Multivariate Regression Techniques for Analyzing Auto-Crash Variables in Nigeria

Olushina Olawale Awe[1*] Mumini Idowu Adarabioyo[2]

1. Department of Mathematics, Obafemi Awolowo University Ile-Ife, Nigeria.
2. Department of Mathematical Sciences, Afe Babalola University, Ado Ekiti, Nigeria.

*E-mail of the corresponding author: olawaleawe@yahoo.co.uk

Abstract

It is unequivocally indisputable that motor vehicle accidents have increasingly become a major cause of concern for highway safety engineers and transportation agencies in Nigeria over the last few decades. This great concern has led to so many research activities, in which multivariate statistical analysis is inevitable. In this paper, we explore some regression models to capture the interconnectedness among accident related variables in Nigeria. We find that all the five variables considered are highly interrelated over the past decade, resulting in a high risk of mortality due to auto-crash rate. The result of our analysis, using an appropriate statistical software, also reveals that the simple regression models capture the relationships among the variables more than the multiple regression model considered.

Key Words: Multivariate Model, Analyzing, Regression, Data, Accident, Rate.

1. Introduction

Multivariate techniques and statistical tests are needed to analyze data in many areas of human endeavor in order to provide descriptive and inferential procedures which we can use to detect behavioral patterns or test hypotheses about parameters of interest. Controversy has continued to trail the exact number of deaths recorded yearly through road accident in Nigeria with World Health Organization(WHO), the National Union of Road Transport Workers(NURTW) and the Federal road Safety Commission of Nigeria(FRSCN) giving conflicting reports. While the international agency claimed that 32,000 died yearly through road accidents in Nigeria, the FRSCN insisted that the country had only recorded between 4000 and 5000 deaths from road accidents in the last three years. The president of the National Union of Road Transport Workers of Nigeria(NURTW) once claimed that, “despite the fact

Journal of Natural Sciences Research www.iiste.org
ISSN 2224-3186 (Paper) ISSN 2225-0921 (Online)
Vol.1, No.1, 2011
that not all deaths and accidents on our roads are officially reported, 8, 672 people were said to have lost their lives to road accidents in Nigeria in 2003, while another 28,215 people sustained different degrees of injuries within the period. The number of people dying as a result of road accident in Nigeria has reached an alarming proportion as accident rates increases towards the end of the year especially as from the month of September (Ojo, 2008). Analysis of the traffic crashes recorded over a five year period of 2000- 2006 showed that 98,494 cases of traffic crashes were recorded out of which 28,366 were fatal and resulted into deaths(FRSCN Report,2009).This revealing statistics shows that Nigeria is placed among the fore front nations experiencing the highest rate of road tragedies in the world. This paper focuses on determining the degree of association between those who are killed in road crashes and variables like the number of vehicles involved, number of accidents recorded, number injured and the particular month the accident occurred. The rest of the paper is organized as follows: section two considers the data and methodology used in the study, section three enumerates the main results, section four is on the discussion and findings from the study, while section five concludes the study. The various analysis performed are presented and labeled as exhibits below the conclusion.

2. Data and Methodology

2.1 Data

Accidents Statistics covering s period of five years were collected (2003-2007) from Lagos State Command of the Federal Road Safety Corps. The data were then summed up according to the particular month the accident occurred, thereby giving us a sample size of twelve. The essence of this is to determine the effect of a particular month in the year on accident situation in Lagos State as the month increases to December.

2.2 Methodology

A simple linear regression equation of the dependent variable on each of the other factors and a multiple regression equation was fitted on all the independent variables. The simple linear regression is a special case of the multiple linear regression(Rencher,2002),so we consider first simple linear regressions of the dependent variable on each of the independent variables.The dependent variable for the analysis is the number of people killed and the independent variables are x_1, x_2, x_3 and x_4 (what each variable represents is given below).

Y = f(X1+ X2+ X3+ X4)---------------------------(1)

The hypothesis tested in the study is that: there is no significant relationship between Number of people killed and the variables x_1, x_2, x_3 and x_4 which could not be explained on the basis of chance alone.

The Multiple linear regressions is defined by:

$$Y_i = \alpha_i + X_i\beta_i + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + \varepsilon_i \text{ -------------------(2)}$$

Where Y_{i_killed} = the number people killed in the accident

$X_{1i\_accident}$ = the number of accidents

$X_{2injured\_}$ = the number of injured persons

$X_{3i\_vehicle}$ =Number of vehicles involved

X_{4i_month} = the particular month the accident occurred.

ε_i is the random error term of the model

After identifying the hypothesis for testing, statistical analysis was performed on all the variables (Y, X1, X2, X3 and X4). The results of the analyses are presented in exhibits 1, 2, 3, 4 and 5.

The simple linear regression is carried out between Y_{i_killed} and each of the independent variables $X_{1i\_accident}, X_{2i\_injured}, X_{3i\_vehicle}$ *and* X_{4i_month}

and the results are displayed in Tables 1, 2, 3 and 4.

2.3 Classical Assumptions of the Linear Regression Model

The assumptions of the linear regression model are stated as follows:

- The model has a linear functional form.
- The independent variables are fixed.
- Independent observations.
- Normality of the residuals.
- Homogeneity of residuals variance.
- No Multicollinearity.
- No autocorrelation of the errors.
- No outlier distortion.

3. Main Results

This section discusses the results of the various regression models fitted to the accident data.

3.1 Linear regression of Y_{i_killed} on $X_{1i\_accident}$.

In the analysis the coefficient of correlation(r) between the two variables is 0.326 and the coefficient of determination (r^2) is 0.1063. r^2 is small that is the amount of variation in the number killed accounted for by the number of accident is 10.63% with probability value of 0.151 greater than alpha (0.05) so the association is not so statistically significant.

The regression equation is

$$Y_{i\_killed} = 1786.116 + 0.559\, X_{1i\_accident} \qquad \text{------------------------------(3)}$$

that is for every unit change in the number of accident, there is a positive 0.559 change in the number of those killed. This is a direct relationship. The model is not significant at P(0.05) as the P-value is 0.301 greater than alpha. See exhibit 1.

3.2 Linear regression of Y_{i_killed} on $X_{2i\_injured}$.

In the analysis, the coefficient of correlation(r) between the two variables is 0.702 and the coefficient of determination (r^2) is 0.493. r^2 is large that is the amount of variation in the number killed accounted for by the number injured is 49.3% with probability value of 0.011 less than alpha (0.05) so the association is statistically significant.

The regression equation is Y_{i_killed} =*1005.283 +1.674* $X_{2i\_injured}$ --------------(4)

that is for every unit change in the number injured; there is a positive 1.674 change in the number of those killed. This is a direct relationship. The model is significant at P(0.05) as the P-value is 0.011less than alpha. See Exhibit 2.

3.3. Linear regression of Y_{i_killed} on $X_{3i\_vehicle}$.

In the analysis the coefficient of correlation(r) between the two variables is 0.705 and the coefficient of determination (r^2) is 0.443. r^2 is large that is the amount of variation in the number killed accounted for by the number of vehicle involved is 44.3% with probability value of 0.011 less than alpha (0.05) so the association is statistically significant.

The regression equation is

Y_{i_killed} =*845.674 +0.688* $X_{3i\_vehicle}$ -------------------------(5)

that is for every unit change in the number of vehicle, there is a positive 0.688 change in the number of those killed. This is a direct relationship. The model is significant at P(0.05) as the P-value is 0.011 less than alpha. See exhibit 3.

3.4 Linear regression of Y_{i_killed} on X_{4i_month}.

In the analysis the coefficient of correlation(r) between the two variables is 0.675 and the coefficient of determination (r^2) is 0.455. r^2 is large that is the amount of variation in the number killed accounted for by the particular month is 45.5% with probability value of 0.016 less than alpha (0.05) so the association is statistically significant.

The regression equation is

$$Y_{i\_killed} = 2445.132 + 69.318\, X_{3i\_vehicle} \text{ ----------------------------(6)}$$

that is for every unit change in the number of vehicle, there is a positive 69.318 change in the number of those killed. This is a direct relationship. The model is significant at P(0.05) as the P-value is 0.016 less than alpha. See exhibit 4.

3.5 Multiple Linear Regression Analysis of Y_{i_killed} on all the explanatory variables.

In the analysis, the coefficient of correlation(r) between the two variables is 0.0.79 and the coefficient of determination (r^2) is 0.591. r^2 is large, that is the amount of variation in the number killed accounted for by all the independent variables is 59.1% with probability value of 0.135 greater than alpha (0.05) so the association is not statistically significant.

The multiple regression equation is

$$Y_{i\_killed} = 739.489 + 0.075\, X_{1i\_accident} + 0.657\, X_{2i\_injured}$$
$$+ 0.39\, X_{3i\_vehicle} + 15.576\, X_{4i\_month} \; (6)$$

There is positive correlation between Y_{i_killed} and all other independent variables. The P-value of all variables except $X_{1i\_accident}$ are less than alpha and so shows statistically significant relationship. The p-value of $X_{1i\_accident}$ is 0.151 greater than alpha and shows that there is no statistically significant relationship between the number of people who were killed and the number of vehicles involved.

4. Discussion of Findings

Our findings reveal that the Multiple Linear Regression Model fitted is not statistically significant. However, the relationship between each variable and the Y_{i_killed} separately are statistically significant in except for the variable $X_{1i\_accident}$. The variance accounted for by the variable Y_{i_killed} was low in all the variables. The correlation matrix (Exhibit 5) more accurately justifies the hypothesis of positive correlation between all the independent variables and the dependent variable. The correlation of those who were killed with the injured, the number of vehicles and the month the accidents occurred were strongly positive (Exhibits 2, 3 and 4). The implications of these findings is that the more vehicles involved in an accident the more people are killed and as the months approaches December the more people are killed in road accident in Nigeria. The overall probability value of the model is 0.135 which is greater than the alpha value of 0.05, so the model is not relevant. However, there may be many more variables affecting number of people killed in an accident Y_{i_killed} that needs to be explored in further studies.

5.0 Conclusion.

From our analysis, we have seen that the overall model (Multiple Linear Regression Model) fitted for the accident data is not significant, though there is positive and strong correlation between the dependent variable and each of the independent variables. This suggests that there are other variables that actually account for deaths resulting from auto-crash in Lagos State, Nigeria, which if included in the model will make it more relevant. These variables need to be explored to form a more robust model for predicting factors affecting number of people killed as a result of auto-crash in Lagos State, Nigeria.

References

Anyata, B. U.et al (1986); A Case for Increased Investment on Road Usage Education in

Nigeria, Proceedings of the First International Conference Held in University of

Benin, Nigeria.

Alvin.C. Rencher (2002):Methods of Multivariate Analysis.2nd Edition.Bringham Young University. A John Wiley Publications.

Brussels, (2006); Commission of the European Communities Proposal for a Directive of the

European Parliament and of the Council on Road Infrastructures Safety. Management. [SEC(2006) 1231/1232]

Hohnsheid, K. J, (2003): Road Saftey Impact Assessment. Bergisch Gladbach, Bundesanstalt

Strassenwesen. (Internet report)

Reurings M, (2006): Modelling the Number of Road Accidents using Generalized Linear

Models. SWOV, Leidschendan

Rob E. (2005): Accident Prediction Models and Road Safety Assessment (Internet Report)

Slefan. C. (2006): Predictive Model of Injury Accidents on Austrian Motorways. KFV. Vienna.

Vikas Singh, (2006); Statistical Analysis of the Variables Affecting Infant Mortality Rate in

United States. Journal of the Department of Health Services Administration, University of Arkansas Medical Services

Wichert S, (226): Accident Prediction Models For Portuguese Motorways. LNEC, Lisbon

www.makeroadsafe.org

Exhibit 1

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.326[a]	.106	.017	367.27931	.106	1.187	1	10	.301

a. Predictors: (Constant), X1_ACCDT

b. Dependent Variable: Y_KILLED

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1786.116	1023.944		1.744	.112					
	X1_ACCDT	.559	.513	.326	1.090	.301	.326	.326	.326	1.000	1.000

a. Dependent Variable: Y_KILLED

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	160133.4	1	160133.360	1.187	.301[a]
	Residual	1348941	10	134894.089		
	Total	1509074	11			

a. Predictors: (Constant), X1_ACCDT

b. Dependent Variable: Y_KILLED

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2664.2156	3065.2813	2895.7500	120.65479	12
Std. Predicted Value	-1.919	1.405	.000	1.000	12
Standard Error of Predicted Value	106.032	237.487	143.759	44.506	12
Adjusted Predicted Value	2177.2100	3041.1980	2870.2766	226.39478	12
Residual	-564.965	677.78455	.00000	350.18708	12
Std. Residual	-1.538	1.845	.000	.953	12
Stud. Residual	-1.869	2.419	.029	1.143	12
Deleted Residual	-834.152	1164.790	25.47340	509.89718	12
Stud. Deleted Residual	-2.198	3.564	.089	1.438	12
Mahal. Distance	.000	3.682	.917	1.210	12
Cook's Distance	.000	2.103	.286	.618	12
Centered Leverage Value	.000	.335	.083	.110	12

a. Dependent Variable: Y_KILLED

EXHIBIT 2

Model Summary[b]

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.702[a]	.493	.443	276.47418	.493	9.742	1	10	.011

a. Predictors: (Constant), X2_INJURED

b. Dependent Variable: Y_KILLED

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1005.283	610.904		1.646	.131					
	X2_INJURED	1.674	.536	.702	3.121	.011	.702	.702	.702	1.000	1.000

a. Dependent Variable: Y_KILLED

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	744694.5	1	744694.535	9.742	.011[a]
	Residual	764379.7	10	76437.971		
	Total	1509074	11			

a. Predictors: (Constant), X2_INJURED

b. Dependent Variable: Y_KILLED

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2403.0457	3288.5745	2895.7500	260.19128	12
Std. Predicted Value	-1.894	1.510	.000	1.000	12
Standard Error of Predicted Value	79.831	176.882	109.159	29.982	12
Adjusted Predicted Value	2588.7886	3266.6946	2907.2408	243.34302	12
Residual	-420.170	396.56958	.00000	263.60779	12
Std. Residual	-1.520	1.434	.000	.953	12
Stud. Residual	-1.587	1.499	-.017	1.039	12
Deleted Residual	-458.388	432.93729	-11.49076	316.14455	12
Stud. Deleted Residual	-1.741	1.615	-.029	1.086	12
Mahal. Distance	.000	3.586	.917	1.074	12
Cook's Distance	.001	.551	.105	.153	12
Centered Leverage Value	.000	.326	.083	.098	12

a. Dependent Variable: Y_KILLED

Journal of Natural Sciences Research www.iiste.org
ISSN 2224-3186 (Paper) ISSN 2225-0921 (Online)
Vol.1, No.1, 2011

EXHIBIT 3

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.703[a]	.494	.443	276.32907	.494	9.763	1	10	.011

a. Predictors: (Constant), X3_VEHICLE

b. Dependent Variable: Y_KILLED

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	845.674	660.937		1.280	.230					
	X3_VEHICLE	.688	.220	.703	3.125	.011	.703	.703	.703	1.000	1.000

a. Dependent Variable: Y_KILLED

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	745496.7	1	745496.714	9.763	.011[a]
	Residual	763577.5	10	76357.754		
	Total	1509074	11			

a. Predictors: (Constant), X3_VEHICLE

b. Dependent Variable: Y_KILLED

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2314.8884	3233.5776	2895.7500	260.33138	12
Std. Predicted Value	-2.231	1.298	.000	1.000	12
Standard Error of Predicted Value	79.858	202.291	107.818	34.666	12
Adjusted Predicted Value	2522.6218	3470.0000	2919.4845	251.47191	12
Residual	-763.578	220.83029	.00000	263.46943	12
Std. Residual	-2.763	.799	.000	.953	12
Stud. Residual	-3.162	.897	-.036	1.096	12
Deleted Residual	-1000.00	278.41348	-23.73454	350.88265	12
Stud. Deleted Residual	-.951	.888	.239	.498	11
Mahal. Distance	.002	4.978	.917	1.389	12
Cook's Distance	.000	1.548	.192	.452	12
Centered Leverage Value	.000	.453	.083	.126	12

a. Dependent Variable: Y_KILLED

EXHIBIT 4

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.675[a]	.455	.401	286.69806	.455	8.360	1	10	.016

a. Predictors: (Constant), X4_MONTH

b. Dependent Variable: Y_KILLED

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	2445.182	176.450		13.858	.000					
	X4_MONTH	69.318	23.975	.675	2.891	.016	.675	.675	.675	1.000	1.000

a. Dependent Variable: Y_KILLED

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2514.5000	3277.0000	2895.7500	249.93026	12
Std. Predicted Value	-1.525	1.525	.000	1.000	12
Standard Error of Predicted Value	83.626	155.683	114.262	26.497	12
Adjusted Predicted Value	2474.0601	3249.8181	2896.0913	246.51793	12
Residual	-391.091	378.18182	.00000	273.35587	12
Std. Residual	-1.364	1.319	.000	.953	12
Stud. Residual	-1.576	1.498	.000	1.050	12
Deleted Residual	-538.200	487.93985	-.34125	333.19846	12
Stud. Deleted Residual	-1.725	1.614	-.015	1.100	12
Mahal. Distance	.019	2.327	.917	.847	12
Cook's Distance	.002	.520	.114	.156	12
Centered Leverage Value	.002	.212	.083	.077	12

a. Dependent Variable: Y_KILLED

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	687116.5	1	687116.477	8.360	.016[a]
	Residual	821957.8	10	82195.777		
	Total	1509074	11			

a. Predictors: (Constant), X4_MONTH

b. Dependent Variable: Y_KILLED

EXHIBIT 5

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change
1	.769[a]	.591	.357	297.07324	.591	2.525	4	7	.135

a. Predictors: (Constant), X4_MONTH, X1_ACCDT, X3_VEHICLE, X2_INJURED

b. Dependent Variable: Y_KILLED

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	Correlations Zero-order	Partial	Part	Collinearity Statistics Tolerance	VIF
1	(Constant)	739.489	1850.432		.400	.701					
	X1_ACCDT	.075	.478	.044	.158	.879	.326	.059	.038	.752	1.330
	X2_INJURED	.657	2.108	.276	.312	.764	.702	.117	.075	.075	13.384
	X3_VEHICLE	.390	.367	.399	1.064	.323	.703	.373	.257	.417	2.401
	X4_MONTH	15.576	84.752	.152	.184	.859	.675	.069	.044	.086	11.639

a. Dependent Variable: Y_KILLED

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	891306.7	4	222826.671	2.525	.135[a]
	Residual	617767.6	7	88252.509		
	Total	1509074	11			

a. Predictors: (Constant), X4_MONTH, X1_ACCDT, X3_VEHICLE, X2_INJURED

b. Dependent Variable: Y_KILLED

Correlations

		Y_KILLED	X1_ACCDT	X2_INJURED	X3_VEHICLE	X4_MONTH
Pearson Correlation	Y_KILLED	1.000	.326	.702	.703	.675
	X1_ACCDT	.326	1.000	.220	.472	.218
	X2_INJURED	.702	.220	1.000	.683	.955
	X3_VEHICLE	.703	.472	.683	1.000	.627
	X4_MONTH	.675	.218	.955	.627	1.000
Sig. (1-tailed)	Y_KILLED	.	.151	.005	.005	.008
	X1_ACCDT	.151	.	.246	.061	.248
	X2_INJURED	.005	.246	.	.007	.000
	X3_VEHICLE	.005	.061	.007	.	.014
	X4_MONTH	.008	.248	.000	.014	.
N	Y_KILLED	12	12	12	12	12
	X1_ACCDT	12	12	12	12	12
	X2_INJURED	12	12	12	12	12
	X3_VEHICLE	12	12	12	12	12
	X4_MONTH	12	12	12	12	12