Loglinear modelling of cancer patients cases in Nigeria: An exploratory study approach

Odetunmibi, O. A.¹, Adejumo, A. O.², *, O. O. M. Sanni²

¹Department of Computer and Information Sciences, Covenant University, Otta, Nigeria
²Department of Statistics, University of Ilorin, Ilorin, Nigeria.

Email address
oluwole.odetunmibi@covenantuniversity.edu.ng (Odetunmibi, O. A.), aodejumo@unilorin.edu.ng (Adejumo, A. O.)

To cite this article

Abstract

The spread of cancer disease today worldwide is becoming rampant. Curbing the menace that the disease pose to the humanity has been a thing of concern and has put all hands on deck. Those that are health related workers and non-health related workers. The main objectives of this research work are to: test whether treatment Outcome (O) of cancer patients is dependent of Age (A) and Gender (G) from the two hospitals we have; check for the best model among various models that we have from the two locations; compare the result of the two hospitals in order to be able to conclude whether treatment outcome is the same from the two locations; and use the result of the two hospitals to determine what happen to cancer patients in South West region of Nigeria. We observed that Model (AO:GO which uses Age: Outcome of treatment and Gender: Outcome of treatment) has the minimum Akaike Information Criteria (AIC) value from the two hospitals and therefore is accepted to be the best model. We also observed that Age and Gender are individually independent of treatment outcome of cancer patients from the two hospitals. We can therefore conclude that the treatment outcomes from the two hospitals are the same and this implies that South-West region of Nigeria has the same treatment outcome for cancer patients.

Keywords
Cancer, Loglinear, Exploration, Likelihood Ratio, Treatment Outcome

1. Introduction

Loglinear models are special cases of generalized linear models (GLM, which include regression and anova models) to better treat case of dichotomous and categorical variables. Loglinear analysis deals with association of categorical or grouped data, looking at all levels of possible main and interaction effects, comparing the saturated model with reduced models, with the primary purpose being to find the most parsimonious model which can account for cell frequencies in a table, that is, log linear model analysis is a non dependent procedure for accounting for the distribution of cases in a cross tabulation of categorical variables. Log linear analysis is a type of multi-way frequency analysis (Agresti, 2002; Adejumo, 2005).

Let \( Y = (Y_1, Y_2, Y_3, \ldots, Y_D) \) be categorical variables. Then a rectangular \((N \times D)\) data matrix consisting of \( N \) observations on \( Y \) can be re-arrange as a \( D \)-way contingency table with cells defined by joint levels of the variables. Let \( n_{ij...t} \) denote the frequency for a cell \( Y = (i, j, \ldots, t) \) and \( n = (n_{ij...t}) \). Suppose that \( Y \) has a multinomial distribution with an unknown parameter \( \Theta = \{\Theta_{ij...t}\} \), where \( \Theta_{ij...t} \geq 0 \) and \( \sum \Theta_{ij...t} = 1 \).

The log-linear model is expressed in the form

\[
\log \Theta = X\lambda
\]

Where \( X \) is a \( D \times r \) design matrix and \( \lambda \) is an \( r \times 1 \) parameter vector.

When \( Y \) has a Poisson distribution, the log-linear model is re-written by:

\[
\log M = X\lambda
\]
Where M = (mij⋯t = NΘij⋯t) is the vector of expected frequencies.

The data examined in this research work is focusing on a major disease, called cancer. When there is low oxygen in the cells and low alkalinity, the body’s cells become abnormal and abnormal multiplication of white blood cells always denote toxicity, imbalance, poison and foreign invasion of that which is unnatural or that which should not be in the body. Excessive white blood cell count is almost always indicative of cancer (WHO, 2005).

Cancer is a term used for diseases in which abnormal cells divide without control and are able to invade other tissues. Cancer cell can spread to other parts of the body through the blood and lymph system.

Cancer is not just one disease but many diseases. There are more than 100 different types of cancer. Most cancers are named for the organ or type of cell in which they start.

Although, this research work is not generally based on the causes of cancer, but having an understanding of what cause it as it has been stated above will go a long way in dealing with its spread.

However, this research work is more concerned about some factors that affect the outcome of treatment of cancer patients in Nigeria.

2. Methodology

The log-linear model which is used in analysis of contingency tables is a generalized linear models for counted data and the variety of associations and interaction terms in log-linear models can easily be describe by goodness of fit tests. The methodology of log-linear model for analysis of contingency tables is described in many book such as Bishop et al. (1975); Everitt (1977); Agresti (2002); Adejumo, 2005 and so on.

Consider an I x J contingency table. The log-linear model is represented by:

\[ \log(M_{ij}) = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_{ij}^{12} \]  

For all i and j, under the constraints of the \( \lambda \) term to sum to zero over any subscript such as:

\[ \sum_{i=1}^{I} \lambda_i^1 = 0, \quad \sum_{j=1}^{J} \lambda_j^2 = 0, \quad \sum_{i=1}^{I} \lambda_{ij}^{12} = \sum_{j=1}^{J} \lambda_{ij}^{12} = 0 \]  

The log-linear model given above is called the saturated model or full model for the statistical dependency between \( Y_1 \) and \( Y_2 \).

By analogy with analysis of variance models, we define the overall mean by:

\[ \lambda_0 = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \log M_{ij} \]  

The main effects of \( Y_1 \) and \( Y_2 \) by

\[ \lambda_i^1 = \frac{1}{J} \sum_{j=1}^{J} \log M_{ij} - \lambda_0 \]  

\[ \lambda_j^2 = \frac{1}{I} \sum_{i=1}^{I} \log M_{ij} - \lambda_0 \]  

And the two-factor effect between \( Y_1 \) and \( Y_2 \) by

\[ \lambda_{ij}^{12} = \log M_{ij} = (\lambda_i^1 + \lambda_j^2) - \lambda_0 \]  

Then the main and two-factor effects are determined by the odds and odds ratio, and can be written by:

\[ \lambda_i^1 = \frac{1}{IJ} \sum_{j=1}^{J} \sum_{i=1}^{I} \log \frac{M_{ij}}{M_{i+j}}, \]  

\[ \lambda_j^2 = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \log \frac{M_{ij}}{M_{i+j}}, \]  

And

\[ \lambda_{ij}^{12} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \log \frac{M_{ij} M_{i+j}}{M_{i} M_{j}}, \]  

For the independence model that \( Y_1 \) is statistically independent of \( Y_2 \), the cell probability \( M_{ij} \) can be factorized into the product of marginal probabilities \( M_{i+j} \) and \( M_{i+j} \), that is,

\( M_{ij} = M_{i+j} M_{i+j} \),

Where \( M_{i+j} = \sum_{i=1}^{I} M_{ij} \) and \( M_{i+j} = \sum_{j=1}^{J} M_{ij} \). Then the two-factor effect is:

\[ \lambda_{ij}^{12} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \log \frac{M_{ij} M_{i+j} M_{i+j}}{M_{i} M_{j} M_{i+j}} = 0, \]  

So that the log-linear model for the independence model is expressed by:

\[ \log M_{ij} = \lambda_0 + \lambda_i^1 + \lambda_j^2, \]  

for all i and j (11)

For an I x J x K contingency table, the saturated log-linear model for the contingency table is:

\[ \log M_{ijk} = \lambda_0 + \lambda_i^1 + \lambda_j^2 + \lambda_k^3 + \lambda_{ij}^{12} + \lambda_{jk}^{23} + \lambda_{ik}^{31} + \lambda_{ijk}^{123} \]  

for all i, j, and k.

The \( \lambda \) terms satisfy the constraints:

\[ \sum_{i=1}^{I} \lambda_i^1 = \sum_{j=1}^{J} \lambda_j^2 = \sum_{k=1}^{K} \lambda_k^3 = 0, \]  

\[ \sum_{i=1}^{I} \lambda_{ij}^{12} = \sum_{j=1}^{J} \lambda_{ij}^{12} = \cdots = \sum_{k=1}^{K} \lambda_{ijk}^{123} = 0, \]  

for all i, j, and k.
\[
\sum_{i=1}^{I} \lambda_{0}^{123} = \sum_{j=1}^{J} \lambda_{1}^{123} = \sum_{k=1}^{K} \lambda_{2}^{123} = 0, \quad (15)
\]

We define the \( \lambda \) terms as follows:
The overall mean is given by:
\[
\lambda_{0} = \frac{1}{JK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \log M_{ijk}. \quad (16)
\]

The main effects of \( Y_1 \), \( Y_2 \), and \( Y_3 \) are:
\[
\lambda_{i}^{1} = \frac{1}{IK} \sum_{j=1}^{J} \sum_{k=1}^{K} \log M_{ijk} - \lambda_{0} \quad (17)
\]
\[
\lambda_{j}^{2} = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} \log M_{ijk} - \lambda_{0} \quad (18)
\]
\[
\lambda_{k}^{3} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \log M_{ijk} - \lambda_{0} \quad (19)
\]

Each interaction effect is given by:
\[
\lambda_{ij}^{12} = \frac{1}{K} \sum_{k=1}^{K} \log M_{ijk} - (\lambda_{i}^{1} + \lambda_{j}^{2}) - \lambda_{0} \quad (20)
\]
\[
\lambda_{ik}^{13} = \frac{1}{J} \sum_{j=1}^{J} \log M_{ijk} - (\lambda_{i}^{1} + \lambda_{k}^{3}) - \lambda_{0} \quad (21)
\]
\[
\lambda_{jk}^{23} = \frac{1}{I} \sum_{i=1}^{I} \log M_{ijk} - (\lambda_{j}^{2} + \lambda_{k}^{3}) - \lambda_{0} \quad (22)
\]

and,
\[
\lambda_{ijk}^{123} = \log M_{ijk} - (\lambda_{i}^{1} + \lambda_{j}^{2} + \lambda_{k}^{3}) - (\lambda_{i}^{1} + \lambda_{j}^{2} + \lambda_{k}^{3}) - \lambda_{0} \quad (23)
\]

### 2.1. Model Selection in Categorical Data Analysis

Given a data set, several competing models may be ranked according to their Akaike Information Criteria (AIC) value with the one having the lowest AIC value being the best (See Galindo-Garre and Vermunt, 2005; Vermunt, 1997).

Suppose we denote two variables by \( a_0 \) and \( a_1 \) and suppose they have \( c_0 \) and \( c_1 \) categories respectively, we assumed that each variable \( a_i \) takes values \( a_i = 1, \ldots, c_i \).

Let \( P(a_0, a_1) \) be the probability that the variables \( a_0 \) and \( a_1 \) take values \( l_0 \) and \( l_1 \) respectively, and let \( n(a_0, a_1) \) be the corresponding cell frequency. If we denote the sample size by \( n \) then:
\[
\sum_{a_0=1}^{c_0} \sum_{a_1=1}^{c_1} P(a_0, a_1) = 1 \quad (24)
\]
\[
\sum_{a_0=1}^{c_0} \sum_{a_1=1}^{c_1} n(a_0, a_1) = n
\]

And the multinomial distribution probability mass function can be obtained by:
\[
M \{ (n(a_0, a_1), n(a_0, a_1)) \} \propto \prod_{n=1}^{1} n(a_0, a_1) \prod_{n=1}^{1} n^\lambda \prod_{n=1}^{1} P(a_0, a_1)^{a_0a_1} \quad (25)
\]

When equation above is regarded as a function of \( P(a_0, a_1) \), then it is called likelihood function.

By ignoring \( P(a_0, a_1) \), the log likelihood is given by:
\[
\log \{ P(a_0, a_1) \} = \sum_{n=1}^{c_0} \sum_{n=1}^{c_1} n(a_0, a_1) \log P(a_0, a_1) \quad (26)
\]

The alternative model having no restriction is represented as MODEL (0):
\[
P(a_0, a_1) = b(a_0, a_1) \quad (27)
\]

Where \( b(a_0, a_1) = \sum_{a_0=1}^{c_0} P(a_0, a_1) \) and
\[
b(a_0, a_1) = \sum_{a_0=1}^{c_0} P(a_0, a_1)
\]

Substituting this, it is seen that this model has only one constraint:
\[
\sum_{a_0=1}^{c_0} b(a_0, a_1) = \sum_{a_0=1}^{c_0} b(a_0, a_1) = 1 \quad (28)
\]

The alternative model having no restriction is represented as model (1):
\[
P(a_0, a_1) = b(a_0, a_1) \quad (29)
\]

Substituting this we have:
\[
\sum_{a_0=1}^{c_0} \sum_{a_1=1}^{c_1} b(a_0, a_1) = 1 \quad (30)
\]

It is also seen that this model has only one constraint.

We then find the maximum likelihood estimators (MLE’s) for each model.

By substitution, the log likelihood for Model (0) is obtained as:
\[
a_0(b(a_0, a_1)) = \sum_{a_0=1}^{c_0} \sum_{a_1=1}^{c_1} n(a_0, a_1) b(a_0, a_1) = 1 \quad (31)
\]

If we let \( \log b(a_0, a_1) = A \)

Then:
\[
A = \sum_{a_0=1}^{c_0} \sum_{a_1=1}^{c_1} n(a_0, a_1) b(a_0, a_1) + \sum_{a_0=1}^{c_0} \sum_{a_1=1}^{c_1} n(a_0, a_1) \log b(a_0, a_1)
\]

Using langrange’s multiple \( \lambda_i, \lambda_2, \ldots \), we have
\[ A = \sum_{a_0} n(a_0, \cdot) \log b(a_0, \cdot) + \sum_{a_1} n(\cdot, a_1) \log b(\cdot, a_1) - \lambda_1 \sum_{a_0} b(a_0, \cdot) - 1 - \lambda_2 \sum_{a_1} b(\cdot, a_1) \]

Differentiating with respect to \( b(a_0, \cdot) \) and \( b(\cdot, a_1) \), \( \lambda_1 \) and \( \lambda_2 \) we get:

\[
\frac{\partial A}{\partial b(a_0, \cdot)} = \frac{n(a_0, \cdot)}{b(a_0, \cdot)} - \lambda_1
\]

(33)

\[
\frac{\partial A}{\partial b(\cdot, a_1)} = \frac{n(\cdot, a_1)}{b(\cdot, a_1)} - \lambda_2
\]

(34)

\[
\frac{\partial^2 A}{\partial \lambda_1^2} = \sum_{a_0} b(a_0, \cdot) - 1
\]

(35)

\[
\frac{\partial^2 A}{\partial \lambda_2^2} = \sum_{a_1} b(\cdot, a_1) - 1
\]

(36)

Equating the derivatives to zero and solving, we have:

\[
n(a_0, \cdot) = \lambda_1 b(a_0, \cdot)
\]

\[
n(\cdot, a_1) = \lambda_2 b(\cdot, a_1)
\]

(37)

\[
\sum_{a_0} b(a_0, \cdot) = 1
\]

\[
\sum_{a_1} b(\cdot, a_1) = 1
\]

(38)

Then, from the two constrain the number of free parameter is equal to \((c_0 - 1) + (c_1 - 1)\), and hence the corresponding AIC is given by:

\[
AIC(0) = -2 \sum_{a_0} \sum_{a_1} n(a_0, a_1) \log \frac{n(a_0, a_1) n(a_0, \cdot) n(\cdot, a_1)}{n^2} + 2(c_0 + c_1 - 2)
\]

(43)

Also doing the same thing for MODEL (1) we have AIC to be:

\[
AIC(1) = -2 \sum_{a_0} \sum_{a_1} n(a_0, a_1) \log \frac{n(a_0, a_1) n}{n^2} + 2(c_0 c_1 - 1)
\]

(44)

The data under consideration, the main factors are Age (A) with five levels, Gender (G) has two levels, and Treatment Outcome (O) having two categories as well. The following loglinear models are considered (see Clogg and Eliason, 1987; and Christensen, 1997).

Model 1: When no association exist between the variables (Model A.G.O), we have

\[
\log(m_{ijk}) = \lambda + \lambda_i^A + \lambda_j^G + \lambda_k^O
\]

(45)

Model 2: Two ways association between just two variables (Model AGO), the model is

\[
\log(m_{ijkl}) = \lambda + \lambda_i^A + \lambda_j^G + \lambda_k^O + \lambda_{ij}^{AG}
\]

(46)

Model 3: Two ways association between each variable with Outcome (Model AO.GO), the model is

\[
\log(m_{ijkl}) = \lambda + \lambda_i^A + \lambda_j^G + \lambda_k^O + \lambda_{ik}^{AO} + \lambda_{jk}^{GO}
\]

(47)

Model 4: Two ways association between the variables (Model AG AO.GO), the model is

\[
\log(m_{ijkl}) = \lambda + \lambda_i^A + \lambda_j^G + \lambda_k^O + \lambda_{ij}^{AG} + \lambda_{ik}^{AO} + \lambda_{jk}^{GO}
\]

(48)

Model 5: Three way association between the variables (Model AGO), the model is also called Saturated model and is

\[
\log(m_{ijkl}) = \lambda + \lambda_i^A + \lambda_j^G + \lambda_k^O + \lambda_{ij}^{AG} + \lambda_{ik}^{AO} + \lambda_{jk}^{GO} + \lambda_{ijk}^{AGO}
\]

(49)

Models in Equations (45) to (49) can be fitted using any of the following methods for the Maximum likelihood (ML) estimation; Newton-Raphson, Fishers Scoring or Iterative proportional fitting (IPF). However, in this work, Fishers scoring iterative method will be used (Christensen, 1997; Adejumon, 2005; Adejumon and Arabi, 2013). The equation of the Fisher-scoring algorithm is

\[
(X^T W^{(k)} X) \hat{\beta}^{(k+1)} = X^T W^{(k)} Z^{(k)}
\]

This is the likelihood equation of a generalized linear model with the response vector \( Z^{(k)} \) and a random error covariate matrix \( (W^{(k)})^{-1} \). If rank(\( X \))=p holds, we obtain the ML estimate \( \hat{\lambda} \) as the limit of
$\lambda^{(k+1)} = (X'^tW^{(k)}X)^{-1}X'^tW^{(k)}Z^{(k)}$.

An algorithm was written in R programme to evaluate the process described and the parameters, log-likelihood ratio statistic ($G^2$) as well as the Akaike information criteria (AIC) are obtained for each loglinear models stated above.

### 2.2. Analysis, Results and Discussion

The data on cancer patients who received treatment over the period of four years and nine months from Eko Hospital Lagos and Obafemi Awolowo University Teaching Hospital Ile-Ife (OAUTH).

A total number of 565 patients were recorded from Eko Hospital Lagos while a total of 750 patients were recorded from OAUTH Ile-Ife.

#### Table 1. likelihood ratio test for hierarchical log-linear models

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>df</th>
<th>Delta(Dev)</th>
<th>Delta(df)</th>
<th>P(&gt;DeltaDev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 4</td>
<td>13.587365</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3</td>
<td>10.618090</td>
<td>9</td>
<td>2.969274</td>
<td>4</td>
<td>0.56298</td>
</tr>
<tr>
<td>Model 2</td>
<td>3.916764</td>
<td>8</td>
<td>6.701327</td>
<td>1</td>
<td>0.00963</td>
</tr>
<tr>
<td>Model 1</td>
<td>0.596732</td>
<td>4</td>
<td>3.320032</td>
<td>4</td>
<td>0.50576</td>
</tr>
<tr>
<td>Saturated</td>
<td>0.000000</td>
<td>0</td>
<td>0.596732</td>
<td>4</td>
<td>0.96343</td>
</tr>
</tbody>
</table>

#### Table 2. Aic values for the models

<table>
<thead>
<tr>
<th>MODEL</th>
<th>G$^2$</th>
<th>AIC</th>
<th>P-VALUE</th>
<th>Df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.596732</td>
<td>122.05</td>
<td>0.9634261</td>
<td>4</td>
</tr>
<tr>
<td>Model 2</td>
<td>3.9168</td>
<td>117.37</td>
<td>0.8645541</td>
<td>8</td>
</tr>
<tr>
<td>Model 3</td>
<td>10.618</td>
<td>122.67</td>
<td>0.3027975</td>
<td>9</td>
</tr>
<tr>
<td>Model 4</td>
<td>13.587</td>
<td>127.04</td>
<td>0.405316</td>
<td>13</td>
</tr>
<tr>
<td>Saturated</td>
<td>0.000000</td>
<td>129.45</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Table 3. Lr test for hierarchical log-linear models

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>df</th>
<th>Delta(Dev)</th>
<th>Delta(df)</th>
<th>P(&gt;DeltaDev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 4</td>
<td>23.720853</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3</td>
<td>19.273721</td>
<td>9</td>
<td>4.447132</td>
<td>4</td>
<td>0.34886</td>
</tr>
<tr>
<td>Model 2</td>
<td>8.626868</td>
<td>8</td>
<td>10.646853</td>
<td>1</td>
<td>0.00110</td>
</tr>
<tr>
<td>Model 1</td>
<td>3.644471</td>
<td>4</td>
<td>4.982397</td>
<td>4</td>
<td>0.28911</td>
</tr>
<tr>
<td>Saturated</td>
<td>0.000000</td>
<td>0</td>
<td>3.644471</td>
<td>4</td>
<td>0.45625</td>
</tr>
</tbody>
</table>

#### Table 4. Aic Values for the Models

<table>
<thead>
<tr>
<th>MODEL</th>
<th>G$^2$</th>
<th>AIC</th>
<th>P-VALUE</th>
<th>Df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>3.644471</td>
<td>136.42</td>
<td>0.4562538</td>
<td>4</td>
</tr>
<tr>
<td>Model 2</td>
<td>8.626868</td>
<td>129.05</td>
<td>0.3747434</td>
<td>8</td>
</tr>
<tr>
<td>Model 3</td>
<td>19.273721</td>
<td>137.70</td>
<td>0.16296352</td>
<td>9</td>
</tr>
<tr>
<td>Model 4</td>
<td>23.720853</td>
<td>134.14</td>
<td>0.07379699</td>
<td>13</td>
</tr>
<tr>
<td>Saturated</td>
<td>0.000000</td>
<td>136.42</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2 gives the values of AIC for each model and compared the values of each.

Tables 1 and 3 show the hierarchical models on treatment outcomes of cancer patients from Eko Hospital Lagos and Obafemi Awolowo University Teaching Hospital (OAUTH) Ile-Ife and it also compares the model with the degree of freedom. The differences in model 1 and model 2 measures the distance of the best fit of model 2 from the best fit of model 1. This also applies to other models compared as shown in the tables.

### 3. Conclusion

The results we have from the analysis carried out from the data collected from each hospitals, using both the AICs and Log likelihood Ratios, give us the same result, accepting model AO.GO to be the best model and having each variable to be independent of one another, we can therefore concluded that the treatment outcomes from the two hospitals are the same and also that south-west region of the country has the same treatment outcome for cancer patients.

### References


