STATISTICAL DATA ANALYSIS CA660

Garima Singh, Kiran Negi, Yangchen Dolkar Sherpa
Student IDs: 19211010,19210510,19210377
Group 52

ABSTRACT

Road Accident incidents around the world have increased over the years. To improve the condition, many researchers in the area have contributed by suggesting various corrective measures. This study is to estimate contributing parameters in road accidents by using log linear statistical method. The Statistical analysis to study the road accidents provides a clear picture of the correlation between various contributing factors to the condition. Log linear method due to its immense statistical power helps with the exploratory analysis on the data. We have used the log linear statistical method for our exploratory analysis because it is the best suited method for our dataset. The method is used to inspect two or more categorical variables. Such method is also the best pick in the scenarios where there is no direct divergence among variables.

KEYWORDS
Road Accidents Analysis, Log linear statistical method, Frequency distribution, Backward process elimination

1. INTRODUCTION

Road accidents are one of the most significant reasons of death and injury around the world. This is an alarming fact which needs to be taken care of. The professionals in road safety target to bring down the accidents count by analysing the road accidents statistics. The given dataset that we have used for our study consists of such factors that would enable us to provide insights of the contributing factors in road accidents and thus affect the accident severity. Many researches have been done in the area one such example is a research paper on the temporal stability of factors affecting driver-injury severities in single vehicle crashes. The paper analysed the single vehicle crash data in Chicago, Illinois for a 9-year period, separated models of driver injuries severities to detect undiscovered diversity. The study used multinomial logit model for the various factors taken into consideration to detect the temporal stability of individual variables to evaluate injury-severity probabilities.12[Behnood, A. and Mannering, F.L., 2015]

2. RELATED WORK

The paper talks about the significance of injury reduction technologies on the basis of factual assessment of multiple interactions that factors such as road condition, vehicle involved and the concerning human involvement have on the severity of the crash. 1[Savolainen, P.T., Mannering, F.L., Lord, D. and Quddus, M.A., 2011] The paper used Palm distribution of various combinations of weather and road conditions and measured it with the dispersal of same conditions as observed during accident. 2[Malin, F., Norros, I. and Innamaa, S., 2019] The paper talks about the traffic congestion due to the road accidents. The paper applied hazard based duration models to statistically measure the duration it takes to report, attend, and clear the traffic congestion due to such events on roads. 3[Nam, D. and Mannering, F., 2000] The Paper proposed a Bayesian spatial joint model to predict the road accidents in road network using road segments and intersections in road network in urban regions.4[Zeng, Q. and Huang, H., 2014] The paper applied the K- function methods on traffic accident data to evaluate the risk corresponding to false positive detection using Planar space to measure a network constrained event. 5[Yamada, I. and Thill, J.C., 2004] The paper demonstrated the evolution of the number of traffic accident deaths for each age group in France from 1979 to 2013. 6[Gicquel, L., Ordonneau, P., Blot, E., Toillon, C., Ingrand, P. and Romo, L., 2017] The paper proposed multivariate Poisson-lognormal (MVPLN) to evaluate time and weather conditions impact on the frequency of multiple crash types in the city of Edmonton. 7[El-Basyouny, K., Barua, S. and Islam, M.T., 2014] The paper applied log linear analysis and found notable correlation between the driver age, severity, manner of collision, alcohol involvement and road conditions. 13[Abdel-Aty, M.A., Chen, C.L. and Schott, J.R., 1998]

3. DATASET AND EXPLORATORY ANALYSIS

The dataset for our analysis of road accidents were obtained from government website divided from years 2009 to 2018. All files contained data with attributes of Date/Time of accident happened, road/weather conditions at that time interval, casualties in number and gender and type/number of vehicles engaged.

On preliminary analysis done on our reported accident data, we observed that there were different but fixed levels for each attribute, such as classification of roads (Motorway, A(M), A, B, C and Unclassified), road surface conditions (Dry, Wet/Damn, Flood, Frost/Ice and Snow), lighting conditions (Daylight and Darkness), weather conditions (wind, rain, fog, etc.), type of car (car, taxi, bus, etc), casualty severity (slight, serious and fatal), sex of casualty (male and female). These attributes were combined with unique reference number and location of each accident occurred.

Initial cleaning was performed using Google refine tool wherein null values were omitted while date/time attributes were refined. The levels discussed above for each attribute were inconsistent for different years, which needed to be uniform among all. After refining all the data, the datasets were combined in R.

4. HYPOTHESES & RESEARCH QUESTIONS

Following hypotheses have been taken into consideration for our study:

- Relationship between Gender and Accident casualty
- Relationship between road condition and Accident casualty
- Relationship between time and Accident casualty

As part of our study, we have targeted answering below research questions

1. Estimation of parameters corresponding to road accidents.
2. To determine the significance of the parameters used as part of road accident analysis by using two techniques of Log linear model named the backward elimination process.
3. To justify Log linear model as the best fit for our dataset.

5. Methods used and why

We have applied Log linear method since we need to analyse more than two variables as part of our analysis. To measure the difference between expected and observed frequencies we used likelihood ratio chi square:

$$\chi^2 = 2\sum[O_t ln(O_t/E_t)].$$

Where E_t is expected frequency and O_t is observed frequency. Log (expected cell frequency) is used to determine main impact of parameters and higher order interactions.

$$logm_{tuvw} = \mu + \lambda_t^G + \lambda_u^L + \lambda_v^T + \lambda_w^C, t = 1,2; u = 1,2; v = 1,2; w = 1,2$$

$$\sum_t \lambda_t^G = \sum_u \lambda_u^L = \sum_v \lambda_v^T = \sum_w \lambda_w^C = 0$$

$$logm_{tuvw} = \mu + \lambda_t^G + \lambda_u^L + \lambda_v^T + \lambda_w^C + \lambda_{tu}^{GL} + \lambda_{tv}^{GT} + \lambda_{tw}^{GC} + \lambda_{uw}^{LT} + \lambda_{uw}^{LC} + \lambda_{vw}^{TC}, t = 1,2; u = 1,2; v = 1,2; w = 1,2$$

Sum to zero identifiability conditions.

$$\sum_t \lambda_t^G = \sum_u \lambda_u^L = \sum_v \lambda_v^T = \sum_w \lambda_w^C = 0$$

$$\sum_t \lambda_{tu}^{GL} = \sum_u \lambda_{tu}^{GL} = \sum_t \lambda_{tv}^{GT} = \sum_v \lambda_{tv}^{GT} = 0$$

$$\sum_t \lambda_{tw}^{GC} = \sum_w \lambda_{tw}^{GC} = \sum_u \lambda_{uw}^{LT} = \sum_v \lambda_{uw}^{LT} = 0$$

$$\sum_u \lambda_{uw}^{LC} = \sum_w \lambda_{uw}^{LC} = \sum_v \lambda_{vw}^{TC} = \sum_w \lambda_{vw}^{TC} = 0$$

We have measured risk ratio, also known as relative risk to calculate probabilities in two groups. It is the ratio having a positive result in the two concerned groups. It depicts the outcome is possible in both the groups. The interpretations of the results we found are as follows: We have used log linear analysis to analyse the relation between variables and we used SPSS 1.0.0.86 Statistics tool. We have provided our findings below in results and findings section. We have used the backward elimination process to determine the significance of the parameters used. The results are provided in below section.

6. RESULTS AND FINDINGS

6.1. Statistical Analysis

We created two-way contingency tables for each of the three factors and calculated the chi square and p value.

Frequency Distribution of Casualty and Gender

Gender/ Casualty	Fatal/serious	Slight	Total
Male	2234	12825	15059
Female	951	9623	10574
Total	3185	22448	25633

χ 2 = 194.7886 p -value = 2.87E-44

We see that male drivers have higher proportion of fatal casualties (14.83%) than female drivers (8.99%) similarly female drivers have higher proportion of slight casualties (92%) than male drivers (85.17%). The null hypothesis of the relation between gender and casualty is rejected at p<0.05 hence casualty is highly related to Gender.

Frequency Distribution of Casualty and Road Condition

Road Condition /Casualty	Fatal/serious	Slight	Total
Good	2363	16662	19025

Bad	822	5786	6608
total	3185	22448	25633

χ 2 = 0.001621 p -value = 0.967888

We see that the chances of fatal casualties are almost similar when the road condition is good (12.42%) or bad (12.43%) similarly the chances of slight casualties is similar when the road condition is good (87.58%) or bad (87.56%). The null hypothesis for road condition is accepted at p>0.05 hence casualty is not related to road condition.

Frequency Distribution of Casualty and Day /Night Condition

Time/Casualty	Fatal/Serious	Slight	Total
Daytime	2056	15789	17845
Night-time	1129	6659	7788
Total	3185	22448	25633

χ 2 = 44.12 p -value = 3.11E-11

The chances of fatal injuries are higher for Night-time (14.49%) and slightly lower for Daytime (11.76%). The null hypothesis is rejected at p<0.05. Hence Casualty is related to the time.
Therefore, while gender and time are significant factors for casualty severity of accident, the road condition is not.

Relative Risk,Odds Ratio and 95% Confidence Interval for accident casualty

	Relative Risk	Odds Ratio	95% Confidence Interval for Odds ratio
Gender	1.65	1.76	1.63-1.91
Road Condition	1	1	0.92-1.09
Time	0.79	0.77	0.71-0.83

Odds ratio is a measure of the association between the two groups. If the odds ratio is 1 it is equally likely for both events to occur. If the odds ratio is greater than 1 then the occurrence of the event is more likely in the first group similarly less than one implies that the event is less likely to occur in the first group.
We can see that after calculating risk factor for comparison between the two of the group instead of one, which is similar to odds ratio in terms of value. That is, a value equal to one suggests that casualty severity may happen in both cases as seen with Road condition, greater than one suggests that casualty may happen in first group more as seen for Gender and less than one as seen for time.

6.2. Log-Linear Analysis

Through log linear approach we are analysing the relationship between the different variables.

The K factor represents the number of interactions between variables

K-way	DF	Chi-square	p-value	Number of Iterations
1	15	27919.765	0.000	0
2	11	507.608	0.000	2
3	5	7.625	0.186	3
4	1	0.295	0.587	2

From the table above, the chi square for model without interaction of the four variables is obtained from K=4 the hypothesis that GTLC=0 is tested. K=3 indicates the model without the last two variables, the hypothesis that GTL=GTC, GLC, TLC=0 is tested, since p>0.05 the hypothesis is accepted for K=4 and K=3. K=2 indicates the model without the last 3 variables i.e. the hypothesis that GT=GL=…=0 is tested and K=1 indicates the model without any effect. Since p<0.05 for K=1 and K=2 we see that the effects were significant. Hence, we will use the interactions between 1 and 2 variables for partial associations table.

Effect	df	Partial Chi-Square	p-value
CasualtySeverity*Gender	1	188.374	.000
CasualtySeverity*RoadCondition	1	.267	.605
Gender*RoadCondition	1	2.604	.107
CasualtySeverity*Time	1	30.235	.000
Gender*Time	1	147.738	.000
RoadCondition*Time	1	99.138	.000
CasualtySeverity	1	16293.946	.000

Gender	1	788.793	.000
RoadCondition	1	6275.560	.000
Time	1	4053.858	.000

The interaction parameters for CasualtySeverity*Gender, CasualtySeverity*Time, Time*Gender were significant since $p<0.05$ while it was not significant for CasualtySeverity*RoadCondition, Gender* Road Condition To determine the significance of the different parameters backward elimination was used:

Step		Deleted Effects	Chi-Square	df	p-value
0	1	CasualtySeverity*Gender*RoadCondition*Time	0.295	1	0.587
1	1	CasualtySeverity*Gender*RoadCondition	3.235	1	0.072
	2	CasualtySeverity*Gender*Time	3.553	1	0.059
	3	CasualtySeverity*RoadCondition*Time	0.437	1	0.509
	4	Gender*RoadCondition*Time	0.766	1	0.381
2	1	CasualtySeverity*Gender*RoadCondition	3.042	1	0.081
	2	CasualtySeverity*Gender*Time	3.493	1	0.062
	3	Gender*RoadCondition*Time	0.682	1	0.409
3	1	CasualtySeverity*Gender*RoadCondition	3.118	1	0.077
	2	CasualtySeverity*Gender*Time	3.486	1	0.062
	3	RoadCondition*Time	99.591	1	0
4	1	CasualtySeverity*Gender*Time	3.095	1	0.079
	2	RoadCondition*Time	99.139	1	0
	3	CasualtySeverity*RoadCondition	0.269	1	0.604
	4	Gender*RoadCondition	2.605	1	0.107
5	1	CasualtySeverity*Gender*Time	3.094	1	0.079
	2	RoadCondition*Time	98.898	1	0
	3	Gender*RoadCondition	2.483	1	0.115
6	1	CasualtySeverity*Gender*Time	3.094	1	0.079
	2	RoadCondition*Time	102.07	1	0
7	1	RoadCondition*Time	102.07	1	0
	2	CasualtySeverity*Gender	188.253	1	0
	3	CasualtySeverity*Time	29.996	1	0
	4	Gender*Time	150.815	1	0

Deleted Effect is the change in the Chi-Square after the effect is deleted from the model. In order to get the best fit model, the non-significant interaction parameters have been removed.

$$logm_{tuvw} = \mu + \lambda_t^G + \lambda_u^L + \lambda_v^T + \lambda_w^C + \lambda_{tv}^{GT} + \lambda_{tw}^{GC} + \lambda_{uv}^{LT} + \lambda_{vw}^{TC}$$

We see that CasualtySeverity*RoadCondition, Gender*RoadCondition of the second order interaction variables is insignificant. The remaining second order interaction terms are all significant.

7. CONCLUSIONS

We took three hypotheses into consideration: Relationship between Gender and Accident casualty, Relationship between road condition and Accident casualty, Relationship between time and Accident casualty. Only our second hypothesis observed false and rest are true. Except for Road condition all other taken into consideration have direct impact on the road accident. Through backward process elimination we found that the interaction between parameters CasualtySeverity * Gender, CasualtySeverity*Time, Time*Gender were significant whereas CasualtySeverity *RoadCondition, Gender * Road Condition were not significant. We have applied log linear analysis as part of our study due to the fact that unlike other analysis models, log linear analysis provide more control over the interaction between the variables in context. Another reason for choosing this kind of analysis is that it is appropriate in scenarios where there is no clear distinction between explanatory and response variables. This is what makes log linear analysis different and a better choice over others.

Effect	Estimate	Std. Error	Z	p-value	95% Confidence Interval	
					Lower Bound	Upper Bound
CasualtySeverity (C)*Gender (G)* RoadCondition (L)*Time (T)	-.007	.012	-.546	.585	-.031	.017
C*G*L	.023	.012	1.901	.057	-.001	.048
C*G*T	-.018	.012	-1.462	.144	-.042	.006
C*L*T	.010	.012	.832	.406	-.014	.034
G*L*T	-.012	.012	-.982	.326	-.036	.012
C*G	.135	.012	10.904	.000	.110	.159
C*L	-.006	.012	-.495	.621	-.030	.018
G*L	.009	.012	.732	.464	-.015	.033
C*T	-.051	.012	-4.147	.000	-.075	-.027
G*T	-.096	.012	-7.803	.000	-.120	-.072
L*T	.086	.012	6.985	.000	.062	.110
C	-.997	.012	-80.82	.000	-1.021	-.973
G	.319	.012	25.873	.000	.295	.343
L	.496	.012	40.228	.000	.472	.520
T	.360	.012	29.145	.000	.335	.384

8. REFERENCES

1. Savolainen, P.T., Mannering, F.L., Lord, D. and Quddus, M.A., 2011. The statistical analysis of highway crash-injury severity: a review and assessment of methodological alternatives. *Accident Analysis & Prevention, 43(5)*, pp.1666-1676.

2. Malin, F., Norros, I. and Innamaa, S., 2019. Accident risk of road and weather conditions on different road types. *Accident Analysis & Prevention, 122,* pp.181-188.

3. Nam, D. and Mannering, F., 2000. An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice, 34(2)*, pp.85-102.

4. Zeng, Q. and Huang, H., 2014. Bayesian spatial joint modeling of traffic crashes on an urban road network. *Accident Analysis & Prevention, 67,* pp.105-112.

5. Yamada, I. and Thill, J.C., 2004. Comparison of planar and network K-functions in traffic accident analysis. *Journal of Transport Geography, 12(2),* pp.149-158.

6. Gicquel, L., Ordonneau, P., Blot, E., Toillon, C., Ingrand, P. and Romo, L., 2017. Description of various factors contributing to traffic accidents in youth and measures proposed to alleviate recurrence. *Frontiers in psychiatry, 8,* p.94.

7. El-Basyouny, K., Barua, S. and Islam, M.T., 2014. Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson

lognormal models. *Accident Analysis & Prevention, 73*, pp.91-99.

8. Thomas, I., 1996. Spatial data aggregation: exploratory analysis of road accidents. *Accident Analysis & Prevention, 28(2)*, pp.251-264.

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

10. Kumar, C.N., Parida, M. and Jain, S.S., 2013. Poisson family regression techniques for prediction of crash counts using Bayesian inference. *Procedia-Social and Behavioral Sciences, 104(2),* pp.982-991.

11. Chong, M.M., Abraham, A. and Paprzycki, M., 2004. Traffic accident analysis using decision trees and neural networks. *arXiv preprint cs/0405050.*

12. Behnood, A. and Mannering, F.L., 2015. The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: some empirical evidence. *Analytic methods in accident research*, *8*, pp.7-32.

13. Abdel-Aty, M.A., Chen, C.L. and Schott, J.R., 1998. An assessment of the effect of driver age on traffic accident involvement using log-linear models. *Accident Analysis & Prevention*, 30(6), pp.851-861.