MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

Identification of an accident prediction model for red light camera analysis

Abstract

The purpose of this article was to develop an accident prediction model for motor vehicle crashes occurring within Miami-Dade County, Florida during 2008-2011.

Motor vehicle crash data were extracted from the Florida Department of Motor Vehicle and Highway Safety dataset for 40 intersections within Miami-Dade County, Florida for development of an accident prediction model. Each intersection was matched at least one of 20 red light camera (RLC) sites using selected geometric variables. In addition, each intersection examined was at least 2 miles away from any RLC site. The dependent variable examined was the number of injury crashes occurring at each intersection between 2008 and 2011. Poisson, negative binomial, and gamma model distributions were compared using the Pearson's chi square $\left(\chi^2\right)$, scaled deviance (G^2), and Akaike Information Criterion (AIC) goodness of fit tests. Our analysis indicated that the negative binomial distribution was the best fit among the three models. Inspection of the observed data also suggested that the outcome variable's distribution was over dispersed. This study provided guidance on the use of goodness of fit testing (GOF) statistics for Poisson, negative binomial, and gamma models which will allow other researchers to evaluate different models.

Keywords: accident prediction model, empirical bayes, red light cameras, motor vehicle crashes, goodness of fit

Volume 4 Issue 3 - 2016

Anthoni L, Nasar U Ahmed
Department of Epidemiology, Florida International University, Florida

Correspondence: Nasar U Ahmed, Department of Epidemiology, Robert Stempel College of Public Health, Florida International University, AHC5-468 Miami, Florida 33199, Florida, Email ahmedn@fiu.edu

Received: June 16, 2016 | **Published:** July 27, 2016

Abbreviations: NHTSA, national highway traffic safety administration; IIHS, insurance institute for highway safety; RLC, red light camera, RTM, regression to the mean; SPF, safety performance function; AADT, annual average daily traffic; GOF, goodness of fit testing; AIC, akaike information criterion; DF, degrees of freedom

Introduction

During 2012, approximately 48% of U.S. crashes occurred at an intersection or were intersection-related, of which over half (53%) were signalized.[1] This indicates an excessive proportion of crashes transpire at signalized intersections considering they constitute only 10% all intersections within the U.S.[2] In addition, crashes at signalized intersections result in considerable numbers of injuries and fatalities. According to the National Highway Traffic Safety Administration (NHTSA), 4,460 fatal crashes and 840,000 injury crashes occurred at a signalized intersection during 2012.[1] Despite national prevention efforts targeting this public health problem, the proportion of fatal crashes occurring at intersections with traffic signals increased 35% between 2000 and 2012.[1,3] Numerous signalized intersection crashes can be attributed to red light running which accounts for 22% of urban collisions and over one-fourth of all injury collisions.[4] According to the U.S. Department of Transportation, approximately 56% of Americans acknowledge running a red light.[5]

The Insurance Institute for Highway Safety (IIHS) estimated 683 persons were killed as the result of a red light running crash and another 133,000 persons were injured during 2012.[6] The IIHS also states that half of those killed in red-light running crashes are not signal violators, but the drivers and pedestrians who were struck.[7] The costs associated with red light running crashes are also significant. An examination of the safety impact of red light running crashes at intersections in the state of Texas found these crash types have a societal cost of $2 billion annually statewide.[8]

Several interventions have been implemented to decrease the risk of red light running crashes, including police enforcement, educational campaigns, and engineering modifications such as signal timing changes. Red light cameras (RLCs), however, are increasingly being used to discourage red light runners and decrease related crashes. Determining whether RLCs are effective is difficult for several reasons.[9] One issue is the phenomenon known as regression to the mean (RTM). Since cameras are typically installed at sites with the highest number of violations and/or crashes instead of random assignment, subsequent reductions in the event analyzed could simply be due to RTM, that is, data falling in line with the average results found in the area, even with or without any intervention implementation. If not accounted for, results may be biased in estimating the benefit of RLCs.[10]

Models that employ an Empirical Bayes analysis allow researchers account for RTM bias by estimating the number of collisions based on crash counts prior to RLC installation at treatment and comparison sites. The Empirical Bayes method requires an accident prediction model (i.e. safety performance function (SPF)) which is a multiple regression formula that fits collision data for comparison intersections to an independent set of variables that may be expected to affect safety such as speed limit or number of straight-through lanes. SPF's are used to assist agencies in network screening processes, that is, identifying sites that may benefit from a safety treatment. In addition, SPFs can be instrumental for countermeasure comparisons, and project evaluations.[11] To properly develop an SPF using motor vehicle crash data, the best fit model must be determined. Although linear regression models can be thought of as a good starting point, most researchers decline to use this statistical method. Previous crash studies have elucidated the problems with linear regression models including a lack of a distribution to sufficiently explain random, discrete, nonnegative, and sporadic events such as motor vehicle accidents.[12] Due to these problems, subsequent crash studies have adopted other models to develop SPF's including 1) Poisson regression, which is used to analyze data that are Poisson distributed and 2) negative binomial regression which accounts for over dispersion. Although these two models possess desirable characteristics to explain motor vehicle

crashes, they are not without limitations. One difficulty is that the two models do not account for under dispersion, where the variance of the data is less than its mean. Although this phenomenon is uncommon in crash analysis, it has been observed by various authors.[13,14] One model that has been proposed to handle under dispersion is the gamma probability count model.[15] This model can handle both over-dispersion and under-dispersion and reduces itself to a Poisson model when the variance is roughly equal to the mean of the number of crashes. Since several types of models are used to develop an SPF, goodness of fit testing can be employed to select the most appropriate distribution. The purpose of this paper was to determine the best fit regression model for the development of an SPF using historical motor vehicle crash data at 40 comparison sites without RLCs.

Materials and methods

The Poisson regression model is usually thought of as the starting point in developing an SPF since crash data are routinely Poisson distributed.[13] Poisson regression models are suited for motor vehicle crash analysis for several reasons, including analyzing events that occur randomly and independently over time[16] along with handling smaller sample sizes than linear regression.[17] In a poisson regression model, the probability of intersection i having y_i crashes per period is given by

$$P(y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i^!} \qquad i = 0, 1, 2, ...$$

where;

$P(y_i)$ = probability of roadway i having y_i crashes/period,

y_i = number of crashes for roadway i/period, and

λ_i = expected number of crashes per period, $E(y_i)$, also known as the Poisson parameter for roadway i.

The relationship between independent variables and expected number of crashes per period is a log-linear model of the following form:

$$Ln(\lambda_i) = \beta X_i \text{ or } \lambda_i = \exp(\beta X_i)$$

where;

ln = natural logarithm

β = vector of regression parameters

X_i = a vector of explanatory variables

The model coefficients are estimated through maximum likelihood methods. The likelihood function for the Poisson regression model is given as:

$$L(\beta) = \prod_{i=1}^{n} \frac{\left[\exp\left[-\exp(\beta X_i)\right]\right]\left[\exp(\beta X_i)\right]^{yi}}{y_i!} \qquad i = 0, 1, 2, ...$$

Poisson regression models assume equality of the mean and variance, which on occasion, is not found in crash data. Studies have shown that accident data can be over dispersed, that is, the variance exceeds the mean.[16] When over dispersion exists, it tends to underestimate the variance of the model coefficients.[18] To account for over dispersion, a negative binomial distribution is used as an alternative to the Poisson model. The negative binomial distribution introduces an over dispersion parameter which corrects for the variance and mean difference. As the over dispersion parameter approaches zero, the negative binomial distribution converges into a Poisson distribution. The probability function for the negative binomial regression model is given below:

$$P(y_i) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{y_i!\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\lambda_i}\right)^{1/\alpha}\left(\frac{\lambda_i}{\left(\frac{1}{\alpha}\right)+\lambda_i}\right)^{y_i} \qquad i = 0, 1, 2, ...$$

where;

$\Gamma(.)$ = gamma function

y_i = number of crashes per period for intersection i and,

α = overdispersion parameter

Considering n number of crashes, the likelihood function is given by:

$$L(\lambda_i) = \prod_{i=1}^{n} \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{y_i!\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\lambda_i}\right)^{\frac{1}{\alpha}}\left(\frac{\lambda_i}{\left(\frac{1}{\alpha}\right)+\lambda_i}\right)^{y_i} \qquad i = 0, 1, 2, ...$$

The primary advantage of the negative binomial distribution over the poisson distribution is that the overdispersion parameter provides increased flexibility into the modeling of the variance function, allowing the variance to differ from the mean. Thus, the negative binomial model can be an appropriate model to address these challenges. A limitation, however, of both Poisson and negative binomial models is its inability to handle under dispersion,[19] that is, when the mean exceeds the variance. As a result, gamma models have been proposed to handle under dispersed crash data.[13,20] The gamma probability model can be given as:

$$\Pr\left[y_i = j\right] = Gam(\alpha j, \lambda_i) - Gam(\alpha j + \alpha, \lambda_i); \qquad i = 0, 1, 2, ...$$

where;

$\lambda_i = \exp(\beta' X_i)$

$Gam(\alpha j, \lambda_i) = 1$, if j = 0, or $\frac{1}{\Gamma(\alpha j)}\int_0^{\lambda_i} u^{\alpha j - 1} e^{-u} du$, if j > 0, j =0, 1, …

The dispersion parameter is α, as in the negative binomial model. The value of α determines whether there is overdispersion, under dispersion, or equi dispersion. If α > 1, there is evidence of under dispersion. In contrast, if α < 1, there is overdispersion, and lastly, equi dispersion if α = 1, which reduces itself to a Poisson model. The conditional mean function and cumulative distribution function for the gamma probability model can be found in Oh et al.,[13]

Data description

Forty intersections within Miami-Dade County, Florida were selected for development of the SPF. Each intersection selected had been previously matched to at least one of 20 intersections with RLC's with respect to selected geometric and daily traffic variables. These variables included the intersection's annual average daily traffic (AADT) across all approaches and the total number of lanes and average speed limits for the intersection's major and minor roads. In addition, each selected intersection was at least 2 miles away from any

Citation: Anthoni L, Ahmed NU. Identification of an accident prediction model for red light camera analysis. *Biom Biostat Int J.* 2016;4(3):94–99. DOI: 10.15406/bbij.2016.04.00095

RLC site. Crash records for the selected intersections were extracted from the Florida Department of Motor Vehicle and Highway Safety dataset. Crashes were selected using several criteria:

1) The crash occurred between 2008 and 2011.
2) The crash occurred within 150 feet of the intersection
3) The crash resulted in at least one injury or fatality
4) The accident did not result in solely pedestrian or bicyclist injuries/fatalities.

The dependent variable was the number of injury crashes.

Goodness of fit testing (GOF) was used to determine the best fit model. GOF uses the properties of a hypothesized distribution to determine whether observed data can be generated from a given distribution.[21,22] Widely used GOF test statistics include the Pearson's chi square $\left(\chi^2\right)$ and scaled deviance (G^2).

As described in Ye et al.,[23] the Pearson's chi square value is calculated as:

$$\chi^2 = \sum_{i=1}^{n} \left[y_i - u_i / \sigma_i \right]^2$$

where;

y_i is the observed data,

u_i is the true mean from the model,

and σ_i is the error and is usually represented by the standard deviation of y_i.

The scaled deviance value is computed as twice the difference between the log likelihoods under the alternative and null model. A third test, Akaike Information Criterion (AIC) is also commonly used to measure model GOF. The model is defined as:

AIC = [-2 log (likelihood) + 2p],

Whereas likelihood is the probability of the data given a model and p is the number of parameters in the model. Lower AIC values indicates a better model fit of the data.[22,24] These three tests were used to select the most appropriate SPF model. SAS 9.2 was used to develop the Poisson, negative binomial, and gamma models using the generalized linear model (GENMOD) procedure. The GENMOD procedure for each distribution produced Pearson's chi square, scaled deviance, log-likelihood, and AIC values, which were subsequently compared to select the best model fit.

Results and discussion

Intersection characteristics: Descriptive characteristics for the 40 comparison sites are displayed in Table 1. Independent variables included the intersection's log [mean AADT] across all approaches, and mean speed limit & number of lanes for the intersection's major and minor roads, along with their standard deviation and 95% confidence intervals.

Results of the Poisson regression model are shown in Table 2 below. For the Poisson model, log [mean AADT], mean speed limit (minor road), number of lanes (minor road), were found to be statistically significant at α = 0.05. In contrast, the negative binomial model indicated, as shown in Table 3, log [Mean AADT] and mean speed limit (minor road) were the only statistically significant variables. The negative binomial model's over dispersion parameter value was 0.16, 95% CI (0.09, 0.29). Since the confidence interval did not overlap zero, thus indicates that over dispersion was present in the crash data, that is, the variance exceeded the mean.

The gamma model was then estimated to test for under dispersion and as shown in Table 4, the dispersion parameter (α) was estimated to be 0.23. In addition, the 95% CI did not overlap one indicating, as in the negative binomial model that over dispersion was present. The gamma model's significant variables included the log [mean AADT] and speed limit (minor road).

All variables for each model were then examined for multicollinearity by removing the least significant variable. For all three models, the least significant variable was number of lanes (major road). After removing the covariate for each model, all non-significant variables remained; indicating multicollinearity likely was not present.

Model goodness of fit: The model GOF for the Poisson, negative binomial, and gamma distributions were measured using the scaled deviance, Pearson Chi-squared Statistic, and AIC tests. The ratios of the scaled deviance and Pearson Chi-Square values to the model's degrees of freedom (DF) were then calculated to determine GOF with values close to 1 suggesting a good fit. All GOF test results are presented in Table 5. The negative binomial model achieved a Scaled Deviance/DF of 1.22 and a Pearson Chi-Squared/DF ratio of 1.09. In contrast, the Poisson model resulted in a scaled deviance/DF and chi square/DF ratios of 5.17 and 5.06, respectively. The log likelihood ratio for the two models resulted in a chi square value of 76.14 suggesting that the negative binomial distribution was a better fitting model. The gamma distributed model's scaled deviance/DF and log-likelihood values were similar to that of the negative binomial model, however, the gamma model's Pearson Chi-Square/DF ratio was only 0.24. In addition, the AIC value for the gamma model was slightly higher in comparison to the negative binomial model. Based on Table 5 results and evidence of over dispersion in Tables 3 & 4, the negative binomial model provided the best fit for developing the SPF.

Discussion: We considered three different regression models using motor vehicle crash data at 40 comparison intersections to develop an SPF for Empirical Bayes analysis. The regression models examined were Poisson, negative binomial, and gamma distributions. We fit each of these models to crash data from 2008–2011 in which the outcome variable was the count of injury crashes. GOF measurements indicated that the negative binomial distribution provided the best fit among the three models examined. Inspection of the observed data also suggested that the outcome variable's distribution was over dispersed, indicating that the negative binomial model was better suited to handle over dispersed data compared to the Poisson and gamma distributions. Similarly, the gamma model's parameter estimates indicated that over dispersion, and not under dispersion was present.

The negative binomial distribution is especially useful for count data whose variance exceeds the sample mean. In vehicle crash data, counts frequently depart from the Poisson distribution due to larger frequencies of extreme observations resulting in a greater variance compared to the mean, resulting in over-dispersion,[25] which was evident in our analysis. Although under dispersion can occasionally occur when analyzing motor vehicle crash data, it was not present according to our results.

Citation: Anthoni L, Ahmed NU. Identification of an accident prediction model for red light camera analysis. *Biom Biostat Int J.* 2016;4(3):94–99. DOI: 10.15406/bbij.2016.04.00095

A limitation of this analysis was the small number of injury crashes at each site. This was expected since injury crashes are infrequent. Approximately 29% of all crashes in the United States results in at least one injury/fatality.[1] In this study, two or three additional crashes may had influenced the results if few sites (4 – 5 intersections) were examined, however, by selecting a larger number of comparison sites this impact was reduced. Other possibilities to further improve the model fit would be to increase the number of crashes by examining additional intersections or using a longer study period. If using a longer study period however, one must be aware that any changes made to a site (i.e., increased number of lanes, law changes) during the period of analysis may be more likely, rendering the results of that site invalid.

Conclusion: The negative binomial model is currently one of the most common type of model employed in vehicle crash analysis.[23] On some occasions, however, the Poisson model can also be a suitable model. Gamma distributed regression models, although relatively new to vehicle crash analysis, is being seen as an alternative to both the Poisson and negative binomial models. Crash frequency data can present several issues in terms of data characteristics, thus new methodological approaches are constantly being introduced.[19] Thus, future studies can be conducted to examine vehicle crash data using novel statistical approaches.

Table 1 Intersection Characteristics

Comparison intersections n=40	Mean 2008 -2011	Standard deviation	95% Confidence interval (C.I.)
Mean AADT (1000's)	65.78	21.58	(58.87, 72.68)
Number of Lanes – Major Road	4.9	1.22	(4.51, 5.29)
Number of Lanes – Minor Road	3.68	1.05	(3.34, 4.01)
Speed Limit – Major Road	40.56	1.92	(39.95, 41.18)
Speed Limit – Minor Road	37.43	3.56	(36.29, 38.58)

Table 2 Poisson Regression Parameter Estimates

Parameter	Estimate	Standard Error	95% C.I.	P-value
Intercept	-10.52	1.47	(-13.39, -7.64)	< 0.01
Log Mean AADT	0.99	0.16	(0.68, 1.29)	< 0.01
Speed Limit – Major Road	0.01	0.02	(-0.03, 0.05)	0.59
Speed Limit – Minor Road	0.06	0.01	(0.03, 0.09)	< 0.01
Street Lanes – Major Road	0.06	0.04	(-0.02, 0.15)	0.16
Street Lanes – Minor Road	-0.12	0.05	(-0.23, -0.02)	0.02

Table 3 Negative Binomial Regression Parameter Estimates

Parameter	Estimate	Standard error	95% C.I.	P-value
Intercept	-12.12	3.39	(-18.76, -5.49)	< 0.01
Log Mean AADT	1.02	0.34	(0.35, 1.69)	< 0.01
Speed Limit – Major Road	0.04	0.05	(-0.05, 0.13)	0.41
Speed Limit – Minor Road	0.07	0.03	(0.02, 0.12)	0.01
Street Lanes – Major Road	0.06	0.09	(-0.12, 0.23)	0.54
Street Lanes – Minor Road	-0.15	0.11	(-0.37, 0.06)	0.16
Dispersion Parameter	0.16	0.05	(0.09, 0.29)	

Citation: Anthoni L, Ahmed NU. Identification of an accident prediction model for red light camera analysis. *Biom Biostat Int J.* 2016;4(3):94–99. DOI: 10.15406/bbij.2016.04.00095

Table 4 Gamma Regression Parameter Estimates

Parameter	Estimate	Standard error	95% C.I.	P-value
Intercept	-12.69	3.59	(-19.73, -5.67)	< 0.01
Log Mean AADT	1.02	0.36	(0.32, 1.72)	< 0.01
Speed Limit – Major Road	0.05	0.05	(-0.05, 0.15)	0.32
Speed Limit – Minor Road	0.07	0.03	(0.02, 0.13)	0.01
Street Lanes – Major Road	0.06	0.09	(-0.12, 0.25)	0.51
Street Lanes – Minor Road	-0.16	0.11	(-0.38, 0.06)	0.15
Dispersion Parameter	0.23	0.05	(0.15, 0.35)	

Table 5 Results of Model Goodness of Fit Tests

Distributions	Scaled deviance/DF	Pearson chi-square value/DF	AIC	Log likelihood
Negative Binomial	1.22	1.09	303.92	-144.96
Poisson	5.17	5.06	378.05	-183.03
Gamma	1.22	0.24	304.53	-145.27

Conclusion

This study provided guidance on the use of GOF statistics for Poisson, negative binomial, and gamma models which will allow other researchers to evaluate different models. Our results suggest the importance of comparing different probability distributions when modeling crash frequency data, particularly when over dispersion and under dispersion may exist.

Acknowledgment

We would like to thank Dr. Hafiz Khan who provided insight and expertise for the models selected for this paper and for comments that greatly improved the manuscript.

Conflict of interest

None.

References

1. http://www–nrd.nhtsa.dot.gov/pubs/812032.pdf.
2. http://safety.fhwa.dot.gov/intersection/signalized/presentations/sign_int_pps051508/short/.
3. http://www–nrd.nhtsa.dot.gov/Pubs/TSF2000.pdf.
4. Retting RA, Williams AF, Preusser DF, et al. Classifying urban crashes for countermeasure development. *Accid Anal Prev*. 1995;27(3):283–294.
5. Romano E, Tippetts AS, Voas R. Fatal red light crashes: The role of race and ethnicity. *Accid Anal Prev*. 2005;37(3):453–460.
6. http://www.ihs.org/iihs/topics/t/red–light–running/topicoverview.
7. http://www.iihs.org/externaldata/srdata/docs/sr4201.pdf.
8. Bonneson J, Zimmerman K. Federal Highway Administration. Development of guidelines for identifying and treating locations with a red light running problem. *Report Number FHWA/TX–05/0–4196–2*. 2004.
9. Shin K, Washington S. The impact of red light cameras on safety in Arizona. *Accident Analysis and Prevention*. 2007;39(6):1212–1221.
10. Retting RA, Ferguson SA, Hakkert AS. Effects of red light cameras on violations and crashes: A review of the international literature. *Traffic Inj Prev*. 2003;4(1):17–23.
11. US Department of Transportation, Federal Highway Administration. Revised assessment of economic impacts of implementing minimum levels of pavement marking retroreflectivity. *Report Number: FHWA–SA–10–016*. 2001.
12. Miaou SP, Lum H. Modeling vehicle accidents and highway geometric design relationships. *Accid Anal Prev*. 1993;25(6):689–709.
13. Oh J, Washington SP, Nam D. Accident prediction model for railway–highway interfaces. *Accident Analysis and Prevention*. 2006;38(2):346–356.
14. Cameron AC, Trivedi PK. *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, MA, India. 1988;1–370.
15. Winklemann R, Zimmermann KF. Recent developments in count data modelling: Theory and application. *Journal of Economic Surveys*. 1995;9(1):1–20.
16. Karlaftis MG, Golias I. Effects of road geometry and traffic volumes on rural roadway accident rates. *Accid Anal Prev*. 2002;34(3):357–365.
17. Jovanis PP, Chang HL. Modeling the relationship of accidents to miles traveled. Transportation Research Record. 1986;1068:42–51.
18. Abdel–Aty MA, Radwan AE. Modeling traffic accident occurrence and involvement. *Accid Anal Prev*. 2000;32(5):633–642.

Citation: Anthoni L, Ahmed NU. Identification of an accident prediction model for red light camera analysis. *Biom Biostat Int J*. 2016;4(3):94–99. DOI: 10.15406/bbij.2016.04.00095

19. Lord D, Mannering F. The statistical analysis of crash–frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*. 2010;44(5):291–305.

20. Winklemann R. Duration dependence and dispersion in count–data models. *Journal of Business & Economic Statistics*. 1995;13(4):467–474.

21. Read TRC, Cressie N. *Goodness–of–Fit Statistics for Discrete Multivariate Data*. Springer Series in Statistics, New York. 1988.

22. Khan HMR, Saxena A, Rana S, et al. Bayesian method for modeling male breast cancer survival data. *Asian Pac J Cancer Prev*. 2014;15(2):663–669.

23. Ye Z, Zhang Y, Lord D. Goodness–of–fit testing for accident models with low means. *Accident Analysis and Prevention*. 2013;61:78–86.

24. http://www.hindawi.com/journals/tswj/aip/604581/.

25. Hu MC, Pavlicova M, Nunes EV. Zero–inflated and hurdle models of count data with extra zeros: examples from an HIV–risk reduction intervention trial. *Am J Drug Alcohol Abuse*. 2011;37(5):367–375.

Citation: Anthoni L, Ahmed NU. Identification of an accident prediction model for red light camera analysis. *Biom Biostat Int J.* 2016;4(3):94–99.
DOI: 10.15406/bbij.2016.04.00095