2016

Car Crash Conundrum

Mohammad Sadra Sharifi
Utah State University, sadra.sharifi@gmail.com

David Tate
Utah State University

Spenser Tingey
Utah State University

Car Crash Conundrum

Course: Categorical Data Analysis

STAT 5120

Prepared by:
Mohammad Sadra Sharifi, David Tate, Spenser Tingey

Logan, Utah
United States

Contents

List of Tables

Executive Summary

The following report is a compilation of injury traffic crashes analysis using logistic regression. The purpose of this study is to use real world data collected in Orange County, California to learn how crash characteristic relate to probability of injury crashes. The data used in this project involves crashes that occurred in 1998 on six Orange County freeways including Interstates 5 and 405, and State Routes 22, 55, 57 and 91. This dataset involves some information about crash typology. The real world data was processed and potential dependent variables were identified using explanatory analysis. Then, processed data were imported to SAS to estimate logistic regression coefficients. Also, several logistic regression models concentrating on different dependent variable interactions were fitted. Finally, the best model was selected using deviance as goodness-of-fit measure. The final model gives following results: Crashes involving speeding and alcohol usage cause to higher probability of injury than crashes due to other causes. Crashes on the weekend cause to higher probability of injury than crashes on weekdays. Crashes off the road cause to higher probability of injury than crashes that occur on the road. Also, Highway 91 was identified as the highest risky highway for injury crashes comparing other highways which involved in this study.

Introduction

Safety analysis is one of the most important branches of traffic engineering. Designing roads without proper safety level can cause injury crashes on the roadways (Baratian et al., 2014). In some developing countries, more people have been killed in highway crashes than have in all of the wars in which the nation has been involved. Also, many people die from vehicles crashes in developed countries too. In the year 2000, 41,821 people were killed in accidents on U.S highways and a there was a total of 6,394,000 police reported crashes. Preventing accidents is one of the most important tasks of traffic engineers and it is necessary for them to study, analyze, and predict accidents with suitable tools. Applied statistical techniques are a common tool used to develop models that widely used in many Transportation Engineering applications (for example see Asgari et al., 2014; Asgari and Jin, 2015; Asgari and Jin, 2016a; Asgari and Jin, 2016b; Asgari, 2015; Soltani-Sobh et al, 2016, Khalilikhah et al., 2016, Zolghadri et al., 2013, Zolghadri et al., 2016). The main goal of this project is to analyze the factors that impact on probability of injury crashes using real data set. Because of categorical nature of variables which can impact on injury crashes, logistic regression will be used to identify the most important factors which affect on the probability of injury crashes.

Data Description and Methods

The data used in this project involves crashes that occurred in 1998 on six Orange County, California freeways including Interstates 5 and 405, and State Routes 22, 55, 57 and 91. These are crashes that are based on police reports. The crash data were obtained from the Traffic Accident Surveillance and Analysis System (TASAS) maintained by the California Department of Transportation (Caltrans). For calendar year 1998, 9,341 collisions involving vehicles are recorded in the database for these six major highways. After implementation of the filtering and cleaning, a sample of 1,191 collisions was generated. This represents 12.8% of the total collisions on the six major Orange County freeways.

This dataset involves some information about crash typology. Crash typology is defined according to three primary crash characteristics: 1- crash type 2- crash location 3- crash severity. Crash type is defined based on the type of collision (for example rear end, sideswipe, or hit object), the number vehicles involved, and the movement of these vehicles prior to the crash. Crash location is defined based on the location of the primary collision (for example left lane, interior lanes, right lane, right shoulder area, and off-road beyond right shoulder area) and crash severity is defined in terms of injuries and property damage only crashes. The variables and their definitions in the raw data set are shown in the following table.

Table 1. The variables and their definitions in the raw data set

hour	Hour of the day
route	Highway number on which crash occurred
cause	Cause of crash (alcohol, speeding, other)
dayofwk	Day of the week
type	Auto-auto, auto-pedestrian, other
numvehs	Number of vehicles in crash
dry	Dry or wet road surface
xrgt50c	Median volume/occupancy right lane
vleftmuc	Mean volume left lane
vmidmuc	Mean volume middle lane
vrgtmuz	Mean volume right lane
acctype6	Accident type (rear-end, weaving, etc.)
locatn5	On-road, off-road
segment	Daylight, dusk, dark

The processed data were imported to SAS to fit logistic regression model. Procgenmodstatement in SAS was used to estimate the coefficients of the model.The hypothesis in this model that we are interested to test is that what variables are associated with the injury crashes simultaneously and the assumption is that the log odds of injury crashes change linearly with respect to dependent variables.

Results

Exploratory Analysis

A correlation matrix was computed to assess pairwise correlations between significant explanatory factors, and thereby determine which factors may be confounding. Using the p-value given in SAS, we were able to determine if any two variables have a statistically significant correlation. In the following table, a 1 entry denotes correlation and a 0 denotes no correlation (significance level .05):

Table 2. Correlation matrix

	hour	route	pmloop	cause	dayofwk	type	numvehs	dry	numvehs4	xrgt50c	vleftmuc	vmidmuc	vrgtmuz	acctype6	locatn5	segment	segment5	segment3
hour																		
route	1																	
pmloop	0	1																
cause	0	0	0															
dayofwk	0	0	0	0														
type	1	0	0	1	0													
numvehs	1	0	0	1	0	1												
dry	1	0	1	1	1	1	1											
numvehs4	1	0	0	1	0	1	1	1										
xrgt50c	0	1	0	1	0	1	1	1	1									
vleftmuc	0	1	0	0	1	1	1	1	1	1								
vmidmuc	0	1	1	0	0	1	1	1	1	0	1							
vrgtmuz	0	1	1	0	0	1	1	1	1	1	1	1						
acctype6	1	0	0	1	0	1	1	1	1	1	1	1	1					
locatn5	0	0	0	0	0	0	1	0	1	0	1	1	0	1				
segment	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0			
segment5	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1		
segment3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	

This matrix helped us determine which factors to include in the model so that there would be no confounding factors. We used route, cause, dayofwk, and locatn5 of which no pair has a significant correlation. We needed to process these data so that we could import them into SAS. The following table shows how these variables are coded in our model.

Table 3.Nominal variables

Cause	alcohol	1 if cause is alcohol, 0 otherwise
	speeding	1 if cause is speeding, 0 otherwise
	other	1 if cause is other, 0 otherwise
Location	off-road	1 if location is off-road, 0 otherwise
	on-road	1 if location is on-road, 0 otherwise
Route	H5	1 if accident took place on highway 5, 0 otherwise
	H22	1 if accident took place on highway 22, 0 otherwise
	H55	1 if accident took place on highway 55, 0 otherwise
	H57	1 if accident took place on highway 57, 0 otherwise
	H91	1 if accident took place on highway 91, 0 otherwise
	H405	1 if accident took place on highway 405, 0 otherwise
Day of Week	Weekend	1 if accident occurred on weekend, 0 otherwise
	Weekday	1 if accident occurred on weekday, 0 otherwise
Outcome variable	Outcome	1 if injury occurred, 0 if only property damage occurred

Further Analysis

Next, we wanted to find a final model for the data and determine which interactions (if any) are significant. We performed model comparisons using the model deviance and computing the chi-square test statistic and corresponding p-value. Since we have two nominal categorical variables, cause and route, when we do an interaction involving one of these terms, we consider all pairwise interactions between each dummy variable and the other factor. For example, for cause*offroad, there are two interaction terms, alcohol*offroad and speeding*offroad. The following table summarizes the statistics relevant to the different models we considered:

Table 4. Different models chi-square statistics

Model	Log Likelihood	df	χ^2	p-value
base	-680.0167			
base+cause*locatn	-672.0112	2	16.011	.0003
base+cause*route	-675.8797	10	8.274	.6021
base+cause*dayofwk	-678.2648	2	3.5038	.1734
base+locatn*route	-679.7346	5	2.5642	.7668
base+locatn*dayofwk	-679.8303	1	.3728	.5415
base+route*dayofwk	-677.5907	5	4.852	.4341

Based on the chi-square tests above, we decided to include the interaction between cause and locatn in our final model. Then there are two additional terms in the model, alcohol*offroad and speeding*offroad. The following table summarizes the estimated coefficients for each term in the model, along with confidence intervals and significance tests:

Table 5. Coefficients estimation

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.3367	0.2322	-2.7917	-1.8817	101.30	<.0001
h5	1	0.5610	0.2071	0.1550	0.9670	7.33	0.0068
h22	1	0.3687	0.2546	-0.1303	0.8677	2.10	0.1476
h55	1	0.3796	0.1962	-0.0048	0.7641	3.75	0.0529
h57	1	0.4289	0.2579	-0.0765	0.9343	2.77	0.0962
h91	1	0.5875	0.2561	0.0855	1.0895	5.26	0.0218
alcohol	1	1.8266	0.4497	0.9452	2.7081	16.50	<.0001
speeding	1	0.9429	0.2012	0.5485	1.3373	21.96	<.0001
weekend	1	0.3556	0.1385	0.0841	0.6272	6.59	0.0102
offroad	1	1.4692	0.2459	0.9871	1.9512	35.68	<.0001
speeding*offroad	1	-1.1198	0.3484	-1.8028	-0.4369	10.33	0.0013
alcohol*offroad	1	-1.9392	0.6228	-3.1598	-0.7186	9.70	0.0018
Scale	0	1.0000	0.0000	1.0000	1.0000		

The final model, with the estimated regression coefficients is:

$$logit(\hat{\pi}) = -2.3367 + .5610X_{H5} + .3687X_{H22} + .3796X_{H55} + .4289X_{H57} + .5875X_{H91} + 1.8266X_{alcohol} + .9429X_{speeding} + .3556X_{weekend} + 1.4692X_{offroad} - 1.1198X_{speeding*offroad} - 1.9392X_{alcohol*offroad}$$

The estimated coefficients on the explanatory factors in the above model represent the estimated difference in log odds of injury for presence vs. absence of the corresponding factor. For example, the coefficient on the alcohol term is 1.8266, meaning the difference in odds of injury for alcohol-related crashes vs. non-alcohol-related crashes is $e^{1.8266} = 6.2127$.Also, negative sign for interaction coefficient shows that the impact of speeding and alcohol is lower than other causes (baseline group) on injury crashes in offroad segment.

Conclusions; what conditions are more dangerous?

Based on The values and sign of coefficients of our final model we can conclude that:

- Crashes involving speeding and alcohol usage have a higher probability of injury than crashes due to other causes.
- Crashes on the weekend have a higher probability of injury than crashes on weekdays.
- Crashes off the road have a higher probability of injury than crashes that occur on the road.
- Highway 91 and Highway 5 were identified as the riskiest highways for injury crashes comparing other highways which were involved in this study.

To analyze the effect of the interaction of cause and location, we can look at the percentage of crashes involving injury for each cause controlling for off-road and on-road separately. The following tables summarize the percentage of crashes which involved injury for each factor in the model in descending order. From these tables we can observe that percentage of injury accidents in offroad location are modified by cause variable. Specifically, other cause has the main contribution in offroad injury crashes. So, the negative sign of interaction terms can be justified with this analysis. Also, using other tables we can justify the sign of other coefficients. For example, percentage of injury crashes for weekends is higher than percentage of injury crashes for weekdays. So, this confirms the positive sign for log odds for weekend variable.

Table 6. Percentage of crashes which involve injury for each factor

Location	Off-road	40.15%
	On-road	24.70%
Cause	Alcohol	45.10%
	Speeding	29.50%
	Other	23.68%
Weekend/Weekday	Weekend	33.82%
	Weekday	25.10%
Highway	91 Riverside/Artesia freeways	34.58%
	5 San Diego/Santa Ana freeways	33.62%
	22 Garden Grove freeway	29.17%
	57 Orange freeway	27.73%
	55 Costa Mesa Freeway	27.33%
	405 San Diego freeway	21.79%
Cause-Location	Alcohol-off road	40.74%
	Alcohol-on road	50%
	Speeding-off road	38.10%
	Speeding-on road	28.37%
	Speeding off road	41.18%
	Other	13.58%

References

Asgari, H., 2015. On the impacts of telecommuting over daily activity/travel patterns: A comprehensive investigation through different telecommuting patterns. PhD dissertation, Florida International University.

Asgari, H, Jin, X., 2015. Towards a Comprehensive Telecommuting Analysis Framework; Setting the Conceptual Outline. Transportation Research Record 2496, 1-9.

Asgari, H, Jin, X., 2016a. Examining the impacts of telecommuting on the time-use of nonmandatory activities. Proceedings of 95th Transportation Research Board Annual Meeting, Washington DC.

Asgari, H., Jin, X., 2016b. Investigation of commute departure time to understand the impacts of part-day telecommuting on the temporal displacement of commute travel. Proceedings of 14th World Conference on Transport Research, Shanghai, China.

Asgari, H., Jin, X., Mohseni, A., 2014. Choice, Frequency, and Engagement - A Framework for Telecommuting Behavior Analysis and Modeling. Transportation Research Record 2413, 101-109.

Baratian-Ghorghi, F., Huaguo, H., Shaw, J., 2014. Overview of wrong-way driving fatal crashes in the United States. Institute of Transportation Engineers. ITE Journal, 84(8), 41-47.

Khalilikhah, M., Habibian, M., Heaslip, K., 2016. Acceptability of increasing petrol price as a TDM pricing policy: A case study in Tehran. Transport Policy 45, 136-144.

Soltani-Sobh, A., Heaslip, K., Bosworth, R., Barnes, R., Song, Z., 2016. Do natural gas vehicle miles traveled? An aggregate time-series analysis. Proceedings of the 95th annual meeting of Transportation Research Board, Washington DC.

Zolghadri, N., Halling, M., Barr, P., Petroff, S., 2013. Identification of truck types using strain sensors include co-located strain gauges. Structures Congress, 363-375.

Zolghadri, N., Halling, M., Barr, P., Petroff, S., 2016. Field verification of simplified bridge weigh-in-motion techniques. Journal of Bridge Engineering, 10.1061/(ASCE)BE.1943-5592.0000930.