Special Proceedings of 20th Annual Conference of SSCA
held at Pondicherry University, Puducherry during January 29-31, 2018; pp 35-43

Detection of Outliers in Categorical Data using Model Based Diagnostics

T.P. Sripriya and M.R. Srinivasan
Department of Statistics, University of Madras, Chepauk, Chennai

Final Version Received on September 10, 2018

Abstract

Detection of outliers is an important and interesting problem in data analysis. However, detecting outliers in categorical data poses additional difficulties due to polarization of cell counts. Generally, residual based analysis is used to detect outliers in categorical data. The present study considers model based approach to detect outliers in $I \times J$ contingency table. The procedure deals with fitting a Poisson Log-Linear Model for the count data and examine different types of residuals supplemented by boxplot in identifying the outlying cells. The robustness of the model is investigated through a simulation study along with applications to real datasets.

Key words: Log-Linear Model, Diagnostics, Residuals, Boxplot, Outlier(s).

1. Introduction

In every statistical data analysis, surprising observations can occur which deviate strongly from the remaining observations or the assumed model. On the other hand, they might simply be measurement or reporting error. Regardless of the origin of observations, it is commonly dealt as "outliers". Unlike in metric case, there exists no clarity in the definition of outliers for categorical data as the cells are purely frequency or counts of a contingency table. Outliers are only vaguely described as such cell frequencies which deviate markedly from the expected value or cause a significant lack of fit.

Many classical statistical methods are extremely sensitive even to slight deviations from usual distributional assumptions. Until now research on outliers in $I \times J$ contingency tables has been restricted mainly to the independence model [Feinberg (1969), Brown (1974), Fuchs and Kenett (1980), Kotze and Hawkins (1984), Simonoff (1988), Lee and Yick (1999)]. Andersen (1992) proposed cook's distance for contingency tables and examined the studentized residuals for each cell using Goodman RC model. Kuhnt (2004) described a procedure to identify outliers based on the tails of the Poisson distribution and declared a cell as outlier if the actual count falls in the tails of the distribution.

Rapallo (2012) studied the pattern of outliers by fitting log-linear model and tests the goodness of fit to specify the notion of outlier with the use of algebraic statistics. Kuhnt et al. (2014) detected outliers through subsets of cell counts called minimal patterns for the

Corresponding Author: M.R. Srinivasan
E-mail: mrsrin8@gmail.com

independence model viz., OMP, OMPC and OLTCS algorithm. Grizzle, Starmer and Koch (1969) proposed fitting models to contingency tables by fitting a linear regression model to the logarithm of the observed counts using weighted least squares. However, this study presents an alternative approach to detect outliers based on the assumption of model independence.

Residuals can be used to determine potential outliers. Salsas *et al*. (1999) differentiate between residuals resulting from perfect values and those resulting from predicted values. Residual based techniques has been widely used to detect outliers in contingency table (Haberman 1973; Gentleman and Wilk 1975a, b; Bradu and Hawkins 1982; Yick and Lee 1998). Thus, residuals play an important role in detecting outliers in two-way contingency tables, and an extensive review is presented in Kateri (2014). Even though, the residual technique has been widely used by the researchers, no specific limit for choosing the maximum residuals exist and the method is more heuristic in nature (Simonoff 2003).

The structure and nature of cell counts in a contingency table play an important role in the data analysis with the cell counts ranging from zero to very high frequencies (Sangeetha *et al*. 2014). Thus the nature and location of frequency in cells could create polarization posing an additional challenge in the detection of outliers. The relevance of sparseness on summary measure and the sensitivity of analysis in 2×2 tables have been discussed by Subbiah and Srinivasan (2008). The prevailing research on the characteristics of $I \times J$ table with cell frequencies n_{IJ} are: Order of the table $I \times J$, numerical issues (aberration/zero width intervals), polarization of cell frequencies, expected cell frequencies less than 5, zero frequencies and computational complexity.

For an $I \times J$ contingency table, the measures of interest are (i) total frequency (N), (ii) order of the table ($I \times J$), (iii) high cell frequencies, (iv) low frequencies and (v) cells with zero frequencies which in turn cause a problem of polarization and this leads to a major issue in the detection of outliers.

In this paper, we propose a new approach to detect potential outliers in two-way contingency table. It deals with fitting a Poisson log-linear model and the usual diagnostics of the model such as residuals helps to detect the outlying cell in $I \times J$ table.

2. Log-linear Poisson Model

In an analysis of a two-way contingency table, the basic interest will be a hypothesis of independence between two categorical variables or a hypothesis of homogeneity, depending on the sampling scheme. Here, we focus on the Poisson Log-Linear independence model for two dimensional contingency tables, with I & $J \geq 3$.

Consider N paired sample observations that are cross-classified in $I \times J$ contingency table, and assumed to be realizations of random variables Y_j , $j = 1, \ldots, N$, from a log-linear Poisson model. This model may be presented as generalized linear models (Agresti 2002) with structural component $E(Y_j) = exp(x_j'\beta) = m_j$, $j= 1, \ldots, N$, where x_j is the j^{th} column of the full rank design matrix X of the model and β, the unknown parameter vector.

Several diagnostics methods such as residuals received considerable attention in linear models and has rapidly extended to other areas and those emulating in log-linear models have

been proposed (Moolgavkar *et al*. 1985, Tsujitani and Koch 1991, Andersen 1992), as well as graphical methods (Genest and Green 1987, Friendly 1994, 1995, 1999), for detecting outlier cells (Fuchs and Kennet 1980, Simonoff 1988). For contingency tables in particular, we considered a residual diagnostic method to detect the outlying cell in $I \times J$ table based on Poisson log linear model.

Residual techniques have been carried out by many researchers in order to identify the outlying cells in a table by considering residuals greater than ±3 and this range is heuristic in nature. In this heuristic approach, outliers are identified irrespective of the polarization of cell frequencies and order of the tables. To overcome this, the box plot of different types of residuals has been considered to identify the outlying cell. The different diagnostic measures considered are,

(i) Standardized residual
(ii) Working residual
(iii) Response residual
(iv) Deviance residual
(v) Pearson residual
(vi) Deleted residual

Thus this procedure provides a systematic approach of identifying outliers under conditions of polarity for varying order of the table. The following section deals with examining the robustness of proposed procedure as envisaged through a simulation study.

3. Simulation Study

The study of over 100 real time datasets available in the literature has shown that polarization is largely observed in tables of order more than 2×2. However, the study considered log-linear Poisson models of order (3×3), (4×4) and (5×4) with N varying from 50 to 350 for detection of outliers and the high dimensional tables may not be suitable for Poisson model and has been studied elsewhere. The cell frequencies of the tables are assumed to follow Mult (N, ($p_{1,}$ $p_{2,\ldots,}$ p_k)) where $p_i \sim U\ (0,1)$; i=1,2,..,k.. and fitted Poisson log-linear model. The nature and behaviour of different types of residuals with contaminating the cells has been observed in the process of diagnostics for outlier detection. Kuhnt (2014) adopted α-outlier region for contamination purpose to identify the inliers and outliers in a table for the simulations. Here, contamination is restricted to single cell at a time and the number of cells to be contaminated are selected using min{r, c} where r and c be the number of rows and columns respectively. Different level of contamination α (10% to 100% of Row total) are considered and repeated 500 times. We examined the consistency of correctly identified cells among six different residuals in this simulation study.

The four different scenarios described below are carried out through a simulation study performed with R and the results are summarized in Table 1- 4.

1. Generate 500 tables of size 3×3 and the total frequency N ranges from 50 to 100. The results reveals that the standardized, response and deviance residuals shows swamping in detecting outliers and deleted residuals performs well in detecting the outliers.
2. Generate 500 tables of size 3×3 and the total frequency N ranges from 100 to 350. The residual analysis shows that standardized, response, deviance and Pearson residuals

detects the outlying cell with low percentage and deleted residual identified the outliers to a greater level.

3. Generate 500 4×4 tables and the total frequency N ranges from 50 to 100. The results reveals that the all the six residuals performed poorly in detecting the outliers.
4. Finally, simulation is carried out by considering 500 tables of size 5×4 and the total frequency N ranges from 50 to 100. The result showed that the all the six residuals performed poorly in detecting the outliers.

The polarization of the cell counts becomes a major issue in the detection of outliers in $I \times J$ contingency tables. Indeed, the use of residuals under the Poisson log-linear model with the use of boxplot turns out to be a good choice in detecting the outlying cells. The limitation of this simulation study is restricted to smaller tables since Poisson model may not be a suitable for higher dimensional tables and modelling such other $I \times J$ tables for detecting outliers is under progress. Further to simulation, the study explored certain well known data to establish the results of simulation.

Table 1: Number of Outliers Detected in 3×3 with N lying between 50 and 100 (out of 500)

Contamination α (in %)	Residuals					
	Standardized	Working	Response	Deviance	Pearson	Deleted
10	03	05	08	07	06	05
20	10	11	11	10	10	198
30	15	14	13	12	12	210
40	17	16	18	16	17	248
50	21	19	24	19	20	256
60	27	23	200	21	24	267
70	31	27	253	205	27	311
80	35	29	256	245	29	340
90	38	31	256	256	33	364
100	39	34	290	288	36	402

Table 2: Number of Outliers Detected in 3×3 with N lying between 100 and 350 (out of 500)

Contamination α (in %)	Residuals					
	Standardized	Working	Response	Deviance	Pearson	Deleted
10	03	02	03	04	05	06
20	06	05	05	05	08	86
30	10	08	08	06	17	135
40	14	13	11	09	21	176
50	17	17	14	15	29	188
60	21	22	32	27	32	230
70	23	27	88	93	39	255
80	99	32	146	114	87	264
90	150	37	162	153	175	299
100	184	39	195	238	198	368

Table 3: Number of Outliers Detected in 4×4 with *N* lying between 50 and 100 (out of 500)

Contamination α (in %)	Residuals					
	Standardized	Working	Response	Deviance	Pearson	Deleted
10	02	04	08	06	05	08
20	06	09	14	12	09	99
30	09	12	25	17	15	160
40	13	15	29	23	18	196
50	16	18	32	29	32	238
60	20	23	37	56	72	260
70	23	26	46	104	92	274
80	28	77	67	118	118	282
90	31	83	86	122	187	344
100	39	92	93	226	236	362

Table 4: Number of Outliers Detected in 5×4 with *N* lying between 50 and 100 (out of 500)

Contamination α (in %)	Residuals					
	Standardized	Working	Response	Deviance	Pearson	Deleted
10	03	03	04	04	07	08
20	11	12	07	08	15	145
30	13	16	14	17	26	160
40	17	22	25	28	37	182
50	24	26	38	39	42	192
60	32	32	52	50	63	220
70	42	46	140	136	152	242
80	52	51	190	188	194	258
90	66	60	242	232	242	294
100	72	68	249	255	249	302

4. Data Analysis

Yick and Lee (1998) considered the artificial data by Simonoff (1988) in identifying outliers and detected three cells (2, 1), (1, 2) and (1, 3) as outliers and the cell (1, 1) being swamped in the perturbation approach. In our method, the six types of residuals detected the same cells (1, 1), (1, 2), (1, 3) and (2, 1) as outlying cells.

Kotze and Hawkins (1984) generated a 14 x 14 table and detected the outlying cells for another generated table from the 14 x 14 table using some weights for eight cells. In our method, the residuals identified the cells (5, 4), (10, 1) and (12, 7) as outliers. Additionally, working and response residuals identified the cells (3, 1), (5, 1), (6, 1) and (11, 13), (10, 10) as outliers respectively.

Bradu and Gabriel (1978) considered the cotton yield of different varieties at a number of centres and analysed for fitting of row and column models in terms of deviation. In our method,

Poisson model fits the data well and detected the cell (4, 4) as outlier in six types of residuals. The boxplot for all the three datasets are presented in the Appendix.

5. Conclusions

Diagnostics in $I \times J$ contingency table has drawn a great deal of attention to the statisticians for many years. The problem of outliers in two-way table can often be serious. However, there is no general agreement among the statisticians about the detection of outliers due to the polarization of cell frequencies in contingency tables, often ignored in all the studies. Such polarized cells in $I \times J$ contingency tables has been examined through the independence Poisson model. The model based approach is proposed as an alternative identification method of outlying cell in categorical data analysis. It deals with fitting a Poisson log-linear model and the usual diagnostic measures such as residuals supplemented by boxplot is used to identify the exact outlying cells. The stability of our proposed methods towards the identification of outliers is examined through a simulation study. The results have revealed that deleted residuals approach identifies the outliers to a greater extent than compared to other residual methods. Moreover, it is evident that the results provides an idea on impact of polarization in the table, and is found to be useful in detecting outliers. Based on the numerical results, we conclude that the alternative technique as a combination of diagnostic measures and boxplot for residuals could be a viable approach in detecting outlier cells in $I \times J$ contingency tables as compared to other existing methods. The results based on fitting of negative binomial, multinomial and such other generalised linear models for fitting categorical data to detect outliers is under investigation.

Acknowledgements

This work was supported by the University Grants Commission-Major Research Project, New Delhi.

References

Agresti, A. (2002). *Categorical Data Analysis (2nd Edition).* Wiley, New York.

Andersen, E.B. (1992). Diagnostics in categorical data analysis. *Journal of the Royal Statistical Society*, **B 54**, 781 - 791.

Barnett, V.D. and Lewis, T. (1994). *Outliers in Statistical Data (3rd Edition).* Wiley, New York.

Bradu, D. and Gabriel R.K. (1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics*, **20(1)**, 47-68.

Bradu, D. and Hawkins, D.M. (1982). Location of multiple outliers in two-way tables using tetrads. *Technometrics*, **24**, 103-108.

Brown, B.M. (1974). Identification of the sources of significance in two-way contingency tables. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **23**, 405-413.

Fienberg, S.E. (1969). Preliminary graphical analysis and quasi-independence for two-way contingency tables. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **18**, 153-168.

Fuchs, C. and Kenett, R. (1980). A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *Journal of the American Statistical Association*, *Theory and Methods Section*, **75,** 395-398.

Genest, C. and Green, P.E.J. (1987). A graphical display of association in two-way contingency tables. *The Statistician*, **36**, 371-380.

Gentleman, J.F. and Wilk, M.B. (1975a). Detecting outliers in a two-way tables: I statistical behavior of residuals. *Technometrics*, **17**, 1-14.

Gentleman, J.F. and Wilk, M.B. (1975b). Detecting outliers: II. Supplementing the direct analysis of residuals. *Biometrics*, **31**, 387-410.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics*, **25**, 489-504.

Haberman, S.J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, **29**, 205-220.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*. Birkhäuser Basel.

Kotze, T.J.vW. and Hawkins, D.M. (1984). The identification of outliers in two-way contingency tables using 2×2 subtables. *Applied Statistics*, **33**, 215-223.

Kuhnt, S. (2004). Outlier identification procedures for contingency tables using maximum likelihood and L1 estimates. *Scandinavian Journal of Statistics*, **31**, 431–442.

Kuhnt, S., Rapallo, F. and Rehage, A. (2014). Outlier detection in contingency tables based on minimal patterns. *Statistics and Computing*, **24**, 481-491.

Lee, A.H. and Yick, J.S. (1999). A perturbation approach to outlier detection in two-way contingency tables. *Australian & New Zealand Journal of Statistics*, **41(3)**, 305–314.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, **89(425)**, 190-200.

Friendly, M. (1995). Conceptual and visual models for categorical data. *The American Statistician*, **49(2)**, 153-160.

Friendly, M. (1999). Extending mosaic displays: marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, **8(3)**, 373-395.

Moolgavkar, S.H., Lustbader, E.D. and Venzon, D.J. (1985). Assessing the adequacy of the logistic regression model for matched case-control studies. *Statistics in Medicine*, **4(4)**, 425-435.

Rapallo, F. (2012). Outliers and patterns of outliers in contingency tables with algebraic statistics. *Scandinavian Journal of Statistics*, **39(4)**, 784–797.

Salsas, P., Guillen, M. and Alemany, R. (1999). Perfect value and outlier detection in logistic binary choice models. *Communications in Statistics: Theory and Methods*, **26(6)**, 1447–1460.

Sangeetha, U., Subbiah, M., Srinivasan, M.R. and Nandram, B. (2014). Sensitivity analysis of bayes factor for categorical data with emphasis on sparse multinomial data. *Journal of Data Science*, **12**, 339-357.

Simonoff, J.S. (1988). Detecting outlying cells in two-way contingency tables via backwards stepping. *Technometrics*, **30(3)**, 339-345.

Simonoff, J.S. (2003). *Analyzing Categorical Data*. Springer Texts in Statistics, Springer-Verlag New York.

Subbiah, M. and Srinivasan, M.R. (2008). Classification of 2×2 sparse data with zero cells. *Statistics & Probability Letters*, **78**, 3212-3215.

Tsujitani, M. and Koch, G.G. (1991). Residual plots for log odds ratio regression models. *Biometrics*, **47(3)**, 1135-1141.

Yick, J.S. and Lee, A.H. (1998). Unmasking outliers in two-way contingency tables. *Computational Statistics and Data Analysis*, **29**, 69-79.

Appendix

Figure 1: Simonoff Data

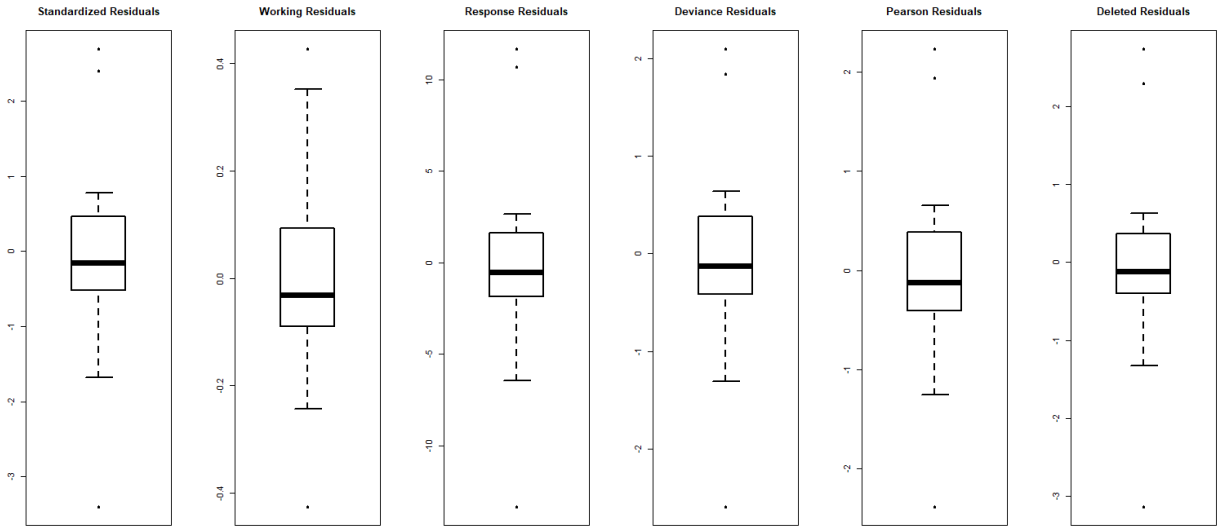

Figure 2: Kotze and Hawkins Data

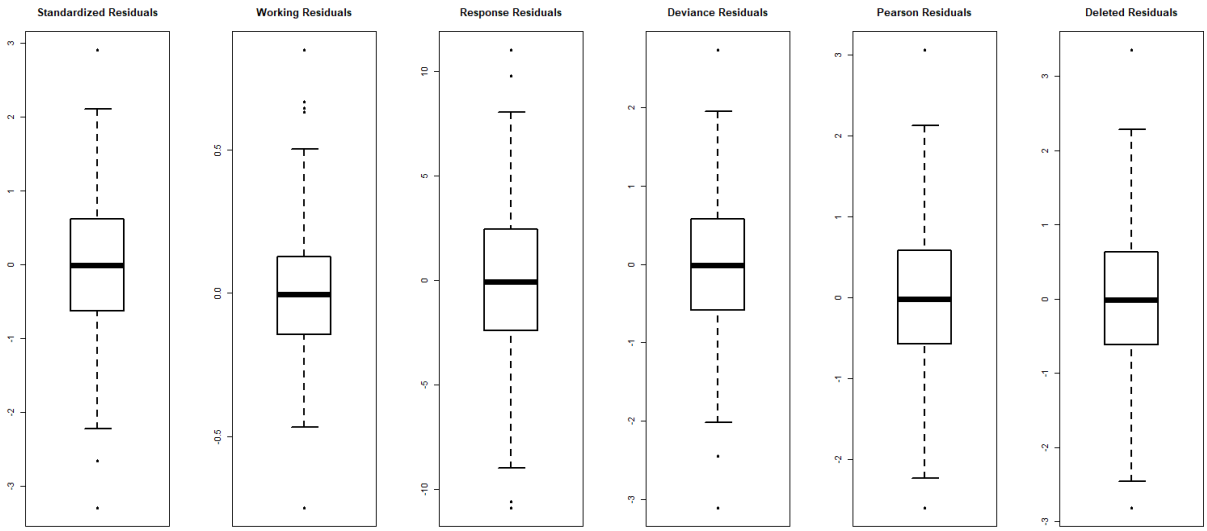

Figure 3: Bradu and Gabriel Data