VOL 2 NO 5: MAY, 2025
AN OPEN ACCESS PEER-REVIEWED JOURNAL
Frontline Professionals Journal 2(5), 16-23

Original Research Article

LOGLINEAR MODELLING OF ROAD ACCIDENTS FATALITIES IN NIGERIA (2018-2023)

Authors: Akinyemi S.G[1], Bada O[1] and Ahmadu A.O[1]

Affiliations: [1]Department of Statistics, Auchi Polytechnic, Auchi

Corresponding Author: seyegbemiga@yahoo.com,

Authors' contributions
This study was a collaborative effort among all authors. Each author reviewed and approved the final version of the manuscript for publication.

Article Information
EISSN 1596-0501
Website:https://frontlineprofessionalsjournal.info
Email: frontlineprofessionalsjournal@gmail.com

CITATION*:* Akinyemi S.G, Bada O and Ahmadu A.O (2025). Loglinear modelling of road accident fatalities in Nigeria (2018-2023)**.** *Frontline Professionals Journal* 2(5), 16-23

ABSTRACT
In recent years, road accidents have claimed many lives. This fatalities on the road are caused by poor road conditions, lack of drivers' licenses, alcohol, and other factors. However, this paper focuses on the number of fatalities caused by road accidents in Nigeria from 2018 to 2023. Data were collected from the office of road safety regarding cases of fatalities from road accidents across the six geopolitical zones from 2018 to 2023. The data were analyzed using log-linear modeling, which helped us select the best model to represent the data. R Statistics was used to analyze and fit various models based on the year, state and geopolitical zones. A Chi-square test for independence was conducted between the Year of Death, State and Geopolitical Zone. The fitted model was selected based on the AIC (Akaike Information Criterion) generated by the R package for categorical modeling. The findings showed that the model with the three factors (Geopolitical Zones, State, and Year) best explains the collected. It was concluded that the years of death by road accident are dependent of the Geopolitical zone. It is recommended that attention should be paid to road accident prevention to avoid unplanned expenses through supplementary budgeting for the unexpected rise in accident casualties.

Key Words: Geopolitical Zones, Akaike Information Criterion (AIC), Loglinear Modelling, Categorical Data, Simulation.

INTRODUCTION
Road accidents worldwide and in particular in Nigeria have become a regular phenomenon. Every day, Nigerians die on the road and many are injured. The essential role of medical institutions emerges during these situations since the initial minutes following an accident are known as the "golden hour" but they are both critical and incredibly important. Many lives could be saved and disabilities prevented by providing immediate treatment to accident victims. Among the accidents resulting from the development of air, land, sea, and river transport, road accidents predominate both in respect of their frequency and seriousness and in terms of human and economic cost. There is no doubt about it; there are far too many accidents on the road today. Even if you are the most careful driver on the road, there is still a chance that someone will slam into your car, causing you at the very least some property damage and may something far worse. This is disheartening indeed because road accidents is becoming a daily occurrence due to road traffic integration being treated as a secondary issue in Nigeria. It is a known fact that the transportation system is the movement of people and goods from one location to another. Transport is performed by various modes such as air, rail, road, water, cable, pipeline and space, Road is an identifiable roadway or path between two or more places. Roads receive different treatments to become suitable for simple movement. The automobile stands as the leading road vehicle because it provides flexible mobility with limited transportation capabilities. The system provides enhanced transportation solutions with reduced budget requirements. Adekoye (2016) lamented that "road accident in Nigeria

claim more lives daily than any other faster because of the multiple lapses in our system contribute to such road mishap, source of which are fatal in some cases" there is nothing you can do that will guarantee you a constantly safe and totally accident free driving experience but there are several things that travelers could do to help minimize the amount of accident that occur on the road every day. First, obey the traffic law that states make prospective drivers pass a test before they get behind the wheel for a reason, automobiles can be very dangerous if they are not operated properly and according to the laws of the road. Speed limits, traffic signs and all other traffic signals should be closely adhered to every single time someone gets behind the wheel. Traffic laws are put in place to product all the travelers who are sharing space on the road. Following traffic signals is an option, those drivers that treat it as such are usually the ones you see pulled over to the side of the road being handed a violation paper by officers. Nigeria Government has been trying all their effort to reduce the rate of road accidents in the country. Investigation have shown that many factors contributed to road accident which are over speeding, bad roads, careless driving by the drivers and faulty vehicles etc.

Categorical Data

Categorical data are data consisting of classification of behavior into number of mutually exclusive responses (Sanni, 2012) and Lawal (2003).

Contingency of Table

Let x and y denote two categorical set of variables, x have r level and y have c levels when we classify subject on both variables there are possible contribution of the classification. The response x and y of a subject randomly chosen from population having probabilities, we display this distribution in a regular table of rectangular one having r rows for the categories x and c columns for the categories y. the cells of the table represent the rc possible outcome, their possibilities are (π_{ij}) where (π_{ij}) denote the probability that (x, y) falls in the cell I row and column j, y when the cell contain frequency counts of column the table is called contingency table as proposed by Karl Pearson (1974). A contingency table having **r** rows and **c** column is referred to as **r** by **c** table the probability distribution (n_{ij}) is the joint distribution x and y. A contingency table May involve three or higher dimensions, which arises when a sample is classified with respect to more than two quantitative variables.

Table 1: General Form of a Two-Dimensional Contingency Table

$variable\ 1$	$1,\ 2,\ .\ .\ .\ .,\ c$ $variable\ 2$	$TOTAL$
1	$n_{11,}\ n_{12,}\ .\ .\ .,\quad n_{1c}$	n_1
2	$n_{21,}n_{22,}.\ .\ .,\quad n_{2c}$	n_2
.	. . . . .	.
.	. . . . .	.
.	. . . . .	.
r	$n_{11,}\ n_{12,}\ .\ .\ .,\quad n_{1c}$	n_r
$TOTAL$	$n_{1,}\ n_{2,}\ .\ .\ .,\quad n_c$	$n_{..}$

where

$$n_i = \sum_j^c 1_{if}^n \qquad (1)$$

$$n_j = \sum_i^r =_{1_{if}^n} \qquad (2)$$

similarly

$$n = \sum_{i=1}^r \sum_{j=1}^c n_{ij} \qquad (3)$$

$$n_{..} = \sum_{i=1}^{r} n_i \times \sum_{j=1}^{c} n_j \quad (4)$$

$n_{..}$ Represents the total number of observations in the sample and is usually denoted by N.

METHODOLOGY

The data used for this research work is secondary data with respect to the six geopolitical zone, state and year. The data were collected from the Federal Road safety Corp in the federal capital territory. It covers the number of incidence of vehicle accidents in Nigeria from 2018 to 2023. However, the method of data analysis adopted for the research is the Log-linear modelling while the R statistics package was adopted for the statistical analysis of data.

Modeling Categorical Data Analysis

The concept of log-linear model analysis in contingency tables is analogous to the concept of variance (ANVOVA) for the continuously distributed factor-response variables. While response observations are assumed to be continuous with underlying normal distributions to in ANOVA, the log-linear analysis assumes that the response observations are counts having Poisson distributions. The log-linear model (Goodman, 1970, 1971, 1972; Bishop et al., 1975; Gilula and Haberman, 1994) formulations are similar to analysis of variance (ANOVA) techniques, except that the models are applied to the natural logarithms to determine the magnitude of interaction between categorical variable. That is common with a statistical model assumed independent observations. The estimated values were of the form.

$$m_{ij} = \frac{x_i x_j}{X}, \qquad j = 1,2 \quad (5)$$

Both under the model of independence of row and column variables and the model of equality of binomial proportions, more generally referred to as a model of homogeneity of proportions. Taking the natural logarithm of both sides of the equation (1) we have

$$\log(m_{ij}) = logx_i + logx_j\, j - logN \quad (6)$$

And thinking in terms $I \times J$ total with I rows and j columns, reveals close similarities of variable notation
Indeed, the addictive form suggests that the parameter M_{ij} be expressed in the form

$$\log(m_{ij}) = \mu + \mu_{1_{(i)}} + \mu^2_{(i)} \quad (7)$$

$$\mu = i \sum_i \sum_j \log m_{ij} \quad (8)$$

$$\mu + \mu 1_{(i)} = i \sum_i \sum_j \log m_{ij} \quad (9)$$

$$\mu + \mu_{2(i)} = i \sum_i \sum_j \log m_{ij} \quad (10)$$

Because $\mu_{1(i)} and\ \mu_{2(j)}$represent derivative from the ground

$$\mu \sum u_{1(i)=} \mu \quad (11)$$

in the case of model homogeneity, proportion $n_1=x_1$ is fixed and thus

$$\log(M_{ij}) = \log(n_{1+}\, n_2) + \log(x_1 + x_2) \quad (12)$$

We think at the interaction term in the independence model, we have

$$\log(M_{ij}) = \mu 1_{(i)} + \mu_{2(j)} + \mu_{12(ij)} \quad (13)$$

$$\sum i = 1_{\mu 12(iy)} = \sum j = 1_{\mu 12(ij)} = 0 \quad (14)$$

Test of Independence

In general, the most important question is whether the qualitative variable forming the table independent of arrow contingency with multinomial are is sampling. The null hypothesis of statistics independence is $H_o = \pi_{ij} = \pi_{i.}\pi_{.j}$ for all i and j. To test H_o, we use the Pearson Chi-square. when we have two or more independent sample with data consisting of frequencies in discrete categories it the comparability of the observed and expected frequencies in two tables o_{ij}(observed). In place of n_j and with e_{ij} expected.

$= n\pi_{ij} = n\pi_{i.}.\pi_{.j}$ place of n_i here e_{ij} is the expected value of o_{ij} under the null hypothesis, usually($\pi_{i..}$) and ($\pi_{.j}$)are known.

Pearson Chi-Square Test

We estimate the expected frequencies by ($e_{ij} = np_i + p_j$), the chi-square statistic is equal.

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Pearson(1900-1992) claimed that replacing e_{ij} by the estimate e_{ij} would appear to affect the distribution of x^2 since there are N = r x c categories for the classification we argued that X^2 would have an asymptotic chi-square distribution with degree of freedom I x j -1 on the contrary, since e_{ij} are determined by estimating ($\pi_{i.}$) and ($\pi_{.j}$) the chi- square distribution has $D.f = (ij-1) - (i-1) - (j-1) = (i-1) = (i-1)(j-1)$ Where o_{ij} is the j_{th} observation in the i_{th} sample e_{ij} is the expected value that can be obtained using
$e_{ij} = \frac{N_{i.} \times N_{.j}}{N}$ Where n_i and n_j are determined in (1), (2) and (3) respectively. Note that $X^2_{(r-1)(c-1)}$ the hypothesis testing of independence is stated as

$H_o = \pi_j = \ vs\ H_1 = \pi_j \ not\ equal\ to\ \pi_i\pi_j$

The decision is taken by considering the fact that if the calculated X^2 is greater than the tabulated $\chi^2, i.e\ X^2 > \chi^2_{1-\alpha,(r-1)(c-1)}$

Table 2: The Chi-Square Test for Independence

	Class 1	**class 2**	**Total**
$\boldsymbol{population\ 1}$	n_{11}	n_{12}	$n_{1.}$
$\boldsymbol{population\ 2}$	$\boldsymbol{n_{21}}$	$\boldsymbol{n_{22}}$	$\boldsymbol{n_{2.}}$
	$\boldsymbol{c_{.1}}$	$\boldsymbol{c_{.2}}$	$\boldsymbol{N = n_1 + n_2}$

Assumption

1. Each sample is random
2. Each sample are mutually independent
3. Each example may be categorized into class 1 or 2

Test Statistics

$$X^2 = \sum_{i=1}\sum_{j=1}\frac{(o_{ij-eij^2})}{e_{ij}}$$

Data Analysis

In this session, we shall analyze the data according to the aims and objectives stated in chapter one.

MODELLING DATA (DEATH OF PERSONS BY ROAD ACCIDENT)

The factors considered are:

I. Geopolitical zones
II. State
III. Year

Table 3: Model for geopolitical zone

Analysis of deviance table
Terms added sequentially (first to last)

	$\boldsymbol{Df}$	$\boldsymbol{Deviance\ Resid.}$	$\boldsymbol{Df}$	$\boldsymbol{Resid, dev}$	$\boldsymbol{P(>\ |chi|)}$
$\boldsymbol{NULL}$			215	67692	
$\boldsymbol{geopol}$	5	14484	210	52808	$< 2.2e-16$

$AIC: 54339S$

Table 4: State model.

Analysis of deviance table
Terms added sequentially (first to last)

	$\boldsymbol{Df}$	$\boldsymbol{Deviance\ Resid.}$	$\boldsymbol{Df}$	$\boldsymbol{Resid, dev}$	$\boldsymbol{P(>\ |chi|)}$
$\boldsymbol{NULL}$			215	67692	
$\boldsymbol{State\ 1}$	5	2658.3	210	65034	$< 2.2e-16$

Table 5: Model for year

Analysis of deviance table
Terms added sequentially (first to last)

	$\boldsymbol{Df}$	$\boldsymbol{Deviance\ Resid.}$	$\boldsymbol{Df}$	$\boldsymbol{Resid, dev}$	$\boldsymbol{P(>\ |chi|)}$
$\boldsymbol{NULL}$			215	67692	
$\boldsymbol{Year\ 1}$	5	12059.0	210	52975	$< 2.2e-16$

AIC: 39642

Table 6: Model of (state (s) and Year (Y))

Analysis of deviance table

Terms added sequentially (first to last)

	Df	$Deviance\ Resid.$	Df	$Resid, dev$	$P(>\|chi\|)$
$NULL$			215	67692	
$State\ 1$	5	2658.3	210	65034	$< 2.2e-16$
$Year\ 1$	5	12059.0	205	52975	$< 2.2e-16$
$Geopol1$	5	14884.2	200	38091	$< 2.2e-16$

AIC: 39642

Table 7: Model of (State, Year, Geopolitical Zone)

Analysis of deviance table

Terms added sequentially (first to last)

	Df	$Deviance\ Resid.$	Df	$Resid, dev$	$P(>\|chi\|)$
$NULL$			215	67692	
$State\ 1$	5	2658.3	210	65034	$< 2.2e-16$
$Year\ 1$	5	12059.0	205	52975	$< 2.2e-16$
$Geopol1$	5	7518.0	180	45457	$< 2.2e-16$

AIC: 47048

Table 8: Model of (State, Year, Geopolitical Zone, State and Year)

Analysis of deviance table

Terms added sequentially (first to last)

	Df	$Deviance\ Resid.$	Df	$Resid, dev$	$P(>\|chi\|)$
$NULL$			215	67692	
$State\ 1$	5	2658.3	210	65034	$< 2.2e-16$
$Year\ 1$	5	12059.0	205	52975	$< 2.2e-16$
$Geopol1$	5	14884.2	200	38091	$< 2.2e-16$
$State1\ year\ 1$	25	7518.0	175	30573	$< 2.2e-16$

Table 9: Model of (State, Year, Geopolitical Zone, State and Year, State and Geopolitical Zone)

Analysis of deviance table

Terms added sequentially (first to last)

	Df	$Deviance\ Resid.$	Df	$Resid, dev$	$P(>\|chi\|)$
$NULL$			215	67692	
$State\ 1$	5	2658.3	210	65034	$< 2.2e-16$
$Year\ 1$	5	12059.0	205	52975	$< 2.2e-16$
$Geopol1$	5	14884.2	200	38091	$< 2.2e-16$
$State1\ year\ 1$	25	7518.0	175	30573	$< 2.2e-16$
$State1\ geopol\ 1$	25	13207.8	150	17365	$< 2.2e-16$

AIC: 19016

Table 10: Model of (S, Y ,G, SY, SG, YG)

$Analysis\ of\ deviance\ table$

$Terms\ added\ sequentially\ (first\ to\ last)$

	Df	$Deviance\ Resid.$	Df	$Resid, dev$	$P(>\|chi\|)$
$NULL$			215	67692	
$State\ 1$	5	2658.3	210	65034	$< 2.2e-16$
$Year\ 1$	5	12059.0	205	52975	$< 2.2e-16$
$Geopol1$	5	14884.2	200	38091	$< 2.2e-16$
$State1\ year\ 1$	25	7518.0	175	30573	$< 2.2e-16$
$State1\ geopol\ 1$	25	13207.8	150	17365	$< 2.2e-16$
$Year\ 1: geopol\ 1$	25	6510.9	125	10854	$< 2.2e-16$

AIC: 1951.1

Table 11: Model of (S, Y, G, SY, SG, YG, SYG)

Analysis of deviance table

Terms added sequentially (first to last)

	Df	$Deviance\ Resid.$	Df	$Resid, dev$	$P(>\ \|chi\|)$
$NULL$			215	67692	
$State$	5	2658.3	210	65034	$< 2.2e-16$
$Year$	5	12059.0	205	52975	$< 2.2e-16$
$Geopol$	5	14884.2	200	38091	$< 2.2e-16$
$State:year$	25	7518.0	175	30573	$< 2.2e-16$
$State:geopol$	25	13207.8	150	17365	$< 2.2e-16$
$Year:geopol$	25	6510.9	125	10854	$< 2.2e-16$
$State:year:geopol$	125	10854.0	0	0	$< 2.2e-16$

$AIC: 1951.1$

Table 12: SUMMARY OF THE MODELS FITTED

Models	*Degree of Freedom*	*G^2 (residual deviance)*	*AIC*	*P value*
Null	215	67692	54339	<0.0001
State	210	65034	66565	<0.0001
Year	210	55633	57164	<0.0001
Geopol	210	52808	54339	<0.0001
S,Y,G	200	38091	39642	<0.0001
S,Y SY	180	45457	47048	<0.0001
S,Y G SY	175	30573	32174	<0.0001
SYG SY, SG	150	17365	19016	<0.0001
S,Y,G SY,SG,YG	125	10854	12555	<0.0001
SYG	0	0	195.1	1

From the label above, the model (S, Y, G SY, SG, and YG SYG) has the smallest AIC (akaike's information criteria). Therefore, model (S, Y, G SY, SG, YG SYG), which is the SATURATED MODEL, according to Akaike and the test of goodness of fit (G^2) is the best model that can fit or explain the data (death of person by road accident).

INTERPRETATION OF FACTORS IN THE CHOSEN MODEL (SATURATED)

The model is given as

$$I(m_{ij}) = \mu + \lambda_i^s + \lambda_j^y\ \lambda_k^g + \lambda_{ij}^{SY} + \lambda_{ik}^{SG} + \lambda_{jk}^{YG} + \lambda_{ijk}^{SYG}$$

I = 1…6 states j= 1…6 years k= 1…6 zones/GP zones

μ is the overall mean of death of persons by road accidents

λ_i^s Is the main effect of state i

λ_j^y Is the main effect of year j

λ_k^g is the main effect of geopolitical zone k

λ_{ij}^{SY} Is the interaction effect ij between states an year

λ_{ik}^{SG} Is the interaction effect ik between geopolitical zone and year

λ_{jk}^{YG} Is the interaction effect jk between geopolitical zone and year

λ_{ijk}^{SYG} Is the interaction effect among state, geopolitical zone and year

TEST OF INDEPENDENCE BETWEEN YEARS AND GEOPOLITICAL ZONES

Table 13: TABLE OF OBSERVED COUNT

GEO–POLITICAL ZONES	2018	2019	2020	2021	2022	2023
South–south	4447	1038	1401	971	1982	795
South-east	1684	634	441	644	437	829
South-west	8377	2223	2753	2106	4332	1803
North-west	3224	1228	1386	1491	2678	2756
North-east	2020	866	899	1115	1290	1523
North-central	2753	1708	1281	2487	1759	1685
TOTAL	22395	7697	8161	8712	12172	9391

$$e_{ij} = \frac{N_i . N_j}{N}$$

Table 14: TABLE OF EXPECTED COUNT

GEO–POLITICAL ZONES	2018	2019	2020	2021	2022	2023
South – south	3466.07	1185.440	1256.902	1357.473	1921.778	1446.339
South-east	1521.82	520.483	551.860	596.017	843.782	635.034
South-west	7038.40	2407.222	2552.337	2756.561	3902.470	2937.017
North-west	4159.99	1422.773	1508.543	1629.248	2306.531	1735.906
North-east	2513.99	859.818	911.650	984.596	1393.894	1049.051
North-central	3804.72	1301.264	1379.709	1490.106	2109.546	1587.65

HYPOTHESIS

$H_o = \pi_{ij} = \pi_{i.}.\pi_{.j}\ vs\ \ H_1 = not\ H_o$

DECISION RULE

Reject $H_o\ if\ X^2$ calculated is greater than $X^2_{(r-1),(c-1),(1-\alpha)}$ OR when p-value is less than α otherwise we do not reject H_σ

TEST STATISTICS

Pearson's chi-squared test
Data: zone
χ^2 = 4375.086, DF =25, p value<2.2e-16

DISCUSSION: We reject H_O
Since p-value (<0.0001) is less than α=0.05, we have to reject the null hypothesis that Y are not independent of the geopolitical zones. Loglinear analysis was performed i.e model selection, to determine the model that can best explain the death (death of persons by road accidents). The analysis showed that the model containing the three factors (geopolitical zones, state and year) appears to be the best. This was achieved by using Alkaike's information criteria and test of the goodness of fit (G^2). The model is of the form

$$\boldsymbol{i(m_{ij}) = \mu + \lambda_i^S + \lambda_j^Y + \lambda_K^G + \lambda_{ij}^{SY} + \lambda_{ik}^{SG} + \lambda_{jk}^{YG} + \lambda_{ijk}^{SYG}}$$

I=1…6states j=1…6 years k=1…6Gzones
For test of independence, the analysis showed that the years are not independent of the geopolitical zones.

CONCLUSION

From the data analyzed, the model that contain the three factors (Geopolitical zones, state and year) is the best to explain the data (death of persons by road accident). The analysis of independence also shows that years are not independent of the geopolitical zones. It is recommended that attention should be paid to road accident prevention to avoid unplanned expenses through supplementary budgeting for the unexpected rise in accident casualties.

REFRENCES

Adekoye, J (2016), Statistical Analysis of Traffic Accident In Nigeria: A Case Study of Federal Road Safety Corps, (Project of the Department of Mathematics and Statistics The School of Natural Science Joseph Ayo Babalola University Ikeji-Arakeji, Osun State.

Bishop Y. M. M., Fienberg E. F., Holland P. W. (1975), Discrete multivariate analysis, MIT Press, Cambridge, Massachusetts.

Gilula, Z., & Haberman, S. J. (1994). Conditional log-linear models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*(426), 645-656.

Lawal B (2003) categorical data analysis. Lawrence Erlbaum associates, Inc. ISBN 0-8058 -4605-0.

Christensen, R. 1997. Log-linear Models and logistic regression, springer-verlag Inc. New York, new York, USA.

Collet, David (2003). Modeling survival data in medical research, second edition, champman & Hall/CRC. ISBN 1-58488-325-1.

Everitt, B.S. 1977. The analysis of contingency tables. John wiley & sons, Inc. New York, New York, USA.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, *65*(329), 226-256.

Goodman, L. A. (1971). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. *Journal of the American statistical Association*, *66*(334), 339-344.

Goodman, L. A. (1972). A general model for the analysis of surveys. *American journal of sociology*, *77*(6), 1035-1086.

Knoke, D and P.J. Burke 1980. Log-linear models, sage publications, Inc. Newberry park, California, USA.

McCullagh, Peter; Nelder, john (1989). Generalized linear models, second edtion. Champman & Hall/CRC. ISBN 0412317605.

Sanni, O.O. (2012) elementary categorical data analysis. STA 449 lecture note, (university of Illorin unpublished.

William B. King (2011). R Tutoorial on Loglinear analysis (coastal Carolina University).