한국통계학회논문집 제3권 2호
1996년 8월 pp. 69-76

Influence Functions on χ^2 Statistic in Contingency Tables[1)]

Honggie Kim[2)] and Hee-Sook Lee[3)]

Abstract

In a two-way contingency table, the analyst is most interested in the hypotheses of either homogeneity or independence. For testing this as a null hypothesis, Pearson's χ^2 statistic is most commonly used in practice. Once the null hypothesis is rejected, he will further search for cells which caused the rejection of the null hypothesis. For this purpose, so called cell χ^2 components are used. In this paper, we derive the influence function of an observation to the χ^2 statistic, with which cells with high influence can be identified.

1. Introduction

Contingency tables are summarized forms of categorical data arising in many research areas such as social science and humanities. Analyses of these contingency tables were very primitive until 1960's.

Starting from 1960's, theoretical developments such as log-linear model, correspondence analysis have been helping researchers have better understanding of their valuable data. In spite of the excellence of these advanced statistical methods, the theoretical complexities have been obstacles to the researchers who obtain the raw data and want the last bit of information contained in the contingency tables.

In an analysis of a two-way contingency table, the first interest will be a hypothesis of independence between two categorical variables which the rows and columns of contingency tables consist of, or that of homogeneity among rows of the contingency tables, depending on the sampling scheme.

The most popular statistic for testing either of these hypotheses as a null one is Pearson's χ^2 statistic. The theory is well introduced in most elementary statistical texts. Once the null

1 This paper was supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1994.
2 Assistant Professor, Department of Statistics, Chungnam National University, Daejon, 305-764, Korea.
3 Lecturer, Department of Computer & Information Engineering, Kongju National Junior College, Chungnam, 314-060, Korea.

hypothesis is rejected, the next interest of a sophisticated analyst will be investigation of cells which highly contributed to the rejection of the null hypothesis. For this purpose, cell χ^2 components are used.

A cell χ^2 component is the square of the difference between the observed and expected cell frequencies divided by the expected cell frequency. There have been numerous researches on χ^2 statistic and its components. Among them are Irwin (1949), Kimball (1954), Kastenbaum (1960), and Kass (1980).

The idea of influence function is first introduced by Hampel (1974). Cook and Weisberg (1980) used this technique in detection of outliers in regression. Critchley (1985) studied influence in principal component analysis, and Campbell (1978) obtained some interesting results on influence in discriminant analysis. Kim (1992) derived influence functions in correspondence analysis, which has been extended to multiple correspondence analysis in Kim (1994).

By applying Hampel's idea and treating the χ^2 statistic as a multiplication of matrices, we will derive the influence of an observation to the χ^2 statistic as a function.

2. Influence Functions

Let $N=\{n_{ij}\}$ be an $(I\times J)$ contingency table with n_{i+} $(i=1,\cdots,I)$ being the i^{th} row total, n_{+j} $(j=1,\cdots,J)$ being the j^{th} column total and n being the total frequencies in N. Under the null hypothesis of independence or homogeneity, the expected cell frequency is given by

$$e_{ij}=\frac{n_{i+}\cdot n_{+j}}{n} \quad , \quad i=1,\cdots,I \quad , \quad j=1,\cdots,J \quad ,$$

and the Pearson's χ^2 statistic is then

$$X^2=\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} \quad .$$

The estimated probability matrix P is obtained by dividing the entries of N by n. Let

$$r_i=\frac{n_{i+}}{n} \quad , \quad i=1,\cdots,I \qquad \text{and} \qquad c_j=\frac{n_{+j}}{n} \quad , \quad j=1,\cdots,J$$

be the estimated marginal probabilities. Consider the two vectors $r=\{r_i\}$ and $c=\{c_j\}$. Letting D_r and D_c be diagonal matrices with r and c as their diagonals, the χ^2 statistic is

given by (Greenacre, 1984)

$$X^2 = n\,trace\{D_r^{-1}(P-rc^t)D_c^{-1}(P-rc^t)^t\} \ . \tag{1}$$

We will regard the probability matrix P as known to derive an influence function. When P is in fact an estimated probability matrix, the influence function is called an empirical influence function, which is an estimated influence function. Of course, the latter will be the one we can use in practice.

Define an $(I\times J)$ random matrix Y so that its $(i,j)^{th}$ element is 1 and others are 0 when a randomly chosen subject is classified into i^{th} row category and j^{th} column category. Let Y have a distribution F, which is multinomial $M(1, P)$. Now, we can see that the probability matrix P is a functional evaluated on F. That is,

$$P = EY = \int Y dF \ . \tag{2}$$

Also X^2 given by (1) is a functional evaluated on F, since X^2 is a function of P.

Given i and j, let y_{ij} be an $(I\times J)$ matrix where the $(i,j)^{th}$ element is 1 and the others are 0. That is, y_{ij} is a realization of the random matrix Y. To measure the influence of an observation y_{ij} on $X^2 = T(F)$, we use the influence function (IF) of Hampel (1974) which is defined as :

$$IF(X^2, y_{ij}) = \lim_{\varepsilon\to 0}[\ T(F_\varepsilon) - T(F)]\ /\varepsilon$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_{y_{ij}}$, is a perturbation of F by $\delta_{y_{ij}}$, a measure with point mass one at y_{ij}.

Perturbing F produces a perturbation of P, hence perturbation of X^2. If we let X^2_ε be the perturbation of X^2, the influence of the observation y_{ij} on X^2 can be measured by

$$IF(X^2, y_{ij}) = \lim_{\varepsilon\to 0}[\ X^2_\varepsilon - X^2]\ /\varepsilon \ . \tag{3}$$

To find X^2_ε, we replace P, r, c, D_r^{-1} and D_c^{-1} in (1) with the corresponding perturbations, P_ε , r_ε , c_ε , $(D_r^{-1})_\varepsilon$ and $(D_c^{-1})_\varepsilon$, with the subscript ε meaning a perturbation.

The perturbation of P is the functional evaluated on F_ε as given by (2). That is,

$$\begin{aligned} P_\varepsilon &= \int Y dF_\varepsilon \\ &= \int Y d\,[\ (1-\varepsilon)F+\varepsilon\delta_{y_{ij}}] \\ &= (1-\varepsilon)\int Y dF+\varepsilon\int Y d\delta_{y_{ij}} \\ &= (1-\varepsilon)P+\varepsilon y_{ij}\ . \end{aligned}$$

Then the perturbations of r is

$$\begin{aligned} r_\varepsilon &= P_\varepsilon 1 \\ &= (1-\varepsilon)r+\varepsilon y_i\ , \end{aligned}$$

where y_i is an $(I\times 1)$ unit vector with i^{th} element 1. And

$$\begin{aligned} c_\varepsilon &= P_\varepsilon^t\, 1 \\ &= (1-\varepsilon)c+\varepsilon y_j\ , \end{aligned}$$

where y_j is a $(J\times 1)$ unit vector with j^{th} element 1.

Consequently,

$$P_\varepsilon - r_\varepsilon c_\varepsilon^t = (1-\varepsilon)(P-rc^t)+\varepsilon(y_{ij}+rc^t-ry_j^t-y_ic^t)+O(\varepsilon^2)\ .$$

As in Kim (1992),

$$\begin{aligned} (D_r^{-1})_\varepsilon &= [\ diag(r_\varepsilon)]^{-1} \\ &= (1+\varepsilon)D_r^{-1}-\varepsilon\frac{1}{r_i^2}\,diag(y_i)+O(\varepsilon^2) \end{aligned}$$

and

$$(D_c^{-1})_\varepsilon = (1+\varepsilon)D_c^{-1}-\varepsilon\frac{1}{c_j^2}\,diag(y_j)+O(\varepsilon^2)\ .$$

If we let

$$M = D_r^{-1}(P-rc^t)D_c^{-1}(P-rc^t)^t\ ,$$

the χ^2 statistic given in (1) is

$$X^2 = n\ trace(M)\ .$$

The influence function given by (3) becomes

$$IF(X^2, y_{ij}) = \lim_{\varepsilon \to 0} [\ n\ trace(M_\varepsilon) - n\ trace(M)]\ /\varepsilon \tag{4}$$

where $M_\varepsilon = (D_r^{-1})_\varepsilon (P_\varepsilon - r_\varepsilon c_\varepsilon^t)(D_c^{-1})_\varepsilon (P_\varepsilon - r_\varepsilon c_\varepsilon^t)^t$.

Expanding M_ε gives

$$M_\varepsilon = M + \varepsilon(A_1 + A_2 + A_3 + A_4) + O(\varepsilon^2)\ ,$$

where

$$\begin{aligned}
A_1 &= -\frac{1}{r_i^2}\, diag(y_i)(P - rc^t)D_c^{-1}(P - rc^t)^t \\
A_2 &= D_r^{-1}(y_{ij} + rc^t - ry_j^t - y_i c^t)D_c^{-1}(P - rc^t)^t \\
A_3 &= -\frac{1}{c_j^2}\, D_r^{-1}(P - rc^t)diag(y_j)(P - rc^t)^t \\
A_4 &= D_r^{-1}(P - rc^t)D_c^{-1}(y_{ij}^t + cr^t - y_j r^t - cy_i^t)\ .
\end{aligned}$$

Note that A_1 , A_2 , A_3 and A_4 are all $(I \times I)$ square matrices.
The influence function given by (4) now becomes,

$$\begin{aligned}
IF(X^2, y_{ij}) &= n \lim_{\varepsilon \to 0}[\ trace(M_\varepsilon) - trace(M)]\ /\varepsilon \\
&= n \lim_{\varepsilon \to 0}[\ trace(M_\varepsilon - M)]\ /\varepsilon \\
&= n\ trace(A_1 + A_2 + A_3 + A_4) \\
&= n[\ trace(A_1) + trace(A_2) + trace(A_3) + trace(A_4)]\ .
\end{aligned}$$

Through a careful matrix algebra, we can obtain

$$\begin{aligned}
trace(A_1) &= -\frac{1}{r_i} \sum_{j=1}^{J} \frac{(P_{ij} - r_i c_j)^2}{r_i c_j}\ , \\
trace(A_2) &= trace(A_4) = \frac{P_{ij} - r_i c_j}{r_i c_j}\ , \\
trace(A_3) &= -\frac{1}{c_j} \sum_{i=1}^{I} \frac{(P_{ij} - r_i c_j)^2}{r_i c_j}\ .
\end{aligned}$$

Hence, the influence function will be

$$IF(X^2, y_{ij}) = 2n\frac{P_{ij} - r_i c_j}{r_i c_j} - \frac{n}{r_i} \sum_{j=1}^{J} \frac{(P_{ij} - r_i c_j)^2}{r_i c_j} - \frac{n}{c_j} \sum_{i=1}^{I} \frac{(P_{ij} - r_i c_j)^2}{r_i c_j}\ .$$

This function will measure the instantaneous rate of change in χ^2 statistic when an observation is added to the contingency table. The true change in χ^2 statistic when an observation is added to the contingency table can now be estimated by

$$IF(X^2, y_{ij}) \times \frac{1}{n+1} \quad , \tag{5}$$

since $IF(X^2, y_{ij})$ plays the role of $f'(x)$ and $\frac{1}{n+1}$ plays the role of Δx in differential calculus. For more detail, refer Cook and Weisberg (1982).

3. Numerical Example

Table 1 contains 8×5 contingency table taken from Guttman (1971). It represent 1554 Israeli adults cross-classified according to their types of principal worries (rows), and country of origin (columns). The data are also used by Greenacre (1984) to illustrate correspondence analysis. The χ^2 statistic computed from this contingency table is 120.44.

Table 2 shows the estimated changes in χ^2 statistic computed through (5) along with the true changes, which are obtained recomputing the χ^2 statistic after adding an observation to each cell after cell.

As an example, when we use frequency 62 instead of 61 for the cell (1,1), the χ^2 statistic decreases by 0.0817, which can be estimated by 0.0924.

The product moment correlation between the estimated changes and the true changes is 0.9337. Without two cells, cell (5,3) and cell (5,5), the product moment correlation becomes 0.9881. This proves the adequacy of our influence function as a measure of instantaneous rate of change in χ^2 statistic when an observation is added to the contingency table. If we extend this result to the whole observations in a cell, we may be able to suggest a measure competitive to the widely used cell χ^2 components.

4. Comments

Note that we used the estimated probability matrix. That is, we used empirical influence function. We thank the referee for his/her useful comments to improve this paper.

Table 1. Principal worries of Israeli adults. Description of categories of variable B, country of origin, is given at the foot of the table.

	Country of origin (B)				
Principal worry (A)	1	2	3	4	5
Enlisted relative (1)	61	104	8	22	5
Sabotage (2)	70	117	9	24	7
Military situation (3)	97	218	12	28	14
Political situation (4)	32	118	6	28	7
Economic situation (5)	4	11	1	2	1
Other (6)	81	128	14	52	12
More than one worry (7)	20	42	2	6	0
Personal economics (8)	104	48	14	16	9

1: From Asia or Africa
2: From Europe or America
3: From Israel and their father from Asia or Africa
4: From Israel and their father from Europe or America
5: From Israel and their father from Israel

Table 2. Estimated changes and true changes in χ^2 statistic when an observation is added.

row \ col	1	2	3	4	5
1	-0.0924	0.0040	-0.2168	-0.1951	-0.6731
	-0.0817	0.0086	-0.1164	-0.1598	-0.5423
2	-0.0681	-0.0121	-0.2314	-0.2676	-0.3414
	-0.0588	-0.0079	-0.1438	-0.2363	-0.2337
3	-0.4025	0.2527	-0.5999	-0.8215	0.0267
	-0.3959	0.2545	-0.5443	-0.7996	0.0812
4	-1.0953	0.2979	-0.7135	0.3510	-0.1078
	-1.0780	0.3012	-0.5997	0.3825	0.0155
5	-0.7640	0.1910	0.3317	-0.3237	0.8414
	-0.6160	0.2279	1.4030	0.0684	2.1041
6	-0.2886	-0.3339	0.1510	1.0017	0.2310
	-0.2806	-0.3295	0.2110	1.0160	0.3017
7	-0.2819	0.2583	-0.8164	-0.6806	-2.1466
	-0.2476	0.2678	-0.4967	-0.5659	-1.7222
8	1.1400	-1.4115	0.9965	-1.0065	0.2224
	1.1437	-1.3994	1.0747	-0.9634	0.3341

REFERENCES

[1] Campbell, N.A. (1978). The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics*, Vol. 27, 251-258.

[2] Cook, R.D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression, *Technometris*, Vol. 22, 495-508.

[3] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, London.

[4] Critchley, F. (1985). Influence in principal components analysis, *Biometrika*, Vol. 72, 627-636.

[5] Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*, Academic Press, New York.

[6] Guttman, L. (1974). Measurement as structural theory, *Psychometrika*, Vol. 36, 329-347.

[7] Hampel, F. (1974). The influence curve and its role in robust estimation, *Journal of American Statistical Association*, Vol. 69, 383-393.

[8] Irwin, J.O. (1949). A note on the subdivision of χ^2 into components, *Biometrika*, Vol. 36, 130-134.

[9] Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, Vol. 29, 119-127.

[10] Kastenbaum, M.A. (1960). A note on the additive partitioning of chi-square in contingency tables, *Biometrics*, Vol. 16, 416-422.

[11] Kim, H. (1992). Measures of influence in correspondence analysis, *Journal of Statistical Computation and Simulation*, Vol. 40, 201-217.

[12] Kim, H. (1994). Influence functions in multiple correspondence analysis, *The Korean Journal of Applied Statistics*, Vol. 7, 69-74.

[13] Kimball, A.W. (1954). Short-cut formulars for the exact partitioning of χ^2 in contingency tables, *Biometrics*, Vol. 10, 452-458.