. As demonstrated in Figure 4 there was insufficient evidence to reject the data error normality and homoscedasticity assumption for the log-transformed data (p=0.479) and thus a least squares GLM could be applied to the data.



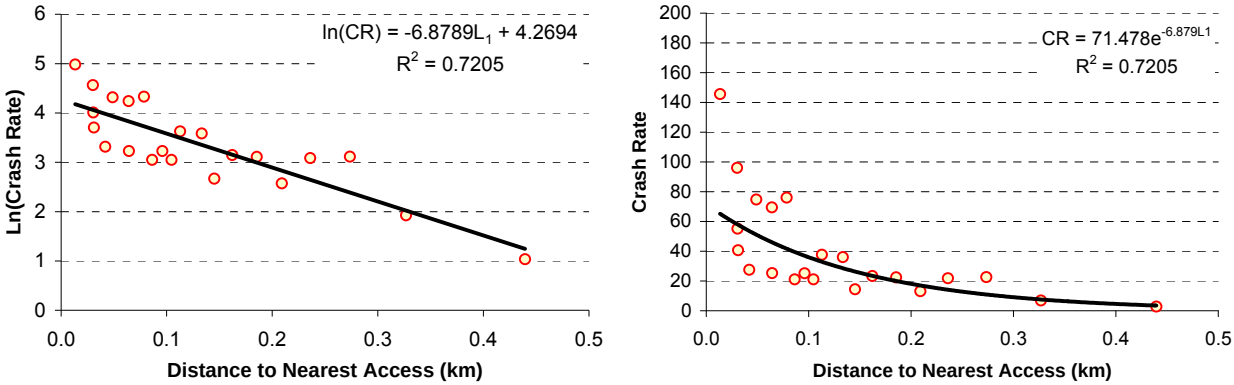
(a) Normality Test on Crash Rate

(b) Normality Test on Log-transformed Crash Rate

Figure 4. Test of Normality for Crash Rate Data.

A robust linear regression was applied to the data to derive the model parameters and remove outlier data. This procedure dampens the effect of observations that would be highly influential if least squares were used [9]. The robust linear regression fit uses an iteratively re-weighted least squares algorithm, with the weights at each iteration calculated by applying the bi-square function to the residuals from the previous iteration. This Matlab algorithm gives lower weight to points that appear to be outliers. Data that should be disregarded are given a weight of zero. Consequently, the regression model is less sensitive to outliers in the data as compared with ordinary least squares regression. Data observations with zero weights were removed from the analysis (in this case a single observation was removed).

The results of the analysis demonstrate a statistically significant model (F=51.56 and p<<0.0005) with an R^2 of 0.72. The intercept and L_1 coefficients are statistically significant (p<<0.0005 and p<<0.0005, respectively) with values of 4.269 and -6.879, respectively.



Regression Statistics	
Multiple R	0.849
R^2	0.721
R^2 (Adj)	0.707
SE	0.479
Obs.	22

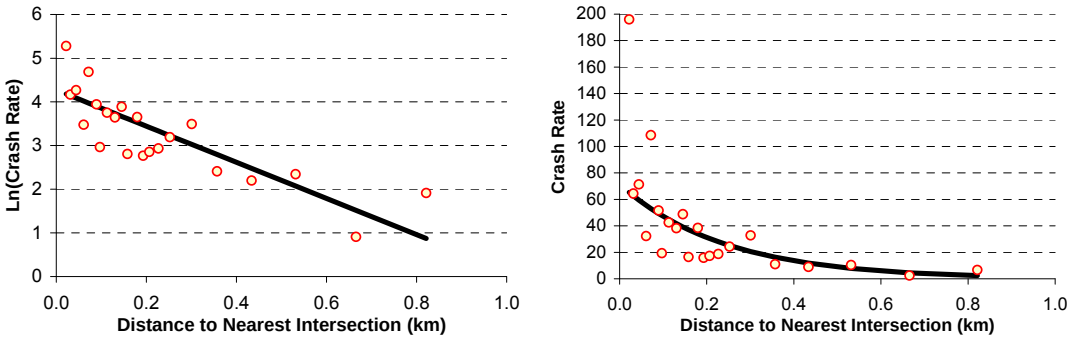
ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regressior	1	11.819	11.819	51.563	0.000
Residual	20	4.584	0.229		
Total	21	16.403			

	Coeff.	*SE*	*t Stat*	*P-value*	*Low 95%*	*Up 95%*
Intercept	4.269	0.163	26.222	0.000	3.930	4.609
L_1	-6.879	0.958	-7.181	0.000	-8.877	-4.881

Figure 5. Crash Prediction Model Considering Nearest Access Point.

Similarly, a regression model was fit to the data considering the independent variable as the distance to the first intersection. A similar robust regression was applied to the data to derive the model intercept and slope. Given that the intercept confidence limits included the value of intercept of the first model, the intercept was kept constant in both models. A regression was then performed to estimate the optimum slope. The model is significant (F=111.44 and p<<0.0005) with an R^2 of 0.85. The slope of the line is significant (p<<0.0005) with a value of -4.135.



Regression Statistics	
Multiple R	0.921
R^2	0.848
R^2 (Adj)	0.801
SE	0.558
Obs.	22

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regressior	1	36.552	36.552	117.436	0.000
Residual	21	6.536	0.311		
Total	22	43.088			

	Coeff.	*SE*	*t Stat*	*P-value*	*Low 95%*	*Up 95%*
Intercept	4.269	#N/A	#N/A	#N/A	#N/A	#N/A
L_2	-4.135	0.382	-10.837	0.000	-4.928	-3.341

Figure 6. Crash Prediction Model Considering Nearest Intersection.

In summary, the final models that were developed are of the form

$$\mathrm{E}(C) = \frac{\left(365 \times 5\right)^{p}}{10^{6}} \cdot \exp\left(\beta_0 + \beta_1 L_1 + \ln(L_2) + p\ln(V)\right), \text{ or} \quad [12]$$

$$\mathrm{E}(C) = \frac{\left(365 \times 5\right)^{p}}{10^{6}} \exp\left(\beta_0\right) \cdot \exp\left(\beta_1 L_1\right) \cdot L_2 V^{p} = \gamma \cdot \exp\left(\beta_1 L_1\right) \cdot L_2 V^{p}. \quad [13]$$

The expected number of crashes in a single year (C'') can be computed by adjusting the model intercept by the (p-1)×ln(5) as

$$\mathrm{E}\left(C''\right) = \frac{365^{p}}{10^{6}} \cdot \exp\left(\beta_0 + (p-1)\ln 5 + \beta_1 L_1 + \ln(L_2) + p\ln(V)\right). \quad [14]$$

The expected crash rate for a single year in million vehicle kilometers where the traffic volume is raised to the exponent p (CR') can be computed as

$$\mathrm{E}\left(CR'\right) = \exp\left(\beta_0 + (p-1)\ln 5 + \beta_1 L_1\right). \quad [15]$$

The expected crash rate in vehicle-miles traveled (VMT) considering an exponent of 1.0 (CR'') is computed as

$$\mathrm{E}\left(CR''\right) = \exp\left(\beta_0 + (p-1)\ln 5 + \beta_1 L_1\right) \times 1.6\left(365V\right)^{p-1}, \text{ or}$$

$$\mathrm{E}\left(CR''\right) = \exp\left(\beta_0 + \ln 1.6 + (p-1)\ln 5 + (p-1)\ln 365 + \beta_1 L_1 + (p-1)\ln V\right). \quad [16]$$

A number of researchers have argued for the need to calibrate the exposure AADT exponent. Consequently, a sensitivity analysis was conducted to study the impact of alternative exponents on the CPM predictions. Exponent values ranging from 0.6 to 1.2 were evaluated

based on values reported in the literature [7], as illustrated in Figure 7. The results clearly indicate that the crash predictions increase as the exponent increases, however the variation in crash predictions as a function of V and L_1 remains fairly consistent.

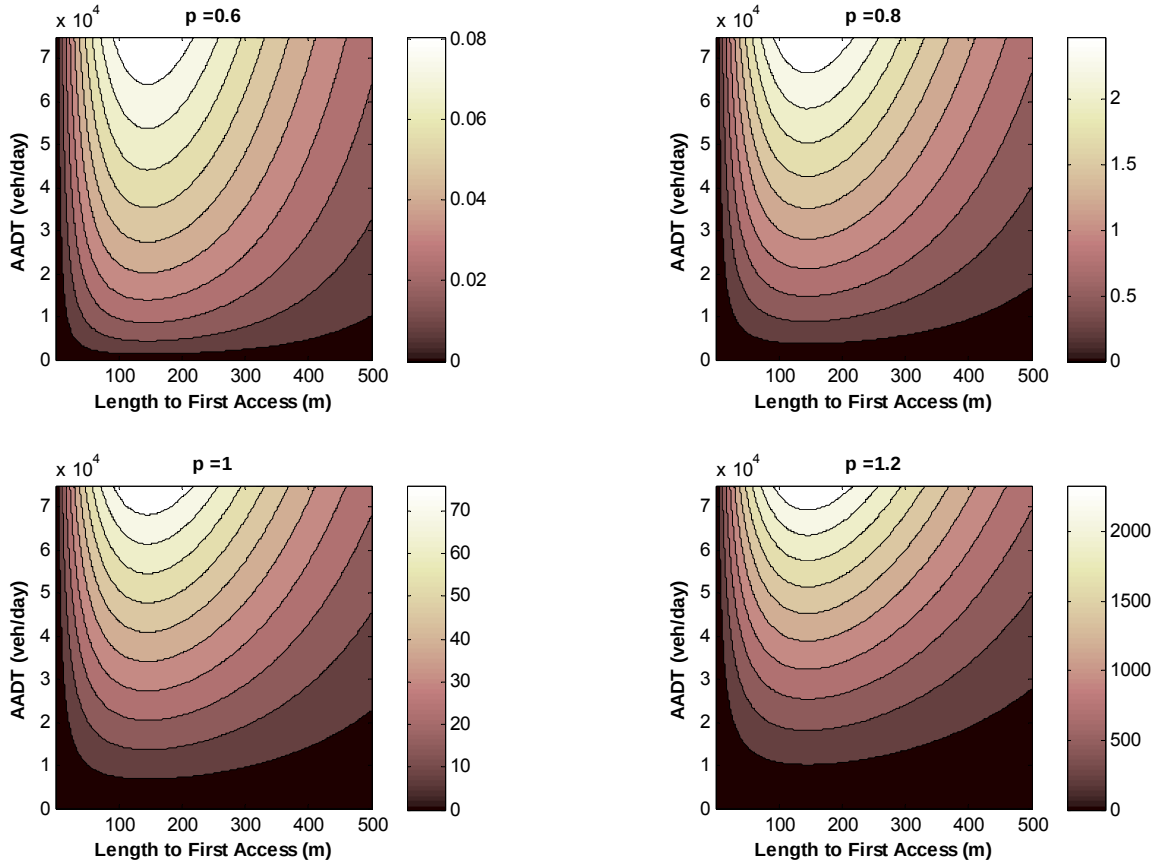


Figure 7. Variation in Expected Crashes as a Function of AADT Exponent.

It should be noted that for a constant AADT the expected crashes initially increases as the length of the spacing between the freeway ramp and the first access road section increases before decreasing again, as demonstrated in Figure 8. The maximum expected crashes occurs at an access road spacing of approximately 150 m (500 ft). The observed behavior might appear to be counter intuitive at first glance, however can be explained by the fact that as the study section increases the expected number of crashes per unit distance decreases, as illustrated in Figure 9 and Figure 10, while the level of exposure increases. Initially, the rate of increase in the level of exposure exceeds the rate of decrease in the crash rate producing an increase in the number of crashes. Consequently, decisions should be made using either a crash rate or the expected number of crashes per unit distance, as illustrated in Figure 10. Noteworthy is the fact that the expected number of crashes for an access road spacing of 30 and 150 m is highlighted to demonstrate the current criteria for access road spacing.

The average number of crashes across all 186 study sections was 2.45 crashes/year with an average AADT of 19,456 and an average access road spacing of 169 m (550 ft). The expected number of crashes derived from the model for the same AADT and access road spacing is estimated by the linear regression model at 2.43 crashes/year (highlighted in Figure 8) and thus demonstrating the validity of the model results. Alternatively, the both the NB and Poisson models over-estimate the expected number of crashes to be 2.96 and 2.66 crashes/year, respectively.

L (m)	L (ft)	AADT (veh/day)										
		5000	10000	15000	20000	25000	30000	35000	40000	45000	50000	75000
0.0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15.3	50	0.19	0.35	0.49	0.63	0.76	0.89	1.02	1.14	1.26	1.38	1.96
30.6	100	0.34	0.62	0.88	1.13	1.37	1.61	1.83	2.06	2.28	2.49	3.53
45.9	150	0.46	0.84	1.20	1.53	1.86	2.17	2.48	2.78	3.08	3.37	4.77
61.2	200	0.56	1.01	1.44	1.84	2.23	2.61	2.98	3.34	3.69	4.04	5.73
76.5	250	0.63	1.14	1.62	2.07	2.51	2.93	3.35	3.76	4.16	4.55	6.45
91.7	300	0.68	1.23	1.75	2.24	2.71	3.17	3.62	4.06	4.49	4.92	6.97
107.0	350	0.71	1.29	1.83	2.35	2.85	3.33	3.80	4.26	4.72	5.17	7.32
122.3	400	0.73	1.33	1.89	2.42	2.93	3.43	3.91	4.39	4.86	5.32	7.54
137.6	450	0.74	1.35	1.91	2.45	2.97	3.47	3.96	4.45	4.92	5.39	7.63
152.9	500	0.74	1.35	1.91	2.45	2.97	3.47	3.97	4.45	4.92	5.39	7.64
168.2	550	0.74	1.34	1.89	2.43	2.94	3.44	3.93	4.41	4.88	5.34	7.57
183.5	600	0.72	1.31	1.86	2.38	2.89	3.38	3.86	4.33	4.79	5.24	7.43
198.8	650	0.71	1.28	1.82	2.33	2.82	3.30	3.76	4.22	4.67	5.12	7.25
214.1	700	0.68	1.24	1.76	2.26	2.73	3.20	3.65	4.09	4.53	4.96	7.03
229.4	750	0.66	1.20	1.70	2.18	2.64	3.08	3.52	3.95	4.37	4.79	6.78
244.6	800	0.63	1.15	1.63	2.09	2.53	2.96	3.38	3.79	4.20	4.60	6.52
259.9	850	0.61	1.10	1.56	2.00	2.42	2.83	3.24	3.63	4.02	4.40	6.23
275.2	900	0.58	1.05	1.49	1.91	2.31	2.70	3.09	3.46	3.83	4.19	5.94
290.5	950	0.55	1.00	1.41	1.81	2.20	2.57	2.93	3.29	3.64	3.99	5.65
305.8	1000	0.52	0.95	1.34	1.72	2.08	2.43	2.78	3.12	3.45	3.78	5.36
321.1	1050	0.49	0.89	1.27	1.62	1.97	2.30	2.63	2.95	3.26	3.57	5.06
336.4	1100	0.46	0.84	1.20	1.53	1.86	2.17	2.48	2.78	3.08	3.37	4.78
351.7	1150	0.44	0.79	1.13	1.44	1.75	2.04	2.33	2.62	2.90	3.17	4.50
367.0	1200	0.41	0.75	1.06	1.36	1.64	1.92	2.19	2.46	2.72	2.98	4.23
382.3	1250	0.39	0.70	0.99	1.27	1.54	1.80	2.06	2.31	2.55	2.80	3.96
397.5	1300	0.36	0.66	0.93	1.19	1.44	1.69	1.93	2.16	2.39	2.62	3.71
412.8	1350	0.34	0.61	0.87	1.11	1.35	1.58	1.80	2.02	2.24	2.45	3.47
428.1	1400	0.32	0.57	0.81	1.04	1.26	1.47	1.68	1.89	2.09	2.29	3.24
443.4	1450	0.29	0.53	0.76	0.97	1.17	1.37	1.57	1.76	1.95	2.13	3.02
458.7	1500	0.27	0.50	0.71	0.90	1.09	1.28	1.46	1.64	1.81	1.99	2.82

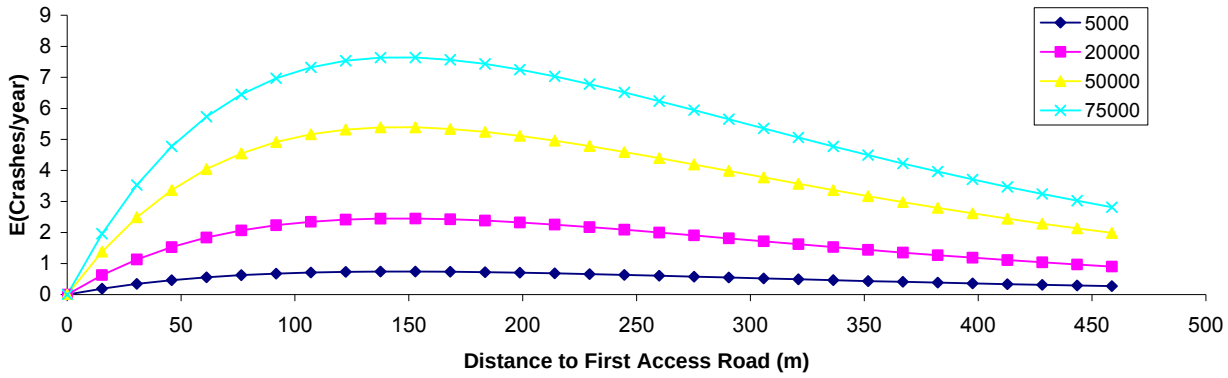


Figure 8. Variation in the Expected Number of Yearly Crashes as a Function of the Access Section Length and AADT.

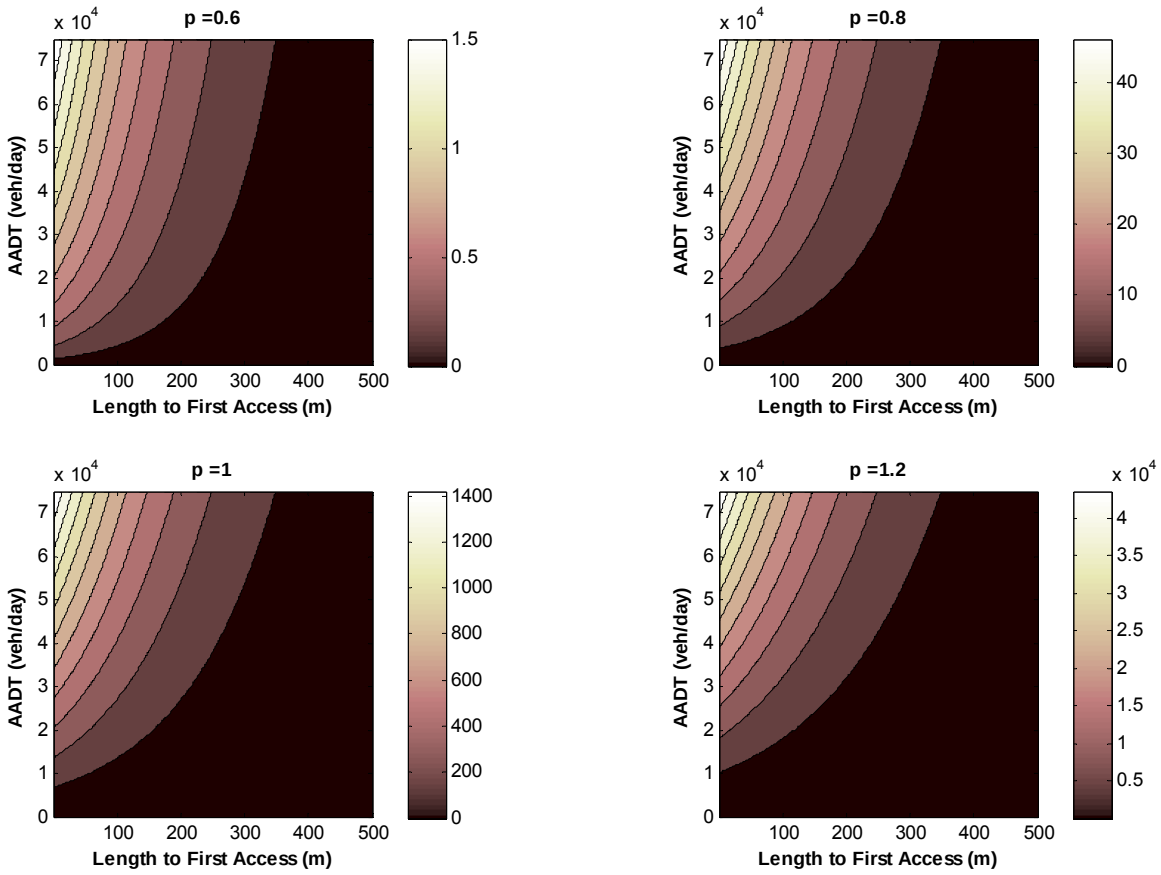


Figure 9. Variation in Expected Crashes/Km as a Function of AADT Exponent.

L (m)	L (ft)	AADT (veh/day)										
		5000	10000	15000	20000	25000	30000	35000	40000	45000	50000	75000
0.0	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
15.3	50	12.49	22.67	32.14	41.16	49.88	58.35	66.62	74.73	82.70	90.55	128.35
30.6	100	11.25	20.42	28.94	37.07	44.91	52.54	59.99	67.29	74.47	81.54	115.57
45.9	150	10.13	18.38	26.06	33.38	40.44	47.31	54.02	60.60	67.06	73.42	104.07
61.2	200	9.12	16.56	23.47	30.06	36.42	42.60	48.64	54.57	60.38	66.11	93.71
76.5	250	8.21	14.91	21.13	27.06	32.79	38.36	43.80	49.13	54.37	59.53	84.39
91.7	300	7.39	13.42	19.03	24.37	29.53	34.54	39.44	44.24	48.96	53.61	75.99
107.0	350	6.66	12.09	17.13	21.95	26.59	31.11	35.52	39.84	44.09	48.27	68.42
122.3	400	6.00	10.88	15.43	19.76	23.94	28.01	31.98	35.88	39.70	43.47	61.61
137.6	450	5.40	9.80	13.89	17.79	21.56	25.22	28.80	32.31	35.75	39.14	55.48
152.9	500	4.86	8.83	12.51	16.02	19.41	22.71	25.93	29.09	32.19	35.25	49.96
168.2	550	4.38	7.95	11.26	14.43	17.48	20.45	23.35	26.19	28.99	31.74	44.99
183.5	600	3.94	7.16	10.14	12.99	15.74	18.42	21.03	23.59	26.10	28.58	40.51
198.8	650	3.55	6.44	9.13	11.70	14.18	16.58	18.93	21.24	23.51	25.74	36.48
214.1	700	3.20	5.80	8.23	10.53	12.76	14.93	17.05	19.13	21.17	23.17	32.85
229.4	750	2.88	5.23	7.41	9.49	11.49	13.45	15.35	17.22	19.06	20.87	29.58
244.6	800	2.59	4.71	6.67	8.54	10.35	12.11	13.83	15.51	17.16	18.79	26.63
259.9	850	2.33	4.24	6.01	7.69	9.32	10.90	12.45	13.96	15.45	16.92	23.98
275.2	900	2.10	3.82	5.41	6.93	8.39	9.82	11.21	12.58	13.92	15.24	21.60
290.5	950	1.89	3.44	4.87	6.24	7.56	8.84	10.09	11.32	12.53	13.72	19.45
305.8	1000	1.70	3.09	4.38	5.62	6.81	7.96	9.09	10.20	11.28	12.35	17.51
321.1	1050	1.53	2.79	3.95	5.06	6.13	7.17	8.19	9.18	10.16	11.12	15.77
336.4	1100	1.38	2.51	3.56	4.55	5.52	6.46	7.37	8.27	9.15	10.02	14.20
351.7	1150	1.24	2.26	3.20	4.10	4.97	5.81	6.64	7.44	8.24	9.02	12.79
367.0	1200	1.12	2.03	2.88	3.69	4.47	5.23	5.98	6.70	7.42	8.12	11.51
382.3	1250	1.01	1.83	2.60	3.33	4.03	4.71	5.38	6.04	6.68	7.31	10.37
397.5	1300	0.91	1.65	2.34	2.99	3.63	4.24	4.85	5.44	6.02	6.59	9.34
412.8	1350	0.82	1.49	2.11	2.70	3.27	3.82	4.36	4.89	5.42	5.93	8.41
428.1	1400	0.74	1.34	1.90	2.43	2.94	3.44	3.93	4.41	4.88	5.34	7.57
443.4	1450	0.66	1.20	1.71	2.19	2.65	3.10	3.54	3.97	4.39	4.81	6.82
458.7	1500	0.60	1.08	1.54	1.97	2.39	2.79	3.19	3.57	3.96	4.33	6.14

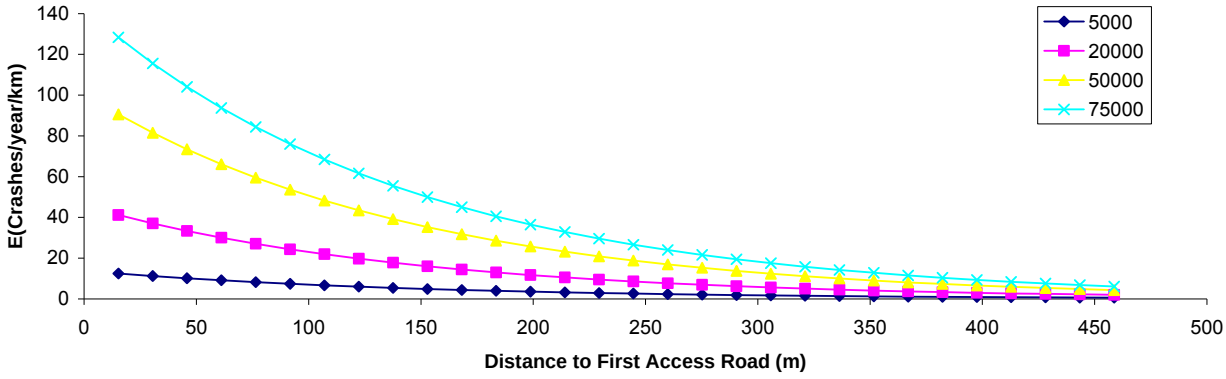


Figure 10. Variation in the Expected Number of Yearly Crashes per Kilometer as a Function of the Access Section Length and AADT.

In an attempt to validate the developed model, the AADT and access road spacing parameters for each of the 186 sites were input to the various models and the expected number of crashes was estimated. A comparison between the observed and estimated crashes revealed a reasonable level of correlation (Pearson correlation coefficient of 0.24) between the observed and linear regression model estimated crash rates, as illustrated in Figure 11. However, a high level of variability is observed in the data. The figure also clearly demonstrates that all models tend to under-estimate the expected number of crashes is slopes ranging from 0.339 to 0.549, as summarized in Table 1. The results of Table 1 demonstrate that the ZIP model produces the least Sum of Squared Error (SSE) between the estimated and observed number of crashes followed by the proposed LRM model. The ZIP model, however, in reducing the SSE results in a slope that is only 0.339 and thus greatest under-estimation error compared to the other models. Alternatively, the Poisson model produces a slope that is closest to 1.0 (0.549), however, the

SSE is 20% higher than the proposed LRM model SSE. Consequently, the proposed LRM model appears to offer the best compromise in terms of SSE and model prediction.

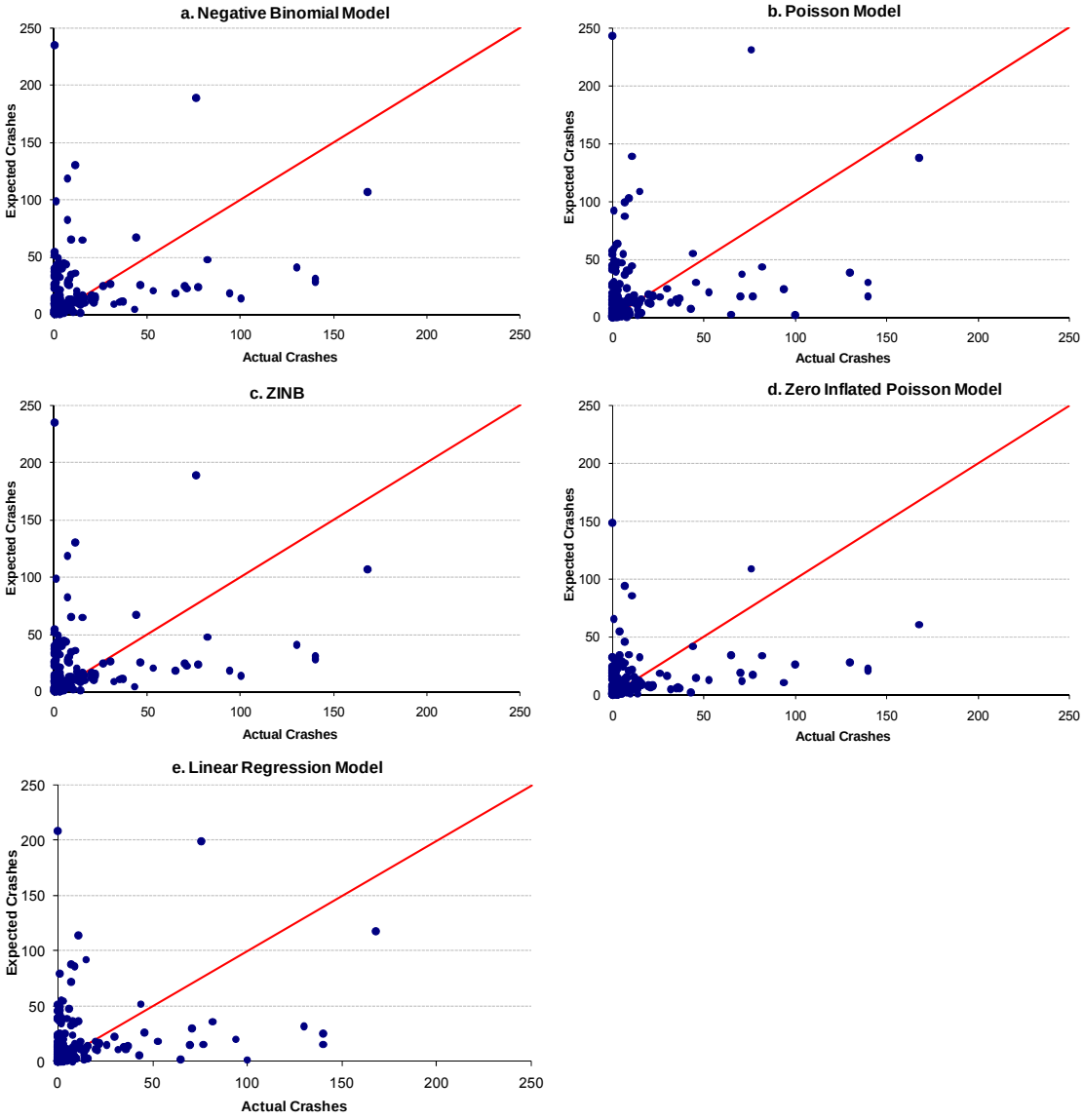


Figure 11. Comparison of Actual and Expected Crashes over a 5-year Period (All 186 Sites).

SUMMARY FINDINGS

A key desire of departments of transportation is to identify the minimum distance from a freeway ramp to provide access to local businesses. The model developed as part of this research effort was utilized to compute the crash rate associated with alternative section spacing, as summarized in **Error! Not a valid bookmark self-reference.**. The results demonstrate an eight-fold decrease in the crash rate over an access road spacing ranging from 0 to 300 m. An increase in the minimum spacing from 90 m (300 ft) to 180 m (600 ft) results in a 50% reduction in the crash rate.

Distance to First Access Road					Distance to First Intersection				
L (ft)	L (m)	Crashes per 10^6 VMT	Relative	Relative	L (ft)	L (m)	Crashes per 10^6 VMT	Relative	Relative
0	0.0	10.07	1.00	8.14	0	0.0	10.07	1.00	3.53
50	15.2	9.07	0.90	7.33	50	15.2	9.46	0.94	3.31
100	30.5	8.17	0.81	6.60	100	30.5	8.88	0.88	3.11
150	45.7	7.35	0.73	5.94	150	45.7	8.34	0.83	2.92
200	61.0	6.62	0.66	5.35	200	61.0	7.83	0.78	2.74
250	76.2	5.96	0.59	4.82	250	76.2	7.35	0.73	2.57
300	91.4	5.37	0.53	4.34	300	91.4	6.90	0.69	2.42
350	106.7	4.83	0.48	3.91	350	106.7	6.48	0.64	2.27
400	121.9	4.35	0.43	3.52	400	121.9	6.08	0.60	2.13
450	137.2	3.92	0.39	3.17	450	137.2	5.71	0.57	2.00
500	152.4	3.53	0.35	2.85	500	152.4	5.36	0.53	1.88
550	167.6	3.18	0.32	2.57	550	167.6	5.04	0.50	1.76
600	182.9	2.86	0.28	2.31	600	182.9	4.73	0.47	1.66
650	198.1	2.58	0.26	2.08	650	198.1	4.44	0.44	1.55
700	213.4	2.32	0.23	1.88	700	213.4	4.17	0.41	1.46
750	228.6	2.09	0.21	1.69	750	228.6	3.91	0.39	1.37
800	243.8	1.88	0.19	1.52	800	243.8	3.67	0.36	1.29
850	259.1	1.69	0.17	1.37	850	259.1	3.45	0.34	1.21
900	274.3	1.53	0.15	1.23	900	274.3	3.24	0.32	1.13
950	289.6	1.37	0.14	1.11	950	289.6	3.04	0.30	1.07
1000	304.8	1.24	0.12	1.00	1000	304.8	2.86	0.28	1.00

Table 2. Impact of Access Road Spacing on Annual Crash Rate (AADT = 20,000 veh/day).

CONCLUSIONS

The paper demonstrates that a least square LRM approach can be applied to crash data to develop crash prediction models. The proposed approach involves creative manipulation of the data to satisfy the least square LRM assumptions; namely normality and homoscedasticity. The approach can be summarized as follows:

(a) Consider the use of an exponential function. This function ensures that the number of crashes equals zero when the exposure is zero; that the number of crashes are always positive; and that the model reverts to a linear function after performing a logarithmic transformation.

(b) Sort and aggregate the data based on the AADT using a variable bin size while ensuring that the second independent variable remains constant across the various bins.

(c) Compute crash adjustment factors to normalize the maximum number of crashes across the various bins.

(d) Perform a logarithmic transformation on the crash adjustment factors to compute the AADT exponent using a LRM while ensuring that the data satisfy the LRM assumptions of normality and homoscedasticity.

(e) Compute crash rates using the AADT exponent that was computed earlier and then sort and aggregate the crash rate data based on the second independent variable using an equally sized bin structure (equal number of observations in each bin).

(f) Compute the average dependent and independent variable for each bin.

(g) Perform a logarithmic transformation of the data and ensure normality and homoscedasticity to develop the final crash prediction model.

The proposed approach was tested and validated using data from 186 access road sections in the state of Virginia. The approach was demonstrated to be superior to traditional negative binomial models because it is not influenced (through data aggregation) by the prevalence of the large number of zero observations that are typical of crash data.

Further testing of the proposed approach on other datasets is needed to validate the proposed approach. Furthermore, a sensitivity analysis of the sensitivity of results on different binning approaches for task (b) on the model outcomes is required.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support from the Virginia Department of Transportation. The authors also acknowledge the work of Ivy Gorman, Pengfei Li, and Dhruv Dua in extracting the data from the HTRIS and GIS Integrator. Furthermore the authors are grateful to the input and feedback received from Eugene Arnold, Stephen Read, Travis Bridewell, Hari Sripathi, and Stephen Bates.

REFERENCES

Ivan, J. (2004). "New approach for including traffic volumes in crash rate analysis and forecasting." Transportation Research Record **1897**: 134-141.

Lord, D., S. Washington, et al. (2004). Statistical Challenges with Modeling Motor Vehicle Crashes: Understanding the Implications of Alternative Approaches, Center for Transportation Safety, Texas Transportation Institute.

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

Medina Flintsch, A., H. Rakha, et al. (2008). Safety Impacts of Access Control Standards on Crossroads in the Vicinity of Highway Interchanges. Submitted to 87th Transportation Research Board Annual Meeting, Washington D.C., TRB.

Montgomery, D. C., E. A. Peck, et al. (2001). Introduction to Linear Regression Analysis.

Qin, X., J. Ivan, et al. (2004). "Selecting exposure measures in crash rate prediction for two-lane highway segments." Accident Analysis & Prevention **36**(2): 183-191.

Sawalha, Z. and T. Sayed (2006). "Traffic accident modeling: some statistical issues." Canadian Journal of Civil Engineering **33**: 1115-1124.

Shankar, V., J. Milton, et al. (1997). "Modeling accident frequency as zero-altered probability processes: an empirical inquiry." Accident Analysis & Prevention **29**(6): 829-837.

Shankar, V. N., G. F. Ulfarsson, et al. (2003). "Modeling crashes involving pedestrians and motorized traffic." Safety Science **41**(7): 627-640.