Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data*

Walter R. Mebane, Jr.† Jasjeet S. Sekhon‡

July 24, 2003

*Earlier versions of this paper were presented in seminars at Harvard University, Washington University and Binghamton University–SUNY, at the 2002 Annual Meeting of the American Political Science Association, the 2002 Political Methodology Summer Meeting, and the 2002 Annual Meeting of the Midwest Political Science Association, and significantly different versions of some parts were presented at the 2001 Joint Statistical Meetings, and at the 2001 Political Methodology Summer Meeting. We thank Jonathan Wand for contributions to earlier versions of this work, Todd Rice and Lamarck, Inc., for generous support and provision of computing resources, John Jackson for giving us his FORTRAN code and Poland data, and Gary King for helpful comments. The authors share equal responsibility for all errors.

†Professor, Department of Government, Cornell University. 217 White Hall, Ithaca, NY 14853–4601 (Phone: 607-255-2868; Fax: 607-255-4530; E-mail: wrm1@cornell.edu).

‡Assistant Professor, Department of Government, Harvard University. 34 Kirkland Street, Cambridge, MA 02138 (Phone: 617-496-2426; Fax: 617-496-5149; E-mail: jasjeet_sekhon@harvard.edu).

Abstract

Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data

We develop a robust estimator—the hyperbolic tangent (tanh) estimator—for overdispersed multinomial regression models of count data. The tanh estimator provides accurate estimates and reliable inferences even when the specified model is not good for as much as half of the data. Seriously ill-fitted counts—outliers—are identified as part of the estimation. A Monte Carlo sampling experiment shows that the tanh estimator produces good results at practical sample sizes even when ten percent of the data are generated by a significantly different process. The experiment shows that, with contaminated data, estimation fails using four other estimators: the nonrobust maximum likelihood estimator, the additive logistic model and two SUR models. Using the tanh estimator to analyze data from Florida for the 2000 presidential election matches well-known features of the election that the other four estimators fail to capture. In an analysis of data from the 1993 Polish parliamentary election, the tanh estimator gives sharper inferences than does a previously proposed heteroscedastic SUR model.

Introduction

Regression models for vectors of counts are commonly used in a variety of substantive fields. Count models have been used in international relations (Schrodt 1995) and to analyze domestic political violence (Wang, Dixon, Muller, and Seligson 1993). Other social science applications include research on labor relations (Card 1990), the relationship between patents and R&D (Hausman, Hall, and Griliches 1984), and models of household fertility decisions (Famoye and Wang 1997). Recent work analyzing counts in political science includes studies of child care services (Bratton and Ray 2002), gender in legislatures (McDonagh 2002), gender and educational outcomes (Keiser, Wilkins, Meier, and Holland 2002), negative campaigning (Kahn and Kenney 2002; Lau and Pomper 2002), and votes (Canes-Wrone, Brady, and Cogan 2002; Monroe and Rose 2002).

In most of these cases the most natural model for the counts is the basic multinomial regression model (e.g. Cameron and Trivedi 1998, 270; McCullagh and Nelder 1989, 164–174). Counts of this kind measure the distribution of events among a finite set of alternatives, where each event generates one outcome. For vote counts, the alternatives are the candidates or parties that are competing for a particular office, and the multinomial model is relevant when each voter casts one vote. The model does not examine each individual separately but instead analyzes aggregates that correspond to the unit of observation. For vote counts the aggregates are usually legally defined voting districts, such as precincts, or larger units such as legislative districts, counties or provinces. Observations in this model measure the number of individuals in each unit who choose each alternative.

A multinomial model treats the number of individuals in each observational unit as fixed, and estimation focuses on how the proportion expected to choose each alternative depends on the regressors. Each expected proportion corresponds to the probability of making each choice according to the multinomial model. Usually these probabilities are defined as logistic functions of linear combinations of the regressors (see equation (1) below), and the problem is to estimate values for the unknown coefficient parameters in those linear combinations. In the basic model the probabilities and the total of the counts for each observation are both necessary to define the statistical distribution of the data, including the mean and the variance. One of the most important reasons to use a multinomial model is that the counts are heteroscedastic: the variance of the counts and consequently statistical properties of parameter estimates, such as the estimates' standard errors, depend on both the probabilities and the observation totals. The situation is analogous to the reasons why one should use a logit or probit model and not ordinary least squares (the linear probability model) with binary choice data. Unfortunately, recent analyses of count data in political science, such as Bratton and Ray (2002), Canes-Wrone et al. (2002), Keiser et al. (2002), Kahn and Kenney (2002), Lau and Pomper (2002), McDonagh (2002) and Monroe and Rose (2002), reduce the counts to percentages or proportions and ignore heteroscedasticity. As we shall illustrate in a sampling experiment, ignoring heteroscedasticity generally results in incorrect statistical inferences.

In practice the basic model has proved to be inadequate for vote counts. A problem that has been widely recognized is that aggregate vote data usually exhibit greater variability than the basic multinomial model can account for. In the basic multinomial model, the mean and the variance are determined by the same parameters. A common theme in several recently proposed models is to introduce additional parameters to allow the variance to be greater than the basic model would allow. Indeed, Katz and King (1999), Jackson (2002) and Tomz, Tucker, and Wittenberg (2002) all allow not only the variance of each vote but also the covariances between votes for different candidates to differ from what the basic multinomial model specifies. Katz and King (1999) introduce an "additive logistic" (AL) model for vote proportions by transforming the proportions into multivariate logits and then assuming that the logits for each voting district are distributed according to a multivariate-t distribution.[1] Tomz et al. (2002) describe a similar, seemingly unrelated regression (SUR) model for vote proportions, except assuming that the logits have a multivariate normal distribution. Katz and King (1999) and Tomz et al. (2002) ignore heteroscedasticity and

[1]Katz and King (1999) also allow a party not to have a candidate on the ballot in some districts.

assume that the vote proportions are homoscedastic. Jackson (2002) defines a SUR model that uses the covariance matrix that the multinomial distribution specifies for the logits, in order to account for heteroscedasticity, but adds to that matrix an unrestricted covariance matrix. Jackson (2002) also assumes multivariate normality.

Some extension to allow extra variability relative to the basic multinomial model is certainly necessary with vote data, but that is not enough to accommodate the striking irregularities that often occur in elections. A more general problem, and in a sense a prior problem, is that a single model may not be valid for all of the counts in the data. One well-known example is the vote in Florida for the 2000 U.S. presidential election. Wand, Shotts, Sekhon, Mebane, Herron, and Brady (2001) demonstrate that the vote for Reform party candidate Pat Buchanan in Palm Beach County was produced by processes substantially unlike the processes that generated his vote throughout the rest of Florida. Indeed, Wand et al. (2001) show that vote counts for Buchanan in many counties across the various states of the U.S. were produced by processes unlike those that occurred in most of the counties in each state.

Even if only a small fraction of the data are generated by a different model—perhaps only a single observation—estimation that assumes that all the data are good may produce seriously incorrect results. Given our weak theories and messy data, it is often doubtful that a single model is valid for all of the data. In any event, none of the examples of count data analysis cited at the beginning of this article do anything to detect whether the model is valid for all of the data or to protect against the chance that some of the data are discrepant.

The problem that a specified model may be good for only some of the observed counts is prior to the problem of extra variability in the sense that apparent departures from the basic multinomial model may reflect the failure of the model to hold for a subset of the observations, while it is fine for the others. In that case it is better to identify the part of the data for which the model is good and to separate those observations from the others. In other words, it is better to isolate the observations that are outliers relative to the specified model and not let them distort the analysis. It is possible for several outliers to distort an estimator to such an extent that the distorted data appear to be the norm and not the exceptions. Indeed, in such cases observations for which the model is correct may appear to be the outliers. This is the problem of masking (Atkinson 1986). For these reasons it does not work to try to identify the outliers one point at a time. It is necessary to have a method that locates all the outliers at once.

In this article we introduce a robust estimator for the multinomial model that provides accurate estimates and reliable inferences even when the model of interest is not a good model for a significant minority of the data.

We allow for extra variability relative to the basic multinomial model in the form of overdispersion (McCullagh and Nelder 1989, 174). This means that the covariance matrix of the basic multinomial model is multiplied by a positive constant that is greater than 1.0, so that the model asserts that there is more variability than occurs in the basic model. Overdispersion occurs whenever the choice probabilities vary across the individuals in each observational unit, but clusters of individuals within each unit have similar probabilities. For example, voters with different levels of income may differ in their voting preferences, but we observe only the average income in each district. We do not observe the income variation across individuals, but we do observe that income varies across voting districts. Consequently, we are able to assess how the probability of voting for different candidates varies across districts as a function of average district income. In such cases, the vote probability estimated for a given candidate in a voting district is the mean of the probabilities of voters in that district. The overdispersion parameter measures the variability of the individual vote probabilities in each district around that district's mean probability. The dispersion parameter increases as the individual probabilities vary more within each district. Johnson, Kotz, and Kemp (1993, 141) explicitly formulate the simplest form of a clustering mechanism that implies overdispersion for the binomial case (see also McCullagh and Nelder 1989, 125). Overdispersion is inevitable with aggregate vote data, because district-level variables always fail to capture traits that vary across voters in each district which affect the choices they make.[2]

[2]Underdispersion, where the covariance matrix is multiplied by a constant is less than

The estimator we develop produces correct results with high efficiency if the specified model in fact is a good approximation for the processes that produced most of the observed counts. This is to say that in the extreme case of completely correct specification, the robust estimator and maximum likelihood (ML) estimation of the multinomial model both produce consistent estimates, but the robust method is less efficient. On the other hand, as we shall illustrate, when the model is not correct for a fraction of the data, the robust estimates will continue to be good while ML estimates will in general be wrong, sometimes grossly wrong. There is no need to identify in advance the subset of the data for which the model is a good approximation. The ill-fitted counts—the outliers—are identified as part of the robust estimation procedure. The counts to which the model does not apply are effectively omitted from the analysis and have no effect either on the estimates of the coefficient parameters or on estimates of the coefficients' estimation error.

The method we introduce generalizes the robust estimator for overdispersed binomial regression models that was introduced by Wand et al. (2001).[3] The generalization is difficult and requires us to develop new methods because the counts for the different choices are negatively correlated in the multinomial model. Negative correlations arise because the multinomial model conditions on the total count for each observational unit. A way to understand the negative correlation is to think of the votes as arriving one at a time. If a vote goes for one candidate, it cannot go for any other candidate. So if one candidate's share of the votes goes up, the other candidates' shares go down because their counts remain the same while the total increases. This competition among the choice alternatives implies the negative correlations. Because each candidate attracts votes in proportion to the choice probability for that candidate, the negative correlations are functions of the choice probabilities. Many good

---

1.0, is allowed in our model but rarely occurs in practice. Underdispersion arises when the individual choice probabilities tend to be similar within each observational unit but different across units. Johnson et al. (1993, 138–139) describe a simple form of such clustering.

[3]In addition to extending the model, we also correct an error in the method Wand et al. (2001) used to estimate the standard errors of the parameter estimates.

robust estimation methods exist for uncorrelated observations but the problem of correlated data is much more difficult.

In order to produce uncorrelated residuals, we use a new approach based on a formal orthogonal decomposition of the multinomial distribution's covariance matrix. The method applies to count data some of the statistical theory of qualitative and quantitative robustness that has been developed to fulfill the three desirable features outlined by Huber (1981, 5–17): the method has reasonably good efficiency when the model assumed for the data is correct; small deviations from the model assumptions (which may mean large deviations in a small fraction of the data) impair the model's performance only slightly; and "somewhat larger deviations from the model should not cause a catastrophe" (Huber 1981, 5). Our work also responds to Western's (1995) call for robust estimation to be used with generalized linear models. Victoria-Feser and Ronchetti (1997) rigorously demonstrate that ML estimation of the basic multinomial model is not robust and develop an estimator for contaminated multinomial data, although their estimator makes no provision for overdispersion.

Katz and King's (1999) AL model also can produce good point estimates for coefficient parameters if the specified model is not good for a fraction of the data. This model treats the discrepant data by fattening the tails of the multivariate-t distribution so that larger residuals are more likely according to the model: the distribution's degrees-of-freedom (DF) parameter gets smaller as the amount of differently generated data increases or as the differently generated data's discrepancy from the rest of the data increases (cf. Lange, Little, and Taylor 1989). As we demonstrate, however, this method does not produce correct standard errors and therefore does not support making correct statistical inferences. The two SUR models, which assume the data are multivariate normal, lack any way effectively to downweight discrepant observations, and therefore they generally produce wrong results when the specified model is not valid for some observations, as does the ML estimator for the basic model.

We begin with a brief description of the overdispersed multinomial regression model and our new robust estimation method. Then we present the results of a Monte Carlo sampling experiment that demonstrates that the method produces accurate parameter estimates and supports correct statistical inferences even when the data are contaminated with counts that are generated by a significantly different process. The study also shows that the method correctly flags the contaminated observations and hence provides an accurate method for outlier detection. The AL model, on the other hand, produces accurate coefficient point estimates but incorrect standard errors, while the nonrobust ML estimator and the SUR models fail even to produce accurate point estimates. We then use our method to analyze two sets of data. We analyze Florida vote data from the 2000 presidential election, extending the binomial (Buchanan versus the rest) model results of Wand et al. (2001) to an analysis of five categories of presidential candidates: Buchanan, Nader, Gore, Bush and "other." We also improve the set of regressors, in particular taking the Cuban-American population explicitly into account. Then we use our method to estimate the specification for the 1993 Polish parliamentary election that was introduced by Jackson (2002).

Robust Estimation of an Overdispersed Multinomial Model

We use the overdispersed multinomial model for $J \geq 2$ outcome categories defined and motivated by McCullagh and Nelder (1989, 174). Let $i = 1, \ldots, n$ index an observed vector of J counts $y_i = (y_{i1}, \ldots, y_{iJ})'$, and let $m_i = \sum_{j=1}^{J} y_{ij}$ denote the total of the counts for observation i. Given probability p_{ij}, the expected value of y_{ij} is $Ey_{ij} = m_i p_{ij}$. Let $p_i = (p_{i1}, \ldots, p_{iJ})'$ denote the vector of probabilities for observation i. $P_i = \text{diag}(p_i)$ is a $J \times J$ diagonal matrix containing the probabilities. The covariance matrix for observation i is:

$$E[(y_i - m_i p_i)(y_i - m_i p_i)'] = \sigma^2 m_i (P_i - p_i p_i') ,$$

with $\sigma^2 > 0$ (McCullagh and Nelder 1989, 174, eqn. 5.17). Heteroscedasticity is apparent in the covariance matrix's dependence on both m_i and p_i. If $\sigma^2 = 1$ then the covariance is the same as in the basic multinomial model, but if $\sigma^2 > 1$ then there is overdispersion. The probabilities p_{ij} are functions of observed data vectors x_{ij} and unknown coefficient parameter vectors β_j. In particular, p_{ij} is a logistic function of J linear predictors $\mu_{ij} = x'_{ij}\beta_j$:

$$p_{ij} = \frac{\exp(\mu_{ij})}{\sum_{k=1}^{J} \exp(\mu_{ik})}. \tag{1}$$

When x_{ij} is constant across j, a commonly used identifying assumption is $\beta_J = 0$: J is said to be the reference category. We gather the K unknown coefficients into a vector denoted β.

To estimate the model we use robust estimators for σ^2 and β: the least quartile difference (LQD) estimator (Croux, Rousseeuw, and Hossjer 1994; Rousseeuw and Croux 1993) for $\sigma = \sqrt{\sigma^2}$ and, given the estimate for σ, the hyperbolic tangent (tanh) estimator (Hampel, Rousseeuw, and Ronchetti 1981; Hampel, Ronchetti, Rousseeuw, and Stahel 1986, 160–166) for β. Here in the main text we sketch the main features of the estimation method. In the Appendix we present the further details required to use the approach with overdispersed multinomial data.

The estimator has two key robustness properties. First, both the LQD and tanh estimators have the highest possible breakdown point for a regression model. The finite sample breakdown point of an estimator is the smallest proportion of the observations one needs to replace in order to produce estimates that are arbitrarily far from the parameter values that generated the original data (Donoho and Huber 1983). The concept of breakdown point that in a strict technical sense applies to the current estimation problem is more complicated, for instance to take into account the fact that asymptotic properties of the estimator under regularity conditions are generally of interest (Hampel et al. 1986, 96–98), but the intuition behind the more general concept remains the same: even large perturbations in a fraction of the data should not affect the estimator's performance. The LQD and tanh estimators have a breakdown point of 1/2. The nonrobust ML, AL and the SUR estimators all have finite sample breakdown points of $1/(nJ)$—asymptotically zero.

The second important robustness property concerns the degree to which perturbations of the data affect the variability of parameter estimates. The tanh estimator is optimal in the sense that it minimizes the asymptotic variance of the estimates for a given upper bound on how sensitive the asymptotic variance is to a change in the distribution of the data (Hampel et al. 1981). The existence of such an upper bound implies that the tanh estimator has a finite rejection point, which means that an observation that has a sufficiently large residual may receive zero weight and hence not affect the parameter estimates at all. Given the value of σ, tanh estimators are by construction the most efficient possible estimators of β that may put zero weight on some observations (Hampel et al. 1986, 166).

Under a wide range of conditions in which the data deviate to some extent from the specified model, the tanh estimator is asymptotically normal with covariance matrix given in general by the "sandwich" formula derived by Huber (1967, 231; 1981, 133) and in particular by the matrix $\hat{\Sigma}_{\hat{\beta}}$ that we define in the Appendix. In the special case where the model is exactly correct for a majority of the data but the rest of the data are generated by some other process, the tanh estimator is consistent for the model's parameters. The fact that consistency holds when the model is correct for at least half of the data is an implication of the tanh estimator's high breakdown point. Huber (1981, 127–132) proves the general result for M-estimators that applies in this case. In this special case the tanh estimator typically puts zero weight on the data that are generated by the alternative processes, such that two other familiar covariance matrix estimators are also expected to be correct: the inverse of a weighted Hessian matrix and the inverse of a weighted outer product of the gradient (OPG). We define those matrices in the Appendix, denoted respectively $\hat{\Sigma}_{\mathrm{G}:\hat{\beta}}$ and $\hat{\Sigma}_{\mathrm{I}:\hat{\beta}}$.

A point of departure for our methods is the fact that given any estimated probabilities $\hat{p}_{ij}$, the J residuals $\hat{r}_{ij} = (y_{ij} - m_i\hat{p}_{ij})$ for each i always sum to zero. This result follows from the fact that the multinomial model treats the sum m_i of the counts for each observation i as given, so that each vector of counts y_i has only $J-1$ independent elements. That same feature of conditioning on the total implies that, like the counts, the simple residuals $\hat{r}_{ij}$ are negatively correlated with one another. We use a formal Cholesky decomposition of the multinomial covariance matrix, which is an orthogonal decomposition method derived by Tanabe and Sagae (1992), to produce uncorrelated residuals for each observation, denoted $r_i^{\perp}$. By construction, the J-th value $r_{iJ}^{\perp}$ is zero, so that the first $J-1$ values in $r_i^{\perp}$ contain all the information. Dividing each of the $J-1$ nonzero orthogonalized residuals by its respective standard deviation, which Tanabe and Sagae (1992) also derive, we obtain a set of normalized values, denoted $\hat{r}^*_{ij}$, that have a normal distribution with variance σ^2 if m_i is sufficiently large and the model is correctly specified for all the data. This normalization adjusts for the heteroscedasticity associated with both m_i and the probabilities p_i.

If the model is appropriate for only a majority of the data and the values of the model's parameters are known, then for the counts that were generated by the alternative processes, the residuals $\hat{r}^*_{ij}$ computed using those parameters are typically large relative to the variance σ^2. Ideally, information from those counts would not be used to estimate the parameters of the model that applies to most of the data. The robust estimators we use approximate that ideal behavior. For a given model specification—i.e., a set of observed counts y_{ij}, regressors x_{ij} and linear predictor functional forms μ_{ij}—the estimators find the parameter values that best characterize most of the data while downweighting information that is associated with normalized residuals that are larger than one would expect to observe in a sample of normal variates.

Our estimator produces a vector of $J-1$ weights for each observation, $w_i = (w_{i1}, \dots, w_{iJ-1})'$, with $w_{ij} \in [0, 1]$. The value 1 indicates that the tanh estimator is giving full weight to the orthogonal component of the data corresponding to $\hat{r}^*_{ij}$, and the value 0 indicates that the estimator is completely excluding information from that component.

Further details regarding the robust estimator are in the Appendix. To summarize briefly here, after using the formal Cholesky decomposition to reduce the multivariate robustness problem to a collection of uncorrelated problems, we use the optimizing evolutionary program called GENOUD (Sekhon and Mebane 1998) to find the LQD estimates. Then we compute the tanh parameter estimates via a weighted Newton algorithm and estimate the asymptotic covariance matrix. The use we make of the LQD and tanh estimators is novel, but we have nothing to add to the statistical understanding of those estimators per se. The statistical properties of those estimators are well established in the statistical literature, as are the properties of asymptotic covariance matrices for M-estimators (Carroll and Ruppert 1988, 209–213; Huber 1967; White 1994), which we also apply.

A Monte Carlo Sampling Experiment

To assess the performance of the robust tanh estimator, the nonrobust ML estimator and the AL, SUR and heteroscedastic SUR models under a range of conditions, we conduct a Monte Carlo sampling experiment using six different types of simulated data. We examine six different experimental conditions, replicating each condition 1,000 times. In each replication we generate observations consisting of four counts ($J = 4$) with $m_i = 10,000$. We conduct the experiment for $n = 50$ and $n = 100$ observations. The linear predictors have the same functional form for all conditions. Each of the first $J - 1$ predictors includes a single, simulated regressor, denoted x_i, and a constant, while the J-th predictor is set to zero:

$$\mu_{ij} = \begin{cases} \beta_{j0} + \beta_{j1}x_i, & j = 1, \ldots, J-1, \\ 0, & j = J. \end{cases}$$

In each experimental condition the regressor is constant across choice categories $j = 1, \ldots, J-1$. The conditions differ by having different values for the regressor, the coefficients or the dispersion.

Table 1 lays out the overall design of the experiment. The first four experimental conditions all have the same linear predictors. The regressor is normally distributed with mean one and variance one. The regressor values are the same in every replication. For all the linear predictors, $j = 1, \ldots, J-1$, the coefficient parameters are $\beta_{j0} = -1$ and $\beta_{j1} = 1$. With this specification the expected outcome probability is approximately the same for all four categories: $p_{i1} = p_{i2} = p_{i3} = 0.2442$ and $p_{i4} = 0.2673$. Experimental condition 1 features uncontaminated multinomial data with no overdispersion, i.e., $\sigma^2 = 1$. Condition 2 is the same except that it includes overdispersion. We used the cluster-sampling model (McCullagh and Nelder 1989, 174) to generate counts for which $\sigma^2 = 5.5$. Conditions 3 and 4 have ten percent of the data generated by a different process from the rest of the data. In condition 3, ten percent of the condition 1's observations are perturbed in such a way that the constant parameters in their linear predictors are approximately $\beta_{10} = -2.099$ and $\beta_{30} = -0.489$. The other four parameters are the same for all observations. Condition 4 is the same as condition 3 except with the same kind of overdispersion as in condition 2.

*** Table 1 about here ***

Experimental conditions 5 and 6 feature ten percent contamination with skewed outcome probabilities. For ninety percent of the observations the regressors are again normally distributed with mean one and variance one, but the constant parameter values are $\beta_{10} = -3.5$, $\beta_{20} = -3$ and $\beta_{30} = -1$, and $\beta_{11} = \beta_{21} = \beta_{31} = 1$. The expected outcome probabilities are approximately $p_{i1} = 0.0366$, $p_{i2} = 0.0603$, $p_{i3} = 0.4453$ and $p_{i4} = 0.4578$. The remaining ten percent of the observations have regressor values that are normally distributed with a mean of $-.5$ and a variance of 4. The parameters for these contaminated observations are $\beta_{10} = \beta_{20} = 0.001$, $\beta_{30} = 2.000$, $\beta_{11} = \beta_{21} = -2.000$ and $\beta_{31} = -1.000$. The values of the regressor are constant across replications. Unlike the first four conditions, in conditions 5 and 6 the counts that are contaminated because they are generated according to different parameter values are also associated with regressors that have a different mean and variance than the regressors associated with the balance of the data. These high-variance regressors have high leverage (Carroll and Ruppert 1988, 31–33), which means that nonrobust estimated regression lines should be induced to pass near the contaminated observations.

For each replication we use all five models to compute estimates for the coefficient parameters. The nonrobust ML estimates use the multinomial model likelihood. In the absence of contamination, such ML estimates are consistent for the coefficient parameters whether or not there is overdispersion. For the ML estimates we compute confidence intervals using both the nonrobust inverse Hessian matrix alone and the nonrobust inverse Hessian multiplied by the usual nonrobust estimate of dispersion (McCullagh and Nelder 1989, 175). For the tanh estimates we compute confidence intervals based on the Huber-White sandwich, the inverse weighted Hessian and the inverse OPG estimators. For the AL model we compute confidence intervals based on the estimate of the asymptotic covariance matrix computed by inverting the Hessian matrix of the model's log likelihood function. For the SUR model we use the same FGLS approach as do Tomz et al. (2002).[4] For the heteroscedastic SUR model we use

[4]To estimate the SUR model we used the **R** package `systemfit` (version 0.5-6), avail-

the covariance matrix estimate given by Jackson (2002, 54, eq. 13).[5] We compute symmetric confidence intervals using ordinates of the normal distribution and standard errors computed as the square root of the diagonal of each covariance matrix estimate.

Table 2 summarizes the results for $n = 100$, pooling over all the coefficient parameters. The second column reports the root mean squared error (RMSE) of the coefficient estimates compared to the values used to generate all (conditions 1 and 2) or most (conditions 3–6) of the data. The results illustrate that the tanh estimator gives accurate point estimates even when there is contaminated data: the RMSE is small in every condition. The nonrobust ML estimator need not give accurate point estimates given contaminated data: the RMSE is small when there is no contamination or only slight contamination but very large when there is serious contamination. In particular, in experimental conditions 5 and 6 the RMSE is 1.11. The SUR and heteroscedastic SUR models give point estimates similar to the nonrobust ML estimator.[6] The AL model has accurate point estimates in every condition, although the RMSE of the estimates is larger than the RMSE using the tanh estimator. Similar results (not shown) occur when $n = 50$.

---

able from the Comprehensive R Archive Network (CRAN, `http://cran.r-project.org/`). Standard errors are corrected for degrees-of-freedom.

[5] To estimate the heteroscedastic SUR model we used FORTRAN code originally written by John Jackson and slightly modified by us to suit our simulated data.

[6] The heteroscedastic SUR model fails to converge for a number of replications in experimental condition 5 with $n = 100$ and in conditions 1, 3 and 5 with $n = 50$. Convergence fails because the estimated covariance matrix becomes indefinite. This occurs because the model Jackson (2002) defines features a shortcut that implies that a component of the error variance is double counted. The matrices he denotes Σ_{v_i} are computed using the observed sample proportions, not the values predicted by the model. This means that the error variances and covariances that result from the model's not perfectly reproducing the observed proportions affect both Σ_{v_i} and the expected variance-covariance matrix of the residuals that he computes (Jackson 2002, 64, eq. A3).

*** Table 2 about here ***

Table 2 also shows that the estimated covariance matrices of the tanh estimates support confidence intervals that are approximately correct even when there is contamination. The coverage results in the table report the proportion of replications in which the nominal 90% or 95% confidence interval contain the parameter value that generated all (conditions 1 and 2) or most (conditions 3–6) of the data. With no contamination and no overdispersion (condition 1), the intervals based on the Huber-White sandwich estimator under cover by about three ($n = 100$) percent (for $n = 50$, four percent). The intervals based on the weighted Hessian do slightly better in this condition, and the intervals based on the weighted OPG are even more accurate, under covering by one percent or less. The results with overdispersion (condition 2) are similar. In the other experimental conditions the coverage of the sandwich and weighted Hessian estimators typically improves by about one percent, while the weighted OPG intervals continue to be basically accurate. All three interval estimators have reasonably good coverage even with contaminated data.

In contrast, Table 2 shows that nonrobust confidence intervals are essentially worthless when there is contamination. Both of the nonrobust ML interval estimators produce correct coverage when there is neither contamination nor overdispersion (condition 1). When there is overdispersion but not contamination, correct coverage occurs only when the estimator takes the dispersion into account (condition 2). Given contamination and the symmetric outcome probabilities (conditions 3 and 4), the intervals that are based on ignoring overdispersion include the target values in less than one-third of the replications. The intervals that take overdispersion into account almost always include the target values, because the intervals are too wide. Given contamination and the asymmetric outcome probabilities (conditions 5 and 6), we have the spectacular result that the intervals (with or without the dispersion correction) *never* include the target parameter values. The heteroscedastic SUR model performs similarly, with slightly greater degradation at the smaller sample size. The SUR model never gives correct coverage.

Table 2 shows that the accurate point estimates of the AL model are not matched by accurate confidence intervals. With no contamination and no overdispersion (condition 1), the intervals under cover on average by three or four percent for $n = 100$ (by five or six percent for $n = 50$). With overdispersion (condition 2) the under coverage typically worsens by one or two percent. With contamination (conditions 3–6) coverage performance degrades further, with under coverage ranging from six to ten percent for $n = 100$ (from seven to twelve percent for $n = 50$). The detailed results for each parameter show that Table 2 understates how inaccurate the AL model's confidence intervals are. For instance in condition 1, with parameters ordered as in Table 3, for $n = 100$ the nominal 90% intervals have coverages 0.91, 0.81, 0.91, 0.80, 0.92 and 0.81, and nominal 95% intervals have coverages 0.95, 0.89, 0.96, 0.88, 0.96 and 0.89.

The estimation results with contamination warrant detailed examination. Table 3 shows results for condition 3 (symmetric probabilities, 10% contamination and no overdispersion), with $n = 100$. We report the means and RMSEs of the estimates over replications and coverage results for the estimated confidence intervals. The tanh point estimates are accurate and the tanh intervals have good coverage for all parameters. The AL model has accurate point estimates but mostly incorrect confidence intervals. Intercept parameter intervals are correct or under cover only slightly, but intervals for the other coefficients under cover by as much as twenty percent.

*** Tables 3 about here ***

The contamination of ten percent of the data causes serious problems for the nonrobust estimators. Four of the nonrobust ML parameter estimates are biased: β_{10}, β_{30}, β_{11} and β_{31}. The confidence interval estimates for those parameters utterly fail to cover the target values. Coverage for the estimates that ignore overdispersion ranges from zero to 0.001—overdispersion should be ignored because there is no overdispersion in this condition. Notice that the estimates for β_{20} and β_{21} lack bias, and the overdispersion-ignoring confidence interval estimates for those parameters are accurate. These results reflect the success of our experimental manipulation, which sought to leave the estimates for these parameters undistorted. The nonrobust ML interval estimates that try to accommodate overdispersion all fail to have accurate coverage because they are too wide. The inaccuracy of the parameter estimates generates a biased—too large—estimate for the overdispersion, producing excessively large estimated standard errors. Results with the SUR and heteroscedastic SUR models are similar. Detailed results for $n = 50$ and condition 4 (not shown) are similar.

Table 4 shows detailed results for condition 5 (asymmetric probabilities, 10% contamination and no overdispersion), with $n = 50$. For the tanh estimator the results are similar to those for condition 3. The coverage results for the sandwich and weighted Hessian confidence intervals are slightly worse. The AL model again has accurate point estimates and incorrect confidence intervals. All the nonrobust estimates are seriously biased. Indeed, for β_{11} and β_{21} the mean nonrobust ML estimate has the opposite sign from the parameter values that generated 90 percent of the data. Interestingly, the two parameters that have incorrect signs are significantly different from zero according to confidence intervals constructed using the nonrobust ML covariance matrix. The confidence interval estimates from the nonrobust estimator utterly fail to cover the parameter values that generated 90 percent of the data: the intervals *never* include those values. Results with the SUR and heteroscedastic SUR models are similar. Detailed results for condition 6 (not shown) are similar.

*** Table 4 about here ***

The tanh weights w_{ij} correctly identify the contaminated observations. For the uncontaminated observations the weights have a median of 1 over all six conditions for $n = 100$ (mean 0.994, standard deviation 0.043), and for $n = 50$ the median is also 1 (mean 0.994, standard deviation 0.041).[7] For the contaminated observations, in conditions 3 through 6, the median weight is 0 for $n = 100$ (mean 0.029, standard deviation 0.143), and for $n = 50$ the median is also 0 (mean 0.039, standard deviation 0.171).[8]

---

[7] All six conditions have the same median. By condition, the means and standard deviations for $n = 100$ are 0.988 and 0.062, 0.988 and 0.061, 0.997 and 0.026, 0.997 and 0.026, 0.997 and 0.027, 0.996 and 0.033. For $n = 50$ they are 0.989 and 0.059, 0.989 and 0.057, 0.997 and 0.025, 0.997 and 0.025, 0.997 and 0.026, 0.996 and 0.034.

[8] All four conditions have the same median. By condition, the means and standard devi-

To summarize, the sampling experiment shows that the robust estimator performs well under a wide variety of circumstances: with or without contamination; with or without overdispersion; with symmetric or with highly skewed choice probabilities. Even when some data are contaminated and have high leverage regressors, point estimates for coefficient parameters are accurate and precise, and confidence interval estimates are accurate. In contrast, contamination in part of the data generally destroys the nonrobust estimators. When there is contamination, the nonrobust ML estimator produces estimates that exhibit substantial bias, including incorrectly signed coefficient values. The nonrobust SUR and heteroscedastic SUR models fail similarly. The nonrobust confidence intervals are untrustworthy and useless: recall that in two experimental conditions (5 and 6) the intervals fail to cover the target parameters even one time in 1000 replications.

The AL model typically produces accurate point estimates but incorrect estimates of the sampling error and hence incorrect statistical inferences. The model fails even when there is no contamination because it ignores heteroscedasticity. The SUR model of Tomz et al. (2002) also has this defect. If the multinomial model holds, then the AL model does not satisfy a regularity condition necessary for asymptotically normal estimation (e.g. White 1994, 92, Assumption 3.2'): the counts become normal as the m_i values get large, so that the true value of the multivariate-t distribution's DF parameter goes to infinity. An indication of this occurs in the data for experimental condition 1, where for $n = 100$ the median estimate for the DF is 20.8, but 48 of the 1,000 estimates are greater than 10^6. For $n = 50$, the median DF is 47.4, but 109 estimates are greater than 10^6.

If there is substantial contamination, then the AL model fails to produce asymptotically normal estimates because the DF value becomes too small. If DF < 2, the distribution lacks a finite variance, and if DF < 1, the distribution lacks a finite mean. In such cases ML estimates are not asymptotically normal. DF values less than 2.0 or less than 1.0 occur frequently in the experiment data. In conditions 3 through 6 for $n = 100$, the median DF

---

ations for $n = 100$ are 0.000 and 0.000, 0.035 and 0.135, 0.009 and 0.061, 0.074 and 0.238. For $n = 50$ they are 0.000 and 0.000, 0.026 and 0.115, 0.019 and 0.092, 0.111 and 0.296.

estimates are respectively 1.2, 1.6, 0.96 and 1.1, and for $n = 50$ the median values are 1.2, 1.5, 0.94 and 1.1. Lange et al. (1989, 884) acknowledge that excessively small DF values may occur, observing that "the t model is not well suited to data with extreme outliers."

Florida in 2000

For the first example using real data we consider votes cast for president in the 2000 election in Florida. We compare the tanh estimator to the nonrobust ML estimator and the AL, SUR and heteroscedastic SUR models. Statistical inferences based on the tanh estimator match important features of the election that the other four estimators miss. As in the sampling experiment, the point estimates for the AL and tanh are similar, but statistical inferences based on the estimators differ substantially. Also as in the experiment, differences between the tanh and the other estimators are greater. The tanh results replicate and extend the key results of Wand et al. (2001) regarding the effects of Palm Beach County's butterfly ballot.

We judge the substantive results of the estimators in light of two key features of the 2000 presidential election in Florida. First, during 2000 Gore launched a mobilization drive throughout Florida that brought many voters into the electorate for the first time (e.g. Bonner and Barbanel 2000). About 40 percent of blacks voting in Florida in the 2000 election were new voters (Mintz and Keating 2000). This mobilization suggests that changes in Democratic registration between the 1996 and 2000 elections ought to be important for the Gore vote. Second, not only did the Elián González episode provoke an extremely negative reaction to Gore among many Cuban-Americans, especially in Miami (e.g. Forero and Barringer 2000; Toobin 2001, 149), but Cuban-Americans tend to be strongly Republican (Alvarez and Bedolla 2003; DeSipio 1996; Moreno 1997). For instance, while on the whole Miami-Dade County favored Gore over Bush by 53 percent to 46 percent, in that county Bush received 75.8 percent of the two-party vote in census tracts in which 50 percent or more of the population is Cuban-American.[9]

[9] The percentage is computed from data in Florida Legislative Staff (2001).

We analyze the number of votes cast in Florida in 2000 for presidential candidates Buchanan, Nader, Gore, Bush and a residual category consisting of votes for all of the other candidates. The candidates in the residual category include Harry Browne (Libertarian), Howard Phillips (Constitution Party), John Hagelin (Natural Law Party) and any other candidate listed on the ballot as well as any write-in candidates. We ignore undervotes (no apparent vote recorded on the ballot), overvotes (votes for more than one presidential candidate on a single ballot) and other spoiled ballots. We use all five estimators assessed in the sampling experiment to analyze county-level data from Florida's 67 counties. Using an improved set of regressors, the tanh estimates replicate the basic findings of Wand et al. (2001) regarding the vote for Buchanan in Palm Beach County. The tanh estimates also show that having a higher proportion of Cuban-Americans in a county produced more support for Bush than for Gore, and that increases in Democratic registration produced more votes for Gore. The other estimators fail in various ways to produce these results.

For the county-level model, we use linear predictors μ_{ij} that are functions of presidential vote proportions in the 1996 election, changes in party registration proportions from 1996 to 2000, the proportion of the population in each county in the 2000 Census that is of Cuban national origin, and a principal component computed using the same nine demographic variables that were used in Wand et al. (2001, 796–797).[10] The idea is that the vote for a party's candidate in the previous presidential election is a proxy for the interests, party sentiments and local party and other organization in each county, while the collection of demographic variables picks up changes during the intervening time period. The party registration variables for each county should provide sharper measures of the political changes

[10]The demographic variables are: the 2000 Census of Population and Housing proportions of county population in each of four Census Bureau race categories (White, Black, Asian and Pacific Islander, and American Indian or Alaska Native), 2000 proportion Hispanic, 2000 population density (i.e., 2000 population/1990 square miles), 2000 population, 1990 proportion of population with college degree, and 1989 median household money income. See Wand et al. (2001, 796) for sources.

than the demographics alone do, and so their inclusion represents an important substantive improvement over Wand et al. (2001). We use a principal component instead of the separate demographic variables to enhance the efficiency and interpretability of the other coefficients—the demographic variables are nearly aliased (McCullagh and Nelder 1989, 61–62) with previous vote, party registration and Cuban population.

With $J = 5$, the linear predictors may be written as follows.

$$\mu_{ij} = \begin{cases} \beta_{j0} + \beta_{j1}\mathrm{V96}_{ij} + \beta_{j2}\Delta\mathrm{R00}_{ij} + \beta_{j3}\mathrm{Cuban}_i + \beta_{j4}\mathrm{PC}_{ij}, & j = 1, \ldots, 4, \\ 0, & j = 5. \end{cases} \tag{2}$$

The correspondence between candidates and categories is Buchanan ($j = 1$), Nader ($j = 2$), Gore ($j = 3$), Bush ($j = 4$) and Other ($j = 5$). There are 20 unknown coefficient parameters, $\beta = (\beta_{10}, \ldots, \beta_{44})'$. The $\mathrm{V96}_{ij}$ variables measure the proportion of each county's votes for various presidential candidates in 1996, out of all valid votes cast. $\mathrm{V96}_{i1}$ is the proportion for Ross Perot (Reform), $\mathrm{V96}_{i2}$ is the sum of the proportion for Nader (Green) and the proportion for Bill Clinton (Democrat),[11] $\mathrm{V96}_{i3}$ is the proportion for Clinton, and $\mathrm{V96}_{i4}$ is the proportion for Bob Dole (Republican). The $\Delta\mathrm{R00}_{ij}$ variables measure changes from 1996 to 2000 in party registration. $\Delta\mathrm{R00}_{i1}$ and $\Delta\mathrm{R00}_{i4}$ are both the change in the proportion Republican among registered voters in county i,[12] and $\Delta\mathrm{R00}_{i3}$ is the change in the proportion Democratic. Because Green Party registration in Florida in 1996 was so rare as to be uninformative for voting behavior in 2000, we reduce $\Delta\mathrm{R00}_{i2}$ to simply the proportion Green among registered voters in 2000. Cuban_i denotes the proportion Cuban-American in county i. Applying the same method used by Wand et al. (2001, 797), each PC_{ij} variable is the

[11] Alternative specifications in which $\mathrm{V96}_{i2}$ includes only the 1996 proportion for Nader fit the data worse than does the definition we use here.

[12] Defining $\Delta\mathrm{R00}_{i1}$ as the change in Reform party registration produces a worse fit to the data. In light of many Reform party members' resistance to the Buchanan takeover of the party in Florida (e.g. Garvey 2000), a weak relationship between Reform registration and support for Buchanan is not surprising.

first principal component of the set of standardized residuals produced by regressing each demographic variable on a constant, $\mathrm{V96}_{ij}$, $\Delta\mathrm{R00}_{ij}$ and Cuban_i. The principal components are computed separately for each linear predictor.[13] Hence the principal components vary across the predictors.

Table 5 presents the tanh estimation results, with sandwich standard errors. Votes in 2000 are significantly related to the 1996 election results for all of the candidates except Nader. Changes in voter registration between 1996 and 2000 matter for all of the candidates. The estimated effect of Democratic registration changes on votes for Gore (β_{32}) is not significantly less than the estimated effect of Republican registration changes on votes for Bush (β_{42}). The proportion Cuban-American has significant effects in the linear predictors for Buchanan, Gore and Bush. For Buchanan the effect is large and negative (-5.73), while the effects are positive for the other two candidates, larger for Bush (2.75) than for Gore (2.03). The discrepancy between the estimates for Gore and Bush, which is larger than two standard errors, represents a significant tendency for Cuban-Americans to support Bush more than Gore, net of previous voting history or current partisanship.

*** Table 5 about here ***

Figure 1 summarizes the results for all five models in a graphical form that facilitates comparisons across models. For each model's estimate of each coefficient, the Figure plots the value of $\hat{\beta}_{jk}$ and the usual 95% confidence interval ($\hat{\beta}_{jk} \pm 1.96$ times standard error). The nonrobust ML estimator's standard errors take overdispersion into account.[14] The nonrobust

---

[13]In each case we standardize each of the nine vectors of residuals to have variance equal to 1.0 before computing the principal component.

[14]Other model parameters: Nonrobust ML: $\hat{\sigma} = 8.47$. AL: DF = 10.5, $\hat{\sigma}^2_{11} = 0.177$, $\hat{\sigma}^2_{22} = 0.156$, $\hat{\sigma}^2_{33} = 0.110$, $\hat{\sigma}^2_{44} = 0.099$, $\hat{\sigma}^2_{21} = 0.061$, $\hat{\sigma}^2_{31} = 0.090$, $\hat{\sigma}^2_{32} = 0.103$, $\hat{\sigma}^2_{41} = 0.085$, $\hat{\sigma}^2_{42} = 0.099$, $\hat{\sigma}^2_{43} = 0.102$. SUR: $\hat{\sigma}^2_{11} = 0.249$, $\hat{\sigma}^2_{22} = 0.174$, $\hat{\sigma}^2_{33} = 0.131$, $\hat{\sigma}^2_{44} = 0.116$, $\hat{\sigma}^2_{21} = 0.053$, $\hat{\sigma}^2_{31} = 0.109$, $\hat{\sigma}^2_{32} = 0.114$, $\hat{\sigma}^2_{41} = 0.104$, $\hat{\sigma}^2_{42} = 0.100$, $\hat{\sigma}^2_{43} = 0.119$. Heteroscedastic SUR: $\hat{\sigma}^2_{11} = 0.251$, $\hat{\sigma}^2_{22} = 0.170$, $\hat{\sigma}^2_{33} = 0.126$, $\hat{\sigma}^2_{44} = 0.116$, $\hat{\sigma}^2_{21} = 0.055$, $\hat{\sigma}^2_{31} = 0.112$, $\hat{\sigma}^2_{32} = 0.113$, $\hat{\sigma}^2_{41} = 0.105$, $\hat{\sigma}^2_{42} = 0.108$, $\hat{\sigma}^2_{43} = 0.118$.

ML, AL, SUR and heteroscedastic SUR results differ significantly from the tanh results for several coefficients. For instance, all four estimators produce insignificant estimates for the effects of change in Democratic party registration on votes for Gore, and all four significantly underestimate the effect of Green party registration on votes for Nader. The estimate of the effect of changes in Democratic party registration is important not only because of the reported pattern of Cuban-Americans dropping their Democratic registrations in Miami-Dade county, but also because of the mobilization drive Gore launched in 2000 throughout Florida that brought many voters into the electoral system for the first time. The tanh estimate is the only one that supports concluding that these patterns of disenchantment and mobilization had a significant effect on the Gore vote.

*** Figure 1 about here ***

The models convey significantly different impressions about how Cuban-Americans voted. The tanh estimator is the only one to produce a significant effect for Cuban_i on votes for Buchanan. Buchanan's anti-immigrant reputation makes the tanh estimate more plausible than the others. And the tanh estimator is the only one to support an inference that Cuban-Americans were significantly more likely to vote for Bush than for Gore (i.e., an inference that $\beta_{33} \neq \beta_{43}$). The insignificant estimated differences are dubious in light of Cuban-Americans' pro-Republican bias and the particular hostility toward Gore sparked by the González affair.

The nonrobust ML estimator and the SUR model feature one estimate that appears to be significant but with the opposite sign from the tanh estimate: the effect of the principal component on votes for Bush. The AL and heteroscedastic SUR models also estimate a negative value for this parameter, but the estimates those models produce do not appear to be statistically significant and indeed fall within the tanh estimator's 95% confidence interval. The nonrobust ML estimates also have a significant sign reversal for the effect of the principal component on votes for Gore. Even though effects associated with the principal components are not readily interpretable, these results demonstrate that the pattern of sign reversals illustrated in the sampling experiment can occur in practice with real data.

Table 6 lists all the counties that contain a studentized residual of magnitude greater than 4.0, which typically implies that the corresponding count receives a weight of zero in the analysis.[15] To facilitate the presentation each studentized residual is computed after permuting the categories to place the referent candidate in the first position, i.e., in Table 6 all the residuals displayed for each county are in the form $\tilde{r}_{i1}$ (defined in Appendix equation (A-3)). These residuals have the virtue of being readily associated with the candidates.

*** Table 6 about here ***

The residuals reveal that to diagnose the effect Palm Beach County's butterfly ballot had on would-be Gore voters, it is important to focus on the vote for Buchanan. In Palm Beach County the value (22.79) is large for Buchanan while the value for Gore is negative but not large. The number of votes that went to Buchanan by mistake because of the butterfly ballot is a very high proportion of the total number Buchanan received in Palm Beach County. According to Wand et al. (2001), somewhere between 2,000 and 3,000 of Buchanan's 3,411 votes were mistaken would-be Gore votes. But the number is a tiny fraction of Gore's vote total (269,732) there.

The 1993 Polish Parliamentary Election

For the second example we use the tanh estimator to reestimate Jackson's (2002) model for votes in the 1993 election for the lower House of the Polish parliament.[16] For several important parameters, the tanh estimator produces sharper inferences than the heteroscedastic SUR model that Jackson (2002) used to analyze the same data.

We estimate the form specified for the model and estimated by Jackson (2002). Votes are aggregated to the level of the Polish province, or voivodship, producing 49 voivodship

[15]A studentized residual of magnitude greater than 4.0 need not imply that $w_{ij} = 0$ for the j that indexes that count because the residuals in Table 6 have each category permuted to be in the first position, and to fully studentize each residual the residual is divided by the "hat matrix" element $(1 - h_{ij})^{1/2}$ (see Appendix equation (A-3) for details).

[16]Jackson (2002) compares the SUR and heteroscedastic SUR models.

observations. The total number of votes cast in each voivodship (m_i) ranges from 83,840 to 1,323,540. The analysis focuses on the votes cast for six party groupings: the Democratic Union plus Liberal Democratic Congress (UD+KLD); the Democratic Left Alliance (SLD); the Polish Peasants' Party (PSL); the Union of Work (UP); a coalition of Catholic parties; and all other parties (Other). Jackson (2002) and Jackson, Klich, and Poznańska (2003) define and motivate the regressors used in the analysis: the proportion of jobs in new and small private firms; the unemployment rate; the proportion of jobs in state-managed firms; the proportion of people attending church; mean years of schooling; mean age; the proportion of the population who are farmers; the proportion of the population living in villages. The UD+KLD is treated as the reference party, so that except for a dummy variable that indicates the home voivodship of the UD party leader, Hanna Suchocka, all coefficients are zero in this party's linear predictor. Otherwise all the regressors appear in the linear predictors for the other five parties, except that the farmers and villages variables appear only in the linear predictor for PSL. Finally in each of the linear predictors for SLD, PSL and UP there is a dummy variable to indicate the home of the respective party leaders.

We compare the tanh estimates to the estimates obtained using Jackson's (2002) heteroscedastic SUR model. All of the effects that are statistically significant with the heteroscedastic SUR model (see Jackson 2002, Table 1) are also statistically significant with the tanh estimator. But reflecting the sampling experiment result that the heteroscedastic SUR model frequently produces excessively wide confidence intervals, the tanh model estimates a few effects to be significant that the heteroscedastic SUR model does not.

Some of these differences are substantively important. In Jackson (2002, Table 1) the constant is estimated to be large but insignificant for the SLD and the PSL, two post-communist parties (Jackson et al. 2003, 91–92). In Table 7, which reports the tanh results, the estimated constant is significant for the SLD, but it is much smaller in magnitude and not significant for the PSL. Notwithstanding their statistical insignificance, Jackson interprets the constants he estimates for the SLD and the PSL as "reflecting broad dissatisfaction with the consequences of the harsh economic reforms" (Jackson 2002, 55). The tanh estimates suggest that while such an interpretation may hold for the SLD, for the PSL vote support was more purely contingent on voivodship-level factors. Another important difference from the heteroscedastic SUR results is the effect jobs in state-managed firms are estimated to have on votes cast for the UP and for the Other parties. The tanh estimates are significantly positive, but the same effects are not significant in Jackson (2002, Table 1). These results do not challenge Jackson's conclusion that support for the two post-communist parties, SLD and PSL, did not depend on employment in state-managed enterprises (Jackson 2002, 55), but the tanh estimates suggest it would be wrong to conclude that employment in state-managed firms had no significant effect on the election at all. That the UP's support in part depends on such employment resonates with what Jackson et al. (2003, 92) describe as the UP's opposition to privatization.

*** Table 7 about here ***

The sampling experiment shows that the heteroscedastic SUR model usually produces reasonably good coverage results when the model is correct and there is overdispersion. The tanh estimate of $\hat{\sigma}_{\text{tanh}} = 36.2$ indicates a large amount of overdispersion—much larger than the value of $\hat{\sigma}_{\text{tanh}} = 4.4$ estimated among Florida's counties in 2000. So the differences between the tanh and heteroscedastic SUR results must trace either to there being significantly different electoral processes in some voivodships or to some other kind of model misspecification. In the Polish data there are no outliers, meaning that the tanh estimator does not completely reject any observation by giving it a weight (w_{ij}) of zero. But one observation comes close to that status. The studentized residuals $\tilde{r}_{i1}$, computed as in Table 6 with each party successively placed in the first position to facilitate interpretation, show a value of $\tilde{r}_{i1} = 3.91$ for the Catholic parties in Bialystok Voivodship. The next largest value is $\tilde{r}_{i1} = 3.24$ for the SLD in Bydgoszcz Voivodship. Given the ordering of the categories—the SLD is the first party and the Catholic parties are ordered fourth—the weights associated with these residuals are $w_{i4} = 0.30$ for the Bialystok observation and $w_{i1} = 0.55$ for the Bydgoszcz observation. The remaining studentized residuals are all smaller than 3.0.

Nonrobust Estimation Declared Harmful

Nonrobust estimation is very likely to produce misleading results, often grossly misleading results such as seemingly significant coefficient estimates that have the wrong sign. Until recently the amount of computing required to calculate a good robust estimator was perhaps prohibitive, but nowadays the availability of cheap and plentiful computing power makes it feasible to apply robust estimation to a wide range of interesting models and data. Robust estimators with good properties have been available since at least the early 1980s for linear and generalized linear regression models (e.g. Huber 1981; Hampel et al. 1986; Stefanski, Carrol, and Ruppert 1986). Robust estimation software is available for a wide variety of models and data. Many statistical packages include redescending M-estimators for the linear model, including **R**, SAS, S-Plus and STATA. S-Plus offers the most comprehensive robust estimation software library, including routines for time-series data (Martin 1981), individual-level logistic regression (Carroll and Pederson 1993), Poisson regression (Künsch, Stefanski, and Carroll 1989) and covariance matrices (Rousseeuw and van Driessen 1999). Software for our estimator is available from the authors.

The estimator we have introduced in this article extends robust estimation technology effectively to models for count data. The results of the sampling experiment illustrate how erroneous and misleading the results of nonrobust estimation can be. If the regressors associated with them have high leverage, a small proportion of contaminated observations can cause coefficients to be estimated with apparent statistical significance but the wrong sign. Sign reversal due to such high leverage observations is a well known phenomenon in ordinary linear regression models (e.g. Rousseeuw and Leroy 1987, 5). In such cases the residuals from a nonrobust estimation will often not be large for the contaminated observations, so that the reason for the grossly wrong results—and even the fact that the results are wrong—may be masked (e.g. Atkinson 1986). If for no other reason, robust estimation should be used to provide insurance against the seriously misleading conclusions such grossly wrong estimates may appear to support. Even when results as bad as significant sign reversals do not occur, contamination will usually make nonrobust estimates inaccurate or otherwise distort estimates of sampling error variances, leading to incorrect inferences.

The robust estimation method we have introduced provides accurate parameter estimates and is a powerful technology for detecting irregular outcomes. Accurate parameter estimates can be produced, of course, only when the processes that generated most of the data are well approximated by the specified model. In some cases, outliers the estimator detects may be helpful in diagnosing problems with the model such as erroneously omitted variables. For instance, in the Florida data, estimating a model that omits the Cuban-American variable results in very large studentized residuals for Miami-Dade County, even larger than the residual found for Buchanan's vote in Palm Beach County.[17] As we previously observed, with the Cuban-American variable included, no outliers occur for Miami-Dade county. There may be many plausible explanations for an observed anomaly. Robust estimation and outlier detection are inherently part of a strategy of triangulation. Such an approach calls for mobilizing different kinds of knowledge, data and analysis and doing many different kinds of comparisons, often at different levels of observation and analysis. Wand et al. (2001) did that for the vote for Buchanan in Palm Beach County.

The tanh estimator is not the only approach to robust estimation with count data. For instance, the estimator developed by Victoria-Feser and Ronchetti (1997) could possibly be augmented to allow for overdispersion. The estimator proposed by Christmann (1994), using the least median of squares (LMS), could likewise be modified for overdispersion, although the low efficiency of LMS would be a limitation.

More work is needed to verify the estimator's performance with smaller sample sizes and with more complicated forms of contamination than we have examined here. Nonetheless we have great confidence that robust estimation using the tanh estimator is vastly superior to nonrobust estimation. Nonrobust estimation should be avoided whenever possible.

---

[17]In that model the studentized residual is $\tilde{r}_{i1} = -28.4$ for Gore in Miami-Dade and $\tilde{r}_{i1} = 31.3$ for Bush. For Buchanan in Palm Beach County in that model, $\tilde{r}_{i1} = 20.8$.

Appendix: Robust Estimation Method Details

To orthogonalize the residuals we use the formal Cholesky decomposition of the multinomial covariance matrix that was derived by Tanabe and Sagae (1992). The multinomial covariance matrix, $m_i(P_i - p_i p_i')$, has rank $J-1$. Tanabe and Sagae (1992) show that the matrix has a formal decomposition, $m_i(P_i - p_i p_i') = m_i L_i D_i L_i'$, where L_i is a lower triangular matrix (Tanabe and Sagae 1992, 213, eqn. 8), and D_i is a diagonal matrix with diagonal elements d_{ij}, with $d_{iJ} = 0$ (Tanabe and Sagae 1992, 213, eqn. 9). Both L_i and D_i are functions of the probabilities p_i. The covariance matrix may be diagonalized using the inverse of L_i, denoted L_i^{-1} (Tanabe and Sagae 1992, 213, eqn. 10): $m_i L_i^{-1}(P_i - p_i p_i') L_i'^{-1} = m_i D_i$. The diagonalization implies that if the probabilities were known, the residuals could be orthogonalized by multiplying the residual vector by L_i^{-1}, i.e., $r_i^{\perp} = L_i^{-1}(y_i - m_i p_i)$, because

$$E[r_i^{\perp}(r_i^{\perp})'] = L_i^{-1} E[(y_i - m_i p_i)(y_i - m_i p_i)'] L_i'^{-1} .$$

Because the entries in the last row of L_i^{-1} all equal 1, the last (i.e., J-th) element of $r_i^{\perp}$ is always zero. Hence the orthogonalized residuals $r_{ij}^{\perp}$, $j = 1, \ldots, J-1$, contain all the residual information.

We use the estimated probabilities $\hat{p}_{ij} = \exp(\hat{\mu}_{ij}) / \sum_{k=1}^{J} \exp(\hat{\mu}_{ik})$, where $\hat{\mu}_{ij} = x_{ij}' \hat{\beta}_j$ is the estimated linear predictor, to compute estimated inverse Cholesky factor matrices, $\hat{L}_i^{-1}$, and hence orthogonalized residuals $\hat{r}_i^{\perp} = \hat{L}_i^{-1} \hat{r}_i$, where $\hat{r}_i = y_i - m_i \hat{p}_i$. We also use $\hat{p}_{ij}$ to compute estimated Cholesky factors $\hat{d}_{ij}$, which we use to normalize the $J-1$ nontrivial values of $\hat{r}_i^{\perp}$ for each i. The resulting residuals are $\hat{r}_{ij}^{*} = \hat{r}_{ij}^{\perp}(m_i \hat{d}_{ij})^{-1/2}$, $j = 1, \ldots, J-1$ (note that $\hat{r}_{iJ}^{\perp} = 0$). Expansion of $\hat{r}_{ij}^{\perp}$ and $\hat{d}_{ij}$ gives the formula:

$$\hat{r}_{ij}^{*} = \begin{cases} \dfrac{\hat{r}_{i1}}{\sqrt{m_i \hat{p}_{i1}(1 - \hat{p}_{i1})}}, & j = 1 \\[2ex] \dfrac{\hat{r}_{ij} + \left(\sum_{k=1}^{j-1} \hat{r}_{ik}\right) \hat{p}_{ij} / \left[1 - \left(\sum_{k=1}^{j-1} \hat{p}_{ik}\right)\right]}{\sqrt{m_i \hat{p}_{ij} \left[1 - \left(\sum_{k=1}^{j} \hat{p}_{ik}\right)\right] / \left[1 - \left(\sum_{k=1}^{j-1} \hat{p}_{ik}\right)\right]}}, & 1 < j \leq J-1. \end{cases}$$

If the overdispersed multinomial model is correctly specified, then given a consistent estimate for β, a good moment estimator for σ^2 may be defined in terms of the $\hat{r}_{ij}^{*}$ values (compare

McCullagh and Nelder 1989, 168–169). Moreover, if the values $m_i p_{ij}(1-p_{ij})$ are sufficiently large, then the residuals $\hat{r}^*_{ij}$, $j=1,\ldots,J-1$, are approximately normal.[18]

Let the $n(J-1)$ residuals $\hat{r}^*_{ij}$, $i=1,\ldots,n$, $j=1,\ldots,J-1$, be indexed by $\ell=1,\ldots,n(J-1)$. With K being the number of unknown coefficient parameters in the model, define $h_K = \left\lceil \frac{n(J-1)+K}{2} \right\rceil$. We define the LQD estimator in terms of the $\binom{h_K}{2}$ order statistic of the set $\{|\hat{r}^*_{\ell_1} - \hat{r}^*_{\ell_2}| : \ell_1 < \ell_2\}$ of $\binom{n(J-1)}{2}$ absolute differences (Croux et al. 1994):

$$Q^*_{n(J-1)} = \{|\hat{r}^*_{\ell_1} - \hat{r}^*_{\ell_2}| : \ell_1 < \ell_2\}_{\binom{h_K}{2}:\binom{n(J-1)}{2}} \, .$$

The coefficient estimates $\hat{\beta}_{\mathrm{LQD}}$ minimize $Q^*_{n(J-1)}$. Let $\hat{Q}^*_{n(J-1)}$ designate the corresponding minimized value of $Q^*_{n(J-1)}$. The LQD scale estimate is

$$\hat{\sigma}_{\mathrm{LQD}} = \hat{Q}^*_{n(J-1)} \frac{1}{\sqrt{2}\Phi^{-1}(5/8)},$$

where Φ^{-1} is the quantile function for the standard normal distribution (Rousseeuw and Croux 1993, 1277). The approximate normality of the residuals $\hat{r}^*_\ell$ in the case of correct specification justifies the factor $1/[\sqrt{2}\Phi^{-1}(5/8)]$. We use GENOUD (Sekhon and Mebane 1998) to minimize $Q^*_{n(J-1)}$ because $Q^*_{n(J-1)}$ is not differentiable for all values of β and is not globally concave.[19]

The tanh estimator for β is a redescending M-estimator (Huber 1981, 100–103; Hampel et al. 1986, 149–152) based on the function:

$$\psi(u) = \begin{cases} u, & \text{for } 0 \le |u| \le p \\ (A(d-1))^{1/2} \tanh[\frac{1}{2}((d-1)B^2/A)^{1/2}(c-|u|)] \operatorname{sign}(u), & \text{for } p \le |u| \le c \\ 0, & \text{for } c \le |u| \end{cases}$$

where choices of c and d imply values for p, A and B.[20] The value of c is the truncation threshold, and d is the ratio between the *change-of-variance function*—the sensitivity of

[18]The discussion in Wand et al. (2001, 806–807) of the relationship between $m_i p_{ij}(1-p_{ij})$ and the residuals' approximate normality in binomial models also applies if $J > 2$.

[19]We use the **R** package `rgenoud` (version 1.20), available from CRAN.

[20]We use $c = 4.0$ and $d = 5.0$ which imply values $p = 1.8$, $A = 0.86$ and $B = 0.91$ as given

the estimator's asymptotic variance to a change in the data—and the asymptotic variance. The tanh estimator minimizes the asymptotic variance subject to that ratio.[21] Given scale estimate $\hat{\sigma}_{\text{LQD}}$ and trial estimates $\hat{\beta}$, we compute for each i the $J-1$ weights

$$w_{ij} = \begin{cases} \dfrac{\psi(\hat{r}^*_{ij}/\hat{\sigma}_{\text{LQD}})}{\hat{r}^*_{ij}/\hat{\sigma}_{\text{LQD}}}, & \text{for } \hat{r}^*_{ij} \neq 0 \\ 1, & \text{for } \hat{r}^*_{ij} = 0 \,. \end{cases}$$

A normalized residual that has $w_{ij} = 0$ (i.e., $|\hat{r}^*_{ij}/\hat{\sigma}_{\text{LQD}}| \geq c$) is an *outlier.*

To estimate β we use w_i and $\hat{L}_i$ to weight the gradient and the Hessian in a Newton algorithm (Gill, Murray, and Wright 1981, 105). The negative log-likelihood for a multinomial model is $l_i = -(\log p_i)' y_i$, the gradient with respect to μ_i is $\partial l_i/\partial \mu_i = -(y_i - m_i p_i)$, and the Hessian is $\partial^2 l_i/\partial\mu_i\partial\mu_i' = m_i(P_i - p_i p_i')$. The chain rule gives the gradient $(\partial\mu_i'/\partial\beta)(\partial l_i/\partial\mu_i)$ and Hessian $(\partial\mu_i'/\partial\beta)(\partial^2 l_i/\partial\mu_i\partial\mu_i')(\partial\mu_i/\partial\beta')$ with respect to β. Let W_i denote the $J \times J$ diagonal matrix that has $W_{i,jj} = w_{ij}$ for the diagonal values $j = 1, \ldots, J-1$ and $W_{i,JJ} = 1$. The weighted gradient with respect to β, evaluated at $\hat{\beta}$, is

$$\hat{s}_i = -\frac{\partial\hat{\mu}_i'}{\partial\hat{\beta}} \hat{L}_i W_i \hat{L}_i^{-1}(y_i - m_i\hat{p}_i) \,.$$

For the Hessian, we weight the components of the estimated Cholesky factor matrix $\hat{D}_i$ which has diagonal values $\hat{d}_{ij}$. Evaluated at $\hat{\beta}$, the weighted Hessian for the Newton algorithm is

$$G_i^* = m_i \frac{\partial\hat{\mu}_i'}{\partial\hat{\beta}} \hat{L}_i W_i \hat{D}_i W_i \hat{L}_i' \frac{\partial\hat{\mu}_i}{\partial\hat{\beta}'} \,.$$

Each iteration of the Newton algorithm uses steps proportional to

$$b = -\left(\sum_{i=1}^{n} G_i^*\right)^{-1} \left(\hat{\sigma}_{\text{LQD}}^{-1} \sum_{i=1}^{n} \hat{s}_i\right) \,.$$

---

in Table 2 in Hampel et al. (1981, 645). Hampel et al. (1981) use k for the ratio we have denoted by d. Alternatively see Table 2 of (Hampel et al. 1986, 163) where notation r and k is used for the parameters we have denoted by c and d.

[21] For details see Hampel et al. (1981, 645) or Hampel et al. (1986, 160–165).

We alternate rounds of LQD and tanh estimation (compare Huber 1981, 179–192). Each tanh round is a series of Newton optimizations that uses the preceding estimates $\hat{\beta}_{\mathrm{LQD}}$ to start the coefficients and the preceding LQD values $(\hat{r}^*_\ell - \mathrm{med}_\ell\, \hat{r}^*_\ell)/\hat{\sigma}_{\mathrm{LQD}}$ for an initial set of residuals, where $\mathrm{med}_\ell\, \hat{r}^*_\ell$ denotes the median of the $\hat{r}^*_\ell$ values, $\ell = 1, \ldots, n(J-1)$.

To estimate the asymptotic covariance matrix of the tanh coefficient estimates, $\Sigma_{\hat{\beta}}$, we use Huber's (1967, 231; 1981, 133) sandwich estimator. Let $s_i = -(\partial\mu_i'/\partial\beta) L_i W_i L_i^{-1}(y_i - m_i p_i)$ denote the weighted gradient for β known. Note that

$$\begin{aligned}
\partial s_i/\partial\beta' &= (\partial s_i/\partial\mu_i')(\partial\mu_i/\partial\beta') \\
&= \frac{\partial\mu_i'}{\partial\beta}\left[m_i L_i W_i L_i^{-1} L_i D_i L_i' + \frac{\partial(L_i W_i L_i^{-1})}{\partial\mu_i'}(y_i - m_i\hat{p}_i)\right]\frac{\partial\mu_i}{\partial\beta'} \\
&= m_i(\partial\mu_i'/\partial\beta) L_i W_i D_i L_i'(\partial\mu_i/\partial\beta') + z_i
\end{aligned}$$

where $z_i = 0$ if W_i is the identity matrix (no component of observation i is downweighted) and otherwise z_i is small. Hence using the weighted Hessian,

$$\hat{G} = \sum_{i=1}^{n} m_i \frac{\partial\hat{\mu}_i'}{\partial\hat{\beta}} \hat{L}_i W_i \hat{D}_i \hat{L}_i' \frac{\partial\hat{\mu}_i}{\partial\hat{\beta}'}\,, \tag{A-1}$$

and the outer product of the weighted gradient, $\hat{I} = \sum_{i=1}^{n} \hat{s}_i \hat{s}_i'$, the sandwich estimator is $\hat{\Sigma}_{\hat{\beta}} = \hat{G}^{-1}\hat{I}\hat{G}^{-1}$ (see also White 1994, 92).[22] We also consider two other covariance matrix estimators. One is $\hat{\Sigma}_{\mathrm{G}:\hat{\beta}} = \hat{\sigma}^2_{\mathrm{tanh}}\hat{G}^{-1}$, where, with $\hat{\beta}$ used to compute $\hat{r}^*_{ij}$,

$$\hat{\sigma}^2_{\mathrm{tanh}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{J-1}(\hat{r}^*_{ij})^2 w_{ij}}{\left(\sum_{i=1}^{n}\sum_{j=1}^{J-1} w_{ij}\right) - K}\,.$$

The other covariance matrix estimator we consider is $\hat{\Sigma}_{\mathrm{I}:\hat{\beta}} = \hat{I}^{-1}$.

To obtain studentized residuals (Carroll and Ruppert 1988, 31) for outlier diagnostics, we make a weighting adjustment for leverage (which applies to normalized residuals with $w_{ij} > 0$) or for forecasting error (which applies to the residuals with $w_{ij} = 0$). Let V_i denote

[22] For their special case with $J = 2$, Wand et al. (2001, 805) use an incorrect sandwich estimator, namely $\hat{\sigma}^2_{\mathrm{LQD}}\hat{G}^{-1}\hat{I}\hat{G}^{-1}$. The multiplication by $\hat{\sigma}^2_{\mathrm{LQD}}$ is a mistake. If there is overdispersion, that estimator produces variance estimates that are too large.

the $J \times J$ diagonal matrix that has diagonal values $V_{i,jj} = (m_i d_{ij})^{-1/2}$, for $j = 1, \ldots, J-1$, and $V_{i,JJ} = 0$. The first $J-1$ diagonal values of

$$H_i = V_i \hat{L}_i' \frac{\partial \hat{\mu}_i}{\partial \hat{\beta}'} \left(\sum_{i=1}^{n} \frac{\partial \hat{\mu}_i'}{\partial \hat{\beta}} \hat{L}_i V_i W_i V_i \hat{L}_i' \frac{\partial \hat{\mu}_i}{\partial \hat{\beta}'} \right)^{-1} \frac{\partial \hat{\mu}_i'}{\partial \hat{\beta}} \hat{L}_i V_i \tag{A-2}$$

provide robust estimates of the additional weights (compare McCullagh and Nelder 1989, 397; Carroll and Ruppert 1988, 31–34).[23] For $j = 1, \ldots, J-1$, let $h_{ij} = H_{i,jj}$ if $w_{ij} > 0$ and $h_{ij} = -H_{i,jj}$ if $w_{ij} = 0$ (note that $H_{i,JJ} = 0$). The studentized residuals are:

$$\tilde{r}_{ij} = \hat{r}_{ij}^* / \left(\hat{\sigma}_{\text{LQD}} \sqrt{1 - h_{ij}} \right), \qquad j = 1, \ldots, J-1. \tag{A-3}$$

For $J = 2$, $\tilde{r}_{i1}$ is the same as the residual $\tilde{r}_i$ of Wand et al. (2001, 806).

[23] (A-2) assumes that $\beta_J = 0$ so that $\mu_{iJ} = 0$ and the J-th column of $\partial \hat{\mu}_i' / \partial \hat{\beta}$ is zero.

References

Alvarez, R. Michael and Lisa García Bedolla. 2003. "The Foundations of Latino Voter Partisanship: Evidence from the 2000 Election." *Journal of Politics* 65 (Feb.): 31–49.

Atkinson, A. C. 1986. "Masking Unmasked." *Biometrika* 73 (Dec.): 533–541.

Bonner, Raymond and Josh Barbanel. 2000. "Duval County: Democrats Rue Ballot Foul-Up in a 2nd County." *New York Times*. November 17, Internet Edition.

Bratton, Kathleen A. and Loenard P. Ray. 2002. "Descriptive Representation, Policy Outcomes, and Municipal Day-care Coverage in Norway." *American Journal of Political Science* 46 (Apr.): 428–437.

Cameron, A. Colin and Pravin K. Trivedi. 1998. *Regression Analysis of Count Data*. New York: Cambridge University Press.

Canes-Wrone, Brandice, David W. Brady, and John F. Cogan. 2002. "Out of Step, Out of Office: Electoral Accountability and House Members' Voting." *American Political Science Review* 96 (Mar.): 127–140.

Card, David. 1990. "Strikes and Bargaining: A Survey of the Recent Empirical Literature." *American Economic Review* 80 (May): 410–415.

Carroll, Raymond J. and Shane Pederson. 1993. "On Robustness in the Logistic Regression Model." *Journal of the Royal Statistical Society Series B* 55 (3): 693–706.

Carroll, Raymond J. and David Ruppert. 1988. *Transformation and Weighting in Regression*. New York: Chapman & Hall.

Christmann, Andreas. 1994. "Least Median of Weighted Squares in Logistic-regression with Large Strata." *Biometrika* 81 (June): 413–417.

Croux, Christophe, Peter J. Rousseeuw, and Ola Hossjer. 1994. "Generalized S-Estimators." *Journal of the American Statistical Association* 89 (Dec.): 1271–1281.

DeSipio, Louis. 1996. *Counting on the Latino Vote: Latinos as a New Electorate*. Charlottesville: University of Virginia Press.

Donoho, D. L. and Peter J. Huber. 1983. "The Notion of Breakdown Point." In Peter J. Bickel, Kjell A. Doksum, and J. L. Hodges, editors, *A Festschrift for Erich L. Lehmann*, Belmont, CA: Wadsworth. Pages 157–184.

Famoye, Felix and Weiren Wang. 1997. "Modeling Household Fertility Decisions with Generalized Poisson Regression." *Journal of Population Economics* 10 (Aug.): 273–283.

Florida Legislative Staff. 2001. "FREDS 2000 dataset, PlanStatisticsT00.zip." November 16, 2001, URL `http://www.flsenate.gov/senateredistricting/freds_data.cfm`.

Forero, Juan and Felicity Barringer. 2000. "The Elian Gonzalez Case: The Scene; Police Fire Tear Gas as Hundreds of Angry Protesters Take to the Streets in Miami." *New York Times*. April 20, Internet Edition.

Garvey, Megan. 2000. "Bay Buchanan Sees Something Peculiar in Palm Beach Voting." *Los Angeles Times*. November 10, Internet Edition.

Gill, Philip E., Walter Murray, and Margaret H. Wright. 1981. *Practical Optimization*. San Diego: Academic Press.

Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Hampel, Frank R., Peter J. Rousseeuw, and Elvezio Ronchetti. 1981. "The Change-of-Variance Curve and Optimal Redescending M-Estimators." *Journal of the American Statistical Association* 76 (Sep.): 643–648.

Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. "Econometric Models for Count Data with an Application to the Patents R&D Relationship." *Econometrica* 52 (July): 909–938.

Huber, Peter J. 1967. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions." In Lucien M. Le Cam and Jerzy Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1. Berkeley, CA: University of California Press. Pages 221–233.

Huber, Peter J. 1981. *Robust Statistics*. New York: Wiley.

Jackson, John E. 2002. "A Seemingly Unrelated Regression Model for Analyzing Multiparty Elections." *Political Analysis* 10 (1): 49–65.

Jackson, John E., Jacek Klich, and Krystyna Poznańska. 2003. "Democratic Institutions and Economic Reform: The Polish Case." *British Journal of Political Science* 33 (1): 85–108.

Johnson, Norman L., Samuel Kotz, and Adrienne W. Kemp. 1993. *Univariate Discrete Distributions*. New York: Wiley second edition.

Kahn, Kim Fridkin and Patrick J. Kenney. 2002. "The Slant of the News: How Editorial Endorsements Influence Campaign Coverage and Citizens' Views of Candidates." *American Political Science Review* 96 (June): 381–394.

Katz, Jonathan N. and Gary King. 1999. "A Statistical Model for Multiparty Electoral Data." *American Political Science Review* 93 (Mar.): 15–32.

Keiser, Lael R., Vicky M. Wilkins, Kenneth J. Meier, and Catherine A. Holland. 2002. "Lipstick and Logarithms: Gender, Institutional Context, and Representative Bureaucracy." *American Political Science Review* 96 (Sep.): 553–564.

Künsch, H. R., L. A. Stefanski, and R. J. Carroll. 1989. "Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, with Applications to Generalized Linear Models." *Journal of the American Statistical Association* 84 (June): 460–466.

Lange, Kenneth L., Roderick J. A. Little, and Jeremy M. G. Taylor. 1989. "Robust Statistical Modeling Using the t Distribution." *Journal of the American Statistical Association* 84 (Dec.): 881–896.

Lau, Richard R. and Gerald M. Pomper. 2002. "Effectiveness of Negative Campaigning in U.S. Senate Elections." *American Journal of Political Science* 46 (Jan.): 47–66.

Martin, Doug R. 1981. "Robust Methods for Time Series." In D. F. Findley, editor, *Applied Time Series Analysis*, New York: Academic Press. Pages 683–759.

McCullagh, Peter and John A. Nelder. 1989. *Generalized Linear Models*. New York: Chapman & Hall.

McDonagh, Eileen. 2002. "Political Citizenship and Democratization: The Gender Paradox." *American Political Science Review* 96 (Sep.): 535–552.

Mintz, John and Dan Keating. 2000. "Fla. Black Votes Were More Likely Tossed." *Washington Post*. December 3, Internet Edition.

Monroe, Burt L. and Amanda G. Rose. 2002. "Electoral Systems and Unimagined Consequences: Partisan Effects of Districted Proportional Representation." *American Journal of Political Science* 46 (Jan.): 67–89.

Moreno, Dario. 1997. "The Cuban Model: Political Empowerment in Miami." In F. Chris Garcia, editor, *Pursuing Power: Latinos and the Political System*, Notre Dame: University of Notre Dame Press. Pages 208–226.

Rousseeuw, Peter J. and Christophe Croux. 1993. "Alternatives to the Median Absolute Deviation." *Journal of the American Statistical Association* 88 (Dec.): 1273–1283.

Rousseeuw, Peter J. and Annick M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: Wiley.

Rousseeuw, Peter J. and K. van Driessen. 1999. "A Fast Algorithm for the Minimum Covariance Determinant Estimator." *Technometrics* 41 (Aug.): 212–223.

Schrodt, Philip A. 1995. "Event Data in Foreign Policy Analysis." In Laura Neack, Jeanne A. K. Hey, and Patrick J. Haney, editors, *Foreign Policy Analysis: Continuity and Change in Its Second Generation*, Englewood Cliffs, NJ: Prentice Hall. Pages 145–166.

Sekhon, Jasjeet Singh and Walter R. Mebane, Jr. 1998. "Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models." *Political Analysis* 7: 189–203.

Stefanski, Leonard A., Raymond J. Carrol, and David Ruppert. 1986. "Optimally Bounded Score Functions for Generalized Linear Models with Applications to Logistic Regression." *Biometrika* 73 (Aug.): 413–424.

Tanabe, Kunio and Masahiko Sagae. 1992. "An Exact Cholesky Decomposition and the Generalized Inverse of the Variance-Covariance Matrix of the Multinomial Distribution, with Applications." *Journal of the Royal Statistical Society, Series B* 54 (1): 211–219.

Tomz, Michael, Joshua A. Tucker, and Jason Wittenberg. 2002. "An Easy and Accurate Regression Model for Multiparty Electoral Data." *Political Analysis* 10 (1): 66–83.

Toobin, Jeffrey. 2001. *Too Close to Call: The Thirty-six Day Battle to Decide the 2000 Election*. New York: Random House.

Victoria-Feser, Maria-Pia and Elvezio Ronchetti. 1997. "Robust Estimation for Grouped Data." *Journal of the American Statistical Association* 92 (Mar.): 333–340.

Wand, Jonathan, Kenneth Shotts, Jasjeet S. Sekhon, Walter R. Mebane, Jr., Michael Herron, and Henry E. Brady. 2001. "The Butterfly Did It: The Aberrant Vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* 95 (Dec.): 793–810.

Wang, T. Y., William J. Dixon, Edward N. Muller, and Mitchell A. Seligson. 1993. "Inequality and Political Violence Revisited." *American Political Science Review* 87 (Dec.): 979–994.

Western, Bruce. 1995. "Concepts and Suggestions for Robust Regression Analysis." *American Journal of Political Science* 39 (Aug.): 786–817.

White, Halbert. 1994. *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press.

Table 1: Monte Carlo Sampling Experiment Plan

Experimental Condition	Multinomial Probabilities	Contamination	Overdispersion
1	symmetric	none	no
2	symmetric	none	yes
3	symmetric	10%	no
4	symmetric	10%	yes
5	asymmetric	10%	no
6	asymmetric	10%	yes

Note: In each condition there are $J = 4$ categories, $n = 100$ observations and a total of $m_i = 10,000$ counts per observation. The symmetric outcome probabilities, used for the uncontaminated observations in conditions 1–4, have expected values of approximately 0.244, 0.244, 0.244 and 0.267. The asymmetric probabilities, used for the uncontaminated observations in conditions 5 and 6, have expected values of approximately 0.037, 0.060, 0.445 and 0.458.

Table 2: Monte Carlo Sampling Experiment Results Summary, $N = 100$

Experiment Condition	Hyperbolic Tangent: Coeff. RMSE	H-W Coverage 90%	H-W Coverage 95%	Hessian Coverage 90%	Hessian Coverage 95%	OPG Coverage 90%	OPG Coverage 95%
1	0.004	0.863	0.926	0.870	0.935	0.889	0.944
2	0.009	0.858	0.921	0.868	0.927	0.882	0.936
3	0.004	0.889	0.938	0.897	0.946	0.915	0.955
4	0.010	0.870	0.930	0.884	0.939	0.904	0.951
5	0.007	0.888	0.939	0.896	0.947	0.913	0.955
6	0.016	0.866	0.924	0.883	0.937	0.907	0.951

Nonrobust Maximum Likelihood: Coeff. RMSE	Coverage 90%	Coverage 95%	Coverage: $\hat{\sigma}^2$ 90%	Coverage: $\hat{\sigma}^2$ 95%	Additive Logistic: Coeff. RMSE	Coverage 90%	Coverage 95%
0.004	0.898	0.949	0.899	0.948	0.004	0.860	0.921
0.009	0.511	0.593	0.896	0.945	0.010	0.850	0.909
0.030	0.301	0.316	0.999	1.00	0.005	0.800	0.871
0.031	0.184	0.215	0.969	0.998	0.012	0.815	0.885
1.11	0.000	0.000	0.000	0.000	0.009	0.822	0.887
1.11	0.000	0.000	0.000	0.000	0.020	0.820	0.884

SUR: Coeff. RMSE	Coverage 90%	Coverage 95%	Heteroscedastic SUR: Coeff. RMSE	Coverage 90%	Coverage 95%
0.005	0.837	0.902	0.004	0.896	0.942
0.011	0.832	0.893	0.010	0.896	0.944
0.035	0.922	0.963	0.034	0.953	0.979
0.036	0.849	0.935	0.036	0.894	0.967
1.47	0.000	0.000	1.43[a]	0.345[a]	0.359[a]
1.48	0.000	0.000	1.46	0.000	0.000

Note: Based on 1000 replications for each condition. All results except "Coeff. RMSE" are reported to three significant figures.
[a] Using 795 converged replications.

Table 3: Summary for Experiment Condition 3: Symmetric Probabilities, 10% Contamination, No Overdispersion, $N = 100$

	Hyperbolic Tangent								Additive Logistic			
			H-W Coverage		Hessian Coverage		OPG Coverage				Coverage	
Symbol	Mean	RMSE	90%	95%	90%	95%	90%	95%	Mean	RMSE	90%	95%
β_{10}	−1.00	0.004	0.891	0.945	0.902	0.952	0.922	0.964	−1.00	0.005	0.885	0.937
β_{11}	1.00	0.003	0.892	0.944	0.901	0.948	0.917	0.955	1.00	0.005	0.789	0.864
β_{20}	−1.00	0.004	0.891	0.933	0.892	0.942	0.903	0.951	−1.00	0.005	0.823	0.886
β_{21}	1.00	0.004	0.877	0.929	0.881	0.940	0.908	0.955	1.00	0.005	0.709	0.803
β_{30}	−1.00	0.004	0.893	0.945	0.905	0.947	0.919	0.950	−1.00	0.005	0.851	0.908
β_{31}	1.00	0.004	0.890	0.934	0.900	0.946	0.923	0.957	1.00	0.005	0.743	0.829

Nonrobust Maximum Likelihood						SUR				Heteroscedastic SUR			
		Coverage		Coverage: $\hat{\sigma}^2$				Coverage				Coverage	
Mean	RMSE	90%	95%	90%	95%	Mean	RMSE	90%	95%	Mean	RMSE	90%	95%
−1.05	0.050	0.000	0.000	0.997	1.00	−1.07	0.066	1.00	1.00	−1.07	0.066	1.00	1.00
0.980	0.020	0.000	0.000	1.00	1.00	0.962	0.039	0.999	1.00	0.961	0.039	1.00	1.00
−1.00	0.004	0.901	0.947	1.00	1.00	−1.00	0.005	0.908	0.957	−1.00	0.004	0.894	0.939
1.00	0.003	0.907	0.946	1.00	1.00	1.00	0.005	0.754	0.830	1.00	0.003	0.881	0.936
−0.953	0.047	0.000	0.000	0.998	1.00	−0.970	0.031	0.953	1.00	−0.969	0.031	0.963	1.00
1.02	0.018	0.000	0.001	1.00	1.00	1.02	0.019	0.920	0.993	1.02	0.018	0.978	0.998

Note: Based on 1000 replications for each condition. All results except "Coeff. RMSE" are reported to three significant figures.

Table 4: Summary for Experiment Condition 5: Asymmetric Probabilities, 10% Contamination, No Overdispersion, $N = 50$

	Hyperbolic Tangent								Additive Logistic			
			H-W Coverage		Hessian Coverage		OPG Coverage				Coverage	
Symbol	Mean	RMSE	90%	95%	90%	95%	90%	95%	Mean	RMSE	90%	95%
β_{10}	−3.50	0.015	0.851	0.912	0.874	0.925	0.909	0.952	−3.50	0.019	0.738	0.809
β_{11}	0.999	0.010	0.847	0.912	0.889	0.943	0.934	0.968	1.00	0.012	0.775	0.846
β_{20}	−3.00	0.012	0.871	0.918	0.891	0.938	0.913	0.966	−3.00	0.016	0.752	0.832
β_{21}	1.00	0.008	0.863	0.930	0.895	0.948	0.938	0.969	1.00	0.010	0.783	0.852
β_{30}	−1.00	0.005	0.874	0.929	0.899	0.947	0.919	0.961	−1.00	0.006	0.857	0.915
β_{31}	1.00	0.004	0.870	0.924	0.884	0.948	0.931	0.971	1.00	0.005	0.850	0.905

Nonrobust Maximum Likelihood						SUR				Heteroscedastic SUR[a]			
		Coverage		Coverage: $\hat{\sigma}^2$				Coverage				Coverage	
Mean	RMSE	90%	95%	90%	95%	Mean	RMSE	90%	95%	Mean	RMSE	90%	95%
−2.16	1.34	0.000	0.000	0.000	0.000	−1.95	1.55	0.000	0.000	−1.97	1.54	0.237	0.245
−0.299	1.30	0.000	0.000	0.000	0.000	−0.41	1.41	0.000	0.000	−0.395	1.40	0.222	0.225
−1.81	1.19	0.000	0.000	0.000	0.000	−1.52	1.48	0.000	0.000	−1.54	1.46	0.264	0.268
−0.148	1.15	0.000	0.000	0.000	0.000	−0.371	1.37	0.000	0.000	−0.355	1.36	0.235	0.251
−0.525	0.475	0.000	0.000	0.000	0.000	0.102	1.10	0.000	0.000	0.092	1.09	0.264	0.267
0.552	0.448	0.000	0.000	0.000	0.000	0.042	0.958	0.000	0.000	0.054	0.949	0.255	0.262

Note: Based on 1000 replications for each condition. All results except “Coeff. RMSE” are reported to three significant figures.
[a] Results using 952 converged replications.

Table 5: Overdispersed Multinomial Model for 2000 Election Vote Counts, Florida Counties

	Buchanan		Nader	
	Coeff.	SE	Coeff.	SE
Constant	-0.312	0.212	1.05	0.190
1996 Vote Proportion	3.38	1.74	0.323	0.420
Change in Party Registration	14.4	1.46	1270	110
Proportion Cuban	-5.73	2.45	0.284	0.402
Principal Component	-0.0254	0.0257	-0.00686	0.0175

	Gore		Bush	
	Coeff.	SE	Coeff.	SE
Constant	3.37	0.159	4.13	0.137
1996 Vote Proportion	3.13	0.323	1.71	0.275
Change in Party Registration	1.80	0.910	2.04	0.744
Proportion Cuban	2.03	0.340	2.75	0.290
Principal Component	-0.0246	0.0111	0.00383	0.00988

Note: $n = 67$ counties. Entries are tanh estimates and sandwich standard errors.
Dispersion estimates: $\hat{\sigma}_{\mathrm{LQD}} = 5.06$; $\hat{\sigma}_{\mathrm{tanh}} = 4.45$.

Table 6: Outlier Florida Counties in the 2000 Election

County	Buchanan	Nader	Gore	Bush	Other	Total
	Candidate					
Alachua	0.57	−11.84	3.02	3.61	6.09	85,729
Broward	0.20	−0.16	−5.47	5.59	0.11	575,143
Duval	−1.57	−4.64	−4.51	5.54	2.26	264,636
Escambia	0.13	−2.67	5.30	−4.43	0.11	116,648
Leon	−0.48	−7.63	6.94	−4.06	0.99	103,124
Marion	1.33	0.81	3.76	−4.57	3.82	102,956
Martin	−0.54	−0.01	4.28	−4.06	−0.91	62,013
Orange	−0.26	−4.97	2.75	−1.31	0.16	280,125
Palm Beach	22.79	−1.68	−1.16	−0.48	0.91	433,186
Pasco	1.09	4.00	−7.20	5.99	0.50	142,731
Pinellas	1.48	0.29	−9.54	9.07	3.26	398,472
Santa Rosa	−1.03	−0.29	4.14	−4.13	0.37	50,319

Note: Entries are studentized residuals of the form $\tilde{r}_{i1}$, each computed by permuting the categories to place each candidate in the first position. The last column reports m_i.

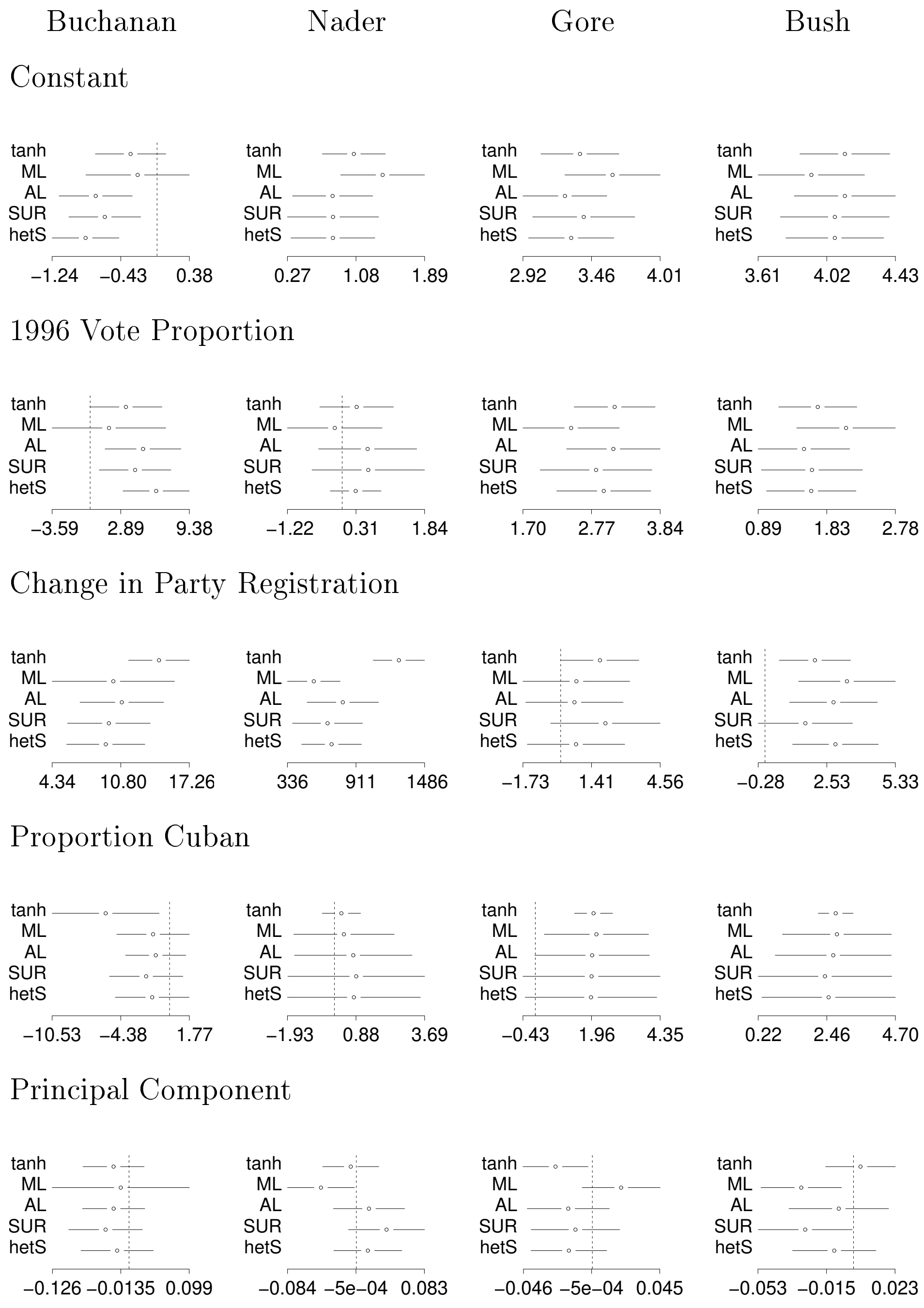
Table 7: Overdispersed Multinomial Model for 1993 Polish Parliamentary Elections

	SLD		PSL		UP	
	Coeff.	SE	Coeff.	SE	Coeff.	SE
Constant	5.87	2.08	2.08	3.66	1.72	2.32
Jobs in New Firms	−6.46	2.28	−7.81	3.13	−3.67	1.54
Unemployment	0.433	0.865	0.808	1.31	0.829	0.796
Jobs in State-enterprises	0.422	0.340	0.470	0.410	0.722	0.238
Church Attendance	−2.09	0.403	−1.54	0.525	−1.78	0.412
Years of Schooling	−0.390	0.092	−0.353	0.112	−0.275	0.074
Age/10	0.037	0.353	0.368	0.645	0.352	0.369
Farmers	—	—	1.99	0.900	—	—
Village	—	—	2.13	0.769	—	—
Party Leader	0.535	0.042	0.947	0.071	0.908	0.075

	Catholic		Other		UD+KLD	
	Coeff.	SE	Coeff.	SE	Coeff.	SE
Constant	−6.71	3.96	2.84	2.02	—	—
Jobs in New Firms	−3.22	2.64	−5.31	1.72	—	—
Unemployment	2.09	1.34	0.638	0.957	—	—
Jobs in State-enterprises	0.476	0.529	0.848	0.313	—	—
Church Attendance	2.49	0.583	0.183	0.406	—	—
Years of Schooling	−0.332	0.155	−0.398	0.086	—	—
Age/10	1.716	0.614	0.501	0.320	—	—
Party Leader	—	—	—	—	0.545	0.086

Note: $n = 49$ voivodships. Entries are tanh estimates and sandwich standard errors. Dispersion estimates: $\hat{\sigma}_{\text{LQD}} = 41.6$; $\hat{\sigma}_{\tanh} = 36.2$.

Figure 1: Five Models for 2000 Election Vote Counts, Florida Counties



Note: Each plot shows the point estimate and 95% confidence interval for the indicated coefficient using each estimator. Where shown, the vertical line marks the value zero. $n = 67$ counties.