A Robust Approach to Categorical Data Analysis

Karen V. SHANE and Jeffrey S. SIMONOFF

Categorical data analysis is typically performed by fitting models to the observed counts in a contingency table using maximum likelihood. An inherent problem with maximum likelihood fits is their sensitivity to outlier cells, ones whose counts are not consistent with the presupposed model. Robust alternatives to maximum likelihood estimation, including least median of chi-squared residuals, least median of weighted squared residuals, and analogous methods using least trimmed functions, are proposed in this article. Equivariance and breakdown properties are discussed. Monte Carlo simulation results and three real examples are used to illustrate the properties of the estimators in practice. In particular, whereas the maximum likelihood estimates break down in the presence of outlying cells, the robust estimators do not as long as the contamination does not exceed the breakdown point. The proposed estimators perform similarly in the simulations; they are competitive with median polish when fitting independence, and generalize easily to other, more complex, models.

Key Words: Breakdown point; Contingency table; Least median of squares; Least trimmed squares.

1. INTRODUCTION

In recent years, a great deal of attention has been paid to the accommodation and identification of unusual observations (outliers) in data. Robust estimators attempt to accommodate outliers by downweighting them in the calculations, while outlier identification methods attempt to identify outliers explicitly; see Barnett and Lewis (1994) for a thorough discussion of such methods. In the context of categorical data, an outlier is a cell—that is, a set of observations rather than a single observation—which deviates greatly from the expected count associated with the parametric model appropriate for the majority of the cells.

Fienberg (1969), Brown (1974), and Fuchs and Kenett (1980) suggested various out-

Karen V. Shane is Adjunct Assistant Professor, Department of Information Systems and Operations Management, Dolan School of Business, Fairfield University, Fairfield, CT 06430-5195 (E-mail: kshane@fair1.fairfield.edu). Jeffrey S. Simonoff is Professor of Statistics, Department of Statistics and Operations Research, Leonard N. Stern School of Business, New York University, 44 West 4th Street, New York, NY 10012-0258 (E-mail: jsimonof@stern.nyu.edu).

Journal of Computational and Graphical Statistics, Volume 10, Number 1, Pages 135–157

lier detection methods for tables under the independence model. Masking (misidentifying outlier cells as nonoutlying) and swamping (misidentifying nonoutlier cells as outlying) are potential problems with the proposed methods. Kotze and Hawkins (1984) proposed a method resistant to masking effects, but it is only applicable to the independence model. Simonoff (1988) proposed a backward stepping method using deleted residuals (residuals where the fit is based on the model with the cell deleted) to identify outliers. This method is applicable to models other than independence and Monte Carlo simulations demonstrated that it is more resistant to masking and swamping effects than those of Fienberg (1969), Brown (1974), and Fuchs and Kenett (1980). Albert (1997) gave a Bayesian approach to outlier detection by examining posterior probabilities to indicate the degree of "outlyingness". A drawback of this approach is the sensitivity to special prespecified hyperparameters used in the method. DuMouchel (1999) proposed fitting an empirical Bayes model to tables with millions of cells (and millions of observations) to try to identify cells with "interestingly large" counts.

Mosteller and Parunak (1985) proposed fitting a robust additive model using median polish on the table of the logarithms of the observed counts in a two-dimensional table as a method of outlier accommodation. Although masking and swamping are apparently addressed using this approach, this technique is not applicable to models other than independence. Hubert (1997) suggested a least absolute residual estimator to robustly fit independence and uniform association loglinear models.

In this article robust estimators of model parameters for categorical data based on the least median of squares (LMS) and least trimmed squares (LTS) regression estimators of Rousseeuw (1984) are proposed and investigated. These estimators have the advantage of being applicable in fitting arbitrary models (not just independence) to multidimensional tables (not just two-dimensional). In the next section the estimators and their properties are described. Section 3 discusses the computational algorithm and its complicating issues. Section 4 summarizes the results of Monte Carlo simulations exploring the properties of the estimators, and Section 5 analyzes several real datasets. Discussion of future work concludes the article.

2. LEAST MEDIAN AND LEAST TRIMMED ESTIMATORS

2.1 DEFINITIONS

Standard (nonrobust) analysis of categorical data typically proceeds based on maximum likelihood estimation. Consider a D-dimensional contingency table with d cells; we will write the table as a $d \times 1$ vector $\mathbf{n} = \{n_k\}$, $k = 1, \ldots, d$. Let $\mathbf{e} = \{e_k\}$ be the $d \times 1$ vector of expected cell counts under a hypothesized model. The expected counts are $e_k = N\pi_k$, where N is the total sample size, $N = \sum_{k=1}^{d} n_k$, and $\boldsymbol{\pi} = \{\pi_k\}$ are the cell probabilities. Since $\boldsymbol{\pi}$ is usually unknown, the estimated expected counts $\hat{\mathbf{e}}$ are the fitted values derived from the observed counts and the model, obtained by estimating the model parameters $\boldsymbol{\beta}$ (a $p \times 1$ vector) with estimates $\hat{\boldsymbol{\beta}}$.

Maximum likelihood is based on minimizing the likelihood ratio goodness-of-fit statistic

$$G^2 = 2\sum_{k=1}^{d} n_k \log(n_k/\hat{e}_k). \tag{2.1}$$

An asymptotically equivalent approach is minimum chi-squared, which is based on minimizing the Pearson chi-squared statistic

$$X^2 = \sum_{k=1}^{d} X_k^2(n_k, \hat{e}_k) \equiv \sum_{k=1}^{d} (n_k - \hat{e}_k)^2/\hat{e}_k. \tag{2.2}$$

See Agresti (1990) for a complete discussion of fitting models to contingency tables; Agresti (1984) discussed the fitting of models to tables with ordered categories.

The robust estimators proposed here are based on using criteria that are related to the chi-squared statistics (2.1) and (2.2), but result in more robust estimators. The focus here will be on median and trimmed criteria (Rousseeuw 1984) related to the Pearson statistic (2.2), although other choices are possible; see Shane (1998) for details. Let $X^2_{(h)}$ denote the hth order statistic of X^2_k. Robust Pearson estimators are defined as the minimizers of the criterion

$$\sum_{k=1}^{d} c_k X^2_{(k)}(n_k, \hat{e}_k), \tag{2.3}$$

where appropriate choice of the vector $\mathbf{c}$ leads to robust estimators. In particular, the Pearson least median of chi-squared residuals (LMCS) estimator, $\hat{\boldsymbol{\beta}}_{X^2\text{LMCS}(h)}$, is the $\boldsymbol{\beta}$ which minimizes (2.3) over all values of $\boldsymbol{\beta}$, where the weight vector $\mathbf{c}$ has $c_h = 1$ and all other entries equal to zero. Although the name "least median squares" would suggest taking h as the median of the values $\{1, \ldots, d\}$, h is in fact chosen to maximize the breakdown point of the estimator, the minimum proportion of outlying cells that can totally disrupt the estimate $\hat{\boldsymbol{\beta}}$ (which results in h somewhat larger than the median). The Pearson least trimmed chi-squared residuals (LTCS) estimator, $\hat{\boldsymbol{\beta}}_{X^2\text{LTCS}(h)}$, minimizes (2.3) where $\mathbf{c}$ satisfies

$$c_k = \begin{cases} 1, & \text{if} \quad k \leq h \\ 0, & \text{if} \quad k > h. \end{cases}$$

An alternative approach to deriving robust estimators is to use a criterion based on weighted least squares residuals. Grizzle, Starmer, and Koch (1969) proposed fitting models to contingency tables by fitting a (linear) regression model to the logarithms of the observed counts using weighted least squares, where the parameter estimates are chosen to minimize

$$\sum_{k=1}^{d} w_k r_k^2(n_k, e_k) \equiv \sum_{k=1}^{d} \hat{e}_k (\log n_k - \log e_k)^2. \tag{2.4}$$

Grizzle et al. (1969) suggested taking $w_k \equiv \hat{e}_k = n_k$ in (2.4) since $\text{var}(\log n_k) \approx e_k^{-1}$, and showed that the resultant estimator is asymptotically equivalent to maximum likelihood

(and is correspondingly nonrobust). A robust version of this estimator minimizes

$$\sum_{k=1}^{d} c_k w_k r_{(k)}^2(n_k, e_k), \tag{2.5}$$

where w_k is a robust preliminary estimate of e_k. Clearly (2.4) is not defined if $n_k = 0$. If too many cells that are consistent with the true model have zero counts, they cannot be distinguished by the robust estimators based on (2.5) from outlier cells that are zero because of contamination leading to excessively small counts. For this reason we assume that the sample size N is large enough so that no uncontaminated cells have zero counts. From a practical point of view, the implication is that robust estimation is very difficult for sparse tables that have many zero counts.

We avoid using weights equal to n_k as proposed by Grizzle et al. (1969), because they might be contaminated, and instead we use a robust set of weights, $\hat{e}_k$. Here the unweighted least median of squares estimate of e_k (using least squares residuals based on logged counts) is used for w_k; that is, $w_k = \hat{e}_k$ where $\hat{e}_k$ comes from using (2.5) with $w_k = 1$. The breakdown result is proved in Theorem 3, in which the weights are treated as being fixed and bounded. The unweighted estimator using $w_k = 1$ in (2.5) satisfies this condition, so the breakdown point given in Theorem 3 applies. As long as contamination does not reach the breakdown point, the estimate $\hat{\boldsymbol{\beta}}$ is bounded and hence all fitted values $\hat{e}_k$ must be bounded. Thus, in the two-step weighted estimator, given the initial unweighted estimates the weights in (2.5) are fixed and bounded, and the breakdown point remains the same.

Once again least median and least trimmed versions of the estimator can be defined by setting $\mathbf{c}$ appropriately. Using wtlnLMS and wtlnLTS to represent weighted logged least median and least trimmed squares, respectively, $\hat{\boldsymbol{\beta}}_{\text{wtlnLMS}(h)}$ and $\hat{\boldsymbol{\beta}}_{\text{wtlnLTS}(h)}$ symbolize these robust estimators.

2.2 PROPERTIES OF THE ESTIMATORS

The regression equivariance and scale equivariance properties of the linear regression LMS and LTS estimators carry over to loglinear models except that they are defined on a log scale. Consider a loglinear model, $\ln \mathbf{e} = \mathbf{Z}\boldsymbol{\beta}$, where $\mathbf{Z}$ is the $d \times p$ matrix defining the model. An estimator $\hat{\boldsymbol{\beta}}$ for a table of counts is called log scale location equivariant (LSLE) if the estimator $\hat{\boldsymbol{\beta}}^*$ on the counts $(an_k), k = 1, \ldots, d, a > 0$ is $\hat{\boldsymbol{\beta}} + (\ln a \ 0 \ \ldots \ 0)^{\mathrm{T}}$, where $(\ln a \ 0 \ \ldots \ 0)^{\mathrm{T}}$ is a p x 1 vector with $p - 1$ zeroes. An estimator $\hat{\boldsymbol{\beta}}$ is called log scale regression equivariant (LSRE) if the estimator $\hat{\boldsymbol{\beta}}^*$ on the counts $(n_k \exp(\mathbf{z}_k^{\mathrm{T}}\mathbf{v})), k = 1, \ldots, d$, is $\hat{\boldsymbol{\beta}} + \mathbf{v}$, where $\mathbf{v}$ is a p x 1 vector.

Lemma 1. *Given a D-dimensional contingency table, the robust Pearson estimator based on minimizing (2.3) and the weighted logged least squares estimator based on minimizing (2.5) are log scale location equivariant for any loglinear model* $\ln \mathbf{e} = \mathbf{Z}\boldsymbol{\beta}$ *that includes an overall grand mean term (usually represented as* μ*).*

Lemma 2. *Given a D-dimensional contingency table, the weighted logged least squares estimator based on minimizing (2.5) is log scale regression equivariant for any*

loglinear model $\ln \mathbf{e} = \mathbf{Z}\beta$ *that includes an overall grand mean term (usually represented as* μ*).*

The appendix outlines the proofs of these lemmas. Note that the robust Pearson estimators are not LSRE.

2.3 BREAKDOWN POINTS

The *breakdown point* of an estimator $\hat{\beta}$ is defined as

$$\varepsilon^*(\mathbf{n}, \hat{\beta}) = \min_{1 \le m \le d} \left\{ m/d ; \sup_{\text{all } \mathbf{n}^*} \| \hat{\beta}^* - \hat{\beta} \| = \infty \right\},$$

where $\hat{\beta}^*$ is the estimator for the contaminated table $\mathbf{n}^*$ with m contaminated cells, $\hat{\beta}$ is the estimator for the original uncontaminated table $\mathbf{n}$ and d is the number of cells in the table. That is, the breakdown point is the minimum proportion of outlier cells that can result in an unbounded estimate of β.

The breakdown points of a LSRE estimator, the maximum likelihood estimator, the weighted logged least squares estimator and the robust Pearson estimator are presented in Theorems 1–4, respectively. In what follows, let G be the maximum number of linearly independent rows in $\mathbf{Z}$ [see Mili and Coakley (1996) for discussion of robust estimation for structured designs of this type]. Let positive contamination define contamination due to excessively large counts in one or more cells of the table, relative to the remaining cells in the table, and let negative contamination define contamination due to excessively small counts in one or more cells of the table, relative to the remaining cells in the table.

Theorem 1. *The breakdown point of a LSRE estimator for the parameters of a loglinear model* $\ln \mathbf{e} = \mathbf{Z}\beta$ *satisfies*

$$\varepsilon^*_{\text{LSRE}} \le \varepsilon^*_{\text{LSREmax}} = \lfloor (d - G + 1)/2 \rfloor / d,$$

where $\varepsilon^*_{\text{LSREmax}}$ *is the maximum breakdown point for a LSRE estimator.*

Theorem 2. *The breakdown point of the maximum likelihood estimator (MLE) for the parameters of a loglinear model* $\ln \mathbf{e} = \mathbf{Z}\beta$ *is*

$$\varepsilon^*_{\text{MLE}} = \frac{1}{d}.$$

Theorem 3. *Define the weighted logged robust estimator,* $\hat{\beta}_{\text{wtln}}$*, to be the minimizer of (2.5), where the weights* $w_k = \hat{e}_k$*, the fitted values from an unweighted LMS or LTS fit to the table of logged counts. The breakdown point of this estimator satisfies*

$$\varepsilon^*_{\text{wtln}}(h_{op}) = \varepsilon^*_{\text{LSREmax}} = \lfloor (d - G + 1)/2 \rfloor / d$$

for $\lfloor (d + G + 1)/2 \rfloor \le h_{op} \le \lfloor (d + G + 2)/2 \rfloor$*, where* h_{op} *is the value of* h *that yields this optimal breakdown. The unweighted estimator taking* $w_k = 1$ *also possesses this breakdown point.*

Theorem 4. *The breakdown point of the robust Pearson estimator when only positive contamination (i.e., excessive counts) exists satisfies*

$$\varepsilon^*_{\text{LSREmax}} = \lfloor (d - G + 1)/2 \rfloor / d \leq \varepsilon^*_{X^2}(h_{op}) \leq \lfloor (d + 1)/2 \rfloor / d$$

for $\lfloor (d + G + 1)/2 \rfloor \leq h_{op} \leq \lfloor (d + G + 2)/2 \rfloor$, *where* h_{op} *is this optimal value of* h.

See the appendix for the outline of the proofs for Theorems 1–4. Note that the theorems apply to both the median and the trimmed versions of the estimators.

It can be shown that the breakdown point for the Pearson chi-squared estimator is the same as that for the weighted logged robust estimator in the case of the independence and uniform association models when only positive contamination is present. For two-way independence and uniform association models with the usual zero sum constraints, G is equal to $d - \min(r, c)$, where r is the number of rows and c is the number of columns. Hence, $\varepsilon^*_{\text{LSREmax}} = \lfloor (\min(r, c) + 1)/2 \rfloor / d$, or roughly half of the smaller of the proportion of cells in one row or one column of the table. That is, if at least half of the cells in a given row or column are outliers, that can lead to breakdown of the entry of β corresponding to that row or column.

In the context of finite sample breakdown, theoretically, the Pearson chi-squared estimator cannot handle any negative contamination. Under positive contamination the value X_k^2 of an outlier cell becomes infinite as $n_k \to \infty$ for bounded $\hat{\beta}$, and hence can be identified and isolated from the fitting, but under negative contamination ($n_k \to 0$) X_k^2 approaches $\hat{e}_k$, which is not necessarily unbounded; thus, the outlier cell can affect the estimates. From a practical point of view, it is reasonable to expect that the robust Pearson estimator can still perform satisfactorily under negative contamination as long as N is large enough so that an outlier cell with a (near-)zero count is unusual enough to be identified (i.e., the "correct" $\hat{e}_k$ is far from zero). Note that in the context of asymptotic breakdown, the probability that a nonoutlier cell has a count of zero approaches zero, and identifying negative contamination is no more difficult than identifying positive contamination.

Theorem 4 states that the maximum breakdown point for the robust Pearson estimator with positive contamination is essentially 50%. This is not very meaningful, however, because it is not clear that it is attainable. It can be shown, for example, that for the independence model and the uniform association model for two-way contingency tables, the breakdown point of the Pearson estimator equals $\varepsilon^*_{\text{LSREmax}} = \lfloor (d - G + 1)/2 \rfloor / d$ (see Shane 1998).

3. COMPUTATIONAL ISSUES

The estimates described in this article are based on minimizing the criteria (2.3) and (2.5). These criteria have many local minima, so standard gradient methods to find the minimum are not feasible. A standard approach to this problem in the robustness literature is to approximate the estimate using elemental subsets (other possible approaches are mentioned in Section 6). The idea is to evaluate the criterion for many different candidate values β, and then choose the value that minimizes the criterion as the final estimate $\hat{\beta}$. The elemental subset method proceeds by fitting the model to a table constructed from a subset of the cells. There must be at least p cells in this elemental subset. A model is fit to

the table by substituting a structural zero for cells not in the elemental subset and using the observed counts for cells in the elemental subset, and the estimation criterion is calculated. This process is repeated over many subsets, with the final estimate corresponding to the minimum value of the criterion over the elemental subsets.

One potential drawback of this method is that loglinear models can be complex, and hence the number of parameters may be large relative to the number of cells. This is problematic since if p, the number of parameters, is relatively large compared to the number of cells, there is a lower probability that the elemental subset method will find an elemental subset with only clean cells. Also, Hawkins (1993b) showed that the elemental subset method for LMS and LTS regression performed poorly compared to ordinary least squares when the number of independent variables was large relative to the sample size. This could be the case for the elemental subset method for the robust methods proposed here. Second, many elemental subsets may not provide a candidate solution due to singularity of the design matrix for the quasi model. Hubert and Rousseeuw (1997) discussed this issue in the context of robust regression with binary and continuous regressors. A third possible problem is that the number of elemental subsets for a model can be very large even for relatively small tables, making exhaustive selection impossible. For example, for a 5×5 table of independence, there are more than two million elemental subsets of size $p = 9$.

Fortunately, in all of the computations performed here (in both simulations and analyses of real tables) none of these possible difficulties posed problems. In all analyses either exhaustive elemental subset selection was performed, where feasible, or a predetermined number of elemental subsets were sampled. In all of the computations, the size of the elemental subset was p, thereby minimizing the chances that their size was too large relative to the number of cells. In cases where exhaustive selection of the subsets was not possible, the properties of the robust estimates were virtually the same whether 500, 1,000 or 1,500 elemental subsets were evaluated. If a subset was singular, another subset was chosen and the specified number of elemental subsets could still easily be examined. In all simulation runs and examples a solution was found without difficulty.

4. SIMULATION RESULTS

In this section, Monte Carlo simulations are described for 5×5 tables with outlier cells in various positions. The independence model and the uniform association model are assumed in several simulation runs and are reviewed separately.

4.1 INDEPENDENCE MODEL

A uniform probability structure was used to simulate 5×5 tables of independence. The sample size was 500. In all of the runs, 100 simulations were evaluated. The breakdown maximizing value of h is 23, as given by the formula for h_{op} in Theorem 3.

Three simulation runs are summarized here: (1) no contamination, (2) positive contamination in the first two cells of the first row of the table, and (3) positive contamination in the first three cells of the first row of the table. Each contaminant adds an extra 500 counts to the table. Figures 1, 2, and 3, respectively, display boxplots showing the distribution of

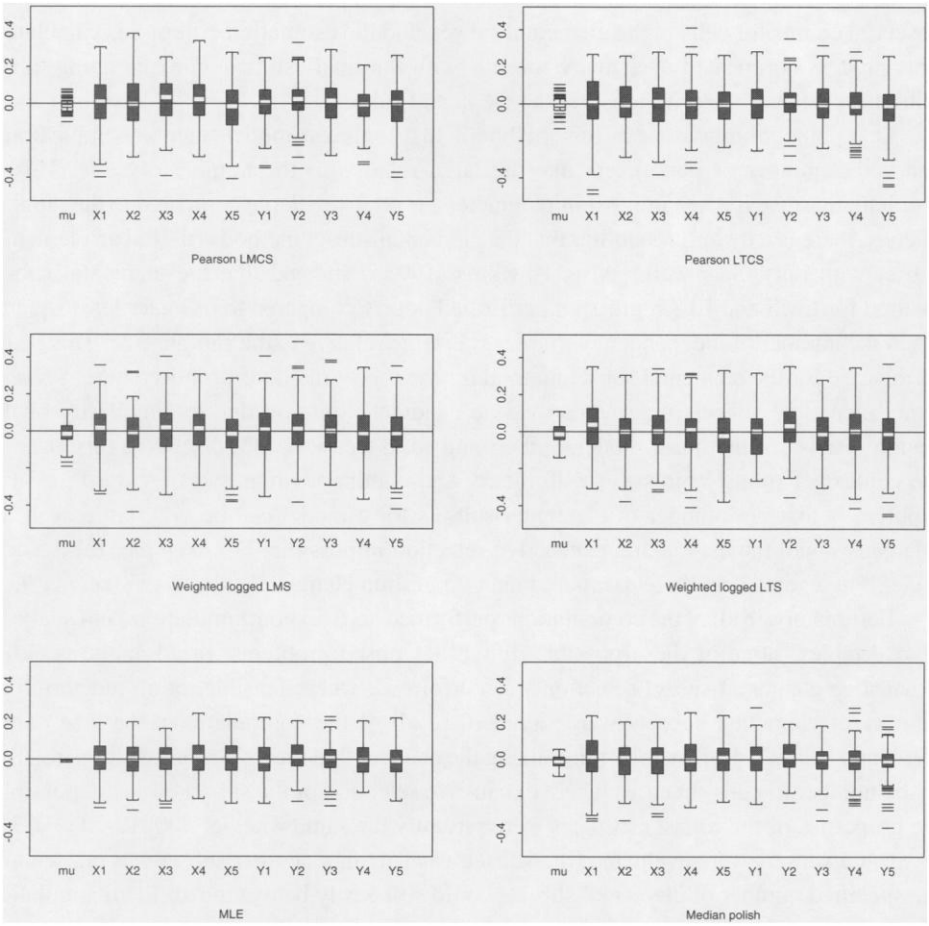

Figure 1. Boxplots of the error of the parameter estimates for simulations with no contamination in 5 × 5 tables for the independence model. Mu refers to the μ parameter, X1, . . . , X5 refer to the λ_i^X parameters, and Y1, . . . , Y5 refer to the λ_j^Y parameters.

the errors of the estimates over 100 simulations. Each chart within a figure contains a boxplot for each parameter, $\mu, \lambda_1^X, \ldots, \lambda_5^X, \lambda_1^Y, \ldots, \lambda_5^Y$, where the model of independence is $\ln e_{ij} = \mu + \lambda_i^X + \lambda_j^Y$.

When no contamination is present (Figure 1) all of the estimators perform similarly. The MLE is only slightly less variable than the other estimators. On average, the estimates are unbiased for all of the objective functions. Note that median polish has slightly less variability than the newly proposed robust estimators. In all of the independence model simulations, median polish performs well and sets a high standard for any robust method for the independence model.

When two contaminants are included (Figure 2) the MLE clearly breaks down, while the robust estimators do not. The robust chi-squared and weighted least squares estimators are slightly less biased, on average, than median polish, especially for λ_1^X (the parameter for the row with the outliers).

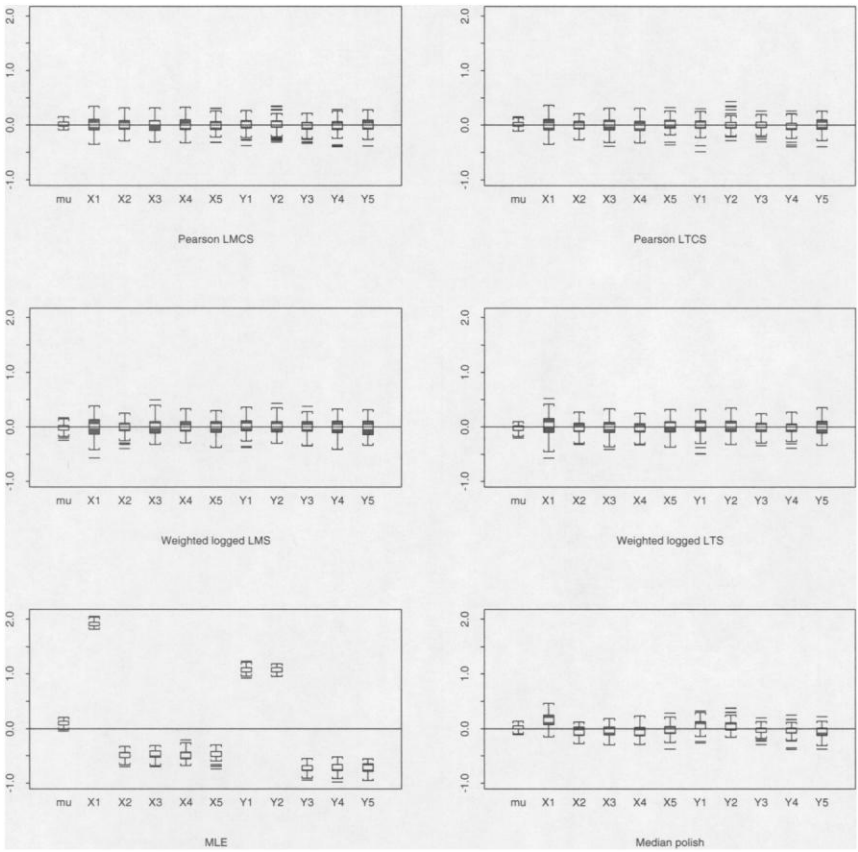

Figure 2. Boxplots of the error of the parameter estimates for simulations with contamination in the first two cells of the first row of 5 × 5 tables for the independence model.

With contamination in three cells of the first row (Figure 3) the robust estimators lose control of the parameters as we would expect based on the breakdown point. Note, however, that not all is lost since the $\lambda_j^Y, j = 1, \ldots, 5$ parameters remain fairly accurate for the robust estimates (while the maximum likelihood estimates of three of these parameters are positively biased).

Monte Carlo simulations were also performed on 8 × 8 tables for independence. The overall boxplot characteristics were similar to those displayed for 5 × 5 tables. For increasingly larger sample sizes, N increasing from 75 to 556, the accuracy of the estimates improved. However, sample sizes larger than that did not noticeably increase the accuracy.

4.2 UNIFORM ASSOCIATION MODEL

Simulations were performed on 5×5 tables of uniform association with row and column scores unit spaced and centered around zero. The sample size for the set of parameters

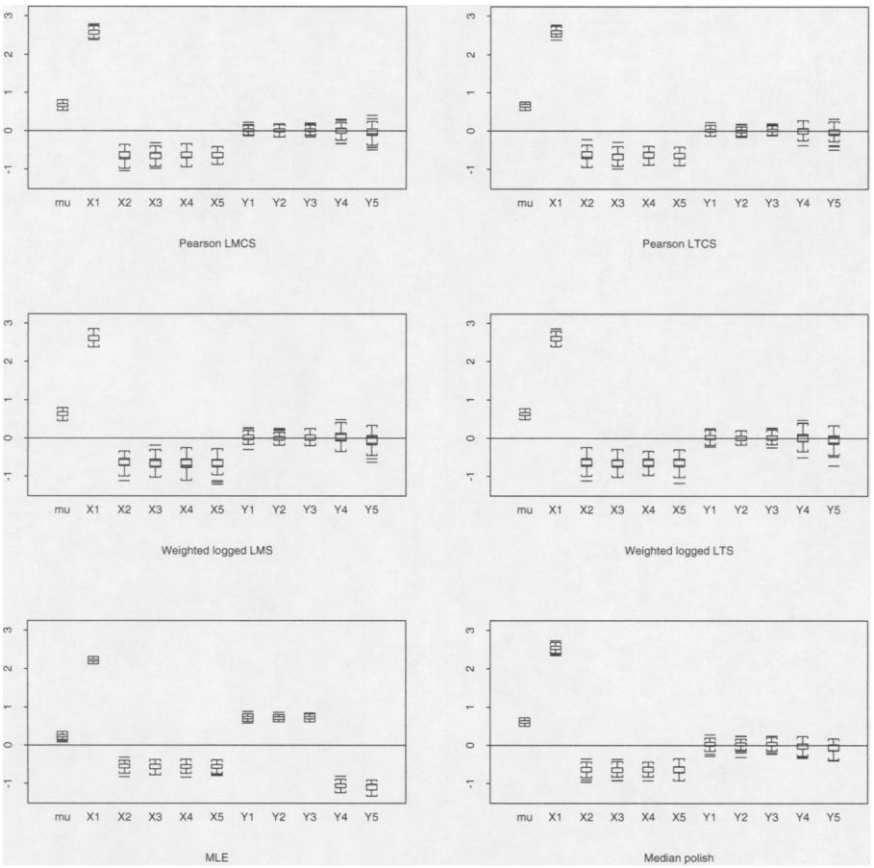

Figure 3. Boxplots of the error of the parameter estimates for simulations with contamination in the first three cells of the first row of 5 × 5 tables for the independence model.

applied is 3,090. The value $h = 23$ was used in the objective functions since it is the optimal value given in Theorem 3. Each chart within a figure contains a boxplot for each parameter $\mu, \lambda_1^X, \ldots, \lambda_5^X, \lambda_1^Y, \ldots, \lambda_5^Y, \gamma$ where the model of uniform association is $\ln e_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \gamma ij$.

When no contamination is present (Figure 4), the MLE is less variable, as expected, than the robust estimators. All of the robust estimators are close to unbiased on average.

When contamination of 500 observations is added to each of the last two cells of the first row (Figure 5) the MLE shows bias in almost all of the parameters. The robust estimators show less bias than the MLE. Even with contamination of only 100 observations (not shown), the robust estimators show less bias than the MLE. All of the robust estimation methods are successful in estimating the association parameter, γ, reasonably well.

Monte Carlo simulations were also performed on 8×8 tables for uniform association. The overall boxplot characteristics were similar to those displayed for 5×5 tables.

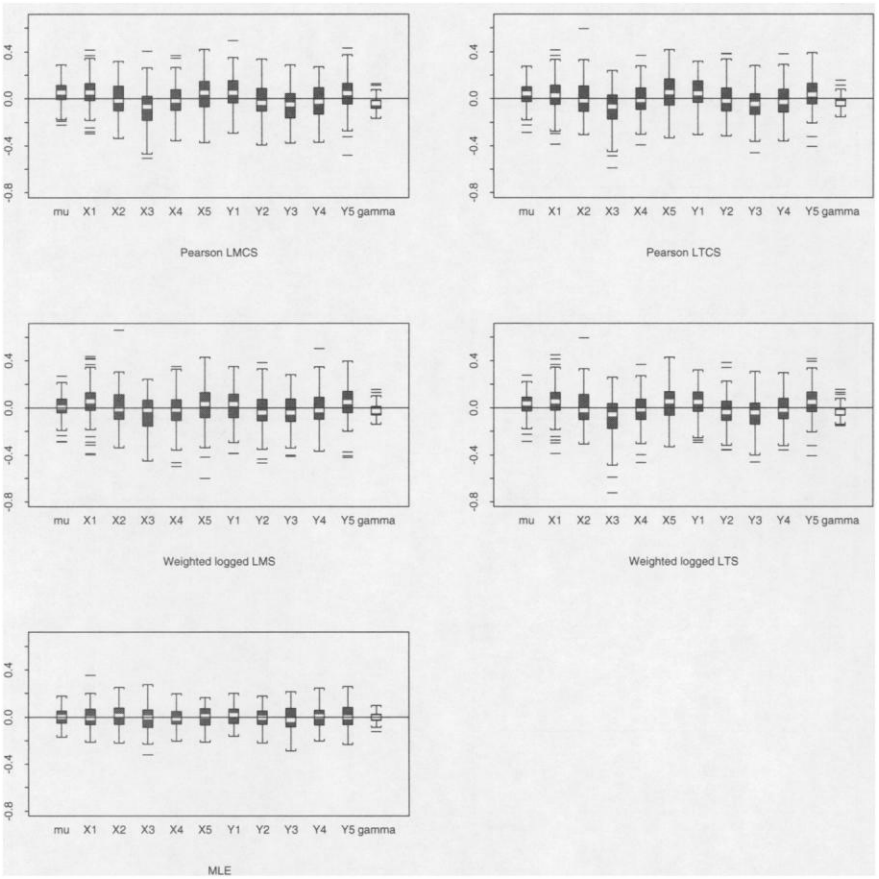

Figure 4. Boxplots of the error of the parameter estimates for simulations with no contamination in 5 × 5 tables for the uniform association model. Gamma refers to the γ association parameter.

5. EXAMPLES

In this section several real datasets are examined to illustrate the use of the methods proposed here.

5.1 ARCHAEOLOGICAL DATA

A portion of the archaeological artifact contingency table given by Mosteller and Parunak (1985), where the closeness of artifacts to permanent water is tabulated, is used as an example to motivate the need for robust fits (Table 1). Mosteller and Paranuk (1985) hypothesized that X, the type of artifact, and Y, the distance to permanent water, are independent. They used three resistant procedures to find outliers in this table.

The resistant methods used by Mosteller and Paranuk consistently found the cell corresponding to (grinding stones, immediate vicinity) cell (3,1) to be a large outlier. Cell (3,1) is noteworthy since, as hypothesized by the archaeologist who provided the data, grinding stones are not portable and hence it is more likely that they will be found near water sites.

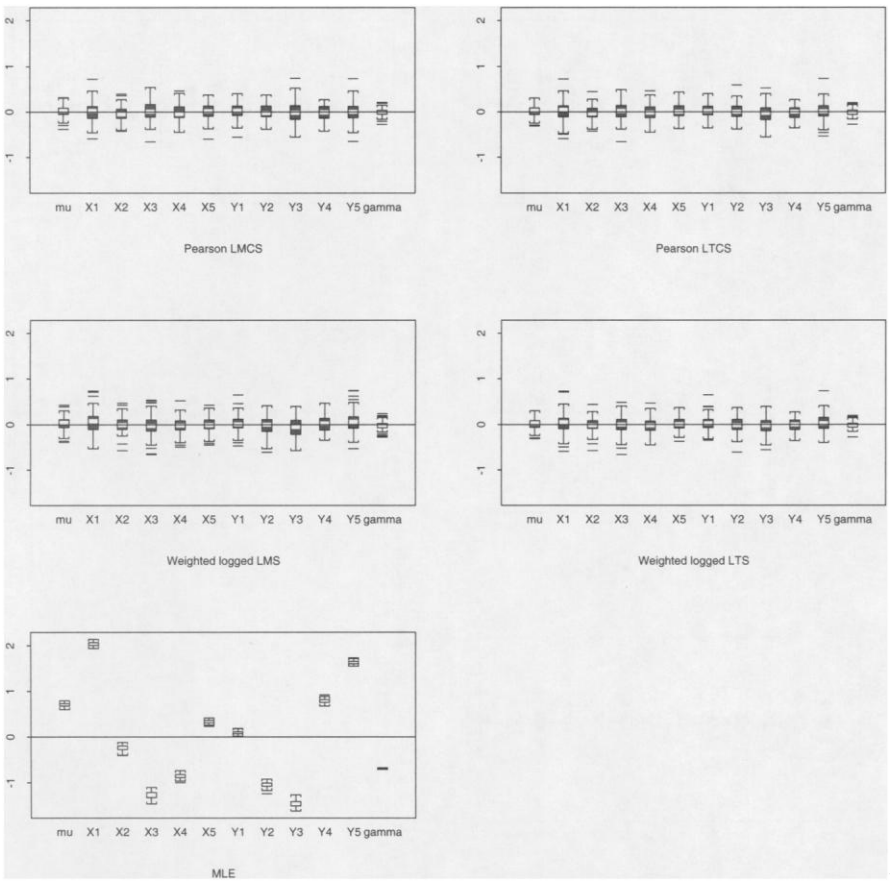

Figure 5. Boxplots of the error of the parameter estimates for simulations with contamination in the last two cells of the first row of 5 × 5 tables for the uniform association model.

Tables 1 and 2 compare the fits ($\hat{e}_i$) and the standardized residuals ($(n_i - \hat{e}_i)/\hat{e}_i^{1/2}$), respectively, for maximum likelihood, X^2LMCS, X^2LTCS, wtlnLTS, and median polish. The results for wtlnLMS were similar to those for wtlnLTS and are not included. Results are also given for the quasi-independence model (Goodman 1968) that takes cell (3,1) as a structural zero; this allows comparison to the results of MLE fitting after identifying a cell as an outlier and effectively removing it from the analysis.

The presumed outlier in (3,1) does not draw the robust fits toward it, and is easily identified (note that while the standardized residual for this cell when fitting using the MLE is the largest in the table, it is not large enough to unambiguously identify the cell as an outlier). Cell (3,4) is also flagged by the robust methods as a potential, less serious, outlier. The MLE and the resistant methods in Mosteller and Paranuk (1985) show no indication that this cell may be peculiar relative to the remaining cells, but it is interesting to note that this cell has the largest standardized residual for nonstructural zero cells under the quasi-independence model.

Table 1. Observed and Fitted Counts for Several Estimation Methods for the 4 × 4 Archaeological Example

		Distance to Water			
Method	*Artifact*	*Immediate vicinity*	*Within 1/4 mi.*	*1/4 to 1/2 mi.*	*1/2 to 1 mi.*
Observed Counts	Drills	2	10	4	2
	Pots	3	8	4	6
	Grinding Stones	13	5	3	9
	Point Fragments	20	36	19	20
MLE	Drills	4.17	6.48	3.29	4.06
	Pots	4.87	7.56	3.84	4.73
	Grinding Stones	6.95	10.79	5.49	6.77
	Point Fragments	22.01	34.18	17.38	21.43
X^2LMCS	Drills	4.21	7.58	4.00	4.21
	Pots	4.21	7.58	4.00	4.21
	Grinding Stones	2.78	5.00	2.64	2.78
	Point Fragments	20.00	36.00	19.00	20.00
X^2LTCS	Drills	4.21	8.00	4.00	4.21
	Pots	4.21	8.00	4.00	4.21
	Grinding Stones	2.63	5.00	2.50	2.63
	Point fragments	20.00	38.00	19.00	20.00
wtlnLTS	Drills	3.70	6.67	4.00	3.70
	Pots	4.44	8.00	4.80	4.44
	Grinding Stones	2.78	5.00	3.00	2.78
	Point Fragments	20.00	36.00	21.60	20.00
Median Polish	Drills	2.67	5.84	3.00	3.77
	Pots	3.60	7.89	4.05	5.09
	Grinding Stones	4.12	9.02	4.64	5.82
	Point Fragments	16.66	36.49	18.74	23.55
*Quasi Independence	Drills	3.36	6.86	3.49	4.30
	Pots	3.92	8.00	4.07	5.02
	Grinding Stones	3.90	7.96	4.05	4.99
	Point Fragments	17.72	36.18	18.40	22.69

*The quasi-independence model sets cell (3,1) to be a structural zero.

5.2 SEMICONDUCTOR DATA

Simonoff (1988) presented a 3 × 5 table of 162 counts giving data on the formation of contact windows in 3.5 μm complementary metal-oxide semiconductor (CMOS) circuits. The table is a cross-classification of spin speed (low, medium, high) by window size (I—not open or printed; II—(0, 2.25); III—[2.25, 2.75); IV—[2.75, 3.25); V—[3.25, ∞)) (Table 3).

Using row scores of $(-1, 0, 1)$ and column scores of $(-2, -1, 0, 1, 2)$, the likelihood ratio chi-squared statistic for the uniform association model on the 3 × 5 table is 11.14 on 7 degrees of freedom. This is not large enough to reject the null hypothesis that the uniform association model holds ($p = .13$), although a row effects model fits significantly better (the test of the adequacy of the uniform association model given the row effects model is $G^2 = 7.42$, $df = 1$, $p = .006$). The maximum likelihood fits for the uniform association model are apparently satisfactory and there is no indication of any outlier cells (see Table

Table 2. Standardized Residuals for Several Estimation Methods for the 4 × 4 Archaeological Example

		Distance to Water			
Method	*Artifact*	*Immediate vicinity*	*Within 1/4 mi.*	*1/4 to 1/2 mi.*	*1/2 to 1 mi.*
MLE	Drills	−1.06	1.38	.39	−1.02
	Pots	−.85	.16	.08	.58
	Grinding Stones	2.29	−1.76	−1.06	.86
	Point Fragments	−.43	.31	.39	−.31
X^2LMCS	Drills	−1.08	.88	.00	−1.08
	Pots	−.59	.15	.00	.87
	Grinding Stones	6.13	.00	.22	3.73
	Point Fragments	.00	.00	.00	.00
X^2LTCS	Drills	−1.08	.71	.00	−1.08
	Pots	−.59	.00	.00	.87
	Grinding Stones	6.39	.00	.32	3.93
	Point Fragments	.00	−.32	.00	.00
wtlnLTS	Drills	−.89	1.29	.00	−.89
	Pots	−.69	.00	−.37	.74
	Grinding Stones	6.13	.00	.00	3.73
	Point Fragments	.00	.00	−.56	.00
Median Polish	Drills	−.41	1.72	.58	−.91
	Pots	−.32	.04	−.03	.40
	Grinding Stones	4.38	−1.34	−.76	1.32
	Point Fragments	.82	−.08	.06	−.73
*Quasi Independence	Drills	−.74	1.20	.28	−1.11
	Pots	−.46	.00	−.03	.44
	Grinding Stones	4.61	−1.05	−.52	1.79
	Point Fragments	.54	−.03	.14	−.57

*The quasi-independence model sets cell (3,1) to be a structural zero.

3 for the standardized residuals). Since the first column of the table came closest to being identified as outliers using the backward stepping procedure, Simonoff (1988) recommended the uniform association model be used to fit the 3 × 4 table resulting from truncating window size I from the 3 × 5 table. Box and Jones (1986) gave this same suggestion. The likelihood ratio chi-squared statistic for uniform association on the 3 × 4 table is 3.14 with 5 degrees of freedom; the model fits the reduced table, and the model fits significantly better to the reduced table than to the full table ($G^2 = 8.00$, $df = 2$, $p = .02$).

Although the 3 × 4 table omitting the first column fits the uniform association model well, the robust methods indicate that in fact only cell (2,1) does not fit the uniform association model well (see Table 3). (The X^2LTCS and the wtlnLTS results were identical to the X^2LMCS result and are not given in Table 3.) The likelihood ratio chi-squared statistic for the quasi-uniform association model on the 3 × 5 table with cell (2,1) deleted is 3.28 on 6 degrees of freedom, which indicates a good fit to the reduced table and a significantly better fit than the model on the full table ($G^2 = 7.86$, $df = 1$, $p = .005$).

The association parameter $\hat{\gamma}$ is roughly equal to .50 for all of the estimation methods.

Table 3. Observed Counts and Standardized Residuals for the Uniform Association Model for the 3 × 5 CMOS Window Formation Example

		Window size				
Method	*Spin speed*	*I*	*II*	*III*	*IV*	*V*
Observed Counts	Low	47	5	6	2	0
	Medium	17	7	10	16	5
	High	12	4	7	15	9
MLE	Low	.64	−.55	.14	−1.11	−.92
	Medium	−1.57	.34	.33	1.39	.67
	High	1.24	.30	−.47	−.57	−.14
X^2LMCS	Low	.00	.00	.91	−.78	−.90
	Medium	−4.08	−.19	.00	.72	.00
	High	.00	.40	.00	.00	.00
wtlnLMS	Low	.00	.00	1.34	−.41	−.71
	Medium	−5.55	−1.12	−.71	.00	.00
	High	.00	.23	.00	.00	.41
*Quasi Uniform Association	Low	−.06	.09	.94	−.62	−.79
	Medium	−4.41	−.34	−.38	.51	.15
	High	.11	.42	−.23	−.20	.10

*The quasi-uniform association model sets cell (2,1) to be a structural zero.

5.3 MISCARRIAGE AND EDUCATION DATA

The data for this three-dimensional table were given in Silverman et al. (1985) in an article relating the employment environment and the educational background for pregnant mothers with spontaneously aborted conceptions (i.e., miscarriage). A miscarriage was classified according to the chromosome makeup of the fetus and reported as normal (NM) or abnormal (AM). The control group for this study included mothers who successfully delivered live babies (at 28 weeks or later) at the three hospitals in the study for the relevant time period. This category is labeled "live birth" (LB) in the tables.

The factors influencing the relationship between miscarriage and employment were education level (less than high school (<HS), high school (HS), or more than high school (>HS)) and payment type (public or private). Public payment referred to mothers on public assistance such as Medicaid, whereas private payment referred to mothers using private health insurance or some other form of private payment.

The hypothesized model is conditional independence of the education level and the type of miscarriage given the payment type. Table 4 gives the observed table. For this table, the theoretical breakdown point is one cell for both the MLE and the robust estimators, as one outlier cell can break down the parameter estimates relating to payment status. Nonetheless, the robust estimators can perform better than the MLE since some large residuals can be trimmed and the size of any contamination may not necessarily cause breakdown.

Table 4 gives fits and standardized residuals for the MLE and Pearson LTCS estimators. The MLE residuals from conditional independence do not indicate the presence of any outlier cells (the largest standardized residual is 2.14 in cell (Public, NM,<HS), see Table 4). A goodness-of-fit test indicates the model does not adequately fit the data ($G^2 = 21.04$,

Table 4. Observed Counts, Fits, and Standardized Residuals (in parentheses) for the Conditional Independence Model for the 2 × 3× 3 Miscarriage and Education Example. LB refers to live birth, NM refers to normal miscarriage, and AN refers to abnormal miscarriage.

			Counts (standardized residuals)					
			Education					
Method	*Payment Status*	*Type of birth*	*<HS*		*HS*		*>HS*	
Observed Counts	Private	LB	24		71		272	
		NM	23		48		156	
		AM	11		32		132	
	Public	LB	794		555		379	
		NM	298		147		98	
		AM	149		73		66	
MLE	Private	LB	27.7	(−.70)	72.1	(−.13)	267.3	(.29)
		NM	17.1	(1.42)	44.6	(.51)	165.3	(−.72)
		AM	13.2	(−.61)	34.4	(−.40)	127.4	(.40)
	Public	LB	838.0	(−1.52)	523.3	(1.38)	366.7	(.64)
		NM	263.3	(2.14)	164.4	(−1.36)	115.2	(−1.60)
		AM	139.7	(.79)	87.2	(−1.52)	61.1	(.63)
X^2LTCS	Private	LB	29.3	(−.98)	70.3	(.09)	266.7	(.32)
		NM	17.7	(1.26)	42.5	(.85)	161.3	(−.42)
		AM	14.5	(−.92)	34.8	(−.47)	132	(0.00)
	Public	LB	1132.5	(−10.06)	555.0	(.00)	379	(.00)
		NM	298	(.00)	146	(.08)	99.7	(−.17)
		AM	149	(.00)	73	(0.00)	49.9	(2.29)
*Quasi Conditional Independence	Private	LB	27.7	(−.70)	72.1	(−.13)	267.3	(.29)
		NM	17.1	(1.42)	44.6	(.51)	165.3	(−.72)
		AM	13.2	(−.61)	34.4	(−.40)	127.4	(.40)
	Public	LB	1087.2	(−8.89)	549.2	(.25)	384.8	(−.30)
		NM	292.1	(.35)	147.5	(−.04)	103.4	(−.53)
		AM	154.9	(−.48)	78.3	(−.59)	54.8	(1.51)

*The quasi-conditional independence model sets cell (2,1,1) to be a structural zero.

$df = 8, p = .007$). After finding robust fits, a large outlier is indicated. The large residual of −10.06 in cell (LB,<HS) for public payment status for the robust LTCS Pearson method is indicative of an unusual count in this cell. The robust fit shows that there are fewer live births than expected for poor, lowly educated mothers—possibly the byproduct of the socioeconomic factor that poorly educated mothers do not receive suitable prenatal care. The logged least squares estimates and the X^2LMCS estimates were slightly different from the chi-squared LTCS estimates but the residual pattern was similar.

Table 4 also gives the fits and standardized residuals for a quasi-conditional independence model where cell (LB, <HS) for public payment status is taken as a structural zero (thereby preventing it from affecting the model fit). This model fits the data well ($G^2 = 7.22$, $df = 7, p = .41$). Since the difference between these two goodness-of-fit statistics is 13.82, the quasi model fits the data significantly better than the MLE on the full table ($df = 1$, $p = .0002$).

6. CONCLUSION

In this article outlier resistant methods for fitting models to contingency tables are proposed and examined. These methods provide a general approach to the robust estimation of parameters for log-linear models in multidimensional tables (unfortunately, the "curse of dimensionality" is present since the breakdown point is a function of the smallest dimension of the table).

Several other parameter estimation methods were considered by Shane (1998). For example, rather than minimize the Pearson residual, the likelihood ratio chi-squared residuals were minimized or the power divergence family residuals with parameter $\lambda = 2/3$ (Read and Cressie 1988) were minimized. Typically, very similar results among the various methods emerged.

Although only the independence, uniform association, and conditional independence models were examined here, the LMCS/LMS and LTCS/LTS techniques can be adapted to other log-linear models as well. In all of the simulations and examples approximate versions of the estimates were found using the (random) elemental subset method, whereby subsets of cells are used to generate values of $\hat{\beta}$, over which the objective function is minimized. For more complex models, more efficient computational procedures may be required. In several articles, Hawkins (1993a, 1994, 1995) discussed a more efficient algorithm for computing least median and least trimmed squares solutions in linear regression (see also Hawkins and Olive 1999). This feasible set solution algorithm might be valuable in this context also.

Other future research involves adjusting the grand mean term in the log-linear model to minimize the criterion further. This idea of intercept tuning originated in robust linear regression in Rousseeuw and Leroy (1987); see also Rousseeuw and Hubert (1998). Adjustments can be made to the final regression equation or, more optimally, at each calculation of the regression equation for each elemental subset. By adjusting the intercept, a smaller LMS or LTS criterion value is achieved since the intercept term is replaced by one which makes the residuals have location zero. Hawkins (1993b) showed that intercept adjustment is at least as good as no intercept tuning for the linear regression model and that the tuning improved efficiency for all sample sizes. Since intercept adjustment has proved helpful for LMS and LTS estimates in linear regression, it is worthwhile to investigate the same for the elemental subset approximation for the robust methods presented here.

Theorems 1–4 concern a worst case breakdown point. Emerson and Hoaglin (1983) proposed a "well placed" breakdown point, which represents the maximum proportion of outliers that will not break down an estimator, and determined that value for median polish. We speculate that similar results for the proposed methods also exist. Other theoretical questions such as the consistency and efficiency of these robust chi-squared and least squares methods remain unanswered. For linear regression, the convergence rate of LMS is known to be only of order $N^{-1/3}$. It is not yet known if this characteristic carries over to the robust methods in this article. Although the simulations did not suggest a difference in convergence rates for the median and trimmed estimators, the convergence rates and asymptotic efficiencies remain to be determined.

One other research extension involves determining the potential use of the value of the objective function. Without further theory about the asymptotics of the robust chi-

squared and least squares estimators, the criterion has no relevance with respect to the robust goodness-of-fit of the model.

S-Plus code to perform the analyses in this article can be obtained from the authors.

APPENDIX

Proof of Lemma 1: Let $\hat{\boldsymbol{\beta}}_{X^2}$ be $\hat{\boldsymbol{\beta}}_{X^2\text{LMCS}}$ or $\hat{\boldsymbol{\beta}}_{X^2\text{LTCS}}$. Let $\hat{\boldsymbol{\beta}}_{X^2} = \hat{\boldsymbol{\beta}} = \{\hat{\mu}, \hat{\lambda}_1^X, \ldots\}$. Let $\mathbf{n}^* = a\mathbf{n}$ be the transformed observed counts and $\hat{\mathbf{e}}^*$ be fitted expected counts associated with $\hat{\boldsymbol{\beta}}^*_{X^2} = \hat{\boldsymbol{\beta}}^*$ from criterion (2.3) with $\mathbf{c}$ set appropriately.

If LSLE holds, then it can be shown that $\sum_{k=1}^d c_k X^2_{(k)}(n_k^*, \hat{e}_k^*) = a\sum c_k X^2_{(k)}(n_k, \hat{e}_k)$.

Since $\hat{\boldsymbol{\beta}}_{X^2}$ minimizes $\sum_1^d c_k X^2_{(k)}(n_k, \hat{e}_k)$, it follows that the X^2 estimator is log scale location equivariant if we can show that $\sum_{k=1}^d c_k X^2_{(k)}(n_k^*, \hat{e}_k^*) = a\sum_1^d c_k X^2_{(k)}(n_k, \hat{e}_k)$ is the smallest possible value that the objective function can assume.

Suppose there exists an estimator $\tilde{\boldsymbol{\beta}} = \{\tilde{\mu}, \tilde{\lambda}_1^X, \ldots\}$ that achieves a smaller objective function for the counts $n_k^* = an_k$; that is, $\sum_{k=1}^d c_k X^2_{(k)}(n_k^*, \tilde{e}_k) < a\sum_1^d c_k X^2_{(k)}(n_k, \hat{e}_k)$. Let $\tilde{\boldsymbol{\beta}}^*$ be the same as $\tilde{\boldsymbol{\beta}}$ except $\tilde{\mu}^* = \tilde{\mu} - \ln a$. The table associated with $\tilde{\boldsymbol{\beta}}^*$ has fitted counts $\tilde{e}_k^* = (1/a)\,\tilde{e}_k$ where $\tilde{e}_k$ are the fitted counts associated with $\tilde{\boldsymbol{\beta}}$. Then the objective function for the original counts using the estimator $\tilde{\boldsymbol{\beta}}^*$ is

$$\begin{aligned}\sum_{k=1}^d c_k X^2_{(k)}(n_k, \tilde{e}_k^*) &= \sum c_k(n_k - \tilde{e}_k/a)^2/(\tilde{e}_k/a)\\ &= (1/a)\sum c_k X^2_{(k)}(n_k^*, \tilde{e}_k)\\ &< \sum c_k X^2_{(k)}(n_k, \hat{e}_k).\end{aligned}$$

This is a contradiction since $\sum_{k=1}^d c_k X^2_{(k)}(n_k, \hat{e}_k)$ corresponds to the minimizing criterion for $\hat{\boldsymbol{\beta}}_{X^2}$ and must be a minimum. Thus, $\sum_{k=1}^d c_k X^2_{(k)}(n_k^*, \hat{e}_k^*) = a\sum c_k X^2_{(k)}(n_k, \hat{e}_k)$ is the smallest value that the objective function can assume for the transformed counts, n_k^*.

This proves that the robust Pearson estimator is LSLE. The proof that the least squares estimator is LSLE is very similar to the above proof. □

Proof of Lemma 2: Since it can be shown that $\hat{\boldsymbol{\beta}}_{\text{wtlnLMS}}$ and $\hat{\boldsymbol{\beta}}_{\text{wtlnLTS}}$ are $\mathcal{D}$ estimators, as defined by Mili and Coakley (1996), by Theorem 4.1 of Mili and Coakley (1993), they are regression equivariant. Log scale equivariance follows by definition. □

Proof of Theorem 1: The proof follows Theorem 3.1 of Mili and Coakley (1996). Let $\hat{\boldsymbol{\beta}}$ be a LSRE estimator. Suppose $\varepsilon^*_{\text{LSREmax}} > \lfloor(d - G + 1)/2\rfloor/d$. Let $m = \lfloor(d - G + 1)/2\rfloor$ where m counts of $\mathbf{n}^*$ take on arbitrary values. This implies $\| \hat{\boldsymbol{\beta}}^* \|$ remains bounded.

Let $\bar{G}$ be the largest subset of cells (or one of the subsets with the largest number of cells, if there is more than one) and let G be the number of cells in $\bar{G}$. Without loss of generality, assume that these G cells in $\bar{G}$ are the first G cells in the table, $n_1, \ldots, n_G$.

(Note that these cells span a $p-1$ dimensional subspace.) We have

$$G = \underset{\boldsymbol{\beta} \neq 0}{\max \text{card}} \left\{k : \mathbf{z}_k^{\mathrm{T}} \boldsymbol{\beta} = 0\right\},$$

where "card" represents cardinality. There will be $H = d - G$ cells remaining. Let $\bar{H}$ be this subset of the remaining cells.

Let $\mathbf{n}^*$ be a table of counts where $n_k^* = n_k \exp(\mathbf{z}_k^{\mathrm{T}}\mathbf{v})$ for m cells of $\bar{H}$ where $\mathbf{v}$ is a p dimensional vector in $\Re^p$ and is orthogonal to the subspace $\bar{G}$. (That is, $\mathbf{z}_k^{\mathrm{T}}\mathbf{v} = 0$ for $k = 1, \ldots, G$ and $\mathbf{z}_k^{\mathrm{T}}\mathbf{v} \neq 0$ for $k = G+1, \ldots, d$.)

Let $\mathbf{n}^{**}$ be a regression transformation of $\mathbf{n}^*$ where every cell in $\mathbf{n}^*$ is divided by $\exp(\mathbf{z}_k^{\mathrm{T}}\mathbf{v})$. Since $\mathbf{z}_k^{\mathrm{T}}\mathbf{v} = 0$ for the first G cells and since the m cells of $\mathbf{n}^*$ have been multiplied and divided by $\exp(\mathbf{z}_k^{\mathrm{T}}\mathbf{v})$, $\mathbf{n}^{**}$ has $d - m - G$ altered cells relative to $\mathbf{n}$.

We have $d - m - G = \lfloor (d-G)/2 \rfloor \leq m$ which implies $\| \hat{\boldsymbol{\beta}}^{**} \|$ remains bounded.

By LSRE, $\| \hat{\boldsymbol{\beta}}^{**} \| = \| \hat{\boldsymbol{\beta}}^{*} - \mathbf{v} \|$. But since $\| \hat{\boldsymbol{\beta}}^{*} \|$ is bounded and since the elements of $\mathbf{v}$ can be made arbitrarily large, this contradicts $\| \hat{\boldsymbol{\beta}}^{**} \|$ remaining bounded. Thus, $\varepsilon^*_{\text{LSREmax}}$ must be less than or equal to $\lfloor (d-G+1)/2 \rfloor / d$. □

Proof of Theorem 2: The proof involves contradicting the boundedness of the norm of the maximum likelihood equations, and is straightforward. □

Proof of Theorem 3: We need to first define a bad fit and a good fit. A bad fit is a set of parameter estimates $\hat{\boldsymbol{\beta}}^*$ for which $\| \hat{\boldsymbol{\beta}}^* \|$ is unbounded; a bad fit fits one or more contaminated cells resulting in bounded residuals for those particular contaminated cells. A good fit is a set of parameter estimates for which $\| \hat{\boldsymbol{\beta}}^* \|$ is bounded; a good fit is one which fits the clean cells of a table resulting in bounded residuals for the clean cells and unbounded residuals for the contaminated cells.

Since the weights are bounded and fixed, we can show for h_{op}, the breakdown point of a bounded fit is greater or equal to $\lfloor (d-G+1)/2 \rfloor / d$. For an unbounded fit, we need to define the least extreme bad fit (defined in the following). The least extreme bad fit compares to the least favorable bad fit in Mili and Coakley (1996). Using this, we can show that any bad fit leads to an unbounded criterion and hence breakdown.

The proof follows that of theorem 5.1 of Mili and Coakley (1996). First we need to show that the breakdown point of $\hat{\boldsymbol{\beta}}_{\text{wtln}}$ cannot be smaller than

$$\frac{\left\lfloor \frac{d-G+1}{2} \right\rfloor}{d}$$

for h_{op} for any table $\mathbf{n}$. Thus, for $m = \left\lfloor \frac{d-G+1}{2} \right\rfloor - 1$ contaminants , we need to show that $\| \hat{\boldsymbol{\beta}}^*_{\text{wtln}} \|$ remains bounded for a table $\mathbf{n}^*$ with m contaminated cells.

Note that $m = \left\lfloor \frac{d-G+1}{2} \right\rfloor - 1 = \left\lfloor \frac{d-G-1}{2} \right\rfloor$. We will prove that m cannot be less than $\left\lfloor \frac{d-G+1}{2} \right\rfloor$ by showing this for $\| \hat{\boldsymbol{\beta}}^*_{\text{wtln}} \|$ bounded and for $\| \hat{\boldsymbol{\beta}}^*_{\text{wtln}} \|$ unbounded.

Let $\| \hat{\boldsymbol{\beta}}^*_{\text{wtln}} \|$ be bounded (i.e., this is not a bad fit).

We have $\mathbf{n}^*$ with $d - m$ original counts and m contaminated counts. If some contaminated counts get arbitrarily large, then $r_k(n_k^*, e_k^*) = \ln n_k^* - \mathbf{z}_k^{\mathrm{T}} \hat{\boldsymbol{\beta}}^*_{\text{wtln}}$ goes to ∞. If some contaminated counts get arbitrarily small, then $r_k(n_k^*, e_k^*) = \ln n_k^* - \mathbf{z}_k^{\mathrm{T}} \hat{\boldsymbol{\beta}}^*_{\text{wtln}}$ goes to $-\infty$. This is summarized in the following table.

Residual Outcomes for Least Squares Residuals

Observed Count	*Bounded fit* $(0 < \hat{e}_k^* < \infty)$
Uncontaminated $0 < n_k^* < \infty$	Bounded residual
Outlier contamination n_k^*	Unbounded residual

This leads to m unbounded residuals and $d - m$ bounded residuals. Since $d - m = \lfloor \frac{d+G+2}{2} \rfloor \geq h_{op}$, the criterion of (2.5) will be bounded for m contaminants. Since this bounded criterion has a good fit associated with it, we do not have breakdown; hence m cannot be less than $\lfloor \frac{d-G+1}{2} \rfloor$.

Now let $\| \hat{\boldsymbol{\beta}}^*_{\text{wtln}} \|$ be unbounded (i.e., this is a bad fit). The remainder of the proof involves finding the least extreme bad fit passing through m contaminants in the table $\mathbf{n}^*$ and showing that the bad fit gives $G + m$ bounded residuals.

The least extreme bad fit is defined as follows: Suppose there are $j = 1, \ldots, B$ possible bad fits, $\hat{\boldsymbol{\beta}}^*_j$. Given bad fit j, let the squared residual be bounded for $k = 1, \ldots, b_j$ and unbounded for $k = b_j + 1, \ldots, d$. Then the least extreme bad fit corresponds to the bad fit $\hat{\boldsymbol{\beta}}^*_j$ for which b_j is a maximum over $j = 1, \ldots, B$.

We will find the least extreme bad fit passing through m contaminants in the table $\mathbf{n}^*$. This is done by picking the smallest possible number of cells to give a possible solution, say $\tilde{\boldsymbol{\beta}}$, to the objective function. Then we add some contaminated cells such that the contamination makes these cells fall on the p-dimensional hyperplane given by $\tilde{\boldsymbol{\beta}}$. We will call this hyperplane, defined by $\ln \tilde{e} = \tilde{\mathbf{Z}}\tilde{\boldsymbol{\beta}}$, $\mathcal{S}^p$. First consider $p - 1$ clean cells from $\bar{G}$, say $\tilde{n}_k^*$, where $k = 1, \ldots, p - 1$. Let $\tilde{\mathbf{z}}_k$ correspond to $\tilde{n}_k$. (For a loglinear model, $\| \tilde{\mathbf{z}}_k \|$ is obviously finite and cannot be contaminated.) The $p - 1$ cells from $\bar{G}$ have $\tilde{\mathbf{z}}_k$ which span the $p - 1$ dimensional subspace $\mathcal{G}^{p-1}$ since $\mathbf{z}_k$ for every cell in $\bar{G}$ can be written as a linear combination of the first $p - 1$ rows of $\mathbf{Z}$. Pick a cell from $\bar{H}$, say n_d^*. Note that by definition, z_d does not lie on $\mathcal{G}^{p-1}$. Thus, the p cells, $\tilde{n}_1^*, \ldots, \tilde{n}_{p-1}^*, n_d^*$ uniquely determine $\mathcal{S}^p$; that is, the hyperplane given by $\tilde{\boldsymbol{\beta}}$.

Let n_d^* be made arbitrarily large or small. Pick any subset of $m - 1$ cells from $\bar{H}$. Let these $m - 1$ cells be $n^*_{d-m+1}, \ldots, n^*_{d-1}$. Adjust these cell counts so that they lie on $\mathcal{S}^p$. Now we have d cells being

$\underbrace{\tilde{n}_1^*, \ldots, \tilde{n}_{p-1}^*}$,	$\underbrace{n_d^*, n^*_{d-m+1}, \ldots, n^*_{d-1}}$,	$\underbrace{n_1^*, \ldots, n^*_{d-(p-1+m)}}$
$p - 1$ clean cells from $\bar{G}$	m contaminants from $\bar{H}$	$d - (p - 1 + m)$ clean cells from $\bar{H}$ and $\bar{G}$ i.e., $(G - p + 1)$ from $\bar{G}$ and $(d - m - G)$ from $\bar{H}$

This bad fit $\tilde{\boldsymbol{\beta}}$ is the least extreme bad fit passing through m contaminants. By construction, the least extreme bad fit solution $\tilde{\boldsymbol{\beta}}$ yields residuals of zero for the m contaminants since they are made to lie on $\mathcal{S}^p$.

The $p - 1$ cells from $\bar{G}$ satisfy $\ln \tilde{n}_k = \ln \tilde{e}_k = \tilde{\mathbf{z}}_k^{\mathrm{T}} \tilde{\boldsymbol{\beta}}$. Since these cells span $\mathcal{G}^{p-1}$, the

remaining $G-p+1$ cells from $\bar{G}$ can be written as a linear combination of the $\tilde{\mathbf{z}}_k$ for these $p-1$ cells; that is, $\mathbf{z}_k = \sum_{j=1}^{p-1} a_{kj}\tilde{\mathbf{z}}_j$ for scalars a_{kj}. Hence, the logged fitted counts for the $G-p+1$ cells are $\ln \hat{e}_k = \mathbf{z}_k^{\mathrm{T}}\tilde{\boldsymbol{\beta}} = \sum_{j=1}^{p-1} a_{kj} \ln \tilde{e}_k$. Since $\tilde{e}_k$ for $k = 1, \ldots, p-1$ is bounded, so are the $\hat{e}_k$ for the other cells in $\bar{G}$. Hence, the least extreme bad fit solution $\tilde{\boldsymbol{\beta}}$ yields bounded residuals for the G cells in $\bar{G}$.

The least extreme bad fit gives a total of $G+m$ bounded residuals. The remaining cells from $\bar{H}$ will have unbounded residuals, by construction, since they are clean cells whose $\mathbf{z}_k$ do not lie on $\mathcal{S}^p$.

Since we have the least extreme bad fit, $G+m$ is the highest number of bounded residuals that any bad fit can have. Since it can be shown that $G+m < \left\lfloor \frac{d+G+1}{2} \right\rfloor = h_{op}$, the criterion in (2.5) will be unbounded at this least extreme bad fit and at any other bad fit (see Shane 1998, lemma 2.2). Hence, the minimum objective function is found at a bounded $\| \hat{\boldsymbol{\beta}}^*_{\text{wtln}} \|$.

We have shown $\varepsilon^*_{\text{wtln}}(h_{op}) \geq \frac{\left\lfloor \frac{d-G+1}{2} \right\rfloor}{d}$.

Since $\hat{\boldsymbol{\beta}}_{\text{wtlnLMS}}$ and $\hat{\boldsymbol{\beta}}_{\text{wtlnLTS}}$ are log scale regression equivariant, by theorem 1, we also have $\varepsilon^*_{\text{LSRE}} \leq \frac{\left\lfloor \frac{d-G+1}{2} \right\rfloor}{d}$. Thus, $\varepsilon^*_{\text{wtln}}(h_{op}) = \frac{\left\lfloor \frac{d-G+1}{2} \right\rfloor}{d}$. □

Proof of Theorem 4: The proof follows that of Theorem 3. We need to show that the breakdown point of $\hat{\boldsymbol{\beta}}_{X^2}$ cannot be smaller than $\lfloor (d-G+1)/2 \rfloor / d$ for h_{op} for any table $\mathbf{n}$.

For $m = \lfloor (d-G-1)/2 \rfloor$, we will prove that m cannot be less than $\lfloor (d-G+1)/2 \rfloor$ by showing this for $\| \hat{\boldsymbol{\beta}}^*_{X^2} \|$ bounded and for $\| \hat{\boldsymbol{\beta}}^*_{X^2} \|$ unbounded.

Let $\| \hat{\boldsymbol{\beta}}^*_{X^2} \|$ be bounded. We have $\mathbf{n}^*$ with $d-m$ original counts and m contaminated counts. If some contaminated counts get arbitrarily large, then we have m unbounded residuals and $d-m$ bounded residuals. Since $d-m = \lfloor (d+G+2)/2 \rfloor \geq h_{op}$, the criterion of (2.3) will be bounded for m contaminants. Since this bounded criterion has a fit with bounded $\| \hat{\boldsymbol{\beta}}^* \|$ associated with it, we do not have breakdown and m cannot be less than $\lfloor (d-G+1)/2 \rfloor$.

Now let $\| \hat{\boldsymbol{\beta}}^*_{X^2} \|$ be unbounded. The remainder of the proof involves finding the least extreme bad fit passing through m contaminants in the table $\mathbf{n}^*$ and showing that this bad fit gives $G+m$ bounded residuals. The next part of the proof is similar to that given above in Theorem 3 for finding the least extreme bad fit passing through m contaminants in the table $\mathbf{n}^*$.

The least extreme bad fit gives a total of $G+m$ bounded residuals which is the largest number of bounded residuals that any bad fit can have. The remaining cells from $\bar{H}$ will have unbounded residuals, by construction. Since it can be shown that $G+m =< \left\lfloor \frac{d+G+1}{2} \right\rfloor = h_{op}$, the criterion in (2.3) will be unbounded at this least extreme bad fit and at any other bad fit. Hence, the minimum objective function is found at a bounded $\| \hat{\boldsymbol{\beta}}^*_{X^2} \|$.

This shows $\varepsilon^*_{X^2}(h_{op}) \geq \frac{\left\lfloor \frac{d-G+1}{2} \right\rfloor}{d}$. The right hand side of the inequality follows since it is easily shown that the maximum breakdown point of a log scale location equivariant estimator is $\lfloor (d+1)/2 \rfloor / d$. (See Shane 1998, theorem 2.5.) □

[Received January 1999. Revised January 2000.]

REFERENCES

Agresti, A. (1984), *Analysis of Ordinal Categorical Data*, New York: Wiley.

——— (1990), *Categorical Data Analysis*, New York: Wiley.

Albert, J. H. (1997), "Bayesian Testing and Estimation of Association in a Two-Way Contingency Table," *Journal of the American Statistical Association*, 92, 685–693.

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

Brown, M. B. (1974), "Identification of the Sources of Significance in Two-Way Contingency Tables," *Applied Statistics*, 23, 405–413.

Box, G. E. P., and Jones, S. (1986), Discussion of "Testing in Industrial Experiments With Ordered Categorical Data," by V. N. Nair, *Technometrics*, 28, 295–301.

DuMouchel, W. (1999), "Bayesian Data Mining in Large Frequency Tables, With an Application to the FDA Spontaneous Reporting System" (with discussion), *The American Statistician*, 53, 177–202.

Emerson, J. D., and Hoaglin, D. C. (1983), "Analysis of Two-Way Tables by Medians," in *Understanding Robust and Exploratory Data Analysis*, edited by D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: Wiley, pp. 166–210.

Fienberg, S. E. (1969), "Preliminary Graphical Analysis and Quasi-Independence for Two-Way Contingency Tables," *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*, 75, 395–398.

Goodman, L. A. (1968), "Quasi-Independence and Interactions in Contingency Tables With or Without Missing Entries," *Journal of the American Statistical Association*, 63, 1091–1131.

Grizzle, J. E., Starmer, F., and Koch, G. G. (1969), "Analysis of Categorical Data by Linear Models," *Biometrics*, 25, 489–504.

Hawkins, D. M. (1993a), "The Feasible Set Algorithm for Least Median of Squares Regression," *Computational Statistics and Data Analysis*, 16, 81–101.

——— (1993b), "The Accuracy of Elemental Set Approximations for Regression," *Journal of the American Statistical Association*, 88, 580–589.

——— (1994), "The Feasible Solution Algorithm for Least Trimmed Squares Regression," *Computational Statistics and Data Analysis*, 17, 185–196.

——— (1995), "Convergence of the Feasible Solution Algorithm for Least Median of Squares Regression," *Computational Statistics and Data Analysis*, 19, 519–538.

Hawkins, D. M., and Olive, D. J. (1999), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics and Data Analysis*, 30, 1–11.

Hubert, M. (1997), "The Breakdown Value of the L_1 Estimator in Contingency Tables," *Statistics and Probability Letters*, 33, 419–425.

Hubert, M., and Rousseeuw, P. J. (1997), "Robust Regression With Both Continuous and Binary Regressors," *Journal of Statistical Planning and Inference*, 57, 153–163.

Kotze, T. J.v W., and Hawkins, D. M. (1984), "The Identification of Outliers in Two-Way Contingency Tables Using 2×2 Subtables," *Applied Statistics*, 33, 215–223.

Mili, L., and Coakley, C. W. (1993), "Robust Estimation in Structured Linear Regression," Technical Report Number 93–13, Virginia Polytechnic Institute and State University.

——— (1996), "Robust Estimation in Structured Linear Regression," *The Annals of Statistics*, 24, 2593–2607.

Mosteller, F., and Parunak, A. (1985), "Identifying Extreme Cells in a Sizable Contingency Table: Probabilistic and Exploratory Approaches," in *Exploring Data Tables, Trends and Shapes*, edited by D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: Wiley, pp. 189–224.

Read, T. R. C., and Cressie, N. A. C. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*, New York: Springer-Verlag.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J., and Hubert, M. (1998), "Recent Developments in PROGRESS," in L_1*-statistical Procedures and Related Topics*, ed. Y. Dodge, Institute of Mathematical Statistics Lecture Notes—Monograph Series, Volume 31, Hayward, CA: IMS, pp. 201–214.

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

Shane, K. V. (1998), *A Robust Approach to Categorical Data Analysis*, unpublished Ph.D. dissertation, Department of Statistics and Operations Research, New York University.

Silverman, J., Kline, J., Hutzler, M., Stein, Z., Warburton, D. (1985), "Maternal Employment and the Chromosomal Characteristics of Spontaneously Aborted Conceptions," *Journal of Occupational Medicine*, 27, 427–438.

Simonoff, J. S. (1988), "Detecting Outlying Cells in Two-Way Contingency Tables via Backwards Stepping," *Technometrics*, 30, 339–345.