

# The breakdown value of the $L_1$ estimator in contingency tables

Mia Hubert \*

University of Antwerp, Belgium

## Abstract

First we derive the maximal breakdown value of regression equivariant estimators in two-way contingency tables under the loglinear independence model. We then prove that the  $L_1$  estimator achieves this maximal breakdown value. Finally, we illustrate how these results can be generalized towards the uniform association model for contingency tables with ordered categories.

## Keywords

$L_1$  regression, breakdown value, contingency tables, uniform association model.

## 1 Introduction

Contingency tables are widely used to analyze categorical data. Here, we consider two-dimensional  $(r, c)$  tables with  $r$  rows and  $c$  columns. This corresponds to the analysis of two categorical variables with  $r$  and  $c$  categories ( $r, c \geq 2$ ). If the total observed frequency  $\tilde{n}$  is large enough, the individual cell entries  $n_{ij}$  are approximately Poisson distributed with mean  $m_{ij}$ . The standard model (see, e.g. Bishop et al., 1975) assumes independence of these expected frequencies  $m_{ij}$ , i.e.

$$m_{ij} = \frac{m_{i.}m_{.j}}{\tilde{n}} \quad (1)$$

for all  $1 \leq i \leq r, 1 \leq j \leq c$ , where  $m_{i.}$  and  $m_{.j}$  denote the marginal frequencies.

---

\*Address for correspondence: Mia Hubert, Department of Mathematics and Computer Science, University of Antwerp, Universiteitsplein 1, B-2610 Antwerp, Belgium

Equivalently to (1), we can write the loglinear *independence* model

$$\log m_{ij} = \mu + \alpha_i + \beta_j \tag{2}$$

with  $\alpha_r = \beta_c = 0$ . The latter constraints are imposed for identifiability of the parameters. Other constraints are also possible, e.g.  $\sum_{k=1}^r \alpha_k = \sum_{l=1}^c \beta_l = 0$ .

If all cell counts  $n_{ij}$  of a given table are positive, estimates of the expected frequencies  $m_{ij}$  can thus be found by applying any regression estimator to the logarithm of the observed counts. The explanatory variables then consist of dummy variables that indicate the position of the cell in the table. If we introduce two sets of dummy variables  $I_{ik}$  and  $J_{jl}$  and keep the same restrictions on the coefficients, we can rewrite (2) as the linear regression model

$$\log n_{ij} = \mu + \sum_{k=1}^{r-1} \alpha_k I_{ik} + \sum_{l=1}^{c-1} \beta_l J_{jl} + \text{error}_{ij}. \tag{3}$$

To estimate the  $p = r + c - 1$  parameters, we could for instance apply the least squares (LS) estimator to model (3), but it is well-known that LS is very vulnerable to outliers. On the other hand a robust estimator will resist extreme values in the table. In the next section we show that the maximal resistance of any estimator that satisfies a natural equivariance property is determined by the dimensions of the table. In Section 3 we prove that the least absolute deviations ( $L_1$ ) estimator attains this maximal protection. In the last section, we extend these results to the uniform association model.

## 2 Maximal breakdown value for two-way tables

A commonly used measure of robustness of an estimator is the breakdown value  $\varepsilon_n^*$  (see Donoho and Huber 1983, Hampel et al. 1986). It says how many observations of a sample  $Z$  of size  $n$  need to be replaced before the estimate can explode. Formally, the breakdown value of any regression estimator  $T(Z) = T(X, \mathbf{y})$  is

$$\varepsilon_n^* = \varepsilon_n^*(T, Z) = \min \left\{ \frac{m}{n}; \sup_{Z'} \|T(Z')\| = \infty \right\}$$

where  $Z' = (X', \mathbf{y}')$  ranges over all data sets obtained by replacing any  $m$  observations of  $Z = (X, \mathbf{y})$  by arbitrary values.

If we apply a regression estimator to (3), the observations thus correspond to  $\log n_{ij}$  and the sample size to  $rc$ , the number of cells in the table. Note that the total observed

frequency  $\tilde{n} = \sum_{i,j} n_{ij}$  does not occur in (3) anymore. Therefore, it makes sense to permit contamination of  $\log n_{ij}$ , since the sum of their antilogarithms is not fixed.

An upper bound of the breakdown value is formulated in Theorem 1, extending the result of Rousseeuw (1984, page 879). It requires the estimator  $T(Z)$  to be regression equivariant, which means that

$$T(X, \mathbf{y} + X\mathbf{v}) = T(X, \mathbf{y}) + \mathbf{v}$$

for any column vector  $\mathbf{v}$ . For a discussion of this and other types of equivariance, see Rousseeuw and Leroy (1987).

**Theorem 1** *The breakdown value  $\varepsilon_{rc}^*$  of regression equivariant estimators in two-way  $(r, c)$  contingency tables under the loglinear independence model (3) is bounded above by*

$$\varepsilon_{rc}^* \leq \varepsilon_{max}^* = \frac{1}{rc} \left[ \frac{\min(r, c) + 1}{2} \right]. \quad (4)$$

**Proof:**

Let  $p = r + c - 1$  denote the number of columns of a design matrix  $X$  in multiple regression. Mili and Coakley (1996) proved that the upper bound of  $\varepsilon_n^*$  equals

$$\varepsilon_{max}^* = \frac{1}{n} \left[ \frac{n - N + 1}{2} \right],$$

with  $n$  the sample size and  $N$  the maximum number of observations  $(\mathbf{x}_i, y_i)$  whose projections  $\mathbf{x}_i$  lie in a  $p - 1$  dimensional hyperplane through the origin.

W.l.o.g. we assume  $r \leq c$ . Then, all projections of observations that do not belong to the first column ( $X \setminus C_1$ ) lie on a hyperplane. This can easily be seen from equation (3) since the points of  $X \setminus C_1$  satisfy  $J_{j1} = 0$ . This implies  $N \geq rc - r$ .

On the other hand, suppose there is a set (denoted by  $S$ ) of  $rc - r + 1$  observations that lie in a hyperplane through the origin. This means that for a certain non-zero vector  $(\mu, \alpha_1, \dots, \alpha_{r-1}, \alpha_r = 0, \beta_1, \dots, \beta_{c-1}, \beta_c = 0)$  these observations follow model (2) with  $\log m_{ij} = 0$ . This loglinear model (2) has the property that the log odds ratio of the expected frequencies of cells that form a rectangle, is zero. Formally, for all pairs of rows  $(i, i')$  and columns  $(j, j')$ ,

$$\log \left( \frac{m_{ij} m_{i'j'}}{m_{i'j} m_{ij'}} \right) = 0$$

hence

$$\log m_{ij} = \log m_{i'j} + \log m_{ij'} - \log m_{i'j'}. \quad (5)$$

A sample of size  $rc - r + 1$  will always fill at least one full row and column of the two-way table, say the last row  $R_r$  and the last column  $C_c$ . This implies that all remaining cells form a rectangle with vertices of  $R_r \cup C_c$  (see Figure 1), and because of (5), will also have a zero entry. It can finally be verified that this is impossible for a non-zero vector.

					0
	*				<u>0</u>
					0
					0
0	<u>0</u>	0	0	0	<u>0</u>

Figure 1: Every cell forms a rectangle with vertices of  $R_r \cup C_c$

We thus conclude  $N = rc - r = n - \min(r, c)$  from which (4) follows.  $\square$

## Remarks

1. The breakdown value is primarily determined by the shape of the table, rather than by the number of cells. Elongated rectangular tables will be more sensitive to outliers than square tables.
2. The result of Theorem 1 can directly be extended to an  $m$ -way table with dimensions  $(c_1, c_2, \dots, c_m)$ . Then again the complement of any cross-section determines a hyperplane. (A cross-section consists of all the observations belonging to a fixed level of any categorical variable.) If  $c_1 \leq c_2 \leq \dots \leq c_m$  we find

$$\varepsilon^* \leq \frac{1}{c_1 c_2 \dots c_m} \left[ \left( \prod_{j=1}^{m-1} c_j + 1 \right) / 2 \right] \sim \frac{1}{2c_m}. \quad (6)$$

3. The result is also valid for other regression models of which the observations can be seen as entries of an  $m$ -way table. Consider for instance a linear regression model with a continuous response and both continuous and categorical regressors. If the  $m$  categorical variables have  $c_1 \leq \dots \leq c_m$  levels, (6) again applies for any regression equivariant regression estimator.

### 3 The $L_1$ estimator

The least absolute deviations ( $L_1$ ) regression estimator minimizes the sum of the absolute residuals, or

$$\hat{\theta}_{L_1} = \operatorname{argmin}_{\theta} \sum_{i=1}^n |r_i(\theta)|.$$

This estimator performs very badly if leverage points (these are outliers in the  $X$ -space) are present in the data. One can even show that the  $L_1$  fit will pass through a leverage point if it is sufficiently far away from the rest of the data (see, e.g., Bloomfield and Steiger 1983). This implies  $\varepsilon_n^* = 1/n$ .

For the model (3) considered here, it is natural not to allow for leverage points. Since all regressors are binary, it can easily be checked whether extreme values occur in the design matrix  $X$ . Consequently, we define the breakdown value as before, but allow  $Z' = (X, \mathbf{y}')$  only to range over all data sets with  $m$  responses out of  $\mathbf{y}$  being changed. Using this definition, the upper bound (4) still applies (see Müller 1995). Theorem 2 shows that this upper bound is achieved by the  $L_1$  estimator.

**Theorem 2** *The  $L_1$  estimator applied to a two-way contingency table under the loglinear model (2) has breakdown value*

$$\varepsilon_{rc}^* = \frac{1}{rc} \left[ \frac{\min(r, c) + 1}{2} \right] \quad (7)$$

*which is maximal.*

**Proof:**

He et al (1990, Theorem 5.3) showed that in any regression model the  $L_1$  estimator satisfies

$$(m + 1)/n \leq \varepsilon_n^* \leq (m + 2)/n, \quad (8)$$

with  $m$  the largest integer such that for any subset  $M$  of  $N = \{1, 2, \dots, n\}$  of size  $m$ , and for any non-zero regression vector  $\theta$ ,

$$\sum_{i \in N \setminus M} |\mathbf{x}_i \theta| > \frac{1}{2} \sum_{i \in N} |\mathbf{x}_i \theta|$$

which can equivalently be formulated as

$$\sum_{i \in M} |\mathbf{x}_i \theta| < \sum_{i \in N \setminus M} |\mathbf{x}_i \theta|. \quad (9)$$

If we denote  $m' = \lceil (\min(r, c) + 1)/2 \rceil - 1$  we thus have to prove that (9) holds for any subset  $M$  of size  $m'$ . From (8) it then follows that  $\varepsilon_{rc}^* \geq (m' + 1)/rc$ . Theorem 1 provides the inverse inequality, which finishes the proof. Note that one can also use a modified theorem with a nonstrict inequality, as in Bradu (1995).

It remains to prove (9) in this case. Assume w.l.o.g.  $3 \leq r \leq c$  (for  $r = 2$ , equation (7) is trivial). Take any subset  $M$  of size  $m' = \lceil (r + 1)/2 \rceil - 1 = \lfloor (r - 1)/2 \rfloor$  out of  $N = \{(1, 1), (1, 2), \dots, (r, c)\}$ . Using the notations of (2), condition (9) can be written as

$$\sum_{(i,j) \in M} |\mu + \alpha_i + \beta_j| = \sum_{(i,j) \in M} |\log m_{ij}| < \sum_{(i,j) \in N \setminus M} |\log m_{ij}| \quad (10)$$

for all non-zero vectors  $(\mu, \alpha_1, \dots, \alpha_{r-1}, \alpha_r = 0, \beta_1, \dots, \beta_{c-1}, \beta_c = 0)$ . To prove this, we consider for each cell  $(i, j)$  of  $M$  *three* cells with the property that

- they do not belong to  $M$ ,
- together with the cell  $(i, j)$  they form the vertices of a *rectangle*,
- all rectangles are distinct.

Let  $M_2$  denote the collection of these added cells. This construction is pictured in Figure 2 for a (5,6) table. The set  $M$  consists of two cells and is marked by stars, whereas  $M_2$  is formed by the different angle symbols. Of course, lots of other configurations are possible, but since the  $L_1$  estimates of model (2) do not depend on the ordering of the regressors, we may assume that all cells of  $M$  belong to the first column and the  $\lfloor (r - 1)/2 \rfloor$  first rows of the table.

*	1]				
*		2]			
[2		2]			
[1	1]				

Figure 2: Construction of the set  $M$  and  $M_2$  on a (5,6) table

From (5) it follows that

$$|\log m_{ij}| \leq |\log m_{i'j}| + |\log m_{ij'}| + |\log m_{i'j'}|.$$

By construction of  $M_2$ , this yields

$$\sum_M |\log m_{ij}| \leq \sum_{M_2} |\log m_{ij}|. \quad (11)$$

Since (10) requires a strict inequality, we will finally show

$$0 < \sum_{N \setminus (M \cup M_2)} |\log m_{ij}|. \quad (12)$$

Adding (11) and (12) yields (10).

As the total number of rows used in the construction of  $M$  and  $M_2$  is  $2[(r-1)/2] < r \leq c$ , the set  $N \setminus (M \cup M_2)$  will contain at least one complete row and column of the table. If we assume  $\sum_{N \setminus (M \cup M_2)} |\log m_{ij}| = 0$ , it follows from the proof of Theorem 1 that this would imply a zero entry for all cells  $(i, j)$ , which is impossible for non-zero vectors. Relation (12) thus has to be true.

□

## 4 The uniform association model

If both variables of the two-dimensional table are ordinal, the independence model (2) can be refined in order to use the extra information contained in the ordering of the categories. A common way is to assign ordered scores  $\{u_i\}$  and  $\{v_j\}$  to the rows and columns of the table, thus  $u_1 < u_2 < \dots < u_r$  and  $v_1 < v_2 < \dots < v_c$ . They reflect a kind of distance between the categories. The *linear-by-linear association* model (see Agresti 1983) is then defined by

$$\log m_{ij} = \mu + \alpha_i + \beta_j + \gamma(u_i - \bar{u})(v_j - \bar{v}). \quad (13)$$

Again we assume the restrictions  $\alpha_r = \beta_c = 0$ . This model contains an extra linear term with parameter  $\gamma$  and the row and column scores as covariates.

If we would allow the scores  $\{u_i\}$  and  $\{v_j\}$  to take on arbitrary values, model (13) could possibly contain leverage points. As we already mentioned in the previous section, the  $L_1$  estimator cannot cope with this kind of outliers. So, intuitively, a positive breakdown value

for the  $L_1$  estimator will only be attained for scores that lie in a certain bounded subspace of  $\mathbb{R}^r$ , resp.  $\mathbb{R}^c$ . A special submodel of (13) satisfying this condition is the *uniform association* model (Goodman 1979) defined by setting

$$u_i = i \text{ and } v_j = j. \quad (14)$$

Since the scores are fixed for a given table, we can again derive the breakdown value under contamination of the response only. In the next theorem we show that the upper bound  $\varepsilon_{max}^*$  derived in Theorem 1 remains valid (even though model (13) has one more parameter than model (2)), and that this upper bound is also reached by the  $L_1$  estimator.

**Theorem 3** *The breakdown value of the  $L_1$  estimator under the uniform association model applied to a two-dimensional  $(r, c)$  table with dimensions satisfying  $3 \leq r$  and  $4 \leq c$  equals (7) which is maximal for any regression equivariant estimator.*

**Proof:**

Assume w.l.o.g.  $r \leq c$  and  $\bar{u} = \bar{v} = 0$ . Moreover we consider  $\gamma \neq 0$ ; otherwise the situation is reduced to the loglinear model (2), for which we already proved this proposition.

To show the upper bound  $\varepsilon_{max}^*$  given by (4) we can use the same reasoning as in the proof of Theorem 1 (after changing the number of columns of  $X$  into  $p = r + c$ ). Since all points of  $X \setminus C_1$  lie on a hyperplane, we again have  $N \geq rc - r$ . Analogously to (5), the uniform association model has the property that the log odds ratio of the expected frequencies of cells that form a rectangle is proportional to  $\gamma$ , in the sense that

$$\log \left( \frac{m_{ij}m_{i'j'}}{m_{i'j}m_{ij'}} \right) = \gamma(i - i')(j - j') \quad (15)$$

for all pairs of rows  $(i, i')$  and columns  $(j, j')$ . If we now take a sample  $S$  of size  $rc - r + 1$  it will certainly contain four observations that form the vertices of a rectangle, e.g. defined by the indexes  $(i_0, i_1)$ , and  $(j_0, j_1)$ . If we assume  $\log m_{ij} = 0$  for all  $(i, j) \in S$  and apply equation (15) to the pairs  $(i_0, i_1)$  and  $(j_0, j_1)$  we get  $\gamma = 0$ , a contradiction.

The proof of the breakdown value (7) of the  $L_1$  estimator is very similar to Theorem 2, so we will use the same notations. Take any subset of cells  $M$  of size  $m' = [(r + 1)/2] - 1$ . Condition (9) applied to the uniform association model becomes

$$\sum_{(i,j) \in M} |\mu + \alpha_i + \beta_j + \gamma u_i v_j| = \sum_{(i,j) \in M} |\log m_{ij}| < \sum_{(i,j) \in N \setminus M} |\log m_{ij}|$$

for all non-zero vectors  $(\mu, \alpha_1, \dots, \alpha_{r-1}, \alpha_r = 0, \beta_1, \dots, \beta_{c-1}, \beta_c = 0, \gamma)$ . Now consider for each cell  $(i, j)$  of  $M$  seven cells such that together with the cell  $(i, j)$  they form the vertices of two rectangles with the *same length and width*. As in Theorem 2, none of the added cells (denoted by the set  $M_2$ ) may belong to  $M$ , and all formed rectangles should be distinct.

From (15) it follows that the log odds ratios formed by two rectangles of the same form are equal. By construction of  $M_2$ , this again yields relation (11). Finally we will prove that we can always construct  $M \cup M_2$  such that also relation (12) holds for all  $\gamma \neq 0$ .

Consider  $3 \leq r$  and  $6 \leq c$ . Denote  $R_M$  (resp.  $C_M$ ) the set of rows (resp. columns) to which at least one cell of  $M$  belongs. Further assume  $\#C_M \leq \#R_M$  (otherwise an analogous construction is possible). Choose for each cell of  $M$  a different row, not in  $R_M$ ; and use these rows only to construct  $M_2$ . The number of columns required in the construction of  $M \cup M_2$  will then be at least  $4 \leq c-1$  and at most four times the maximum number of cells in  $M$  that belong to the same column. (Here, we will not go into details how to construct rectangles of equal form, since this requires a quite tedious analysis.) Some calculations reveal that  $N \setminus (M \cup M_2)$  will now contain four observations that form the vertices of a rectangle. As showed earlier, this implies that the equation  $\sum_{N \setminus (M \cup M_2)} |\log m_{ij}| = 0$  can only be valid for  $\gamma = 0$ .

The remaining smaller tables have to be analyzed individually. Again we will skip these calculations, since, although by different arguments, they all confirm statement (12).  $\square$

### Remark

We end this section by demonstrating that the breakdown value of  $L_1$  at a (3,3) table equals  $1/9$ , which may be compared to the upper bound  $2/9$  provided by Theorem 3. Take  $M = \{(1, 1)\}$  and  $\theta = (2, 2, -1, -1, 2, -1, -1, 3)$ . (Note that  $\theta$  satisfies  $\sum \alpha_i = \sum \beta_j = 0$ .) With  $\bar{u} = \bar{v} = 0$ , we get  $|\log m_{11}| = 9 = \sum_{N \setminus M} |\log m_{ij}|$ . Ellis and Morgenthaler (1992) showed that this equality implies the lowest possible breakdown value  $1/9$ .

## References

- Agresti, A. (1983), A Survey of strategies for modeling cross-classifications having ordered variables, *J. Amer. Statist. Assoc.* **78**, 184-198.
- Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975), *Discrete Multivariate Analysis*,

(Mass.: MIT Press, Cambridge).

Bloomfield, P., and Steiger, P. (1983), *Least Absolute Deviations: Theory, Applications, and Algorithms*, (Birkhauser, Boston).

Bradu, D. (1995), Identification of outliers by means of  $L_1$  regression: safe and unsafe configurations, Technical Report 95/3, University of South Africa (Pretoria, South Africa).

Donoho, D.L., and Huber, P.J. (1983), The notion of breakdown point, in: P. Bickel, K. Doksum, and J.L. Hodges, Jr., eds., *A Festschrift for Erich Lehmann*, (Wadsworth, California).

Ellis, S.P., and Morgenthaler, S. (1992), Leverage and breakdown in  $L_1$  regression, *J. Amer. Statist. Assoc.* **87**, 143-148.

Goodman, L.A. (1979), Simple models for the analysis of association in cross-classifications having ordered categories, *J. Amer. Statist. Assoc.* **76**, 320-334.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics: the Approach based on Influence Functions*, (John Wiley, New York).

He, X., Jurečková, J., Koenker, R., Portnoy, S. (1990), Tail behavior of regression estimators and their breakdown points, *Econometrica* **58**, 1195-1214.

Mili, L., and Coakley, C.W. (1996), Robust estimation in structured linear regression, *Annals of Statistics*, to appear.

Müller, C.H. (1995), Breakdown points for designed experiments, *J. Statist. Plann. Inference* **45**, 413-427.

Rousseeuw, P.J. (1984), Least median of squares regression, *J. Amer. Statist. Assoc.* **79**, 871-880.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, (John Wiley, New York).