Assessing Impact of Lane Width on Safety of the Vehicles

Yogendra Patil (yjpatil@crimson.ua.edu) student at *University of Alabama, Tuscaloosa, AL*

Abstract— **This study analysis the impact of lane width on the safety of the vehicles by using a 10 year crash data at midblock segment of arterial roads in 4 cities of Nebraska. The data set contains segments details measured using Google Earth, such as, lane width, speed limit, presence of shoulders, etc., and the yearly crash frequency reported for different categories, such as, crash severity, driver age, etc. In order to determine the impact of the lane width on the safety of the vehicles, the analysis was carried out in two folds. First, the safest lane width was determined by using frequency of crash data set. After the safest lane width was determined; factors that contribute the most for the safety of the vehicles were analyzed. The entire analysis was implemented by using logistic regression, with a user created dichotomous output variable. Results indicated that lane width of 10ft is safe, because the odds of frequent occurrence of severe crash and damages are low as compared to other lane widths. Factors such as No On Street Parking, and No Central Business District Parking are important variables and should be implemented if one wants to observe fewer crashes and damages at the midblock section of the arterial roads.**

Softwares used – SAS 9.3 + JMP 10

I. APPROACH JUSTIFICATION

Regression methods have become an integral component of any data analysis, where the relationship between the response variable and one or more explanatory variables can be explored. The most common method for analysis is the linear or multiple regression analysis. However, if the problem is formulated in such a way that the outcome is not continuous, such methods are not appropriate. Poisson regression can be used to model for modeling the frequency of crash. But Poisson distribution has a strong assumption of mean equal to variance, which is generally not exhibited by crash data (Jovanis & Chang, 1986; Joshua & Garber, 1990; Jones, Janssen, & Mannering, 1991; Miaou & Lum, 1993). In case of Negative binomial regression or Poisson-gamma regression the equi-dispersion assumption of the Poisson distribution is mostly inappropriate for crash data sets as they are found to be over or under-dispersed (Hauer & Hakkert, 1988; Persaud, 1994; Mountain, Fawaz, & Jarrett, 1996; Johansson, 1996; Vogt & Bared, 1998; Vogt, 1999; Miaou, 2001). Zero inflated models are used when the data has missing values or has far too many zeros than one would expect in the regular processes like Poisson. This would mean that there are some sites which can never have a crash and some sites where no crashes were recorded during the study period. This assumption does not reflect the true crash data generating process (Shankar, Milton, & Mannering, Modeling Accident Frequency as Zero altered Probability Processes: an Empirical Inquiry, 1997; Lee & Mannering, 2002; Kumara & Chin, 2003). Hence for the above stated reasons it was thought that a logistic regression be appropriate to analyze the given crash data. Generally, logistic regression models are well suited for describing behavior of a data set when large number of categorical explanatory variables are present. The traffic data set provided satisfies most of the conditions and assumptions for application of logistic regression.

The following sections summarizes the approach taken to identify the impact of lane width in the safety of the vehicles. A logistic regression analysis was carried out, and significant factors were selected using log-likelihood test. SECTION II describes the problem formulation followed by SECTION III which describes the assumptions undertaken to tackle the complexity in the data set. SECTION IV outlines the methodology implemented to solve the given problem. Finally, SECTION V discusses about results & final conclusions.

II. PROBLEM FORMULATION

The main objective was to analyze the impact of lane width on the safety of the vehicles. The data set provided consisted of 135 variables with 19601 observations each. In order to tackle this problem, the problem was further divided into two sub-modules. *Module* A requires the analyst to first determine the safest lane width, among the given widths (9, 10, 11, 12 ft), by analyzing only frequency of crash type and frequency of crash severity variable. After deciding the safest lane width, *Module* B requires to determine the factors that play significant role in contributing the safety factor for that particular lane-width.

A. *Module* **A**

Given the frequency of occurrence of crash type and crash severity only, determine the relationship between these explanatory variables and the different types of lane width.

B. *Module* **B**

Determine the factors that have significant impact on the safest lane width, i.e. factors that should be given priority before building a midblock segment of arterial roads.

III. ASSUMPTIONS

Before the actual procedure was applied to the given data set certain assumptions were considered.

A. *Assumption* **I**

The output variable, i.e. safety is a dichotomous variable, with safe = 1 and not safe = 0. (*Note*: the output variable was created by the analyst based on the lane-width and was not initially present in the original data set).

B. *Assumption* II

All the variables can be represented into fewer number of discrete levels. In this assumption, each explanatory variable contains discrete number of levels which can be further discretized into fewer levels (<= 3) as the multiple levels seems to be redundant. For example, variable 'Crash Type = Rear End' has more than 8 levels. It was further categorized into 3 levels, viz., Low (for values = 0), Moderate (for values = 1 or 2) and, High (for values > 3). The categorical levels assigned to the variables are NO (which means No or Null values), LO (which means Low values), MED (which means Medium values) or HI (which means High values) (please refer TABLE I). The categorization for each explanatory variable into different levels was performed after observing the distributions in JMP 10 platform.

TABLE I CATEGORIZATION OF EACH VARIABLE

VARIABLE NAME	Categorization
CITY,SECTION,YEAR, SPEED LIMIT, # OF THRU LANES, SHOULDER, MEDIAN, 1st EVENT LEAD to ACCIDENT, ACCIDENT TIME, # OF TRUCK, CONTRIB CIRCUM, DRIVER AGE, GENDER, VEHICLE MOVE.	Eliminated
CRASH TYPE (CT) = BACKING	If = 0 then NO If > 1 then LO

CRASH TYPE (CT) = SIDE SWIPE	If = 0 then NO
	If = 1or2 then MED
	If >= 3 then HI

CRASH TYPE (CT) = SIDE SWIPE OPPOSITE	If = 0 then NO If >= 1 then MED
CRASH TYPE (CT) = ANGLE	If = 0 then NO If >= 1 then MED

CRASH TYPE (CT) = REAR END	If = 0 then NO
	If = 1or2 then MED
	If >= 3 then HI

CRASH TYPE (CT) = HEAD ON	If = 0 then NO If >=1 then MED
CRASH TYPE (CT) = LEFT TURN	If = 0 then NO If >=1 then MED

CRASH SEVERITY (CS) = NON REPORTABLE	If = 0 then NO
	If = 1 then LO
	If = 2 then MED
	If >= 3 then HI

CRASH SEVERITY (CS) = VISIBLE INJURY	If = 0 then NO
	If = 1 then MED
	If >= 2 then HI

CRASH SEVERITY (CS) = POSSIBLE INJURY	If = 0 then NO
	If = 1 then MED
	If >= 2 then HI

CRASH SEVERITY (CS) = DISABLING INJURY	If = 0 then NO If >= 1 then MED
CRASH SEVERITY (CS) = FATAL INJURY	If = 0 then NO If >= 1 then MED

C. *Assumption* III

Certain variables can be eliminated from the entire analysis. Variables such as SPEED, # OF THRU LANES, SHOULDER, MEDIAN were discarded from analysis because they had observations only for one level/category. Variables such as 1st EVENT LEAD to ACCIDENT, ACCIDENT TIME, # OF TRUCK, CONTRIB CIRCUM, DRIVER AGE, GENDER, VEHICLE MOVE were discarded from analysis because the explanatory variables such as CRASH TYPE and CRASH SEVERITY were assumed to be enough for Hypothesis testing process.

IV. SOLUTION

The following sub-sections summarizes the solutions for *Modul*e A and B problem statements.

A. *Solution for Module* A

In order to address the problem stated in *Module* **A**, hypothesis testing procedure was carried out using logistic regression, to decide the safe lane width. During the hypothesis testing, a particular lane width was chosen to be safe against the rest. For example, lane width = 9ft was presumed to be safe against the rest. If the hypothesis holds true, then odds of occurrence of crashes will be very low. To test this hypothesis an output variable (safe) was separately created with only two levels (1/0). The value of the output variable depended on the value of the lane width considered as safe according to the null hypothesis. For example, safe = 1 for all entries that corresponded to lane width = 9 or else not safe = 0. For this analysis all the categories under Crash Type and Crash severity were considered, because for any lane or road or section can be considered safe if and only if the frequency of crash occurrence is very low.

```
Input: L,S,$C_T$,$C_S$,$H_O$
Output: H
BEGIN
    For i ← 1 to 4 do
      Initialization: $H_O$ ← S for $L_i$
        If P(S|$L_i$,$C_T$,$C_S$)← Highest then
           $L_i$ ← S;
           $H_O$ ← true;
           H ← $H_O$;
        End
    End
 Return H
END
```

Figure 1. Pseudo code for Hypothesis testing using logistic regression

Figure 1 summarizes the pseudo code for the hypothesis testing using logistic regression to decide the safe lane width. The input variables consisted of the set of lane width L, the output variable safe S, all the categories of Crash Types C_T and Crash Severity C_S and the null hypothesis H_O. The output variable H, is the hypothesis that holds true, i.e. the null hypothesis that holds true for a particular lane width. So the process begins by considering each lane width individually from the set of lane width L. The null hypothesis H_O at the first iteration states that lane width = 9 ft is safe, while the rest lane widths are not safe. If this null is true then conditional probability of lane with being safe ($S=1$) will be highest given the variables lane width L_i, all the categories of Crash Types C_T and Crash Severity C_S i.e. $P(S|L_i,C_T,C_S) \leftarrow$ Highest. For each iteration, the hypothesis testing is carried for each lane width and, if $P(S|L_i,C_T,C_S)$ was greater as compared to previous one then the previously assumed true hypothesis was discarded and over written by the current one. At the end of all iterations the final true hypothesis was determined. SECTION V summarizes the results and there explanations in more details.

B. *Solution for Module* **B**

After determining the safest lane width from, analysis for *Module* **B** was carried out by using simple logistic regression analysis. The significant explanatory variables were selected based on their p-values, and these variables contributed into the safety of the safest lane width.

V. RESULTS.

A. *Results for Module* **A**

The analysis for *Module* **A** indicated that **Lane width = 10 ft was the safest** as compared to Lane Width = 9, 11, or 12 ft. The log likelihood ratio test for overall model for lane width = 10 ft gave χ^2 value of 129.35 with significant *p*-value (<0.0001). The χ^2 value along with the *p*-value was used to determine whether the model was adequate or not. TABLE II summarizes the log likelihood ratio (L-R) test for significant variables when safe = 1 for lane width = 10 ft. TABLE II indicates that variables Crash Type (CT) = Rear End, Crash Type (CT) = Non Reportable, Crash Severity (CS) = Property Damage, Crash Severity (CS) = Visible Injury, Crash Severity (CS) = Possible Injury have significant p-values. Finally the odds ratio for these variables was used to determine if the null hypothesis holds true or not, i.e. lane width = 10 ft was the safest. TABLE III summarizes the odds-ratio (OR) for the above significant variables.

The odds-ratio can be used to verify the stated hypothesis. In this case, the null hypothesis states that lane width = 10ft was safe (=1) while the rest were not safe (=0). Now consider the odds ratio for variable CS = Visible Injury. The CS = Visible Injury was categorized into 3 levels (Null, Medium, High values). Now in order for the lane width = 10ft to be safe, odds of Visible Injury not to happen (Null values) should be high as compared to rarely to moderately to happen (Medium) and frequently to happen (High). That's what indicated in TABLE III, which states that "odds for safe = 0 (i.e. lane widths = 9, 11, 12 ft not safe) rather than safe = 1 (i.e. lane width = 10 ft to be safe) improves (by 1.27) for a unit rise in the Null values for CS = Possible Injury with respect to High values". Similarly, separate hypothesis testing is carried out for values of different lane width and the odds ratio is observed. From the entire analysis, it is observed that lane width = 10 ft is the safest, lane width = 9ft is safe. For lane width = 11 ft only the significant factor is CS= Fatal Injury, but the odds ratio value is low to conclude the severity. For lane width = 12ft several factors that are regarded as unsafe were seen significant and does there odds ratio. Hence lane width = 12ft was determined to be unsafe for the vehicles.

s

TABLE II. STATISTICAL TEST FOR INDIVIDUAL PREDICTORS

Term	*Npar*	*DF*	(L-R) χ^2	*p*
CT = Rear End	2	2	7.8	<.0001*
CT = Non Reportable	3	3	9.4	<.0001*
CS = Property Damage	3	3	29.3	<.0001*
CS= Visible Injury	2	2	0.21	0.0069
CS = Possible Injury	2	2	57.0	0.0364

TABLE III. ODDS RATIO FOR SIGNIFICANT PARAMETERS FOR LANE WIDTH = 10 ft.

LEVEL1	LEVEL2	ODDS RATIO	*p*	LOWER 95%	UPPER 95%
No_CT = Rear End	Med_CT = Rear End	0.51	<.0001*	0.38	0.67
No_CT = Rear End	Hi_CT = Rear End	0.80	0.0009	0.70	0.91
No_CT = Non Reportable	Lo_CT = Non Reportable	1.32	<.0001*	1.17	1.49
No_CT = Non Reportable	Med_CT = Non Reportable	1.47	0.0002	1.2	1.7
No_CT = Non Reportable	Hi_CT = Non Reportable	1.92	<.0001*	1.4	2.5
No_CS = Property Damage	Lo_CS = Property Damage	1.32	<.0001*	1.1	1.5
No_CS = Property Damage	Med_CS = Property Damage	1.41	0.0004	1.6	1.7
No_CS = Property Damage	Hi_CS = Property Damage	1.31	0.0534	0.99	1.7
No_CS= Visible Injury	Med_CS= Visible Injury	1.3	0.003	1.09	1.5
No_CS = Possible Injury	Med_CS = Possible Injury	1.18	0.0204	1.02	1.37
No_CS = Possible Injury	Hi_CS = Possible Injury	1.27	0.0823	0.96	1.65

The log likelihood ratio test for overall model for lane width = 9 ft gave χ^2 value of 54.12 with significant *p*-value (<0.0001). TABLE IV L-R test for significant variables when safe = 1 for lane width = 9 ft.

TABLE IV. STATISTICAL TEST FOR INDIVIDUAL PREDICTORS

Term	*Npar*	*DF*	(L-R) χ^2	*p*
CT = Backing	1	1	4.1	0.0429
CT = Side Swipe	2	2	14.5	0.0007
CS = Rear End	2	2	10.2	0.006
CS= Non Reportable	3	3	10.5	0.0146

TABLE V. ODDS RATIO FOR SIGNIFICANT PARAMETERS FOR LANE WIDTH = 9 ft.

LEVEL1	LEVEL2	ODDS RATIO	*p*	LOWER 95%	UPPER 95%
No_CT = Rear End	Hi_CT = Rear End	0.38	0.0098	0.18	0.79

Med_CT = Rear End	Hi_CT = Rear End	0.38	0.0017	0.19	0.70
No_CT = Backing	Lo_CT = Backing	2.06	0.0429	1.02	3.76
No_CT = Non Reportable	Lo_CT = Non Reportable	1.44	0.0074	1.1	1.9
No_CT = Non Reportable	Med_CT = Non Reportable	1.74	0.0186	1.1	2.7
No_CS= Visible Injury	Med_CS= Visible Injury	1.3	0.003	1.09	1.5
No_CS = Possible Injury	Med_CS = Possible Injury	1.38	0.0403	1.01	1.9
No_CT = Side Swipe	Hi_CT = Side Swipe	0.17	0.03	0.01	0.8

The log likelihood ratio test for overall model for lane width = 11 ft gave χ^2 value of 4.84 with significant *p*-value (0.0277). TABLE VI L-R test for significant variables when safe = 1 for lane width = 11 ft.

TABLE VI. STATISTICAL TEST FOR INDIVIDUAL PREDICTORS

Term	*Npar*	*DF*	(L-R) χ^2	*p*
CT = Fatal	1	1	4.13	0.0422

TABLE VII. ODDS RATIO FOR SIGNIFICANT PARAMETERS FOR LANE WIDTH = 11 ft.

LEVEL1	LEVEL2	ODDS RATIO	*p*	LOWER 95%	UPPER 95%
No_CS = FATAL	Med_CT = FATAL	0.43	0.0277	0.17	0.91

The log likelihood ratio test for overall model for lane width = 12 ft gave χ^2 value of 105.13 with significant *p*-value (<0.0001). TABLE VIII L-R test for significant variables when safe = 1 for lane width = 12 ft.

TABLE VIII. STATISTICAL TEST FOR INDIVIDUAL PREDICTORS

Term	*Npar*	*DF*	(L-R) χ^2	*p*
CT = Property Damage	3	3	34.6	<.0001*
CT = Rear End	2	2	26.5	<.0001*
CS = Possible Injury	2	2	8.7	0.013
CS= Fatal	1	1	8.16	0.0043

TABLE IX. ODDS RATIO FOR SIGNIFICANT PARAMETERS FOR LANE WIDTH = 9 ft.

LEVEL1	LEVEL2	ODDS RATIO	*p*	LOWER 95%	UPPER 95%
No_CT = Rear End	Med_CT = Rear End	1.21	<.0001*	1.1	1.3
No_CT = Rear End	Hi_CT = Rear End	1.68	<.0001*	1.3	2.7
No_CT = Property Damage	Lo_CT = Property Damage	0.77	<.0001*	0.71	0.85
No_CT = Property Damage	Med_CT = Property Damage	0.76	0.0003	0.66	0.88
No_CT = Property Damage	Hi_CT = Property Damage	0.72	0.0021	0.59	0.89
No_CS = Possible Injury	Med_CS = Possible Injury	0.86	0.0081	0.78	0.96
No_CS = Possible Injury	Hi_CS = Possible Injury	0.81	0.0423	0.67	0.99
No_CS = Fatal Injury	Med_CS = Fatal Injury	0.87	0.0081	1.0	1.5

B. *Results for Module* **B**

A simple logistic regression model was implemented to determine the significant factors that contributed to the safety of vehicles for lane width = 10 ft. The log likelihood ratio test for overall model for safe = 1 when lane width = 10 ft gave χ^2 value of 762.98 with significant *p*-value (<0.0001). The χ^2 value along with the *p*-value was used to determine whether the model was adequate or not. TABLE X summarizes the final results for significant variables that contribute for safety for lane width = 10 ft. TABLE X indicates that variables **On Street Parking[0], Central Business District[0], Segment Length miles and AADTPLn** had significant p-values, although the parameter for AADTPLn nearly equals to zero. Since JMP gave output for Output=0/ Output =1 the sign for each parameter were reversed so as to give result for Output =1/ Output =0.

TABLE X. FINAL MODEL FOR SAFE LANE WIDTH = 10ft.

Term	*β*	*SE(β)*	χ^2	*p*	LOWER95%	UPPER 95%
Intercept	-0.6225	0.067	84.14	<.0001*	-0.4899	-0.7560
On Street Parking [0]	-0.8089	0.044	334.3	<.0001*	-0.7224	-0.8959
Central Business District [0]	0.5805	0.056	103.88	<.0001*	0.6938	0.4705
Sement Length miles	-0.2228	0.070	10.59	0.0011	-0.0933	-0.3609
AADTPLn	-0.0001	0.0	181.54	<.0001*	0.0001	0.0002

VI. DISCUSSIONS AND CONCLUSIONS

The results of this study demonstrate that logistic regression analysis can be used for analysis of impact of lane width on safety of vehicles. After determining the safest lane width using hypothesis testing, data analysis indicated that lane width = 10 ft is the safest among the rest. The reason for this may be due to the fact that the drivers might be driving slowly for small lane width as compared to wider lane widths. But, if this holds true then lane width = 9ft should have been the safest. But after observing the number of data set for lane width = 9ft, the samples were observed to be less that the rest of the lane widths. This may have caused some effect on the final data analysis. After determining the safest lane width (=10ft), logistic regression analysis indicated that variables On Street Parking[0], Central Business District[0], Segment Length miles and AADTPLn contributed the most for the safety of vehicles. This can be interpreted as for lane width = 10 ft to be safe, one should avoid on street parking and Central Business District parking.

In conclusion, these results demonstrate the validation of logistic regression to assess the safety of roads.

VII. LIMITATIONS AND CRITICAL REVIEW

One of the main limitations of this data was the lack of factors other than Crash Frequencies. For example variables similar to On Street Parking, District parking and so on. These variables proved to be useful for determining the factors that contribute in the safety of a particular lane width

The analyst categorized the data set into few levels without stating solid references.