ML-Based Outlier Detection in Contingency Tables: Implementation and Comparative Analysis

Praveen Kumar

November 17, 2023

Contents

Abstract

The detection of outliers in contingency tables is a challenging statistical problem, especially given the inherent polarization of cell counts. This work implements a novel algorithm for outlier detection in I × J contingency tables, leveraging machine learning techniques. The algorithm introduces a pivot element to identify markedly deviant cells as outliers, following a two-step confirmatory procedure.

1 Introduction

In recent years, the spotlight on identifying and accommodating outliers in data analysis has grown significantly. Outliers, whether indicative of errors or accurate yet unexpected observations, offer valuable insights into studied phenomena (Barnett and Lewis, 1994). Defining outliers for categorical data, particularly in contingency tables, where cells represent frequency, lacks clarity. Outliers here are loosely described as cell frequencies deviating markedly from expectations, prompting an exploration of the 'markedly deviant' through a pivotal element.

Traditional statistical methods, sensitive to deviations, often focus on independence in I×J contingency tables. Graphical displays like biplots and mosaic plots assist in identifying outlying cells (Friendly, 2000; Beh and Lombardo, 2014).

The objective of this work is to implement a novel algorithm for outlier detection in contingency tables. This algorithm addresses the challenges posed by the polarization of cell counts and leverages machine learning techniques for improved performance.

2 Literature Review

In metric data, measures such as cook's distance, leverage, DFFITS, DFBETA, etc., are commonly used to identify outliers.

Two primary reasons drive the search for outliers: 1) interest in outliers for their own sake and 2) acknowledgment that outliers can influence the results from the rest of the data. The challenge lies in establishing exact criteria for defining outliers in contingency tables, where there is no clarity in the definition for categorical data, as cells represent frequency or counts.

The presence of one or two outlying observations in a sample can significantly distort data analyses. In the analysis of a two-way contingency table, the focus often lies on a hypothesis of independence between two categorical variables or a hypothesis of homogeneity. Despite advanced statistical techniques, classical Pearson's chi-squared statistics remain widely used.

Sangeetha et al. (2013) proposed the reversal pattern of association (RAP) to understand associations between attributes in higher-dimensional tables. Detecting outlying cells becomes crucial when deviations from the overall association pattern occur, potentially biasing inferences. Polarization of cell frequencies and sparseness in I × J tables are major issues in outlier detection.

The structure and nature of cell counts play a vital role in data analysis, ranging from zero to very high values. Basu and Sarkar (1994) explored a general class of goodness-of-fit test statistics to study disparities in controlling outliers and inliers. Various researchers, including Kuhnt (2004), Rapallo (2012), and Kuhnt et al. (2014), have employed different approaches to identify outliers, from the tails of the Poisson distribution to algebraic statistics.

Residual-based techniques, extensively used for outlier detection in contingency tables, lack a standardized cutoff, making them more heuristic in nature (Simonoff, 2003). Graphical displays, such as association plots and mosaic plots, based on independence of row and column variables, have been employed for visualization. Friendly (1994) discussed the patterns of deviations in mosaic displays in terms of residuals from various models for categorical data.

This study contributes to the existing literature by proposing a two-step confirmatory procedure for detecting outliers in two-way contingency tables, addressing challenges such as polarization and leveraging model-based diagnostics.

3 Detailed Proposal

In the proposed detailed procedure for outlier detection, consider N sample observations cross-classified in an $I \times J(= N)$ contingency table, with $Y_k, k = 1, \ldots, N$ assumed to be realizations of random variables. The initial focus revolves around testing the hypothesis of homogeneity or independence based on the sampling scheme . If the null hypothesis is rejected, attention shifts to investigating cell residuals to identify those deviating significantly from others. A cell is designated as an outlier when the observed frequency markedly deviates from the expected frequency under the null model.

Let n_{ij} be the observed cell frequencies of an $I \times J$ table, $N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ be the total frequency, and $T = \frac{N}{k}$, where $k = IJ$, be the pivot element through which markedly deviant cells are identified as the candidate set of outliers, denoted by the subset S. For an $I \times J$ table, calculate the deviations $D_{ij} = |T - n_{ij}|$ and examine the deviations D_{ij} for each row. If any D_{ij} is markedly deviant from neighboring cells, that particular cell is considered discordant and included in the subset S. The steps involved in the confirmatory procedure are as follows:

Step 1: Given an $I \times J$ table, locate the set of candidate outliers S, using $D_{ij} = |T - n_{ij}|$.

Step 2: Fit a Poisson Log-Linear model for the data with S, considering the nature of the data as count. If the model fits well, proceed to Step 3 otherwise, go to Step 4.

Step 3: Examine different types of residuals associated with the model and detect outliers through a boxplot of residuals.

Step 4: Fit a Negative Binomial model and repeat Step 3.

Residual techniques have traditionally been used by researchers to identify outlying cells by considering residuals greater than ± 3. In this heuristic

approach, outliers are identified without consideration for the polarization of cell frequencies and the order of contingency tables. To address this, a box plot of different types of residuals has been employed to identify outlying cells. The diagnostic measures considered include (i) Response residual, (ii) Deviance residual, (iii) Pearson residual, and (iv) Deleted residual.

This procedure offers a systematic approach to identifying outliers under conditions of polarity for varying orders of the table. The following section delves into examining the robustness of the proposed procedure, as envisioned through a simulation study.

Implementation

4 Process Implementation(PseudoCode)

```
Algorithm Step 1: Locating Candidate Outliers

Input:
  - Contingency table matrix, matrix_Dij

Procedure:
  1. Initialize an empty matrix, z_scores, of the same
     size as Contingency table matrix
  2. For each cell (i, j) in matrix_Dij:
     a. Identify the neighboring cells of (i, j).
     b. Calculate the mean and standard deviation of the
        neighboring cells.
     c. Compute the z-score for the cell (i, j) using
        the formula:
        z_scores[i, j] = (matrix_Dij[i, j] -
           mean_neighbors) / std_neighbors
  3. Define a threshold value ( may need to adjust this
     based on the data).
  4. Identify outliers by comparing the absolute values
     of z_scores against the threshold.
  5. Set colors for True (outlier) and False
     (non-outlier) cells using numpy where function.
  6. Create a table-like plot using Matplotlib.
  7. Plot the table with colored cells, where red
     represents outliers and white represents
     non-outliers.
  8. Display the plot.

Output:
  - Visualization of the contingency table with colored
     cells representing Candidate outliers.
```

```
Model Fitting Procedure:

Input:
  - Contingency table matrix

Procedure:

  1. Create an indicator matrix (matrix) for schools,
     periods, and frequency count.
     - For each school and period combination, set the
       corresponding indicator to 1.
     - Populate the matrix with frequency counts.
  3. Create a DataFrame (df) from the indicator matrix
     with appropriate column headers.
  4. Extract independent variable X (all columns except
     'frequency_count') and dependent variable y
     ('frequency_count').
  5. Fit a Poisson regression model:
     - Add a constant term to X (X_poisson).
     - Use the Poisson GLM (Generalized Linear Model) to
       fit the model (poisson_model).
  6. Fit a Negative Binomial regression model:
     - Add a constant term to X (X_neg_binomial).
     - Use the Negative Binomial GLM to fit the model
       (neg_binomial_model).
```

Poisson log linear Model:

$$\log(\lambda) = \beta_0 + \beta_1 \cdot \text{School}_1 \ldots + \beta_7 \cdot \text{School}_7 + \beta_8 \cdot \text{Period}_1 \ldots + \beta_{14} \cdot \text{Period}_8$$

Here, $\log(\lambda)$ is the log of the expected frequency count, and $\beta_0, \beta_1, \ldots, \beta_{14}$ are the regression coefficients associated with the intercept, schools, and periods.

Negative Binomial Regression Model:

$$\log(\lambda) = \beta_0 + \beta_1 \cdot \text{School}_1 \ldots + \beta_7 \cdot \text{School}_7 + \beta_8 \cdot \text{Period}_1 \ldots + \beta_{14} \cdot \text{Period}_8 + \alpha \cdot \log(\lambda)$$

Here, α is the dispersion parameter that accounts for overdispersion in the Negative Binomial distribution.

5 Implementation and Analysis on Some Datasets

5.1 Yick and Lee Data(1998):

Yick and Lee (1998) considered a study on student enrolment of seven community schools conducted in eight different periods of the year from Northern Territory, Australia

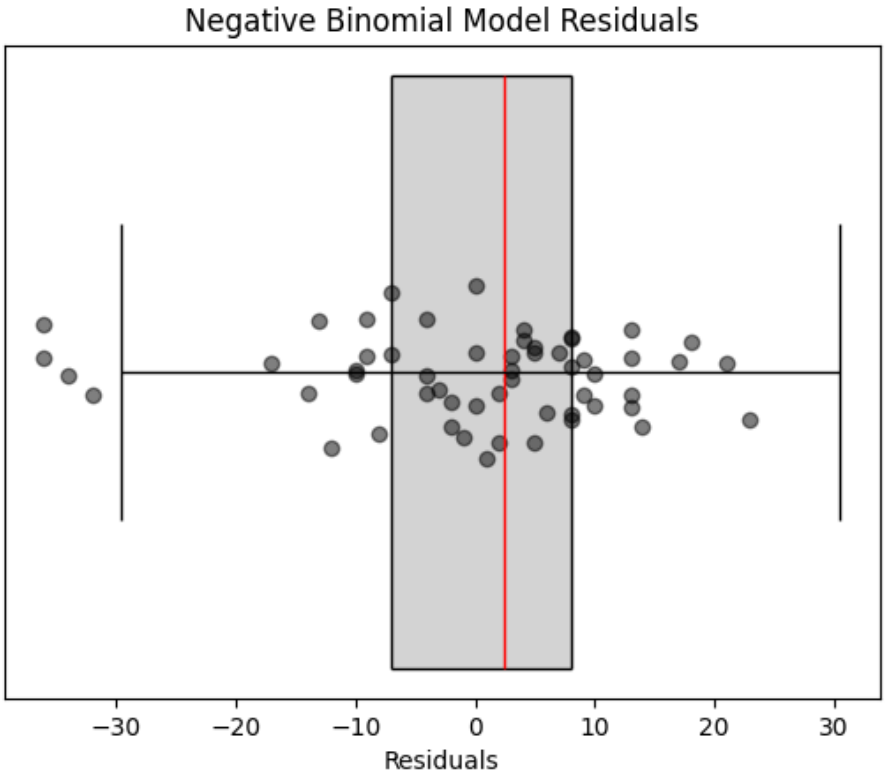
School/Period	1	2	3	4	5	6	7	8
A	93	96	99	99	147	144	87	87
B	138	141	141	201	189	153	135	114
C	42	45	42	48	54	48	45	45
D	63	63	72	66	78	78	82	63
E	60	60	54	51	51	45	39	36
F	174	165	156	156	153	150	156	159
G	78	69	84	78	54	66	78	78

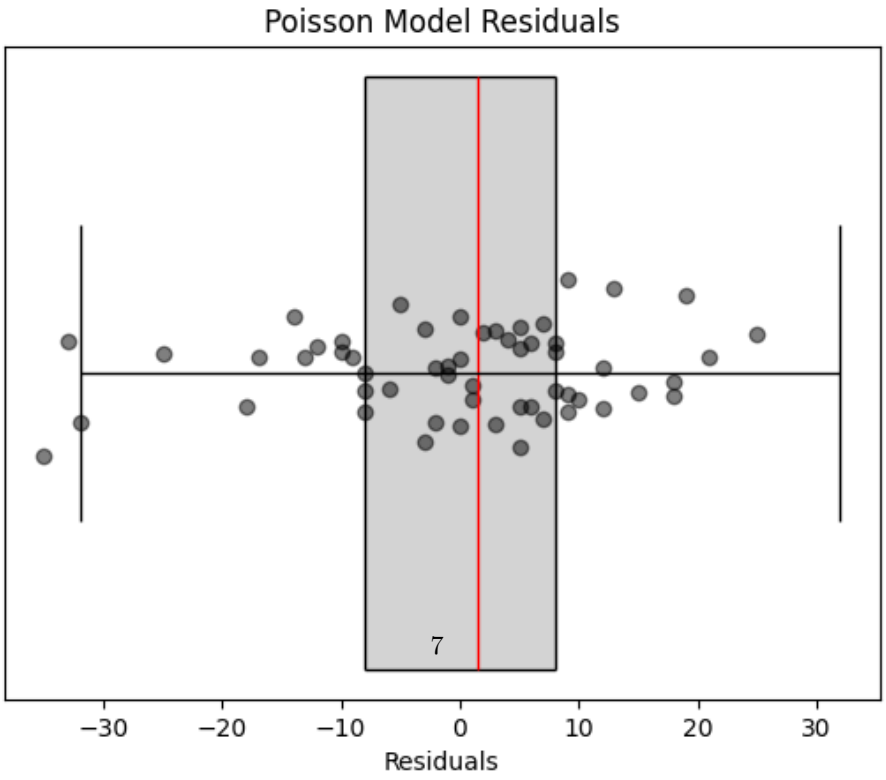
Table 1: School Data for Periods 1-8

5.1.1 Candidate Outliers:

93	96	99	99	147	144	87	87
138	141	141	201	189	153	135	114
42	45	48	54	42	48	54	45
63	72	66	78	78	82	63	63
60	54	51	51	45	39	36	36
174	165	156	156	153	150	156	159
78	69	84	78	54	66	78	78

5.1.2 Model Residuals:

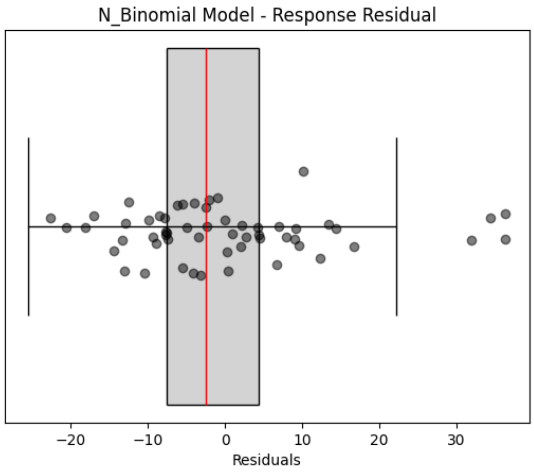




The Negative Binomial Model's Residuals seems to be more tighter and less spread out, suggesting that, this model fits better than poisson log linear for this data set, so we'll take this model for further use.

Now, lets see different Residuals of this model, and get the outlying cells for each of them!

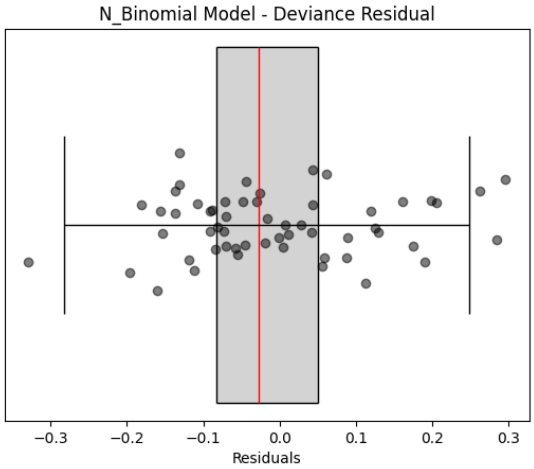
5.1.3 Response Residual



Outliers as per Neg Binomial model Response Residual

Period0	Period1	Period2	Period3	Period4	Period5	Period6	Period7
93	96	99	99	147	144	87	87
138	141	141	201	189	153	135	114
42	45	48	54	42	48	54	45
63	72	66	78	78	82	63	63
60	54	51	51	45	39	36	36
174	165	156	156	153	150	156	159
78	69	84	78	54	66	78	78

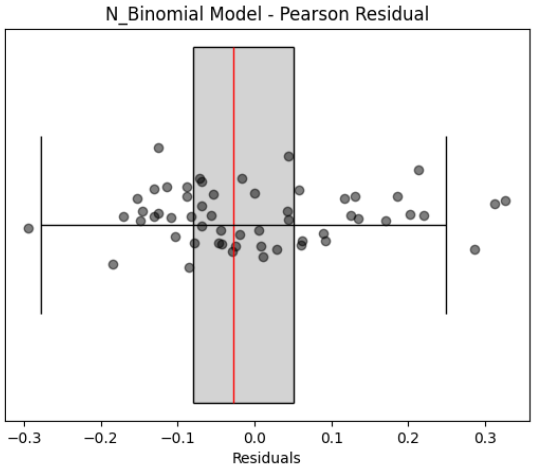
5.1.4 Deviance Residual



Outliers as per Neg Binomial model Deviance Residual

Period0	Period1	Period2	Period3	Period4	Period5	Period6	Period7
93	96	99	99	147	144	87	87
138	141	141	201	189	153	135	114
42	45	48	54	42	48	54	45
63	72	66	78	78	82	63	63
60	54	51	51	45	39	36	36
174	165	156	156	153	150	156	159
78	69	84	78	54	66	78	78

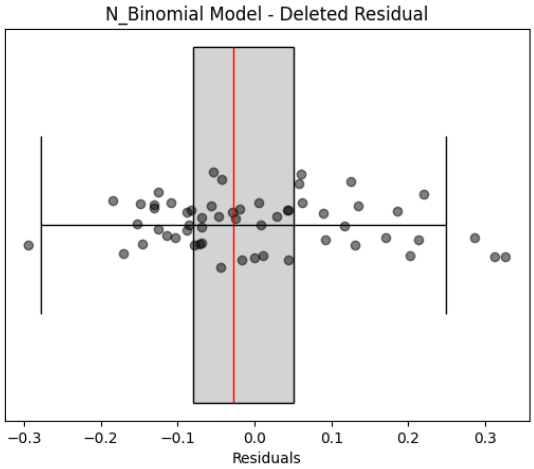
5.1.5 Pearson Residual



Outliers as per Neg Binomial model Pearson Residual

Period0	Period1	Period2	Period3	Period4	Period5	Period6	Period7
93	96	99	99	147	144	87	87
138	141	141	201	189	153	135	114
42	45	48	54	42	48	54	45
63	72	66	78	78	82	63	63
60	54	51	51	45	39	36	36
174	165	156	156	153	150	156	159
78	69	84	78	54	66	78	78

5.1.6 Deleted (Working) Residual



Outliers as per Neg Binomial model Deleted (Working) Residual

Period0	Period1	Period2	Period3	Period4	Period5	Period6	Period7
93	96	99	99	147	144	87	87
138	141	141	201	189	153	135	114
42	45	48	54	42	48	54	45
63	72	66	78	78	82	63	63
60	54	51	51	45	39	36	36
174	165	156	156	153	150	156	159
78	69	84	78	54	66	78	78

In accordance with the procedure outlined in Section (4), deviations from the pivot element led to the identification of the candidate set S = (1, 5), (1, 6), (2, 4), (2, 5) as outliers. Upon further examination using the confirmatory procedure, the Negative Binomial model demonstrated a good fit to the data. The four types of residuals identified (1, 5), (1, 6), (2, 4), and (7, 5) as potential outliers, and their corresponding box plot of residuals is presented in Fig 1. Although the non-anomalous cell (7, 5) was initially identified as an outlier in the residual diagnostic approach, the cells (2, 4) and (2, 5) remained undetected

in the last 3-residuals due to masking during the identification stage. The identification of outliers from the pivot element seems to exhibit resistance to masking effects.

5.2 Bradu and Hawkins (1982):

Similarly, Bradu and Hawkins (1982) considered a study on prevalence rate of men aged 55-64 with hearing levels 16 decibels or more above the audiometric zero for the better ear at 500, 1000, 2000, 3000, 4000, 6000 cycles per second (cps) against the occupational status:

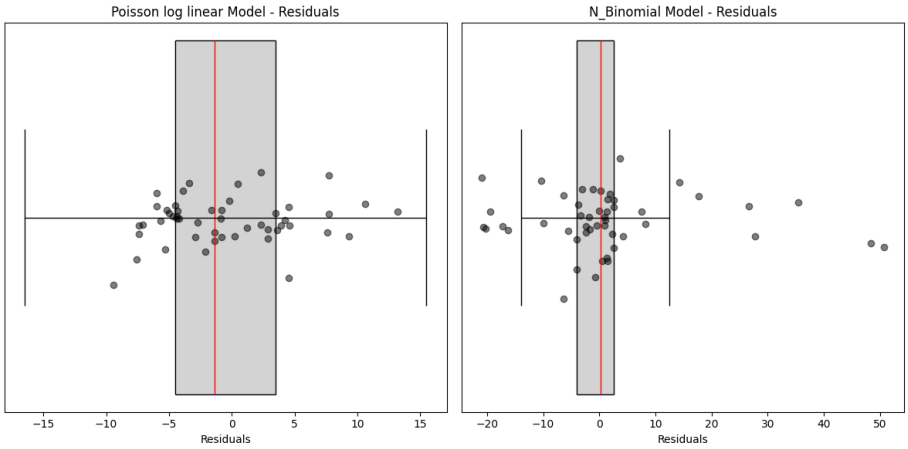
Frequency	Professional, Managerial	Farm	Clerical, Sales	Craftsmen	Operatives	Service	Labourers
Average 500 cps	2.1	6.8	8.4	1.4	14.6	7.9	4.8
1000 cps	1.7	8.1	8.4	1.4	12	3.7	4.5
2000 cps	14.4	14.8	27	30.9	36.5	36.4	31.4
3000 cps	57.4	62.4	37.4	63.3	65.5	65.6	59.8
4000 cps	66.2	81.7	53.3	80.7	79.7	80.8	82.4
6000 cps	75.2	94	74.3	87.9	93.3	87.8	80.5
Normal Speech	4.1	10.2	10.7	5.5	18.1	11.4	6.1

Table 2: Frequency Data for Different Occupational Groups

5.2.1 Candidate Outliers:

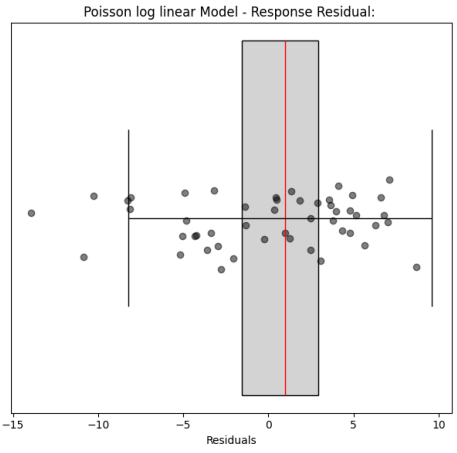
2.1	6.8	8.4	1.4	14.6	7.9	4.8
1.7	8.1	8.4	1.4	12.0	3.7	4.5
14.4	14.8	27.0	30.9	36.5	36.4	31.4
57.4	62.4	37.4	63.3	65.5	65.6	59.8
66.2	81.7	53.3	80.7	79.7	80.8	82.4
75.2	94.0	74.3	87.9	93.3	87.8	80.5
4.1	10.2	10.7	5.5	18.1	11.4	6.1

5.2.2 Model Residuals:



Clearly Poisson fits better.

5.2.3 Response Residual

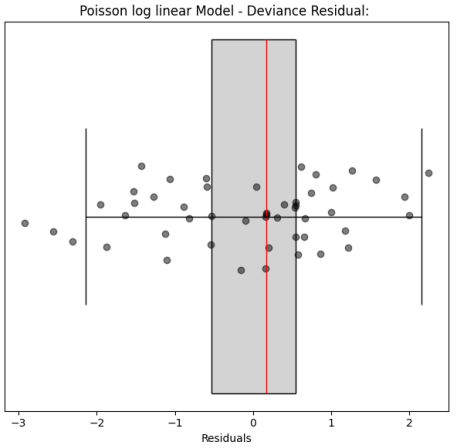


Outliers as per N_Binomial model Response Residual:

Var0	Var1	Var2	Var3	Var4	Var5	Var6
2.1	6.8	8.4	1.4	14.6	7.9	4.8
1.7	8.1	8.4	1.4	12.0	3.7	4.5
14.4	14.8	27.0	30.9	36.5	36.4	31.4
57.4	62.4	37.4	63.3	65.5	65.6	59.8
66.2	81.7	53.3	80.7	79.7	80.8	82.4
75.2	94.0	74.3	87.9	93.3	87.8	80.5
4.1	10.2	10.7	5.5	18.1	11.4	6.1

5.2.4 Deviance Residual

Poisson log linear Model - Deviance Residual:

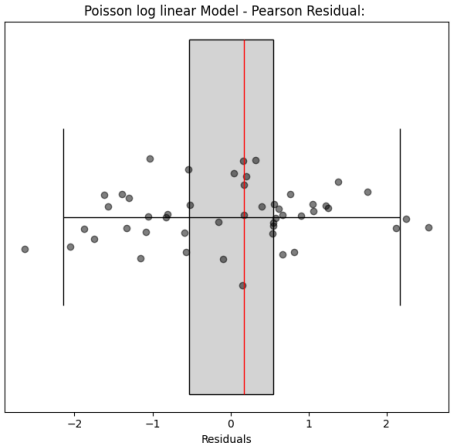


Outliers as per N_Binomial model Deviance Residual:

Var0	Var1	Var2	Var3	Var4	Var5	Var6
2.1	6.8	8.4	1.4	14.6	7.9	4.8
1.7	8.1	8.4	1.4	12.0	3.7	4.5
14.4	14.8	27.0	30.9	36.5	36.4	31.4
57.4	62.4	37.4	63.3	65.5	65.6	59.8
66.2	81.7	53.3	80.7	79.7	80.8	82.4
75.2	94.0	74.3	87.9	93.3	87.8	80.5
4.1	10.2	10.7	5.5	18.1	11.4	6.1

5.2.5 Pearson Residual

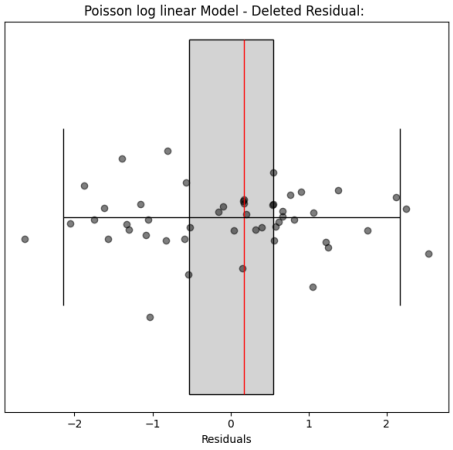
Poisson log linear Model - Pearson Residual:



Outliers as per N_Binomial model Pearson Residual:

Var0	Var1	Var2	Var3	Var4	Var5	Var6
2.1	6.8	8.4	1.4	14.6	7.9	4.8
1.7	8.1	8.4	1.4	12.0	3.7	4.5
14.4	14.8	27.0	30.9	36.5	36.4	31.4
57.4	62.4	37.4	63.3	65.5	65.6	59.8
66.2	81.7	53.3	80.7	79.7	80.8	82.4
75.2	94.0	74.3	87.9	93.3	87.8	80.5
4.1	10.2	10.7	5.5	18.1	11.4	6.1

5.2.6 Deleted (Working) Residual



Outliers as per N_Binomial model Deleted Residual:

Var0	Var1	Var2	Var3	Var4	Var5	Var6
2.1	6.8	8.4	1.4	14.6	7.9	4.8
1.7	8.1	8.4	1.4	12.0	3.7	4.5
14.4	14.8	27.0	30.9	36.5	36.4	31.4
57.4	62.4	37.4	63.3	65.5	65.6	59.8
66.2	81.7	53.3	80.7	79.7	80.8	82.4
75.2	94.0	74.3	87.9	93.3	87.8	80.5
4.1	10.2	10.7	5.5	18.1	11.4	6.1

6 Comparison with Various Methods

In literature, outliers have been identified either using residuals or a suitable cutoff criteria. The following six methods are considered in this comparative study:

1. Haberman (1973) considered residuals greater than 2 as outliers (Method 1).

2. Kuhnt et al. (2014) proposed α-outlier region to detect the outliers, considering cell counts above/below the region as outliers (Method 2).

3. Yick and Lee (1998) proposed perturbation diagnostics to detect outliers (Method 3).

4. Simonoff (1988) considered the cells deviating from the model of independence as outliers (Method 4).

5. Algorithm based on RAP to detect the outliers (Method 5).

6. The Algorithm proposed here (Method 6).

The outcomes of the implementation demonstrate the algorithm's adaptability to different datasets. The examination of various possibilities reveals insights into the algorithm's strengths and potential areas of improvement.

The algorithm was applied to diverse datasets to assess its performance. Results indicate varying effectiveness based on dataset characteristics, highlighting the need for a comprehensive examination.

Data set	Source	Ix J	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
I	Yick and Lee (1998)	7x8	(2,4),(1,5),(2,5), (6,5),(7,5),(1,6), (4,7),(2,8)	(2,4),(1,5), (2,5),(7,5), (1,6)	(1,5),(1,6), (2,4),(2,5)	(1,5),(1,6), (2,4),(7,5)	(2,4),(7,5), (1,6),(4,7)	(1,5),(1,6), (2,4),(2,5)
II	Bradu and Hawkins (1982)	7x7	(3,2),(1,5),(2,5), (7,5)	(1,4),(1,3), (2,3),(3,2), (5,3)	(3,2),(1,5), (2,5),(7,5)	(3,2),(1,5), (2,5),(7,5)	(3,2),(1,5), (7,5),(2,5)	(1,5),(2,5), (3,2)

Table 3: Data Sets and Methods

The results clearly show that method 1 includes all other methods, highlighting the potential of the proposed procedure as a viable alternative for detecting outliers in a two-way contingency table.

7 Drawbacks of the Algorithm

The method has a few challenges worth noting. Firstly, it assumes that the chosen Poisson Log-Linear or Negative Binomial model fits the data perfectly. If the data doesn't quite match what these models expect, the method may not spot outliers accurately.

Another concern is the potential for over fitting, especially when dealing with small amounts of data. The method might get too focused on the specific details of the data.

Additionally, the method relies on deciding what counts as an outlier based on residuals, or what's left over after the model does its work. But here's the catch: the definition of an outlier can change depending on how we look at these leftovers. So, the accuracy of the method in spotting outliers depends on how well we define and stick to these definitions.

Lastly, fitting the Poisson Log-Linear and Negative Binomial models can be computationally demanding, especially with large datasets. The complexity of these models means they require time and a fair amount of computer power. So, using them on big datasets might be slow and resource-intensive in practical terms.

In conclusion, the implemented algorithm showcases promise in outlier detection within contingency tables. The findings provide valuable insights into its adaptability and limitations.