

Confirmatory and Exploratory Data Analyses Using PROC GENMOD: Factors Associated with Red Light Running Crashes

Li wan Chen, LENDIS Corporation, McLean, VA
Forrest Council, Highway Safety Research Center, University of North Carolina
Yusuf Mohamedshah, LENDIS Corporation, McLean, VA

ABSTRACT

Recent studies have shown that more than 200,000 red light running crashes occur annually in the United States. It can be hypothesized that the majority of these accidents result from “driver error,” whether intentional or inadvertent. However, the purpose of this research is the possible contribution of the geometric characteristics of intersections to RLR crash risk -- such factors as entering or cross-street traffic volumes and intersection width. The focus of this current paper is the statistical modeling strategy chosen from many possible strategies to answer the questions using data from the Federal Highway Administration’s the Highway Safety Information System (HSIS). The HSIS contains multi-year, multi-state data on accidents, roadways, and traffic volumes. Poisson and Negative Binomial regression models in PROC GENMOD were utilized to test the hypotheses. Using this example, we are attempting to demonstrate the use of this strategy in terms of when to model the dependent variable for different conditions, how to select predictors if multiple years involved, and how to approach interaction effects. Above all, is the model parsimonious, consistent, or stable? In short, when do we have enough confidence to reject (or fail to reject) the research hypotheses of any association between dependent and independent variables.

INTRODUCTION

It has been shown that at urban signalized intersections, red light running (RLR) crashes have a higher overall severity level than other types of accidents. The possible contribution of the geometric characteristics of the urban four-legged signalized intersections to RLR crash risk was studied [Mohamedshah, Y., et al., 2000]. The current paper is focused on the statistical analyses used in, and subsequent to, that effort. One of the popular modeling strategies is prediction; however, this paper is to explore association. In other words, the estimated parameters are interpreted as the interval effect of geometric variables on RLR accidents [Allison, P. D. , 1999]. The accident and intersection data are extracted

from the HSIS. The four-year RLR accidents occurred during 1993-1996 and the limited set of geometric variables is from 1996 intersection file from California.

The primary hypothesis is that the width of the crossing street and the entering and crossing traffic volume are associated with RLR accidents. The width of the street is defined as the total number of crossing street lanes in this study. A limited number of other geometric factors such as left turn lanes, right turn lanes, traffic control type, and street traffic flow are also examined. In the next section, the data and statistical methodology applied in this study are presented. Conclusion and several aspects of the future studies are also discussed.

DEPENDENT AND INDEPENDENT VARIABLES

The list of the variables used in the analysis was shown in Table 1¹. It is noted that the general accident reporting threshold in California is currently \$500 or personal injury. The count of RLR accidents over a four-year period are restricted to two-vehicle crashes. A given two-vehicle RLR crash was assigned to one of the intersection streets based on the street used by the “at-fault” vehicle (1,859 RLR crashes on cross-street versus 2,850 on mainline). About 8,578 RLR crashes which occurred at “freeway ramps” are excluded from the current study.

PROC GENMOD: POISSON AND NEGATIVE BINOMIAL REGRESSION MODELS

Based on the past studies, Poisson and Negative Binomial (N-B) distributions have been applied to model the occurrence of accidents [i.e., Miaou, S., 1994; Mountain, L., et al. 1998]. Poisson Regression with overdispersion correction was used to explore the association in this

¹. Some of the geometric variables have been transformed. The detailed definition and descriptive statistics are available upon request.

study. Based on these initial results, N-B Regression coefficients were used to quantify the results. The SAS programs for Poisson and N-B regression models and related statistics are presented next.

Poisson regression model (See next program code).

The Poisson Regression distribution(D=P) describes the expected frequency of RLR crashes at particular street in a given time period [SAS® Technical Report P-243, 1993;SAS® *OnlineDoc(TM)*, 1999]. The Pearson chi-square correction (PSCALE) was applied to correct overdispersion in the present study.

An "offset variable" is used to adjust Poisson and N-B models for differential "exposure" in data records (e.g., different lengths of time periods, populations, or the amount of vehicles travel). Past studies have used different variables or scale factors to represent the offset variable depending on the goals of the studies [i.e., Al_Ghamdi, A., 1993; Kleinbaum, D.G., et al.,1988; Miaou, S.,1994]. In current study, we chose not to use an offset variable for two primary reasons. First, the time periods in all records is the same -- four years. Second, and perhaps more important, the most logical offset variable would be Average Annual Daily Traffic (AADT), which would differ between intersection streets. However, since we wished to examine AADT as a possible predictor of RLR crashes, it was inappropriate to use it as an offset. The parameters are estimated using the maximum likelihood method (ML). Below is one of the Poisson regression models:

```
*SIMPLE POISSON REGRESSION MODEL;
DATA MAINLINE;
  SET RLR.MAIN;
  AADTLNE=XSTAADT/XSTLAN3;
  QAADTLNE=SQRT(AADTLNE);
  SQRMAADT=SQRT(ML_AADT); RUN;
PROC GENMOD DATA=MAINLINE;
  MODEL ACC93=QAADTLNE/D=P PSCALE;
RUN;
```

N-B regression model. Efficient estimate for count data can be produced by N-B regression which is a generalization of Poisson Regression models. One of the SAS-PROC GENMOD with N-B regression models is as follows [SAS® Technical Report P-243, 1993;SAS® *OnlineDoc(TM)*, 1999]:

```
*N-B REGRESSION MODEL;
DATA MAINLINE;
  SET RLR.MAIN;
  AADTLNE=XSTAADT/XSTLAN3;
  ML1000=ML_AADT/1000;
  QAADTLNE=SQRT(AADTLNE);
  SQRMAADT=SQRT(ML_AADT); RUN;
PROC GENMOD DATA=NS3TRTMA;
  MODEL ACC936=ML1000 QAADTLNE
  DPREACT1 DPREACT2
  / DIST=NEGBIN LINK=LOG TYPE3
  OBSTATS RESIDUALS LRCI WALDCI;
  MAKE 'OBSTATS' OUT=C; RUN;
```

In PROC GENMOD, the expected RLR crashes is related to a linear predictors through a monotonic differentiable link function g ($LINK=LOG$). X_i is a fixed known vector of explanatory variables, and β is a vector unknown parameters.

$$g(u_i) = X_i^T \beta \tag{1}$$

The log-likelihood function for the N-B distribution are parameterized in terms of the mean and dispersion parameter in Equation(2). SAS-PROC GENMOD procedure applies a ridge-stabilized Newton-Raphson algorithm to maximize the log likelihood function with respect to the regression parameters [SAS® Technical Report P-243, 1993;SAS® *OnlineDoc(TM)*, 1999].

$$L = y_i \log(ku) - (y_i + \frac{1}{k}) \log(1 + ku) + \log\left(\frac{\Gamma(y_i + \frac{1}{k})}{\Gamma(y_i + 1) \Gamma(\frac{1}{k})}\right) \tag{2}$$

Let the resulting unconstrained parameter estimates be β_u and constrained parameters be β_c . Then the likelihood ratio (LR) statistics Equation(3) has an asymptotic chi-square distribution under the hypothesis that the TYPE III contrast is equal to 0, with degrees of freedom equal to the number of parameters associated with the effect (TYPE3 LRCI)[SAS® Technical Report P-243, 1993;SAS® *OnlineDoc(TM)*, 1999].

DIFFERENT CONDITIONS – DIFFERENT MODELS

Given that the goal of the effort was to examine the association of various intersection design factors (geometrics) with RLR crashes, different models for mainlines and cross-streets were developed due to the

different geometric and resulting crash conditions. For example, on average, there were two RLR accidents involving vehicles entering from the 1756 mainline streets (mean=1.623, minimum=0 and maximum=20), and one RLR accident involving a vehicle entering from the 1756 cross-streets (mean=1.059, minimum=0 and maximum=12), 1993-1996. The cross-street's crossing street (i.e., the mainline) is mostly wider than the mainline's crossing street (mode=6-lane versus mode=2-lane). The mainline AADT is higher than cross-street AADT (mean=31,656 versus mean=8,289).

VARIABLE SELECTION AND INTERACTION EFFECT

The variable selection procedure we used for the final models was the result of two factors. First, due to the nature of real-world intersection design, the geometric variables are correlated with each other, i.e., the wider street tends to have higher traffic volume, and higher traffic volume is associated with certain types of traffic control. And each of these geometric factors can be associated with RLR crashes. Thus, in the final selection of variables, we needed to understand both the unique contribution (as a single predictor) and partial contribution (as one of the predictors) of each geometric factor on RLR crashes.

Second, we wished to establish stable models which estimate well regardless of the time period chosen. Note that the hypotheses testing of Poisson or N-B regression coefficients are based on the chi-square test which is proportional to the sample size. The more years, the more cumulated RLR accidents per street, which would lead one to accumulate data across years in all testing. However, to establish the most stable models, instead of using only the four-year accumulated data, the predictors were pre-selected from simple Poisson regression analyses with overdispersion correction based on three different time periods – annual, 2-year, and 4-year data. (Here we used a rule of thumb that any predictor with a p-value less than .25 was considered a potential predictor.²)

After the above analysis process, a series of models is built with potential predictors with four-year data. A potential predictor was dropped due to high probability level. Then, two-way interactions between any two potential predictors were tested one at each step.

² Mainline models reach certain degree of stability within one year and cross-street models within two years.

Since the same strategy was used for both streets, Table 2 only includes the unique contribution of each geometric factor to RLR crashes for the mainline as the entering street. Those variables with high importance levels are shown in bold face type in that table. Table 3 then shows the partial contribution of each selected variable in the final model for the mainline as the entering street. Note that not all of the important variables in Table 2 appear in Table 3 due to the different impacts of unique and partial contributions. In summary, the variables in Table 3 and Table 4 (for the cross-street as entering street) are the more important geometric variables in terms of RLR crashes.

In the database used, the geometric variables are assumed to be the same across 1993-1996. Thus, rather than use statistics concerning point estimates, confidence intervals are used to explain the effects. An example of the interval effects of the important geometric variables to RLR crashes are interpreted as follows. Based on the cross-street model in Table 4 (i.e., cross-street as entering street), the 95 percent confidence interval for the expected change in RLR crashes are (2%, 12%) for each one-lane increase in the crossing street width, if traffic control type, crossing street AADT, and left turn channelization are held constant.

MODEL EVALUATION

Evaluation of the fit of models developed within the N-B regression can be done by using Goodness-of-fit statistics. We use the technique in this study. Overall, the final models have a decent fit to the data³. To ensure the stability of the N-B Regression coefficients, the complete mainline as entering street model was retested based on yearly data. Cross-street as entering street model was confirmed using two-year data. The results indicate the final models appear consistent in terms of Goodness-of-fit statistics and parameters. One of the validation examples is shown in Table 5.

CONCLUSION AND DISCUSSION

Based on the above results, from an engineering or enforcement point of view, the crossing street lanes, mainline AADT, traffic control type, and left turn channelization are found important to RLR crashes in the cross-street model³. The traffic volume and traffic

³The relevant discussion can be found in *Association of Selected Intersection Factors with Red Light Running Crashes* Institute of Transportation Engineers, Institute of Transportation Engineers,

control type are considered important to mainline RLR accidents. Future studies of this topic could be improved by more detailed data. For example, in addition to added data on speed limit and traffic signal timing, additional detail on both the type of left-turn channelization (e.g., painted or raised bar) and the actual crossing street width might prove helpful (we used number of travel lanes as a surrogate).

From a modeling point of view, this multi-year, multi-variable data set has provided interesting opportunities for analysis methodologies. A significant number of Poisson or N-B regressions were used to select and validate the potential predictors. The trade-off is the more the hypotheses testing, the more the cumulated Type I error. However, stability and consistence were considered more important in this study; therefore, the modeling approach is somewhat non-conventional. Confidence intervals, rather than point estimates, were examined due to the nature of the variables in the database. A choice was made to not use a traditional scale factor because the obvious scale factor, AADT, was used as a predictor variable. And finally, the use of multiple models for different time frames provided increased information on predictor variables that would be stable across different time periods.

REFERENCES

- Al_Ghamdi, A. (1993) *Comparison of Accident Rates Using the Likelihood Ratio Testing Technique*. *Transportation Research Record 1401*. TRB, NRC, Washington, D.C., U.S.A.
- Allison, P.D. (1999) *Logistic Regression Using the SAS® System: Theory and Application*. Cary, NC: SAS Institute Inc.
- Council, F.M. and Williams, C.D. (1995) *Highway Safety Information System Guidebook For the California Data Files*, Vol. 1., FHWA, Washington, D.C.
- Kleinbaum, D.G., Kupper, L.L., and Muller, K.E. (1988). *Applied Regression Analysis and Other Multivariable Methods*. PWS-KENT Publishing Company, Boston.
- Maher, J.M , and Summersgill, I. (1996) *A Comprehensive Methodology for the Fitting of Predictive Accident Models*. *Accident Analysis and Prevention*, 28: 281-296.
- Miaou, S. (1994). *The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions*. *Accident Analysis and Prevention*, 26 : 471-482.
- Mohamedshah, Y., Chen, L. W., and Council, F.M..(2000). *Association of Selected Intersection Factors with Red Light Running Crashes*. Unpublished paper to be published by Institute of Transportation Engineers at a later date.
- Mountain, L., Maher, M., and Fawaz, B. (1998).*The Influence of Trend on Estimates of Accidents At Junctions*. *Accident Analysis and Prevention*, 30: 641-649.
- SAS Institute Inc. (1999) *SAS® OnlineDoc(TM)*, Version 7-1, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1993). *SAS® Technical Report P-243, SAS/STAT® Software: The GENMOD Procedure*, Release 6.09, Cary, NC: SAS Institute Inc.

Acknowledgments

This paper is a product of interaction. Valuable comments and tips from Fred Duiven, Mike Griffith, David Harkey, and Dick Stewart make the process possible and wonderful. Dr. Council provides numerous insightful suggestions to this paper. The authors sincerely appreciate the hypothesis from Dr. Sam Tignor. Thank goes to Yusuf Mohamedshah for putting the data together. This paper reflects solely authors' viewpoint.

Contact Information

Li wan Chen
LENDIS Corporation
6300 Georgetown Pike, Rm T-211
McLean, VA 22101
Work Phone: 202-493-3466
Email:li_wan.chen@fhwa.dot.gov.

(2000). Both the details of the traffic signal phasing and speed limit on the approach streets may be potentially important predictor variables. Unfortunately, neither are available in the existing database.

Table 1. Definition of the RLR crashes and geometric intersection variables.

Definition	Category	Mainline as entering street	Cross-Street as entering street
1993 RLR crashes	count	ACC93	ACC93
1994 RLR crashes	count	ACC94	ACC94
1995 RLR crashes	count	ACC95	ACC95
1996 RLR crashes	count	ACC96	ACC96
1993-1994 RLR crashes	count	ACC934	ACC934
1995-1996 RLR crashes	count	ACC956	ACC956
1993-1996 RLR crashes	count	ACC936	ACC936
total crossing street lanes	lanes plus exclusive turn lanes in both directions	XLTOTLNS (1-10 lanes)	XLTOTLNS (2-12 lanes)
AADT on the entering street	average annual daily traffic volume	ML1000 (ML_AADT/1000)	SQRMAADT [SQRT(ML_AADT)]
AADT on the crossing street per lane	AADT per lane	QAADTLNE [SQRT(XSTAADT/XSTLAN3)]	QAADTLNE [SQRT(XSTAADT/XSTLAN3)]
left turn lanes on the entering street	0=no left turn 1 = yes (curbed median, painted, raised bars)	DMLEFT	DMLEFT
right turn lanes on the entering street	0=no 1=yes	DMRIGHT	DMRIGHT
left turn lanes on the crossing street	0=no 1=yes	DXLEFT	DXLEFT
right turn lanes on the crossing street	0=no 1=yes	DXRIGHT	DXRIGHT
left turn lanes interaction between entering and crossing streets	00=no left turn on entering and crossing streets. 01=no left turn entering street with left turn crossing street. 10=no left turn crossing stree with left turn entering street. 11= left turn with entering and crossing streets	MLXL(DMLEFT*DXLEFT)	MLXL(DMLEFT*DXLEFT)
phasing-traffic control type	2-phase multi-phase	DPHASE (multi-phase vs 2-phase)	DPHASE (multi-phase vs 2-phase)
traffic control type	1 = Pre-timed signals 2 = Semi-traffic actuated signals 3 = Full-traffic actuated signals	DPREACT1 (semi vs pre-time) DPREACT2 (full vs pre-time)	DPREACT1 (semi vs pre-time) DPREACT2 (full vs pre-time)
entering street traffic flow	N = 2-way traffic, no left turn permitted P = 2-way traffic, left turn permitted R = 2-way traffic, left turn restricted during peak hours W = 1-way traffic, no left turn Z = Other	DMTRFLO1 (P vs N) DMTRFLO2 (R vs N) DMTRFLO3 (W vs N) DMTRFLO4 (Z vs N)	DMTRFLO1 (P vs N) DMTRFLO2 (R vs N) DMTRFLO3 (W vs N) DMTRFLO4 (Z vs N)
crossing street traffic flow	N = 2-way traffic, no left turn permitted P = 2-way traffic, left turn permitted R = 2-way traffic, left turn restricted during peak hours W = 1-way traffic, no left turn Z = Other	DXTRFLO1 (P vs N) DXTRFLO2 (R vs N) DXTRFLO3 (W vs N) DXTRFLO4 (Z vs N)	DXTRFLO1 (P vs N) DXTRFLO2 (R vs N) DXTRFLO3 (W vs N) DXTRFLO4 (Z vs N)

Table 2. Simple Poisson regressions with overdispersion correction on Mainlane streets.

variables	1993		1994		1995		1996		1993-1994		1995-1996		1993-1996	
	Parameter	P	Parameter	P	Parameter	P	Parameter	P	Parameter	P	Parameter	P	Parameter	P
XLTOTLNS	+	0.23	+	0.40	+	0.05	+	0.20	+	0.20	+	0.05	+	0.06
ML1000	+	0.00	+	<.01	+	<.01	+	<.01	+	<.01	+	<.01	+	<.01
QAADTLNE	+	0.02	+	<.01	+	<.01	+	<.01	+	<.01	+	<.01	+	<.01
DMLEFT	-0.072	0.61	-0.042	0.78	+	0.72	+	0.28	-0.058	0.62	+	0.38	+	0.82
DMRIGHT	-0.007	0.94	+	0.83	+	0.13	+	0.76	+	0.93	+	0.29	+	0.50
DXLEFT	+	0.17	+	0.59	+	0.12	+	0.15	+	0.23	+	0.07	+	0.08
DXRIGHT	+	0.19	+	0.88	+	0.15	+	0.20	+	0.50	+	0.10	+	0.18
DPHASE	-0.065	0.51	+	0.76	+	0.36	+	0.23	-0.027	0.74	+	0.20	+	0.58
DPREACT1	+	0.13	+	0.10	+	0.38	-0.136	0.50	+	0.25	+	0.93	+	0.48
DPREACT2	+	0.01	+	0.53	+	0.08	+	0.35	+	0.01	+	0.11	+	0.01
DMTRFLO1	+	0.82	-0.314	0.63	-0.072	0.79	+	0.89	+	0.59	-0.016	0.94	+	0.79
DMTRFLO2	-0.565	0.37	+	0.02	-0.619	0.30	-0.788	0.25	-0.448	0.38	-0.698	0.18	-0.574	0.19
DMTRFLO3	+	0.16	+	0.28	+	0.90	+	0.15	+	0.02	+	0.32	+	0.05
DMTRFLO4	-0.470	0.61	-0.171	0.76	+	0.51	+	0.15	+	0.72	+	0.20	+	0.32
DXTRFLO1	+	0.53	+	0.76	+	0.91	-0.445	0.34	+	0.83	-0.225	0.60	-0.071	0.86
DXTRFLO2	+	0.26	-0.096	0.87	-0.598	0.65	+	0.77	+	0.37	+	1.00	+	0.60
DXTRFLO3	+	0.58	-0.310	0.82	+	0.72	-0.138	0.78	+	0.81	+	0.98	+	0.88
DXTRFLO4	+	0.89	-0.310	0.82	-0.087	0.95	-20.350	1.00	-0.087	0.94	-1.099	0.46	-0.547	0.61

¹ '+' = the parameter is positive.

Table 3. N-B Regression: RLR Crashes on the Mainline as Entering Street, 1993-1996.

variables	DF	Standard		LR 95%		odds ratio		LR Type 3	
		Estimate	Error	Confidence Limits				Chi-	P
Intercept	1	-0.495	0.127	-0.745	-0.245	0.475	0.782		
ML1000	1	0.013	0.002	0.009	0.018	1.009	1.018	38.300	<.0001
QAADTL	1	0.006	0.002	0.003	0.009	1.003	1.009	15.330	<.0001
DPREACT	1	0.129	0.125	-0.115	0.374	0.891	1.453	1.070	0.301
DPREACT	1	0.328	0.085	0.162	0.494	1.176	1.638	14.840	0.000
Dispersion	1	0.938	0.061	0.823	1.064				

Table 4. N-B Regression: RLR Crashes on the cross-street as Entering Street, 1993-1996.

variables	DF	Standard		LR 95%		odds ratio		LR Type 3	
		Estimate	Error	Confidence Limits				Chi-	P
Intercept	1	-0.702	0.162	-1.020	-0.386	0.361	0.680		
DMLEFT	1	-1.086	0.265	-1.620	-0.577	0.198	0.562	17.99	<.0001
DXLEFT	1	-0.588	0.134	-0.851	-0.326	0.427	0.722	19.19	<.0001
MLXL	1	1.123	0.273	0.598	1.672	1.818	5.322	17.99	<.0001
XLTOTLN	1	0.064	0.024	0.017	0.112	1.017	1.118	7.01	0.0081
SQRMAA	1	0.007	0.001	0.006	0.009	1.006	1.009	73.28	<.0001
DPREACT	1	-0.235	0.145	-0.520	0.048	0.594	1.049	2.64	0.1043
DPREACT	1	0.299	0.101	0.100	0.498	1.106	1.645	8.72	0.0031
Dispersion	1	0.784	0.070	0.655	0.929				

Table 5. Mainline Model Validation, 1993.

Year=1993			
Criterion	DF	Value	Value/DF
Deviance	1751	1330.029	0.760
Scaled	1751	1330.029	0.760
Pearson	1751	1725.487	0.985
Scaled	1751	1725.487	0.985
Log		-1294.114	

Analysis Of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-1.907	0.204	-2.312	-1.510	87.110	<.0001
ML1000	1	0.013	0.003	0.007	0.019	15.980	<.0001
QAADTL	1	0.005	0.002	0.000	0.009	4.390	0.036
DPREACT	1	0.356	0.193	-0.023	0.734	3.380	0.066
DPREACT	1	0.488	0.137	0.224	0.760	13.280	0.000
Dispersion	1	1.244	0.165	0.944	1.594		