Normalization of Accident Prediction Parameters at Unsignalized Intersections Using Skewness and Kurtotic Approach

Nosakhare Kent Oghoyafedo,Jacob Odeh Ehiorobo, Sylvester Obinna Osuji

Abstract – Accident prediction parameters are parameters considered to be the independent variables in the development of an accident prediction model. These prediction parameters are discrete in nature since they cannot assume continuity owing to the fact that they must assume a whole value. For this reason some mathematical models such as Poisson Regression, Negative Binomial Regression, Zero Inflated Regression etc. have been developed specifically these data. It is our bid to use multiple linear regression method for the development of the accident prediction model but the condition of normal distribution must be satisfied.The statistical normality tool employed for normalization of the accident prediction parameters was the skewness and kurtotic approach and further checking the Z-Value which have a satisfying conditional range of ±1.96. The Z-Value is computed using the skewness and kurtotic and standard deviation values. The SPSS software was used in computing the skewness, kurtotic and standard deviation values.The result obtained from the analysis shows that the accident prediction parametersfrom the five selected unsignalized intersection were normally distributed and hence the multiple linear regression method can be adopted in the development of the mathematical model.

Keywords –Accident Prediction Parameters, Kurtosis, Normalization, Prioritization, Skewness, Standard and deviation Z-Value

1 INTRODUCTION

ACCIDENT prediction parameters which are count data are categorized to be discrete random variables which necessitate the used of some specific models such Poisson Regression, Negative Binomial Regression, Zero Inflated Regression etc. These models have been designed because of the dynamic nature of count data.
However, it is intended that in this paper a multiple linear regression method will be used, this is based on the condition that the sample data at the unsignalized intersections are normally distributed using any of the available normality tool.
The assumption of normality is especially critical when constructing reference intervals for variables (Royston, 1991). Normality and other assumptions should be taken seriously, for when these assumptions do not hold, it is impossible to draw accurate and reliable conclusions about reality (Field, 2009; Oztuna, 2006). In large samples (> 30 or 40), the sampling distribution tends to be normal, regardless of the shape of the data (Field, 2009; Elliot et al., 2007).

Lack of symmetry (skewness) and pointiness (kurtosis) are two main ways in which a distribution can deviate from normal. The values for these parameters should be zero in a normal distribution.
An absolute value of the score greater than 1.96 or lesser than -1.96 is significant at $P < 0.05$, while greater than 2.58 or lesser than -2.58 is significant at $P < 0.01$, and greater than 3.29 or lesser than -3.29 is significant at $P < 0.001$. In small samples, values greater or lesser than 1.96 are sufficient to establish normality of the data. However, in large samples (200 or more) with small standard errors, this criterion should be changed to ±2.58 and in very large samples no criterion should be applied (that is, significance tests of skewness and kurtosis should not be used) (Field, 2009).

2 MATERIALS AND METHODS

2.1 The Study Area

The study area is Benin City located in the southern part of Nigeria, Benin City lies between Latitude 6º 14' 00'' North to 6º 21' 00'' North and Longitude 5º 34' 00'' East to 5º 44' 00'' East and with an average elevation of 80 meters above mean sea level. It comprises six local government areas with an estimated population of 1,147, 188 people (NPC 2006). It is 40Km north of the Benin River and 320Km by road east of Lagos. The weather condition in the area is characterized by mainly thunderstorm.

- *Nosakhare Kent Oghoyafedo is currently an Assistant Lecturer in Civil Engineering Department in University of Benin, Nigeria, PMB 1154. E-mail: nosakhare.oghoyafedo@uniben.edu*
- *Jacob Odeh Ehiorobo is currently a Professor in Civil Engineering Department in University of Benin, Nigeria, PMB 1154. E-mail: jacehi@uniben.edu*
- *Sylvester Obinna Osujiis currently an Associate Professor in Civil Engineering Department in University of Benin, Nigeria, PMB 1154. E-mail: Sylvester.osuji@uniben.edu*

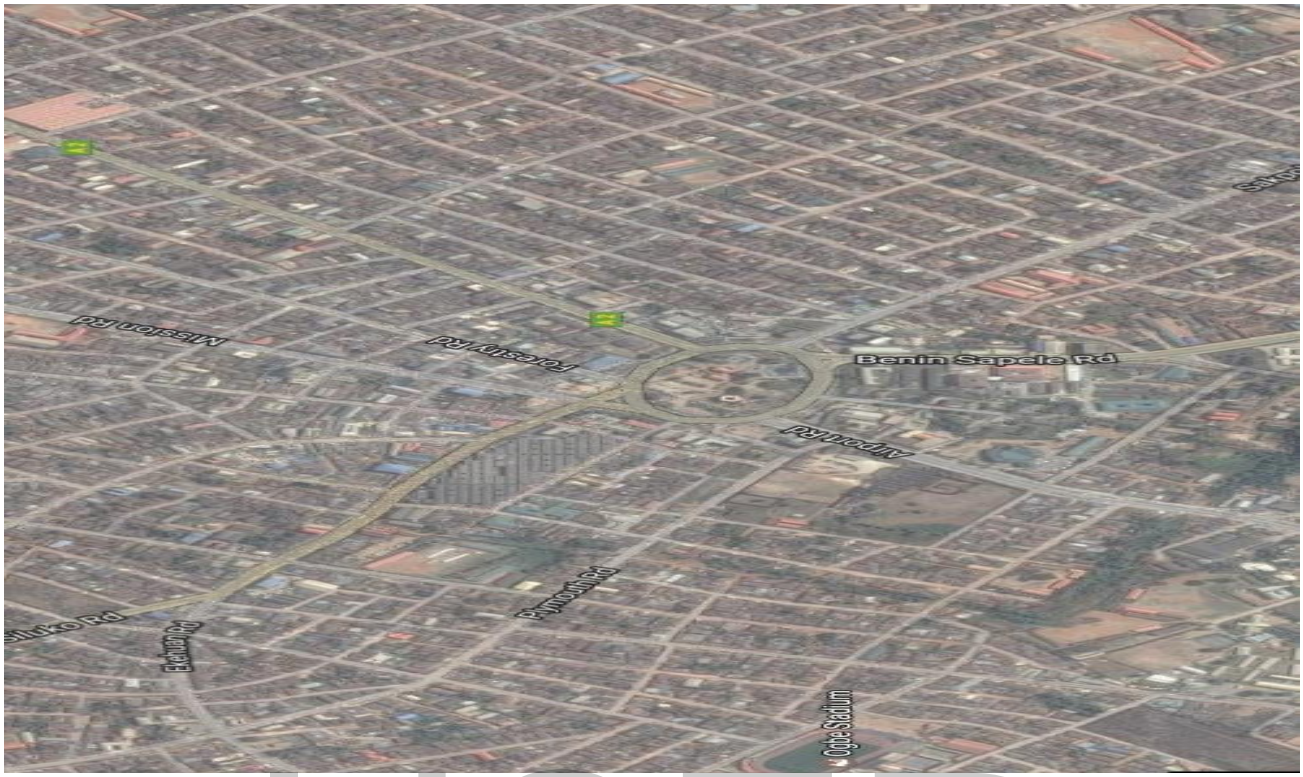

Figure 1: Satellite Earth Imagery Showing Road Network in Benin City

2.2 Data Collection

Selection of sample intersections was made based on stratification by traffic flow and intersection characteristics to ensure that a wide range of flows and intersection characteristics were captured. For each intersection, detailed information regarding accidents, traffic flow, geometric characteristics, traffic characteristics, road way condition, approach speed, lighting, among others were gathered and these are discussed below.

2.2.1 Accident Data at Intersection

Accident data between the following period of years, 2011 - 2015 of the intended selected intersection were collected from the Federal Road Safety Corps head office. The database was compiled from police using a standard accident report form

2.2.2 Traffic Flow Data at Intersection

Traffic flow data collected included vehicle counts from both major and minor road not classified by
type of vehicles and turning movement and spot speeds of vehicles as they approached the intersection area along the major arms. . Traffic counts were conducted during the morning and evening peak periods from 7:00am to 9:00am hours and from 4:00pm to 6:00pm hours, respectively.

2.2.3 Geometric Data AtIntersection

Intersection inventories were carried out to collect information relating to the site details. The information collected included intersection layout, type of major and minor roads (i.e. whether single or dual-carriage way), numbers, type and widths of lanes, types of median or other island, if any, and dimensions. Due to the absence of as built drawings for nearly all the sites, it was not possible to measure the radius of curvature of the entry kerb lines, which is considered important for intersections safety. The width of the minor roads atthe neck of the junctions was measured and used as a proxy for the latter. The site geometric and other traffic variables that were of importance in the modeling process are presented in chapter three.

2.3 Prioritization of Secondary Data

Prioritization involves assigning suitable weights to different factors so as to achieve a desired result.
In this model, the various factors which tend to influence the occurrence of accidents on roads are assigned weights on a scale of 0-10 in such a manner that the factors which tend to increase the probability of the accidents have lower weights. In order to prioritize roads for occurrence of accidents, various factors are considered and the weights assigned to them are given in following below in Table 1. The final weight assigned to each road link is obtained by adding all the individual weights and normalizing the value using maximum weight (in this case 90) that can be assigned. Hence,
Total weight = (Σ Individual Weights) x 100/ 90) (2.1)

Table 1: Factors used in Prioritization with their weights

S/N	Factors affecting occurrence of accident	Possible variation	Rank
1	No of lanes in each direction	4	10
		3	8
		2	6
		1	4
2	Number of vehicles per day`	Less than 1000	10
		Less than 2500	8
		Less than 5000	6
		Greater than 5000	4
3	Width of Road	More than 15m	10
		10.1 – 15m	8
		7.5 – 10.5m	6
		6.1 – 7.5m	4
		Less than 6m	2
4	Presence of Shoulder	Yes	10
		No	4
5	Surface condition of road	Flexible	10
		Rigid	8
6	Drainage condition	Good	8
		Satisfactory	6
		Poor	4
		No Drainage	2
7	Presence of traffic lights	Yes	10
		No	4
8	Provision of median	Yes	10
		No	4
9	Roundabout	Yes	10
		No	4
10	Length of road before	100m	10
		300m	8
		500m	6
		700m	4
		1000m	2
11	Conflict points	24	10
		17	9
		16	8
		13	7
		12	6
		11	5
		7	4
12	Type of Vehicles	Heavy vehicle	10
		Buses/Truck	8
		Car	4
		Two wheelers	1
13	visibility	Good	10
		Average	6
		Poor	4
		Very poor	2

Thus road links with high final weight are less prone to accidents than the road link with low final weight. The classification of roads for occurrence of accidents based on final weights is shown in Table 2.

Table 2: Prioritization Table

Final Normalized weight (%)	Accident prone level
80 – 100	Very Low
60 – 80	Low
40 – 60	Medium
0 – 40	High

2.4 Procedures on how to prioritize Using ArcView

1. Scan the map containing the desired road network and input this image to Arc View for digitizing.
2. Digitize the road network with due considerations for separation of every link and assign id number to every link.
3. Specify the attributes for every road link using the questionnaire provided.
4. Export the road attribute table generated in dbase format so that it can be imported by Arc view.
5. Join the road attribute table to the digitized road map and prioritize the road network for accident occurrence using total weights assigned to every link
6. Accident black spots on a given road network are ranked by result obtained from prioritization.

2.5 Weight Average

The various factors considered for affecting the occurrence of accidents may not have similar effect. Every factor will have a different level of involvement for an accident to take place. For example Presence of Shoulder and AADT cannot be given same weightage because more traffic may be a greater factor in occurrence of accident as compared to whether a shoulder is present on the side of a road or not.

Table 3: Classification of weights

Weight	Accident prone level
1	Very Low
2	Low
3	Medium
4	High
5	Very High

2.6 Test for Skewness and Kurtosis

Skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean. In other words, skewness tells you the amount and direction of skew (departure from horizontal symmetry). The skewness value can be positive or negative, or even undefined. If skewness is 0, the data are perfectly symmetrical, although it is quite unlikely for real-world data. As a general rule of thumb:

i. If skewness is less than -1 or greater than 1, the distribution is highly skewed.
ii. If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
iii. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

The moment coefficient of skewness of a data set is:

$$g1 = \frac{m_3}{m_2^{3/2}} \quad (2)$$

$$m3=\frac{\sum_{i=1}^{n}(\bar{x}-x)^3}{n} \quad (3)$$

$$m2=\frac{\sum_{i=1}^{n}(\bar{x}-x)^2}{n} \quad (4)$$

where;

g_1 = skewness
$\bar{x}$ = mean and
n = sample size,
m_3 = the third moment of the data set,
m_2 = variance.

However, if using the whole population, then g1 above is the measure of skewness; but if using just a sample, the sample skewness is computed by:

$$G_1 = \frac{\sqrt{n(n-1)}}{n-1} g_1 \quad (5)$$

$$Zg_1 = \frac{G_1}{SES} \quad (6)$$

$$SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \quad (7)$$

where;

Zg_1 = Test statistic
G_1 = Sample skewness
n = sample size,
SES = Standard Error of Skewness

Kurtosis is a measure of the "peakedness" of the probability distribution of a random variable. Kurtosis specifies the height and sharpness of the central peak, relative to that of a standard bell curve. As skewness involves the third moment of the distribution, kurtosis involves the fourth moment skewness and in a symmetric distribution both tails increase the kurtosis, unlike skewness where they offset each other. The reference standard is a normal distribution, which has a kurtosis of 3. In token of this, often the excess kurtosis is presented; excess kurtosis is simply kurtosis minus 3.

i. A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈3 (excess ≈0) is called mesokurtic.
ii. A distribution with kurtosis <3 (excess kurtosis <0) is called platykurtic. Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader.
iii. A distribution with kurtosis >3 (excess kurtosis >0) is called leptokurtic. Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper.

The moment coefficient of kurtosis of a data set is computed almost the same way as the coefficient of skewness: just change the exponent 3 to 4 in the formulas:

$$a4 = \frac{m_4}{m_2^2} \quad (8)$$

$$g2 = a4 - 4 \quad (9)$$

$$m4 = \frac{\sum_{i=1}^{n}(\bar{x}-x)^4}{n} \quad (10)$$

$$m2 = \frac{\sum_{i=1}^{n}(\bar{x}-x)^2}{n} \quad (11)$$

where

g2= excess kurtosis
$\bar{x}$ = mean
n = sample size,
m4 = fourth moment of the data set,
m2 = variance.

Again, the excess kurtosis is generally used because the excess kurtosis of a normal distribution is 0. However, if using the whole population, then g2 above is the measure of kurtosis; but if using just a sample, the sample kurtosis is computed by this formula, which comes from

$$G2 = \frac{n-1}{(n-2)(n-3)}[(n+1)g2 + 6)] \quad (12)$$

Also for a sample, there will be the need to calculate standard error of kurtosis (SEK). (i.e dividing the sample excess kurtosis by the standard error of kurtosis (SEK) to get the test statistic, which tells how many standard errors the sample excess kurtosis is from zero).

$$Zg2 = \frac{G_2}{SEK} \quad (13)$$

$$SEK = 2\,(SES)\sqrt{\frac{n^2-1}{(n-3)(n+5)}} \quad (14)$$

2.7 Z Value

If a statistical data set has a normal distribution, it is customary to standardize all the data to obtain standard scores known as z-values or z-scores. The distribution of z-values takes on a standard normal distribution (or Z-distribution).

However if we divide the measure by its standard error; the resulting value is the Z-value which should fall within the range of -1.96 to +1.96. The Z-value is a measure of normality, if the Z-value is less than -1.96 and greater than +1.96; it shows that the sample data consider is not normally distributed. The Z – Value is given by:

$$Z\text{-value} = \frac{\text{Measure}}{\text{Standard Error}} \quad (15)$$

3.0 RESULT AND DISCUSSION

The independent variables are variables that measures and influences the outcome (dependent variable) of a mathematical or statistical model. They are independent of each other. They can constantly and intentionally change to observe their effect on the dependent variable. The Table 4 shows the values of the independent variables as used in this project work.

Table 4: Independent Variables of the Accident Prediction Model

S/N	PREDICTORS	PZ Intsn.	RCC Intsn.	NNPC Intsn.	CJ Intsn.	BBP Intsn.
1	Major Traffic	5942	5906	4986	4110	4016
2	Minor Traffic	660	1000	1684	400	820
3	Turning Traffic	368	324	595	80	300
4	Approach Width (m)	10.46	10.23	10.34	8.42	7.8
5	Number of Legs	3	3	3	3	3
6	Surface Condition	Flexible	Flexible	Flexible	Flexible	Flexible
7	Speed (Km/h)	63.008	56.500	58.900	58.716	45.23
8	Presence of Shoulder	NO	NO	NO	YES	YES
9	Traffic Light	NO	NO	NO	NO	NO
10	Lighting Condition	Very Poor	Very Poor	Very Poor	Very Poor	Very Poor
11	Number of Lanes	3	3	3	3	3

12	Drainage Condition	Poor	Poor	Satisfactory	Satisfactory	No Drainage
13	Frequent Vehicle on the Road	Cars	Cars	Cars	Cars	Cars

3.1 Prioritization value of the Independent Variables

Prioritization involves assigning suitable weights to different factors that influences accident occurrence so as to achieve a desired result. The Table 4 below shows the prioritized value of the accident predictors (independent variables).

Table 4: Prioritized Value of the Independent Variables

S/N	PREDICTORS	PZIntsn.	RCC Intsn.	NNPC Intsn.	CJ Intsn.	BBP Intsn.
1	Major Traffic	4	4	6	6	6
2	Minor Traffic	10	8	8	10	10
3	Turning Traffic	10	10	10	10	10
4	Approach Width	8	8	8	6	6
5	Number of Legs	5	5	5	5	5
6	Surface Condition	10	10	10	10	10
7	Speed	6	6	6	6	6
8	Presence of Shoulder	4	4	4	10	10
9	Traffic Light	4	4	4	4	4
10	Lighting Condition	2	2	2	2	2
11	Number of Lanes	8	8	8	8	8
12	Drainage Condition	4	4	6	6	2
13	Frequent Vehicle on the Road	4	4	4	4	4

3.2 Weightage Point of the Independent Variable

This has to do with ranking of the prioritized values on a scale of 1 – 5 in such a manner that the factors which tend to increase the probability of accidents have higher weights. The final weight assigned to each prioritized values of the independent variables is obtained by adding all the individual weights and normalizing the value using maximum weight (in this case 90) that can be assigned. The table 5 below shows the weightage point of the prioritized values.

Table 5: Weightage Point of the Prioritized Value of the Independent Variables

Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
79	4	2	2	3	4	2	4	4	4	5	3	4	4
69	4	3	2	3	4	3	4	4	4	5	4	3	3
25	4	3	2	4	3	2	3	4	4	5	3	4	3
39	3	2	2	3	4	2	3	2	3	5	3	3	4
57	4	2	2	3	4	2	4	2	4	4	3	5	4

3.3 Normalization of the Weightage Point Using Skewness and Kurtosis Approach

The skewness and kurtosis measure should be as close to zero as possible. In reality, however, data are often skewed and kurtotic. A small departure from zero is no problem as long as the measure is not large compare to their standard error.

The normalization analysis presented below was carried out using SPSS whose procedures has been outlined in previous section. Thirteen prediction parameterswere investigated at five intersections.

The Table 6 represent the case processing summary which indicates the sample size and the percentage of valid, missing and total analysis carried out at each intersection.

Table 6: Case Processing Summary

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
PZ	13	100.0%	0	0.0%	13	100.0%
RCC	13	100.0%	0	0.0%	13	100.0%
NNPC	13	100.0%	0	0.0%	13	100.0%
CJ	13	100.0%	0	0.0%	13	100.0%
BBP	13	100.0%	0	0.0%	13	100.0%

Table 7: PZ Intersection Description output

Descriptives				
			Statistic	Std. Error
PZ	Mean		3.46	.268
	95% Confidence Interval for Mean	Lower Bound	2.88	
		Upper Bound	4.05	
	5% Trimmed Mean		3.46	
	Median		**4.00**	
	Variance		.936	
	Std. Deviation		.967	
	Minimum		2	
	Maximum		5	
	Range		3	
	Interquartile Range		2	
	Skewness		-.525	.616
	Kurtosis		-.784	1.191

$$\text{Z-value} = \frac{\text{Measure}}{\text{Standard Error}}$$

$$\text{Z-value} = \frac{-0.525}{0.616} = -0.852$$

$$\text{Z-value} = \frac{-0.784}{1.191} = -0.658$$

The computation of the Z-values using information in Table 7(– 0.852 and – 0.658), which falls within the range of – 1.96 to + 1.96 shows that the sample data of PZ intersection is normally distributed and are not significantly different than a normal population.

Table 8: RCC Intersection Description output

Descriptives				
			Statistic	Std. Error
RCC	Mean		3.54	.243
	95% Confidence Interval for Mean	Lower Bound	3.01	
		Upper Bound	4.07	
	5% Trimmed Mean		3.54	
	Median		4.00	
	Variance		.769	
	Std. Deviation		.877	
	Minimum		2	
	Maximum		5	
	Range		3	
	Interquartile Range		1	
	Skewness		-.575	.616
	Kurtosis		-.121	1.191

$$\text{Z-value} = \frac{-0.575}{0.616} = -0.933$$

Z-value = $\frac{-0.121}{1.191}$ = – 0.102

The computation of the Z-values using information in Table8 (– 0.933 and – 0.102), which falls within the range of – 1.96 to + 1.96 shows that the sample data of RCC intersection is normally distributed and are not significantly different than a normal population.

Table 9: NNPC Intersection Description output

Descriptives				
			Statistic	Std. Error
RCC	Mean		3.54	.243
	95% Confidence Interval for Mean	Lower Bound	3.01	
		Upper Bound	4.07	
	5% Trimmed Mean		3.54	
	Descriptives			
			Statistic	Std. Error
	Median		4.00	
	Variance		.769	
	Std. Deviation		.877	
	Minimum		2	
	Maximum		5	
	Range		3	
	Interquartile Range		1	
	Skewness		-.575	.616
	Kurtosis		-.121	1.191

Z-value = $\frac{-0.575}{0.616}$ = – 0.933

Z-value = $\frac{-0.121}{1.191}$ = – 0.102

The computation of the Z-values using information in Table9 (– 0.933 and – 0.102), which falls within the rangeof – 1.96 to + 1.96 shows that the sample data of RCC intersection is normally distributed and are not significantly different than a normal population.

Table 10: CJ Intersection Description output

Descriptives				
			Statistic	Std. Error
RCC	Mean		3.08	.265
	95% Confidence Interval for Mean	Lower Bound	2.50	
		Upper Bound	3.65	
	5% Trimmed Mean		3.03	
	Median		3.00	
	Variance		.910	
	Std. Deviation		.954	
	Minimum		2	
	Maximum		5	
	Range		3	
	Interquartile Range		2	
	Skewness		.507	.616
	Kurtosis		-.394	1.191

Z-value = $\frac{-0.507}{0.616}$ = 0.823

Z-value = $\frac{-0.394}{1.191}$ = – 0.331

The computation of the Z-values using information in Table 10 (– 0.823 and – 0.331), which falls within the range of – 1.96 to + 1.96 shows that the sample data of CJ intersection is normally distributed and are not significantly different than a normal population.

Table 11: BBP Intersection Description output

Descriptives				
			Statistic	Std. Error
BBP	Mean		3.38	.311
	95% Confidence Interval for Mean	Lower Bound	2.71	
		Upper Bound	4.06	
	5% Trimmed Mean		3.37	
	Median		4.00	
	Variance		1.256	
	Std. Deviation		1.121	
	Minimum		2	
	Maximum		5	
	Range		3	
	Interquartile Range		2	
	Skewness		-.079	.616
	Kurtosis		-1.387	1.191

$$\text{Z-value} = \frac{-0.079}{0.616} = -0.128$$

$$\text{Z-value} = \frac{-1.387}{1.191} = -1.165$$

The computation of the Z-values using information in Table 11 (– 0.128 and – 1.165), which falls within the range of – 1.96 to + 1.96 shows that the sample data of CJ intersection is normally distributed and are not significantly different than a normal population.

From the normalization analysis above using skewness and kurtotic approach in consideration to the selected intersections with satisfaction to the range condition of ±1.96 shows that the accident prediction parameters at the selected intersections were normally distributed.

Though the skewness and kurtotic values were not exactly zero indicating that the prediction parameters were deviated from the mean.This also indicates that the mean, median and mode were not symmetrical at a particular point. However, since the skewness and kurtotic values were not too deviated from zero, it pose no problem to the accuracy of the result obtained and the inference made as long as the measure is not large compare to their standard error.

4.0 CONCLUSION

The aim of this paper was to normalize the accident prediction parameters at five different selected unsignalized intersections using available statistical tool. Since a multiple linear regression method was to be adopted in developing a mathematical model relating accident rate per year following period of years, 2011 – 2015 to unsignalized intersection parameters which tends to influence the probability of accident occurrence.

However, before the adoption of the multiple linear regression method, the condition for normalization must be satisfied since the intended data are count data.

The analyses carried out using skewness and kurtotic approach as a statistical normality tool indicates that these prediction parameters considered were normally distributed. Furthermore, the use of multiple linear regression method can be considered.

REFERENCES

Elliott A. C., Woodward W. A., Statistical analysis quick reference guidebook with SPSS examples.
1st edition. London: Sage Publications;2007.

Field A. Discovering statistics using SPSS. 3 ed. London: SAGE publications Ltd; 2009. p. 822.

Oztuna D, Elhan A.H., Tuccar E. Investigation of four different normality tests in terms of type 1 error rate and power under
different distributions. Turkish Journal of Medical Sciences.2006;36(3):171-176.

Royston P. Estimating departure from normality. Stat Med. 1991;10(8):1283-1293.