Hacettepe Journal of Mathematics and Statistics
Volume 31 (2002), 83–88

M-ESTIMATORS OF THE UNIFORM ASSOCIATION MODEL IN $R \times C$ CONTINGENCY TABLES

Serpil Aktaş* and Meral Çetin*

Received 05. 06. 2002

Abstract

If all cell counts n_{ij} of a given $R \times C$ contingency table are positive, estimates of the expected frequencies m_{ij} can be found by applying any regression estimator to the logarithm of the observed counts. If an $R \times C$ table contains outlier(s), ordinary least squares estimates will be affected by the outlier(s). Various authors have proposed several robust estimators sensitive to outliers. In this study, robust estimators were applied to an $R \times C$ contingency table with an outlier to obtain the robust parameter estimates instead of the maximum likelihood (ML) estimates, and the results were discussed.

Key Words: Contingency table, Robust regression, Outlier, Uniform association model

1. Introduction

We consider two-dimensional tables with r rows and c columns. If both variables of the two-dimensional table are ordinal, and the scores are assigned to the row and column variables $\{ui\}$ and $\{vj\}$, where $u_1 < u_2 < \cdots < u_r$ and $v_1 < v_2 < \cdots < v_c$, the linear-by-linear association model is defined by Agresti [1] to be,

$$\log m_{ij} = \mu + \alpha_i + \beta_j + \gamma(u_i - \bar{u})(v_j - \bar{v}), \imath = 1, \ldots, r,\ j = 1, \ldots, c \qquad (1)$$

When the scores are $u_i = i$ and $v_j = j$, model (1) is represented by the following *uniform association model*,

$$\log m_{ij} = \mu + \alpha_i + \beta_j + \gamma(u_i v_j),\ i = 1, \ldots, r,\ j = 1, \ldots, c, \qquad (2)$$

where the constrains are $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$.

In model (2), for $\gamma = 0$ we obtain the usual independence model [4]. Since model (2) has one more parameter than the usual independence model, the degree of freedom is $(r-1)(c-1)-1$.

*Hacettepe University, Department of Statistics, Ankara, Turkey.

If all the cell counts n_{ij} of a given table are positive, estimates of the expected frequencies m_{ij} can be found by applying any regression estimator to the logarithm of the observed counts [7]. The explanatory variables then consist of dummy variables. Model (2) can be defined with dummy variables,

$$\log n_{ij} = \mu + \sum_{k=1}^{r-1} \alpha_k I_{ik} + \sum_{l=1}^{c-1} \beta_l J_{lj},\ i = 1, \ldots, r,\ j = 1, \ldots, c, \tag{3}$$

where $I - ik$ and J_{jl} indicate the position of the cell in the table [7]. It is known that outlier(s) affect the results of the analysis [2]. The least squares estimates is very sensitive to outliers, so it is proposed some robust estimates have been proposed for use in the precence of outliers. Robust regression is an alternative to ordinary least squares that can be used when there is evidence that the distribution of the error term is non-normal and/or there are outliers [10]. One class of robust regression estimators is the M-estimators. Huber [5,6] introduced M-estimators that minimizes the objective function of the residual,

$$\min_{\theta} \sum_{i=1}^{n} \rho(e_i), \tag{4}$$

where $\rho(e_i)$ denotes the objective function. The derivative of ρ is denoted by ψ. This may be more convenient when calculating estimators. ψ should be bounded and approximately linear at the origin. If ψ is linear, the M-estimator become the least squares estimator.

Some alternative choices for ψ are given below:

Hubers ψ function [5]:

$$\psi(x) = \begin{cases} k & \text{if } x \geq k \\ x & \text{if } |x| < k \\ -k & \text{if } x \leq -k \end{cases} \tag{5}$$

Tukey's ψ function (biweight),

$$\psi(x) = \begin{cases} 0 & \text{if } |x| > k \\ x(1 - (\frac{x}{k})^2) & \text{if } |x| \leq k \end{cases} \tag{6}$$

Andrew's ψ function (sine),

$$\psi(x) = \begin{cases} \sin(\frac{x}{a}) & \text{if } -\pi a < x < \pi a \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $a = 1.142$ and $k = 1.345$ [8].

In earlier studies, alternative solutions to ML were not mentioned when the $R \times C$ tables included outlier(s). But Hubert [7] proposed a maximal breakdown value of the L_1 estimator in contingency tables. As ML estimators give biased results, we applied the robust regression estimator [9] to an $R \times C$ table to see the influence of the outlier.

2. Numerical Example

Robust estimators and Least Absolute Regression (L_1) were applied to an $R \times C$ contingency table, and parameter estimates were obtained (Çetin and Aktaş, [3]). Both variables in two-way contingency table are ordinal, so we may fit model (2). Age is classified into the six groups $12-14$, $15-19$, $20-24$, $25-34$, $35-54$ and $55+$. Similarly, educational status is classified into the seven groups *illiterate* (1), *literate without diploma* (2), *primary school* (3), *junior high school or equivalent* (4), *high school* (5), *vocational high school* (6), and *universities and other higher educational institutions* (7). The 6×7 contingency table is given below.

Table 1. Unemployment data by educational status and age groups

	Educational Status						
Age group	**1**	**2**	**3**	**4**	**5**	**6**	**7**
12 – 14	1	1	23	2	0	0	0
15 – 19	8	3	112	66	98	43	2
20 – 24	2	0	76	58	165	42	69
25 – 34	4	13	151	49	58	17	54
35 – 54	22	19	161	39	34	18	21
55+	26	4	5	12	2	2	2

Source: SIS Statistical yearbook of Turkey, 1999 [11].

Unemployed persons data, classified by educational status and broad age group, for the year of 1999 in Turkey, are used in this study. We fit regression models to Table 1 and the results are given in Table 2, Table 3 and Table 4. where, a_i indicate the row parameters and b_j the column parameters. Statistically significant parameter estimates are denoted by a star. Standardized residuals are plotted against predicted values for all the estimators given in Figure 1.

3. Conclusion

To see the influence of an outlier, we created an outlier corresponding to the $(3, 5)$ th cell before applying the M-Estimator to the 6×7 unemployment data. Tables 2 and 3 give the parameter estimates and estimated standard errors for the OLS, ML and robust estimates with and without outlier, respectively. In Table 2, all parameter estimates in ML are almost statistically significant, other estimators have less significant estimates. While the age groups $25-34$ and $35-54$ are significant for all estimates, the age group $20-24$ is found to be significant only for Huber and Andrews.

Table 2. Parameter estimates and standard errors of OLS, Huber, Tukey, Andrews and ML without outlier

Parameter Estimates	Ordinary Least Squares	Huber	Tukey	Andrews	Maximum Likelihood
μ	2.4255 ± 2.0037	1.8852 ± 1.8501	2.1797 ± 1.8974	1.6877 ± 1.7731	$5.6306 \pm 0.5196*$
α_1	-2.0775 ± 1.2766	-1.6495 ± 1.1783	-1.9033 ± 1.2095	-1.5792 ± 1.1318	$-3.1317 \pm 0.3966*$
α_2	0.8436 ± 1.0932	1.3900 ± 0.9965	1.0699 ± 1.0277	1.4028 ± 0.9601	-0.0691 ± 0.2813
α_3	1.0520 ± 0.9258	$1.7514 \pm 0.8703 **$	1.3664 ± 0.8898	$1.8806 \pm 0.8449*$	$0.6683 \pm 0.2227*$
α_4	$1.5332 \pm 0.7846 **$	$1.8279 \pm 0.7137*$	$1.6551 \pm 0.7377*$	$1.8460 \pm 0.6862*$	$0.9841 \pm 0.1810*$
α_5	$1.8130 \pm 0.6861*$	$2.0502 \pm 0.6152*$	$1.9004 \pm 0.6374*$	$2.0422 \pm 0.5880*$	$1.3473 \pm 0.1564*$
β_1	-1.0022 ± 1.3505	-0.9776 ± 1.2026	-0.9652 ± 1.2595	-0.8163 ± 1.1569	$-0.1178 \pm 0.0152*$
β_2	-1.4632 ± 1.1905	-1.0761 ± 1.0972	-1.2112 ± 1.1312	-0.7854 ± 1.0590	$-3.3505 \pm 0.3583*$
β_3	1.4423 ± 1.0414	1.4138 ± 0.9508	1.4676 ± 0.9793	$1.5664 \pm 0.9059 **$	$-3.3401 \pm 0.3143*$
β_4	0.7580 ± 0.9087	0.7367 ± 0.8196	0.7658 ± 0.8485	0.8399 ± 0.7800	-0.3146 ± 0.2183
β_5	0.6136 ± 0.8005	0.5520 ± 0.7193	0.5968 ± 0.7520	0.6389 ± 0.6900	$-0.7377 \pm 0.1762*$
β_6	0.0743 ± 0.7279	-0.0436 ± 0.6457	0.0340 ± 0.6741	0.0238 ± 0.6137	0.1254 ± 0.1324
γ	-0.0392 ± 0.0549	-0.0253 ± 0.0496	-0.0332 ± 0.0512	-0.0213 ± 0.0474	$-0.5615 \pm 0.1294*$

$^{*}P < 0.05, \quad ^{**}P < 0.10$

Table 3. Parameter estimates and standard errors of OLS, Huber, Tukey, Andrews and ML with an outlier

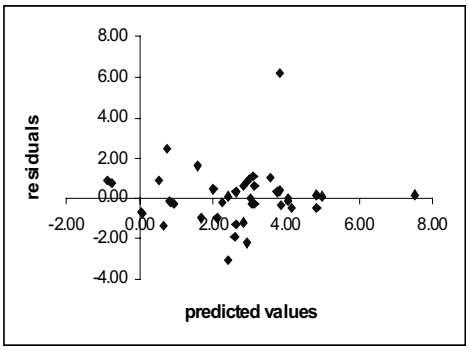
Parameter Estimates	Ordinary Least Squares	Huber	Tukey	Andrews	Maximum Likelihood
μ	2.4812 ± 2.6616	1.8858 ± 2.0177	2.2232 ± 1.8767	1.9865 ± 1.8212	$24.3658 \pm 0.7787*$
α_1	-2.1773 ± 1.6958	-1.6499 ± 1.2851	-1.9091 ± 1.1947	-1.7398 ± 1.1595	$-17.9113 \pm 0.5615*$
α_2	0.7638 ± 1.4522	1.3898 ± 1.0869	1.0604 ± 1.0157	1.2567 ± 0.9823	$-11.4966 \pm 0.4053*$
α_3	1.6907 ± 1.2298	$1.7514 \pm 0.9662 **$	0.9600 ± 0.8937	1.2117 ± 0.8742	$-3.4801 \pm 0.2906*$
α_4	1.4933 ± 1.0423	$1.8279 \pm 0.7784*$	$1.6541 \pm 0.7294*$	$1.7673 \pm 0.7040*$	$-3.9186 \pm 0.2062*$
α_5	$1.7931 \pm 1.9114 **$	$2.0501 \pm 0.6709*$	$1.8963 \pm 0.6300*$	$1.9833 \pm 0.6036*$	$-0.6097 \pm 0.1614*$
β_1	-1.1070 ± 1.7940	-0.9777 ± 1.3224	-0.9528 ± 1.2473	-0.9296 ± 1.1945	$-0.7609 \pm 0.0247*$
β_2	-1.5505 ± 1.5814	-1.0663 ± 1.1966	-1.2654 ± 1.1178	-1.0884 ± 1.0843	$-16.5960 \pm 0.6085*$
β_3	1.3725 ± 1.3833	1.4134 ± 1.0369	1.5013 ± 0.9708	1.5340 ± 0.9358	$-13.4736 \pm 0.4631*$
β_4	0.7056 ± 1.2070	0.7364 ± 0.8938	0.7091 ± 0.8426	0.8105 ± 0.8091	$-8.0211 \pm 0.3179*$
β_5	1.3937 ± 1.0633	0.5518 ± 0.8074	0.2502 ± 0.7724	0.2648 ± 0.7406	$-6.4129 \pm 0.2426*$
β_6	0.0568 ± 0.9670	-0.0441 ± 0.7042	0.0486 ± 0.6710	0.0254 ± 0.6401	$0.4964 \pm 0.1651*$
γ	-0.0442 ± 0.0729	-0.0253 ± 0.0541	-0.0328 ± 0.0506	-0.0271 ± 0.0487	$-2.4301 \pm 0.1409*$

$^{*}P < 0.05, \quad ^{**}P < 0.10$

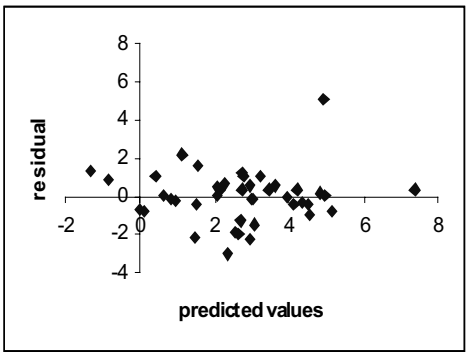
Table 4. Model Results

Without outlier				
	Mean Squares Error	**R^2**	**Coefficient of variation**	**P (Model)**
OLS	1.4789	0.6792	49.3550	0.0002
Huber	1.0925	0.7429	41.1919	0.0001
Tukey	1.1856	0.7181	43.2007	0.0001
Andrews	0.7099	0.7599	32.8498	0.001
With outlier				
	Mean Squares Error	**R^2**	**Coefficient of variation**	**P (Model)**
OLS	2.6095	0.5863	62.6014	0.0033
Huber	1.2994	0.7056	45.3542	0.0001
Tukey	1.1641	0.7227	44.2376	0.0001
Andrews	0.7573	0.7468	35.2856	0.0001

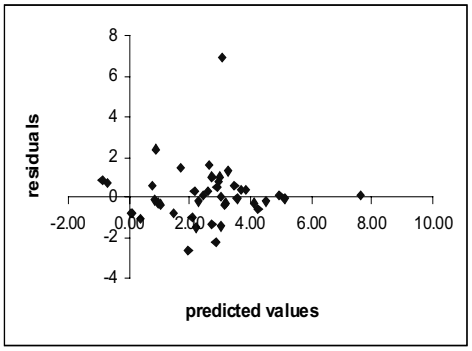
Figure 1. Scatter plots of standardized residuals against predicted values with outlier



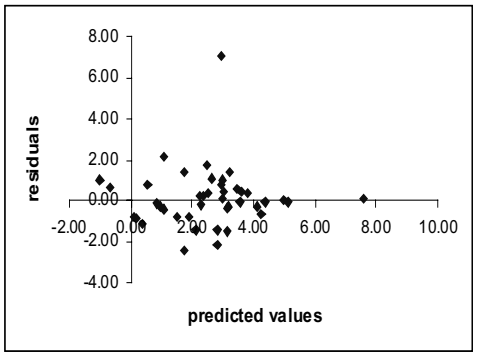
OLS



Huber



Tukey



Andrews

The results were obtained in the presence of an outlier in Table 3. "Primary school" is found to be significant only for Andrews ($P < 0.10$). Unlike the OLS and ML estimators, robust parameter estimates were not affected by the outlier. Standard errors of the OLS estimates are larger when there is an outlier. Note that the highest mean squared error in Table 4 is obtained for the OLS method when an outlier exists. The coefficient of variation is the minimum and R-squared is the maximum for Andrews. In equation (2), γ indicates the association between the row and column variables. In both cases of having outlier and no outlier, we accept the hypothesis that there is no association between education and age, $H_0 : \gamma = 0$.

Standardized residuals are plotted against predicted values given in Figure 1. We can see the influence of outliers for all estimates in Figure 1.

If the data does not contain an outlier, then OLS and robust regression give similar estimates. Robust regression procedures give a weight to each observation and will be able to detect outlier(s); weights close to zero show the outlier(s). Robust estimators detected the $(3, 5)$ th cell as an outlier, giving zero weight in Table 1.

References

[1] Agresti, A. Categorical Data Analysis, John-Wiley, New-York, 1990

[2] Backman, R. C. and Cook R. D. "Outliers", Technometrics **25**, 119–149, 1983.

[3] Çetin, M. and Aktaş, S. $R \times C$ çapraz çizelgelerinde L_1 kestiricisi, İstatistik Sempozyumu 2000, Gazi University, Ankara, 2000.

[4] Goodman, L. A. The analysis of cross-classified data: Independence, quasi-independence and interactions in contingency tables with or without missing entries, JASA **63**, 1091–1131, 1968.

[5] Huber, P. J. Robust regression: Asymptotics, conjectures and Monte Carlo, Ann. Stat. **1**, 799–821, 1973.

[6] Huber, P. J. Robust Statistics, John Wiley & Sons Inc., New York, 308p, 1881.

[7] Hubert, M. The breakdown value of the L_1 estimator in contingency tables, Statistics & Probability Letters **33**, 419–425, 1997.

[8] Hoaglin, D. C., Mosteller F. and Tukey, J. W. Understanding robust and exploratory data analysis, John Wiley & Sons, Inc., New York, 1983.

[9] Rousseeuw, P.J. Robust regression and outlier detection, John-Wiley, New-York, 1984.

[10] Ryan, T. P. Modern Regression Methods, John-Wiley,

[11] SIS, Statistical yearbook in Turkey, 1999. New-York, 1997.