Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data, with an Application to the Elián Effect in Florida[*]

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

May 23, 2002

[*] Previous versions of parts of this paper were presented at the 2002 Annual Meeting of the Midwest Political Science Association, Chicago, Illinois, April 25–28, at the 2001 Joint Statistical Meetings, American Statistical Association, Social Statistics Section, Atlanta, Georgia, August 5–9, and at the 2001 Political Methodology Summer Meeting, Atlanta, Georgia, July 19–21. We thank Jonathan Wand for his contributions to earlier versions of this work, Todd Rice and Lamarck, Inc., for their generous support and provision of computing resources, and Gary King for helpful comments on one of the earlier versions of this paper. The authors share equal responsibility for all errors.

[†] Associate Professor, Department of Government, Cornell University. 121 McGraw 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, with an Application to the Elián Effect in Florida

We develop robust estimation methods for regression models for vectors of counts (overdispersed multinomial models). The methods require only that the model is good for most—not all—of the observed data, and they identify outliers. A Monte Carlo sampling experiment shows that the robust methods can produce consistent parameter estimates and correct statistical inferences even when ten percent of the data are generated by a significantly different process, where nonrobust maximum likelihood estimation fails. We analyze Florida county vote data from the 2000 presidential election, considering votes for five categories of presidential candidates (Buchanan, Nader, Gore, Bush and "other"), focusing on Cuban-Americans' reactions to the Elián González affair. We replicate results regarding Buchanan's vote in Palm Beach County. We use Census tract data within Miami-Dade County to confirm the need to take the Cuban-American population explicitly into account. The analysis illustrates how the robust methods can support triangulation to verify whether a regression specification is adequate.

Introduction

Regression models for vectors of counts are commonly used in a variety of substantive fields. In international relations, count models have been widely used to analyze events of interest (Schrodt 1995). Such regression models have also often been used to examine domestic events including political violence (Wang, Dixon, Muller, and Seligson 1993) and labor relations (Card 1990). Other far-flung social science applications include research into the relationship between patents and R&D (Hausman, Hall, and Griliches 1984) and models of household fertility decisions (Famoye and Wang 1997). There are a multitude of other examples. But the counts political scientists most commonly study are votes.

Political scientists have been busy finding new ways to estimate models for counts of votes in multiparty elections (Honaker, Katz, and King 2002; Jackson 2002; Katz and King 1999; Tomz, Tucker, and Wittenberg 2002). The recently proposed methods all extend the basic multinomial model for count data (e.g. Cameron and Trivedi 1998, 270) in various ways. One common theme of the proposed methods is that they introduce sources of variation other than the usual variability associated with a multinomial model.[1]

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. One prominent example, which we analyze in this paper, is the vote in Florida for the 2000 U.S. presidential election, where many of the observed vote counts are not well represented by the model that describes the majority of the data. We develop a robust estimation method that provides accurate estimates and reliable inferences even when the model of interest is not a good model for a significant portion of the data. Our method helps to minimize the pathological effects of model misspecification, and outliers that the method automatically identifies as a byproduct of the estimation can facilitate diagnosing and correcting misspecification. There is, in general, no reason to believe that a single regression specification will be a good model for all of the counts in a particular application. This is especially true when, as with votes, the counts are aggregates of individual decisions that may be affected by idiosyncratic factors which are difficult if not impossible to observe at the aggregate level.

A single regression model almost certainly does not explain well what occurred throughout Florida in

[1] Katz and King (1999) introduce extra variation by using multivariate t-distributions with low degrees of freedom, instead of multivariate normal distributions, to model the variability of observed counts around their conditional means. With ordinary multinomial sampling, the asymptotic distribution of the counts would be normal. Jackson (2002) introduces extra variation by directly adding a variance component, which may include correlations across the choice categories, while retaining the assumption of asymptotic normality.

the 2000 election. After the voting concluded for the election, a lot of attention focused on Florida, especially on the legal and political fights over the recount efforts (e.g. Dershowitz 2001; Kaplan 2001; Posner 2001; Sunstein and Epstein 2001; Toobin 2001) and on the effects of Palm Beach County's butterfly ballot (e.g. Merzer and the *Miami Herald* 2001; Nichols 2001; Wand, Shotts, Sekhon, Mebane, Herron, and Brady 2001). Somewhat less discussed, although hardly unnoticed, were a number of other extraordinary political factors that contributed to Democratic Party candidate Al Gore's defeat. The 97,426 votes that went to Green Party candidate Ralph Nader in Florida were widely recognized to have helped the Republican, George W. Bush. Less than a one percent swing of votes from Nader to Gore would have erased Bush's margin of victory in Florida's certified election results, which was 537 votes. And there were events that inflamed sentiment against Gore in particular communities of voters. Most noteworthy was the Elián González episode, which provoked an extremely negative reaction among many Cuban-Americans, especially in Miami (e.g. Forero and Barringer 2000; Bragg 2000; Toobin 2001, 149).

The Florida example illustrates the general problem that a single regression model may not be valid for all of the counts in the data. Western (1995) issues a general call for robust estimation to be used with generalized linear models. Wand et al. (2001) demonstrate in particular 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, relative to the collection of regression specifications they use, vote counts for Buchanan in many counties throughout the U.S. seem to have been produced by processes unlike those that occurred in most of the counties in each state. The other unusual events in Florida suggest that dramatic departures from a single model for the state may not be confined to Buchanan's vote or only to Palm Beach County. In view of the typical messiness of elections it may be reasonable to believe that irregularities that thwart the sufficiency of any one regression model are not exceptional but rather the rule.

We introduce and demonstrate the effectiveness of methods that do not require that a single regression model is suitable for all the counts being analyzed. Instead, the methods require only that some one model is a good approximation for the processes that produced most of the observed counts. The methods we use in that sense provide robust estimation of the specified regression model. 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 regression 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 models we use introduce overdispersion to allow the counts to which the regression model applies to vary more than the ordinary multinomial sampling model would specify. The overdispersion can represent roughly the same kinds of extra variability that the assumption by Katz and King (1999) of a *t*-distribution can represent, but it does not allow for extra sources of covariation among counts as occurs in the model of Jackson (2002). The methods we introduce here generalize the methods for robust estimation of overdispersed binomial regression models that were introduced by Wand et al. (2001).[2]

Although robust estimation methods are not generally well known to political scientists (or econometricians), the methods have a long history—going back as far as Laplace (Stigler 1973). Until about 1885 the common wisdom among mathematical statisticians was that one should use the mean or the trimmed-mean and as a consequence least squares with the prior arbitrary removal of offending observations before estimation (Stigler 1973). By 1920 many scholars advocated robust alternatives to the sample mean and least squares (Stigler 1973). Even though robust methods have become far more sophisticated and rigorous since 1920, their use is still relatively rare. This is difficult to explain, but Hampel, Ronchetti, Rousseeuw, and Stahel (1986, 2) state that one historical reason is the great computational power it takes to estimate robust estimators relative to the rather computationally cheap (but more brittle) least squares and maximum likelihood alternatives. But today computational power is abundant, and algorithms exist that allow the difficult optimization problems that robust estimation presents to be solved in a routine manner. Moreover, research in statistical theory has produced rigorously developed models of qualitative and quantitative robustness that support methods that fulfill the three desirable features outlined by Huber (1981, 5–17): they have 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 methods' performance only slightly; and "somewhat larger deviations from the model should not cause a catastrophe" Huber (1981, 5). The robust methods we propose here have these good properties.

We begin with a brief description of the overdispersed multinomial regression model and our new robust estimation methods. Then we present the results of a Monte Carlo sampling experiment that demonstrates that the methods produce consistent parameter estimates and support correct statistical inferences even when the data are contaminated with counts that are generated by a significantly different process. We then

[2] In addition to extending the models from overdispersed binomial to overdispersed multinomial data, we also correct an error in the method Wand et al. (2001) used to estimate the sampling variability of the parameter estimates.

use the methods to analyze Florida vote data from the 2000 presidential election, extending the binomial (Buchanan versus the rest) model results of Wand et al. (2001) to a multinomial analysis of five categories of presidential candidates: Buchanan; Nader; Gore; Bush; and “other.” We especially focus on the huge shift in votes among Cuban-Americans that was evidently motivated by the Elián González affair and its aftermath. We illustrate how our methods indicate the need to take the Cuban-American community explicitly into account in an analysis of votes at the county level, and how the methods confirm the importance of such a specification in an analysis of votes counted at the level of Census tracts within Miami-Dade County. The results from Florida suggest that the outlier detection feature of our methods can support investigations to verify how adequately a chosen regression specification may be capturing the most important aspects of the processes that generated the observed counts.

Robust Estimation of an Overdispersed Multinomial Model

We use the overdispersed multinomial model for $J \geq 2$ outcome categories that is defined and motivated by McCullagh and Nelder (1989, 174). Let $i = 1, \dots, n$ index an observed vector of J counts $y_i = (y_{i1}, \dots, y_{iJ})'$, 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 $p_i = (p_{i1}, \dots, p_{iJ})'$ denote the vector of probabilities for observation i and let $P_i = \text{diag}(p_i)$ be a $J \times J$ diagonal matrix containing the probabilities. The model may depart from a standard multinomial because 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). If $\sigma^2 = 1$ then the covariance is the same as in an ordinary 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 $\mu_{ij} = x_{ij}'\beta_j$:

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

The coefficient parameters are subject to normalizing restrictions that are necessary to uniquely identify the model. Among all the vectors β_j there are K unknown coefficients, which we gather in a vector denoted β.

To estimate the model we extend the approach introduced by Wand et al. (2001), which means that we use robust estimators for σ^2 and β. Here we sketch the main features of the estimation method. Further details appear in the Appendix.

A point of departure for our methods is the fact that given any estimated probabilities $\hat{p}_{ij}$, the J residuals $(y_{ij} - m_i\hat{p}_{ij})$ for each i always sum to zero. This result follows from the basic 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. Hence, using a formal Cholesky decomposition of the multinomial covariance matrix that was derived by Tanabe and Sagae (1992), we can summarize the information the data contain about a model's fit in terms of $J-1$ orthogonalized residuals for each observation. If the model is correctly specified for all the data, then an estimate of the overdispersion parameter σ^2 may be used to studentize the orthogonalized residuals, such that asymptotically they all have a standard normal distribution. If the model is appropriate for only a majority of the data and the values of the model's parameters are known, then orthogonalized and studentized residuals computed using those parameters are typically large for the counts that were generated by the alternative processes. Ideally, the information from those counts would not be used to estimate the parameters of the model that applies to most of the data. The robust estimation methods we use approximate that ideal behavior. For a given regression model specification—i.e., a set of observed counts y_{ij}, regressors x_{ij} and linear predictor functional forms μ_i—the robust estimators find the parameter values that best characterize most of the data while downweighting information that is associated with ortho-studentized residuals that are larger than one would expect to observe in a sample of standard normal variates.

As do Wand et al. (2001), we use the *least quartile difference* (LQD) estimator (Croux, Rousseeuw, and Hossjer 1994; Rousseeuw and Croux 1993) to estimate the scale $\sigma = \sqrt{\sigma^2}$, and given the scale estimate, we use a *hyperbolic tangent* (tanh) estimator (Hampel, Rousseeuw, and Ronchetti 1981; Hampel et al. 1986, 160–166) for β. The estimation method of Wand et al. (2001) corresponds to the special case of our current method when there are only two choice categories. In that case the method produces for each observation a scalar weight valued in the unit interval $[0,1]$, with the value 1 indicating that the tanh estimator is giving full weight to the data from observation i and the value 0 indicating that the estimator is completely excluding information from observation i. In the general case with $J \geq 2$ choice categories, the estimation produces a vector of $J-1$ weights for each observation, $w_i = (w_{i1}, \ldots, w_{iJ-1})'$, with $w_{ij} \in [0,1]$. Each value w_{ij} measures the weight that the tanh estimator is placing on one of the $J-1$ orthogonal components of the residuals for observation i. The weights w_{ij} are functions of the ortho-studentized residuals obtained when the estimated probabilities $\hat{p}_{ij} = \exp(\hat{\mu}_{ij}) / \sum_{k=1}^{J} \exp(\hat{\mu}_{ik})$ are functions of the tanh coefficient estimates $\hat{\beta}$ via the estimated linear predictors $\hat{\mu}_{ij} = x'_{ij}\hat{\beta}_j$.

In terms of technological contributions, the robust estimation and outlier detection methods we introduce here are based on three primary innovations we make that we discuss further in the Appendix. (1) We use the formal Cholesky decomposition to reduce the multivariate robustness problem to a collection of uncorrelated problems. This allows the tanh estimates to be invariant to reordering of the outcome categories—see Aitchison (1986) for a discussion of the multinomial invariance problem. (2) We use an optimizing evolutionary program (Sekhon and Mebane 1998) to solve the difficult optimization problem involved in finding the LQD estimates. (3) Using the formal Cholesky decomposition, we find the functional forms of weighting to use for computing the tanh parameter estimates via a weighted Newton algorithm and for estimating 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 estimator under a range of conditions, we conducted a Monte Carlo sampling experiment using six different types of simulated data. We use the experiments to demonstrate four key points. First, the experiments show that the robust estimator provides consistent point estimates and accurate confidence intervals when the data contain no contamination. Of course, the nonrobust maximum likelihood (ML) estimates to which we compare the robust estimates also provide consistent point estimates and accurate confidence intervals when there is no contamination. Second, we show that when there is subtle contamination (i.e., nonrobust ML estimates at least have the correct sign), the robust estimator is still able to provide consistent point estimates and accurate confidence intervals, but the nonrobust ML point estimates are not consistent nor are the ML confidence intervals accurate for those parameters that are contaminated. Examining subtle contamination is of interest because it allows us to demonstrate our ability to identify outlier observations even when the contamination does not necessarily induce grossly wrong nonrobust estimation results. Third, we show that even when there is very serious contamination, the robust estimator produces consistent point estimates and accurate confidence intervals. This is remarkable because when the simulations have serious contamination, the nonrobust ML estimates include two significant and erroneous sign reversals, *none* of the maximum likelihood estimates are consistent, and the ML confidence intervals *never* include the true parameter values. Fourth, we show that the robust estimator provides consistent point and variance estimates in the presence of overdispersion even when there is contamination.

We examine six different experimental conditions in order to demonstrate the four key substantive points listed above. In each replication of each condition of the experiment, we generate $n = 100$ observations each consisting of four counts ($J = 4$) with $m_i = 10,000$. Each condition is replicated 1000 times. 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_{ij1}, and a constant, while the J-th predictor is set to zero:

$$\mu_{ij} = \begin{cases} \beta_{j0} + \beta_{j1} x_{ij1}, & (j = 1, \ldots, J-1), \\ 0, & (j = J). \end{cases}$$

In the different conditions the coefficients have different values or there is a different value for the overdispersion.

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 identical across the four conditions. The coefficient parameters are $\beta_{j0} = -1$ and $\beta_{j1} = 1$ for all the linear predictors, $j = 1, \ldots, J-1$. With this specification the expected outcome probability p_{ij} is approximately the same for all four categories.[3] The first experimental condition features uncontaminated multinomial data with no overdispersion. The absence of overdispersion means that $\sigma^2 = 1$. The second experimental condition is the same as the first except that it includes overdispersion. We used the cluster-sampling model (McCullagh and Nelder 1989, 174) to generate counts for which $\sigma^2 = 5.484964$. The third and fourth experimental conditions have contamination of ten percent of the data—i.e., ten percent of the data are not generated by the same process as the balance of the data. In the third condition, 90 of the 100 observations in each replication are produced the same way as in the first condition, but 10 of the observations are perturbed in such a way that the constant parameters in their linear predictors are

[3] The precise probabilities expected for the four categories are: 0.2442265, 0.2442265, 0.2442265 and 0.2673204.

approximately $\beta_{10} = -2.099$ and $\beta_{30} = -0.489$. That is, in terms of the parameters, in ten percent of the data β_{10} and β_{30} are distorted, while the other four parameters are the same for all observations. The fourth experimental condition is like the second condition except with 10 observations contaminated in each replication in the same way as in the third condition.

*** Table 1 about here ***

The fifth and sixth experimental conditions feature ten percent contamination with skewed outcome probabilities. For 90 of the 100 observations, the regressors are again normally distributed with mean one and variance one, but the parameter values are not the same as in the other four conditions. For these 90 observations 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 $p_{i1} = 0.037$, $p_{i2} = 0.060$, $p_{i3} = 0.445$ and $p_{i4} = 0.458$.[4] For the remaining 10 observations the regressors are normally distributed with a mean of $-.5$ and a variance of 4. The values of all the regressors are constant across replications. The parameters for the 10 contaminated observations are approximately $\beta_{10} = \beta_{20} = 0.001$, $\beta_{30} = 2.000$, $\beta_{11} = \beta_{21} = -2.000$ and $\beta_{31} = -1.000$. Unlike in the first four conditions, in conditions five and six 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 are likely to have high leverage (Carroll and Ruppert 1988, 31–33) and consequently induce nonrobust estimated regression lines to pass near the contaminated observations.

For each replication we compute both robust (tanh) and nonrobust (ML) 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 tanh estimates we compute confidence intervals based on three different estimates of the asymptotic covariance matrix of the coefficient estimates. The covariance matrix estimates, which are defined in the Appendix, are a Huber-White sandwich estimator (denoted $\hat{\Sigma}_{\hat{\beta}}$), an inverse weighted Hessian estimator ($\hat{\Sigma}_{G:\hat{\beta}}$) and an inverse weighted outer product of the gradient (OPG) estimator ($\hat{\Sigma}_{I:\hat{\beta}}$). We compute symmetric confidence intervals using normal ordinates and standard errors computed as the square root of the diagonal of each covariance matrix estimate. For the nonrobust ML estimates we compute confidence intervals using both the nonrobust inverse Hessian matrix alone and the nonrobust inverse Hessian multiplied by the usual estimate of dispersion (McCullagh and Nelder 1989, 175).

[4] More precisely the probabilities expected for the four categories are 0.03655627, 0.06027109, 0.44534649 and 0.45782615.

Table 2 summarizes the results over 1,000 replications of each experimental condition, pooling over all the coefficient parameters. The second column reports the mean error in the estimates, compared to the true values in the uncontaminated observations (i.e., the mean over the replications of $1'(\hat{\beta}-\beta)$). The third column reports the root mean squared error (RMSE) of the coefficients (i.e., the square root of the mean over the replications of $(\hat{\beta}-\beta)'(\hat{\beta}-\beta)$). For the tanh estimates the mean error is zero to three decimal places, except for condition six where the mean error is -0.001. The RMSE is small for the tanh estimates in every condition. The results illustrate that the robust estimator is consistent even when there is contaminated data. For the nonrobust ML estimates the mean error is zero when there is no contamination but noticeable when contamination is present. For example, with ten percent contamination and asymmetric probabilities the error is very large with a mean of -0.6 and RMSE of 1.1. These results illustrate that the nonrobust ML estimator is not consistent given the contaminated data.

*** 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. Columns four and five in the table report coverage results for nominal 90% and 95% confidence intervals based on the Huber-White sandwich estimator. With no contamination and no overdispersion (experimental condition one), the estimator produces on average correct coverage for both the 90% and 95% intervals; i.e., the proportion of nominal 90% intervals that include the true value for the sandwich intervals is 0.898 and the proportion of 95% intervals that include the true value is 0.950. The remaining columns of the table show that the robust Hessian and OPG based confidence intervals produce similar although slightly worse coverage. In the other experimental conditions the sandwich intervals produce slight undercoverage. For example, actual coverage for the nominal 90% intervals for sandwich estimator ranges from 0.871 to 0.883 while coverage for the nominal 95% intervals ranges from 0.928 to 0.940. As can be seen from the table, the robust Hessian and OPG intervals perform slightly better. In a way that gives more emphasis to large deviations, the RMSE values summarize how much the coverage results deviate from the nominal levels across the six parameters for each experimental condition. By this measure also, the deviations are small. No matter which of the three robust methods of constructing confidence intervals is used, coverage performance of the intervals is good even in the presence of contaminated data.

In contrast, Table 2 shows that nonrobust ML confidence intervals are essentially worthless when there is contamination. Both of the nonrobust covariance matrix estimates produce correct coverage when there is neither contamination nor overdispersion (condition one). When there is overdispersion but not contamination, correct coverage occurs only when the covariance matrix estimator takes the overdispersion into account (condition two). Given contamination and the symmetric outcome probabilities (conditions three and four), the nominal 90% and 95% ML confidence intervals that are based on ignoring overdispersion include the true values in less than one-third of the replications. The ML intervals that take overdispersion into account almost always include the true values, because the intervals are too wide. Given contamination and the asymmetric outcome probabilities (conditions five and six), we have the spectacular result that the ML intervals (with or without the dispersion correction) *never* include the true parameter values. In these two experimental conditions the ML confidence intervals have a type I error rate of 1—they always reject a true null hypothesis.

The results with contamination warrant close examination. We focus here on conditions three (symmetric probabilities, 10% contamination and no overdispersion) and five (asymmetric probabilities, 10% contamination and no overdispersion).[5] Table 3 shows detailed results for experimental condition three. Here we report first the true β parameter values and the mean of the estimates over replications. The means are reported to three significant figures of accuracy. The contamination of ten percent of the data has caused serious problems for the nonrobust estimates. Four of the nonrobust 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 true values. Actual coverage results for the estimates that ignore overdispersion range from zero to 0.001—overdispersion should be ignored because there is no overdispersion in this condition. Notably, 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 ML interval estimates that try to accommodate overdispersion all fail to have accurate coverage because they are too wide. The inconsistency of the parameter estimates generates a biased—too large—estimate for the overdispersion, producing excessively large estimated standard errors. The detailed results for experimental condition four (not shown here) exhibit a similar pattern.

*** Table 3 about here ***

Table 4 shows detailed results for experimental condition five. The nonrobust ML estimates are seriously biased. Indeed, for β_{11} and β_{21} the mean estimate has the opposite sign from the parameter values that generated 90 percent of the data. Interestingly, the two parameters that have the incorrect signs are significant

[5] Detailed results for the other conditions are available from the authors upon request.

according to the confidence intervals constructed using the 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 nonrobust confidence interval estimates *never* include those true values. The detailed results for experimental condition six (not shown here) are similar. In sharp contrast to the ML results for condition five, the robust estimates are very good. The mean robust point estimates are equal to the true parameters to three significant figures, and the RMSE for the point estimates is small. All three of the robust confidence intervals are good although, as mentioned before, they display slight undercoverage—particularly the Huber-White sandwich intervals. But they are remarkably good considering that the ML confidence intervals (using overdispersion or not) *never* include the true parameter values.

*** Table 4 about here ***

The weights w_{ij} correctly flag the contaminated observations as outliers. For uncontaminated observations the weights have a median over all six experimental conditions of 1 (mean 0.994, standard deviation 0.042).[6] For the contaminated observations, in experimental conditions three through six, the median weight is 0 (mean 0.029, standard deviation 0.143).[7]

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. As the statistical theory that motivates the estimator suggests, the estimator is robust to arbitrary kinds of contamination in even a moderately high proportion of the data. Even when some data are contaminated and have high leverage regressors, point estimates for coefficient parameters are consistent and precise, and confidence interval estimates are accurate. Estimated confidence intervals exhibit only a small degree of undercoverage: e.g., a nominal 95% interval actually includes the true value in only 93% or 94% of the Monte Carlo replications. All three of the estimators for the covariance matrix of the robust coefficient estimates perform well, producing very similar results. In contrast, contamination in part of the data generally destroys the nonrobust ML estimator. When there is contamination, the nonrobust estimator produces inconsistent estimates that in our experiments exhibit substantial bias, including incorrectly signed coefficient values. The nonrobust confidence intervals are untrustworthy and useless: recall that in two experimental conditions (five and six) the ML confidence intervals fail to cover the true

[6] All six experimental conditions have the same median. By experimental condition, the means and standard deviations are 0.989 and 0.058, 0.989 and 0.058, 0.996 and 0.032, 0.997 and 0.027, 0.997 and 0.027, 0.995 and 0.038.

[7] All four experimental conditions have the same median. By experimental condition, the means and standard deviations are 0.000 and 0.018, 0.035 and 0.135, 0.009 and 0.063, 0.073 and 0.237.

parameters even one time in 1000 replications.

Florida in 2000

Well before Theresa LePore had any reason to think about a butterfly ballot, there was Elián.[8] At the end of March, 2000, a news story reported a warning from Alex Penelas, the mayor of Miami-Dade County, that Cuban-American voters in the area would blame Gore if the Clinton administration returned six-year-old Elián González to Cuba and hence "civil unrest broke out" (Seelye 2000b). The same story reported that "many of the 40,000 Cuban-Americans who are registered Democrats are changing their party affiliation" (Seelye 2000b). The next day, Gore announced that he thought that "González and several family members should be given permanent residency status in the United States," a position at odds with the Clinton administration but matching a position Bush had adopted in January (Seelye 2000c). Leading Democrats denounced Gore's position (Dao 2000; Nagourney 2000), and in short order events overwhelmed his campaign plan. On April 22, Immigration and Naturalization Service agents and federal marshals removed González from the home of Miami relatives who had been keeping him, after which violence erupted (Forero 2000). Although many Republicans criticized the government action (Van Natta 2000), Cuban-American voters in Miami probably did not need their help to decide to punish Gore, which apparently they did despite late appeals from Gore and running mate Joseph Lieberman (Perez-Pena 2000).[9] In Miami-Dade, a county that on the whole favored Gore over Bush by 53 percent to 46 percent, news reports noted that Bush carried Cuban neighborhoods by 51 percent to 48 percent (Fessenden and Barbanel 2000). During the battle over recounts, one-time Gore supporter Penelas did not help the Democratic effort and indeed seemed to be dealing with the Bush camp (Filkins 2000; Van Natta and Filkins 2000).[10]

We use this background to inform an analysis of the number of votes cast in Florida in 2000 for presidential candidates Buchanan, Nader, Gore, Bush and a residual category which consists of votes for all of the other candidates. The candidates in the residual category include Harry Browne (Libertarian), Howard

[8] Engelhardt (2000) and Merzer and the *Miami Herald* (2001, 40–42) describe how LePore chose the ballot design in order to accommodate the ten presidential candidates that needed to be listed.

[9] Gore's very last appearance of the campaign, on the morning of the election, was before a group of Cuban-Americans in Tampa, Florida (Seelye 2000a). Lieberman had longstanding strong support from Cuban-Americans, including some Cuban exile group leaders (Canedy 2001). Among Democratic Senators, Lieberman received the fourth highest total of contributions from Cuban-Americans during 1979–2000 (Center for Responsive Politics 2001, Appendix).

[10] Reports at the time speculated that one aspect of a deal purportedly made in order to stop the recount in Miami-Dade County was that a new district dominated by Cuban-Americans would be created there during the 2002 reapportionment. Such a district was created (Allen 2002).

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 begin by using robust estimation of the overdispersed multinomial model to analyze county-level data from all of Florida's 67 counties. Using a slightly improved set of regressors, we replicate the basic findings of Wand et al. (2001) regarding the vote for Buchanan in Palm Beach County. Examining the results for the other candidates, especially Gore and Bush, we see that a model that does not include a variable to measure Cuban-American population sizes dramatically fails to explain the vote results in Miami-Dade County. When such a variable is added to the model, Miami-Dade County no longer appears anomalous and several of the estimated coefficients change in plausible ways. We then confirm the strong effect of Cuban-American voting in Miami-Dade County by estimating a model for the number of votes cast for Gore or Bush among Census tracts within the county.

For the initial model for Florida counties, 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, and a principal component computed using the same nine demographic variables that were used in Wand et al. (2001, 796–797).[11] As in Wand et al. (2001), 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 than the demographics alone do, and so their inclusion represents an important substantive improvement over Wand et al. (2001).

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

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

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 16 unknown coefficient parameters, $\beta = (\beta_{10},\ldots,\beta_{43})'$. The

[11] 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 source information.

V96_{ij} variables measure the proportion of each county's votes for various presidential candidates in 1996, out of all valid votes cast. V96_{i1} is the proportion for Perot (Reform), V96_{i2} is the sum of the proportion for Nader (Green) and the proportion for Bill Clinton (Democrat),[12] V96_{i3} is the proportion for Clinton, and V96_{i4} is the proportion for Bob Dole (Republican). The ΔR00_{ij} variables measure changes from 1996 to 2000 in party registration. ΔR00_{i1} and ΔR00_{i4} are both the change in the proportion Republican among registered voters in county i,[13] and ΔR00_{i3} is the change in the proportion Democratic. Because there was no Green registration in Florida in 1996, ΔR00_{i2} reduces to simply the proportion Green among registered voters in 2000. Applying the same method used by Wand et al. (2001, 797), each PC_{ij} variable is the first principal component of the demographic variables, computed using the standardized residuals from regressing each demographic variable on a constant, V96_{ij} and ΔR00_{ij}. The principal components are computed separately for each linear predictor, using the V96_{ij} and ΔR00_{ij} variables for that predictor to residualize the demographic variables. Hence the principal components vary across the predictors.

Table 5 presents the robust estimates for the parameters of the initial model, along with standard errors derived from the sandwich estimator $\hat{\Sigma}_{\hat{\beta}}$. The vote counts for 2000 appear to be significantly related to the 1996 election results for all of the candidates except Nader, and changes in voter registration between 1996 and 2000 appear to matter for all of the candidates except Gore (recall that for Nader the registration variable in effect measures the level of Green registration, because there was no Green registration in 1996). None of the effects of the principal component of the demographic variables are significant at the .05 level according to a conventional two-tailed t-statistic.

*** Table 5 about here ***

Of greater interest than the coefficient estimates with the initial model is the set of outliers. Any ortho-studentized residual value of magnitude greater than 4.0 is an outlier.[14] Table 6 lists all the counties that contain an outlier residual. To facilitate the presentation we have transformed the displayed residuals for all the candidate choices to be in the form $\tilde{r}_{i1}$ (defined in Appendix equation (A-3)); i.e., for the display in Table 6, each ortho-studentized residual is computed after permuting the categories to place the referent candidate in the first position. The displayed residuals are therefore not uncorrelated across candidates but have the

[12] Alternative specifications in which V96_{i2} includes only the 1996 proportion for Nader fit the data significantly worse than does the definition we use here.

[13] Defining ΔR00_{i1} as the change in Reform party registration produces a worse fit to the data. In light of the significant resistance that many Greens put up against the Buchanan takeover of the party in Florida (e.g. Garvey 2000), as happened even more intensely in some other states such as Colorado where John Hagelin, not Buchanan, was on the ballot as the Reform party candidate (Associated Press 2000), a weak relationship between Reform registration and support for Buchanan is not surprising.

[14] Any such value receives a weight $w_{ij} = 0$. See the Appendix for details.

great virtue, to facilitate interpretation, of being readily associated with the votes for a single candidate (which is not true for $\tilde{r}_{ij}$ for $j > 1$).

*** Table 6 about here ***

The first result to notice in Table 6 is the large value (20.76) for Buchanan in Palm Beach County. This result replicates the basic finding of the county-level analysis reported by Wand et al. (2001). Palm Beach County has by far the largest residual value for Buchanan. Reflecting the inclusion of voter registration information in the current analysis but not in the analysis of Wand et al. (2001), the Palm Beach County residual for Buchanan here is smaller than the value of 36.14 reported by Wand et al..

The residual for Gore in Palm Beach County is not an outlier. In order to diagnose the effect the butterfly ballot had on would-be Gore voters, it is important to focus on the vote for Buchanan. This is because the number of votes that went to Buchanan by mistake because of the butterfly ballot is a very high proportion of the total number votes he received, but the number of misguided ballots is only a tiny fraction of Gore's vote total in that county. According to Wand et al. (2001), somewhere between 2,000 and 3,000 of Buchanan's 3,411 votes in Palm Beach County were mistaken would-be Gore votes. Such a number of errors is hard to pick out among the 269,732 votes Gore did receive in Palm Beach County. In Palm Beach County there is a negative residual for Gore but that value has magnitude less than 4.0. In contrast to the results for Buchanan, the results for Gore do not draw attention to Palm Beach County.

The results for Gore do draw attention to Miami-Dade County. Gore's residual of -28.4 in Miami-Dade is exceeded in magnitude only by the residual of 31.3 in the county for Bush. These residuals both dwarf the Palm Beach County result for Buchanan. The magnitude of the raw residuals strongly suggest that the initial model is missing some important aspect of the voting in Miami-Dade County. For Gore in Miami-Dade County the value is $y_{ij} - m_i\hat{p}_{ij} = -50347.9$ while for Bush the value is 54728.8.

The news reports from Florida during the 2000 election period naturally suggest that omission of any variable to measure the distribution of Cuban-American voters among Florida's counties is a strong reason for the initial model's poor performance with the Miami-Dade vote counts. The demographic variables that are combined in the principal component do include a measure of the proportion of each county's population that is Hispanic. But in 2000 only about half of the Hispanic population in Miami-Dade County was of Cuban national origin (Associated Press 2001). Miami-Dade County remains an outlier for the Gore and Bush vote counts even if the Hispanic population variable is included directly in the model.

Hence we specify a model that directly includes a variable that measures the proportion of the population in each county that is of Cuban national origin. The data for this variable come from the 2000 U.S. Census. Using Cuban_i to denote the proportion Cuban-American in county i, the linear predictors for this more fully specified model are

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

Now the principal components are computed separately for each linear predictor, using the V96_{ij}, ΔR00_{ij} and Cuban_i variables for that predictor to residualize the demographic variables. There are 20 unknown coefficient parameters, $\beta = (\beta_{10}, \ldots, \beta_{44})'$.

Table 7 presents the robust estimates for the parameters of the more complete model, along with sandwich standard errors. 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. Indeed, a comparison between the estimated coefficients for the change in party registration variables in the initial model and in the more complete one shows a pattern that matches the news story from early in 2000 that reported that many Cuban-Americans were changing their registration away from being Democrats (Seelye 2000b). In the initial model the coefficient for the change in Democratic party registration in Gore's linear predictor (β_{32}) has an insignificant estimate, but in the more complete model the estimate is not only significantly different from zero but also not significantly less than the estimated coefficient for change in Republican party registration in Bush's linear predictor (β_{42}).[15]

*** Table 7 about here ***

Table 8, which lists all the counties that contain an outlier residual in the more complete model, shows that directly including the effects of the Cuban-American population eliminates the outliers for Gore and

[15] It is worth noting that nonrobust ML estimation with overdispersion taken into account gives substantially different results for several coefficient parameters of interest. For instance, the effects of the 1996 vote and of the proportion Cuban-American do not have significant estimates for Buchanan: $\hat{\beta}_{11} = 1.8$ with SE 2.7 and $\hat{\beta}_{31} = -1.5$ with SE 1.7. The effect of Green party registration in the Nader linear predictor is underestimated by a factor of two: $\hat{\beta}_{22} = 557$ with SE 113. The effect of change in Democratic party registration in Gore's linear predictor is insignificant: $\hat{\beta}_{32} = 0.7$ with SE 1.2. And the estimates of the effects of the principal components in the Gore and Bush linear predictors appear to be significant but with signs opposite those the robust estimator produces.

Bush not only in Miami-Dade County but also in Hillsborough County. For Gore in Miami-Dade County the value of $y_{ij} - m_i\hat{p}_{ij}$ is now -505.8 while for Bush the value is only 77.1. Comparing these raw residuals to the residuals produced by the initial model, one might crudely calibrate the magnitude of Cuban-Americans' punishment of Gore in Miami-Dade County—the magnitude of the Elián effect[16]—as a net swing of roughly 50,000 votes from Gore to Bush. It is provocative to note that of the four remaining negative outliers for Gore, one (Broward) county is among those in which Gore sought to have a manual recount conducted, while the other three (Duval, Pasco, Pinellas) are counties that news reports late in 2000 flagged as counties in which recounts might favor Gore (Fessenden and Barbanel 2000). This may suggest that further analysis of the Florida votes, aimed at eliminating the remaining outliers, should expand the data to include the invalidated votes. Such an extended analysis is beyond the scope of the current paper.

*** Table 8 about here ***

To try to confirm the impression from the county-level analysis that negative voting by Cuban-Americans cost Gore significant support, we examine votes across Census tracts within Miami-Dade County. The available data limit this analysis to an examination of votes for either Gore or Bush. The substantive results for Gore and Bush remain unchanged if we reestimate our county-level analysis looking only at Bush and Gore votes instead of the multiparty analysis we have conducted.

Also in this analysis we lack the data to compute changes in party registration at the Census tracts level, so the model includes only the proportion of registered voters who registered as Democrats in 2000. Several of the demographic variables are also not available as this level of analysis, so in addition to the proportion who have Cuban national origin we include directly the proportion who are Hispanic. We normalize the model by setting the linear predictor for Bush to zero. Robust estimates of the parameters of the resulting overdispersed binomial model appear in Table 9. The proportion Hispanic has a significant negative effect on Gore's vote, but on top of that the proportion Cuban-American has an additional significant negative effect. Notably this analysis produces no outliers among the Census Tracts.[17]

*** Table 9 about here ***

[16] Some may prefer to describe this as the Janet Reno effect, focusing on the Attorney General's decision to take the boy from his Miami relatives by force.

[17] Outliers do occur if the proportion Cuban-American variable is omitted.

Conclusion

One of the most important lessons to take from the Florida analysis is that an outlier may be a signal that an important variable has been left out of the analysis. Robust estimation can prevent a contaminated observation from influencing the coefficient estimates, but robust estimation does not thereby produce correct estimates if the contamination is related to misspecification that significantly affects the balance of the data. While the votes for Gore and Bush in Miami-Dade County do not affect the robust estimates of the initial model, the failure to include a measure of the Cuban-American population produces at least one result that is unlikely to be a correct description of how the voting in Florida typically went. In the initial model, the effect on votes for Gore of changes in Democratic party registration between 1996 and 2000 is estimated to be insignificant ($\hat{\beta}_{32} = 0.992$ with SE$= 1.01$).[18] It is very surprising to see that choices to register for a party are not related to votes for the party's presidential candidate, even more so when the lack of relationship occurs uniquely for the party of the outgoing president. Estimating the more complete model that includes a measure of the proportion who have Cuban nationality in each county largely resolves the problem. In that case we obtain $\hat{\beta}_{32} = 1.80$ with SE$= 0.91$, an estimate that is not only statistically significant but also is not significantly different from the estimated effect changes in Republican party registration had on votes for Bush ($\hat{\beta}_{42} = 2.04$ with SE$= 0.74$).[19] Such a result is much more compatible with our basic understanding of how electoral politics works: people who select an affiliation to a major party when they newly register to vote generally tend to be at least broadly in sympathy with that party's leading candidate, and those who decide to change their registration to disaffiliate with a party usually do not intend to support the party's leading candidate. The contrary result in the initial model is not so much surprising as it is unbelieveable.

Signaling and then perhaps helping to diagnose important model misspecifications is one of the most important functions that the robust estimation and outlier detection methods we have introduced can perform. Robust estimation followed by inspection, investigation and resolution of any outliers should become a routine aspect of regression analysis of count data. One should not expect that the ultimate result of such activity will be that all outliers will be eliminated. There is, in general, no reason to believe that a single generalized-linear regression specification will be a good model for all the counts in a particular data collection, although finding such a specification can be a worthy goal. Especially when, as with votes, the counts are aggregates of individual decisions that may be affected by unobserved idiosyncratic factors, a univer-

[18] The nonrobust ML estimate is $\hat{\beta}_{32} = 1.63$ with SE$= 1.40$.

[19] The nonrobust ML estimates are $\hat{\beta}_{32} = 0.7$ with SE$= 1.2$ (recall footnote 15) and $\hat{\beta}_{42} = 3.4$ with SE$= 1.0$.

sally correct aggregate model may be very difficult if not impossible to define (compare Achen and Shively 1995, 94–115). The conditions for exact aggregation of individual choices into a single model for the aggregated decisions are stringent even when data richer than counts of categorical choices are observed (e.g. Jorgenson, Lau, and Stoker 1982). Therefore it is crucial that the question of adequate model specification and consequently the task of investigating outliers be pursued with a clear sense of the substantive issues the analysis is intended to address and of the possibly limited degree to which the estimation effort can inform the research. To resolve the outliers means to demonstrate that they do not signal problems for the aspects of the estimation that matter for the substantive conclusions. Or if such a demonstration is not possible, as may often occur, resolving the outliers means to give a good characterization of what the problems are likely to be.

The robust estimation methods we have introduced offer an accurate and powerful technology for detecting irregular outcomes. 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.

Using a triangulation approach for Miami-Dade County in the current paper called for gathering additional data to try to measure the effects of particular behavior by many Cuban-Americans in Florida. The extensions of the model confirm the many news reports: While the butterfly ballot was perhaps the final assault that destroyed Gore’s chances of winning on election day in Florida, the turmoil over Elián González had already dealt his campaign a crushing blow.

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 that 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 inconsistent hence leading to incorrect inferences.

The current sampling experiment uses relatively large sample sizes. In particular the count totals m_i are large. Such values are comparable in magnitude to many kinds of vote data, for example, votes in the counties of Florida. But other potential applications of the robust estimator may involve much smaller counts (e.g., the analysis of precinct data in Wand et al. (2001)). 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. Also it is important to extend the robust methods to handle cases where it is suspected that there are variance components that involve additional sources of covariation among choice categories, as in Jackson (2002). Nonetheless we have great confidence that our robust estimation method is vastly superior to nonrobust ML estimation, especially when the robust estimation and outlier detection is accompanied by appropriate efforts to triangulate.

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^{-1}, i.e., $r^{\perp} = L_i^{-1}(y_i - m_i p_i)$, because

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

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

We use the estimated probabilities $\hat{p}_{ij}$ to compute estimated inverse Cholesky factor matrices, $\hat{L}_i^{-1}$, and hence the following orthogonalized residuals:

$$\hat{r}_i^{\perp} = \hat{L}_i^{-1}(y_i - m_i \hat{p}_i).$$

We also use the estimated probabilities to compute estimated Cholesky factors $\hat{d}_{ij}$, which we use to partially studentize the $J-1$ nontrivial values of $\hat{r}_i^{\perp}$ for each i. The resulting studentized residuals are

$$r_{ij}^{*} = \frac{\hat{r}_{ij}^{\perp}}{\sqrt{m_i \hat{d}_{ij}}} \qquad (j = 1, \ldots, J-1)$$

(note that $\hat{r}_{iJ}^{\perp} = 0$). Expansion of $\hat{r}_{ij}^{\perp}$ and $\hat{d}_{ij}$ gives the explicit formula:

$$r_{ij}^{*} = \begin{cases} \dfrac{y_{i1} - m_i \hat{p}_{i1}}{\sqrt{m_i \hat{p}_{i1}(1 - \hat{p}_{i1})}}, & (j = 1) \\[2ex] \dfrac{(y_{ij} - m_i \hat{p}_{ij}) + \left[\sum_{k=1}^{j-1}(y_{ik} - m_i \hat{p}_{ik})\right] \hat{p}_{ij} / \left[1 - \left(\sum_{k=1}^{j-1} \hat{p}_{ik}\right)\right]}{\sqrt{m_i \hat{p}_{ij} \left[1 - \left(\sum_{k=1}^{j} \hat{p}_{ik}\right)\right] / \left[1 - \left(\sum_{k=1}^{j-1} \hat{p}_{ik}\right)\right]}}, & (1 < j \leq J-1) \\[2ex] 0, & j = J. \end{cases}$$

The point of departure for our estimation method is the fact that 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 partially studentized residuals r_{ij}^{*} (compare McCullagh and Nelder 1989, 168–169). Moreover, if the values $m_i p_{ij}(1 - p_{ij})$ are sufficiently large, then the residuals r_{ij}^{*}, $j = 1, \ldots, J-1$, are approximately normal.[20]

[20] The discussion in Wand et al. (2001) of the relationship between the size of $m_i p_{ij}(1 - p_{ij})$ and the approximate normality of

Let the $n(J-1)$ residuals 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 $\{|r^*_{\ell_1} - r^*_{\ell_2}| : \ell_1 < \ell_2\}$ of $\binom{n(J-1)}{2}$ absolute differences. Following Croux et al. (1994) we use

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

to denote that order statistic. For large n, $\binom{h_K}{2}/\binom{n(J-1)}{2}$ is approximately $1/4$, so that $Q^*_{n(J-1)}$ is approximately the first quartile of the set of absolute differences. To implement LQD we choose estimates $\hat{\beta}$ to minimize $Q^*_{n(J-1)}$. Let $\hat{\beta}_{\text{LQD}}$ designate the estimated coefficient vector and let $\hat{Q}^*_{n(J-1)}$ designate the corresponding minimized value of $Q^*_{n(J-1)}$. The LQD scale estimate is

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

where Φ^{-1} is the quantile function for the standard normal distribution (Rousseeuw and Croux 1993, 1277). The approximate normality of the residuals r^*_ℓ in the case of correct specification justifies the factor $1/[\sqrt{2}\Phi^{-1}(5/8)]$. In the binomial case $J=2$, this estimator is the same as the LQD estimator of Wand et al. (2001).

As do Wand et al. (2001), we use GENOUD (Sekhon and Mebane 1998) to optimize the LQD objective function.[21] The LQD objective function, like all quantiles, is not globally concave. Therefore, derivative information, although of some use, cannot be relied upon to find the global optimum. GENOUD combines evolutionary algorithm methods with a derivative-based, quasi-Newton method to solve such unconstrained optimization problems.

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

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

the residuals for a binomial model also applies to multinomial models with $J > 2$.

[21] See `http://jsekhon.fas.harvard.edu/rgenoud/` for an **R** version of the GENOUD software.

where choices of c and d imply unique values for p, A and B.[22] The value of c fixes the truncation threshold. The constant d corresponds to the ratio between the *change-of-variance function*—the sensitivity of 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.[23] Given a scale estimate $\hat{\sigma}_{\text{LQD}}$ and trial estimates $\hat{\beta}$, we compute for each i the $J-1$ partially studentized residuals

$$r_{ij} = r^*_{ij}/\hat{\sigma}_{\text{LQD}} \qquad (j = 1, \ldots, J-1)$$

and weights

$$w_{ij} = \begin{cases} \psi(r_{ij})/r_{ij}, & \text{for } r_{ij} \neq 0 \\ 1, & \text{for } r_{ij} = 0 \,. \end{cases}$$

A studentized residual that has $w_{ij} = 0$ is an *outlier*.

We use the values w_i to weight the gradient and the Hessian matrix in what would otherwise be a standard Newton algorithm (Gill, Murray, and Wright 1981, 105) to estimate β by maximum likelihood if there were no overdispersion. The negative log-likelihood for a multinomial model of y_i with probability vector p_i is $l_i = -(\log p_i)' y_i$, the gradient of l_i with respect to the vector of linear predictors μ_i is $\partial l_i/\partial \mu_i = -(y_i - m_i p_i)$, and the Hessian matrix is $\partial^2 l_i/\partial \mu_i \partial \mu_i' = m_i(P_i - p_i p_i')$. The gradient and Hessian matrix with respect to the coefficient parameters β are computed by using the chain rule to obtain, respectively, gradient $(\partial \mu_i'/\partial \beta)(\partial l_i/\partial \mu_i)$ and Hessian matrix $(\partial \mu_i'/\partial \beta)(\partial^2 l_i/\partial \mu_i \partial \mu_i')(\partial \mu_i/\partial \beta')$. We use the estimates $\hat{p}_i$ and the formal Cholesky decomposition to map the weights onto the appropriate components of the gradient and Hessian matrix. Let W_i denote the diagonal matrix that has $W_{i,jj} = w_{ij}$ for the diagonal values $j = 1, \ldots, J-1$ and $W_{i,JJ} = 1$.[24] For the gradient, we use the Cholesky factor matrix $\hat{L}_i$ to compute $W_i \hat{L}_i^{-1}(y_i - m_i \hat{p}_i)$, which evidently is simply an application of the weights to the orthogonalized residuals. We then use $\hat{L}_i$ to map the result back onto the original dimensions of the model, so that the weighted gradient with respect to β,

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

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

[24] As will immediately become apparent, the value assigned to $W_{i,JJ}$ is practically inconsequential, because it is always multiplied by zero.

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 matrix, 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 matrix 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

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

Because the steps are merely proportional to b, the rescaling by $\hat{\sigma}_{\mathrm{LQD}}$ is not essential.[25]

We alternate complete rounds of LQD and tanh estimation, beginning with the LQD and then using each estimation round's results to start its successor (compare Huber 1981, 179–192). This large iteration between LQD and tanh continues until $\hat{\sigma}_{\mathrm{LQD}}$ and $\hat{\beta}_{\mathrm{LQD}}$ converge over successive LQD rounds, at which point the final tanh estimates are computed. The scale estimate $\hat{\sigma}_{\mathrm{LQD}}$ remains unchanged throughout the Newton iterations for each tanh round. Each tanh round consists of a series of complete Newton optimization sequences, each of which uses the preceding estimates $\hat{\beta}_{\mathrm{LQD}}$ to start the coefficients and uses the preceding LQD values $(r_\ell^* - \mathrm{med}_\ell\, r_\ell^*)/\hat{\sigma}_{\mathrm{LQD}}$ for an initial set of residuals, where $\mathrm{med}_\ell\, r_\ell^*$ denotes the median of the r_ℓ^* values, $\ell = 1, \ldots, n(J-1)$. Each Newton sequence produces a $\hat{\beta}$ that is converged given $\hat{\sigma}_{\mathrm{LQD}}$ and the weights. After each such sequence the weights are updated to match the current coefficient estimates. The series of complete Newton optimizations that comprise each tanh round continues until the weights converge.

To estimate the asymptotic covariance matrix of the tanh coefficient estimates, $\Sigma_{\hat{\beta}}$, we use the sandwich estimator of Huber (1967, 231; 1981, 133). Using the weighted Hessian matrix,[26]

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

[25] Including the rescaling by $\hat{\sigma}_{\mathrm{LQD}}$ improves the performance of our implementation of the Newton algorithm in which the line search initially attempts a step exactly equal to b.

[26] To connect (A-1) to Huber's asymptotic covariance matrix result, let $s_i = -(\partial\mu_i'/\partial\beta)L_iW_iL_i^{-1}(y_i - m_ip_i)$ denote the weighted gradient for β known and, treating L_i and W_i as fixed, note that $\partial s_i/\partial\beta' = (\partial s_i/\partial\mu_i')(\partial\mu_i/\partial\beta') = m_i(\partial\mu_i'/\partial\beta)L_iW_iL_i^{-1}L_iD_iL_i'(\partial\mu_i/\partial\beta')$.

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).[27] In our Monte Carlo sampling study we also examine the performance of two alternative covariance matrix estimators. One is the inverse weighted Hessian multiplied by an estimate of the overdispersion that uses the residuals and weights from the tanh estimation:

$$\hat{\Sigma}_{\mathrm{G}:\hat{\beta}} = \hat{\sigma}^2_{\mathrm{tanh}} \hat{G}^{-1}$$

where, with $\hat{\beta}$ used to compute the partially studentized residuals r^*_{ij},

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

The other covariance matrix estimator we consider is the inverse outer product of the weighted gradient:

$$\hat{\Sigma}_{\mathrm{I}:\hat{\beta}} = \hat{I}^{-1} .$$

To obtain fully studentized residuals for the purpose of outlier diagnostics, we make a weighting adjustment for leverage (which applies to ortho-studentized residuals with $w_{ij} > 0$) or for forecasting error (which applies to the residuals with $w_{ij} = 0$). Let V_i denote the 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

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

provide robust estimates of the additional weights (compare McCullagh and Nelder 1989, 397; Carroll and

[27] 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.

Ruppert 1988, 31–34).[28] 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 fully studentized residuals are:

$$\tilde{r}_{ij} = r_{ij} / \sqrt{1 - h_{ij}} \qquad (j = 1, \ldots, J-1). \tag{A-3}$$

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

[28] The form of (A-2) assumes that the parameters and regressors have been normalized such that the linear predictor for category J is always zero, i.e., $\mu_{iJ} = 0$ for all i, so that the J-th column of $\partial\hat{\mu}_i'/\partial\hat{\beta}$ is zero.

References

Achen, Christopher H. and W. Phillips Shively. 1995. *Cross-level Inference*. Chicago: University of Chicago Press.

Aitchison, John. 1986. *The Statistical Analysis of Compositional Data*. New York: Chapman & Hall.

Allen, Jonathan. 2002. “Florida Remap Gives GOP Edge on Seats.” *Washington Post*. March 22, Internet Edition.

Associated Press. 2000. “Judge Won’t Rule on Reform Party.” *Associated Press*. August 30.

Associated Press. 2001. “Latin American Immigrants Flood Miami.” *Associated Press*. September 7.

Atkinson, A. C. 1986. “Masking Unmasked.” *Biometrika* 73: 533–541.

Bragg, Rick. 2000. “The Voters: Florida, Amid Recount, Learns the Power of One.” *New York Times*. November 9, Internet Edition.

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

Canedy, Dana. 2001. “Cuban Exile Group Fractured as Hard-Liners Quit Board.” *New York Times*. August 8, Internet Edition.

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

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

Center for Responsive Politics. 2001. “The Cuban Connection: Cuban-American Money in U.S. Elections 1979–2000.” Accessed April 20, 2002.
URL `http://www.opensecrets.org/pubs/cubareport/`

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

Dao, James. 2000. “Democrats Criticize Gore For Position on Cuban Boy.” *New York Times*. April 1, Internet Edition.

Dershowitz, Alan M. 2001. *Supreme Injustice: How the High Court Hijacked Election 2000*. New York: Oxford University Press.

Engelhardt, Joel. 2000. “Elections Chief on Firing Line.” *Palm Beach Post*. November 9, Internet Edition.

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

Fessenden, Ford and Josh Barbanel. 2000. “Broad Recount Presents Risks for Both Sides.” *New York Times*. December 9, Internet Edition.

Filkins, Dexter. 2000. “Miami-Dade County: A Mayor, Once Vocal for Gore, Is Silent.” *New York Times*. November 25, Internet Edition.

Forero, Juan. 2000. “Angry Demonstrators Quickly Gather Outside the Gonzalez House in Miami.” *New York Times*. April 23, Internet Edition.

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: 643–648.

Hausman, J., B.H. Hall, and Z. Griliches. 1984. “Econometric Models for Count Data with an Application to the Patents R&D Relationship.” *Econometrica* 52: 909–938.

Honaker, James, Jonathan N. Katz, and Gary King. 2002. "A Fast, Easy and Efficient Estimator for Multiparty Electoral Data." *Political Analysis* 10: 84–100.

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: 49–65.

Jorgenson, Dale W., Lawrence J. Lau, and Thomas M. Stoker. 1982. "The Transcendental Logarithmic Model of Aggregate Consumer Behavior." In R. L. Basmann and G. Rhodes, editors, *Advances in Econometrics*, volume 1. Greenwich, CT: JAI Press. Pages 97–238.

Kaplan, David A. 2001. *The Accidental President*. New York: William Morrow.

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

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

Merzer, Martin and the Staff of the *Miami Herald*. 2001. *The Miami Herald Report: Democracy Held Hostage*. New York: St. Martin's Press.

Nagourney, Adam. 2000. "Mrs. Clinton Opposes Bill Giving Residency to Cuban Boy." *New York Times*. April 2, Internet Edition.

Nichols, John. 2001. *Jews for Buchanan: Did You Hear the One About the Theft of the American Presidency?*. New York: The New Press.

Perez-Pena, Richard. 2000. "The 2000 Campaign: The Democratic Running Mate; Lieberman Appeals for Help From Florida Ethnic Groups." *New York Times*. October 24, Internet Edition.

Posner, Richard A. 2001. *Breaking the Deadlock: The 2000 Election, the Constitution, and the Courts*. New York: Princeton University Press.

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

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

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.

Seelye, Katharine Q. 2000a. “The 2000 Elections: The Vice President; Sleep Waits as Gore Makes His Final Appeals.” *New York Times*. November 8, Internet Edition.

Seelye, Katharine Q. 2000b. “Boy’s Case Could Sway Bush-Gore Contest.” *New York Times*. March 30, Internet Edition.

Seelye, Katharine Q. 2000c. “Gore Supporting Residency Status for Cuban Child.” *New York Times*. March 31, Internet Edition.

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

Stigler, Stephen M. 1973. “Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885−1920.” *Journal of the American Statistical Association* 68: 872–879.

Sunstein, Cass R. and Richard A. Epstein, editors. 2001. *The Vote: Bush, Gore and the Supreme Court*. Chicago: University of Chicago Press.

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: 211–219.

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

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

Van Natta, Jr., Don. 2000. “Bush Criticizes Administration For Removing Boy ‘at Gunpoint’.” *New York Times*. April 23, Internet Edition.

Van Natta, Jr., Don and Dexter Filkins. 2000. “Miami Mayor’s Role a Riddle in Decision to Halt Recount.” *New York Times*. December 1, Internet Edition.

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 (December): 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: 979–994.

Western, Bruce. 1995. “Concepts and Suggestions for Robust Regression Analysis.” *American Journal of Political Science* 39: 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

Tanh Results

Experiment	Coeff.	Coeff.	H-W Sandwich Coverage			Hessian Coverage			OPG Coverage		
Condition	Error	RMSE	90%	95%	RMSE	90%	95%	RMSE	90%	95%	RMSE
1	0.000	0.00385	0.898	0.950	0.00949	0.908	0.955	0.0129	0.924	0.963	0.0265
2	0.000	0.00896	0.873	0.928	0.0281	0.878	0.935	0.0228	0.898	0.943	0.00582
3	0.000	0.0102	0.883	0.940	0.0190	0.894	0.946	0.0108	0.914	0.956	0.0156
4	0.000	0.00947	0.878	0.933	0.0227	0.887	0.939	0.0147	0.902	0.948	0.0100
5	0.000	0.00635	0.880	0.937	0.0205	0.889	0.945	0.0153	0.910	0.955	0.0160
6	-0.001	0.0150	0.871	0.932	0.0299	0.885	0.941	0.0178	0.905	0.951	0.0107

Nonrobust Maximum Likelihood Results

Experiment	Coeff.	Coeff.	Coverage using $\sigma^2 = 1$			Coverage using $\hat{\sigma}^2$		
Condition	Error	RMSE	90%	95%	RMSE	90%	95%	RMSE
1	0.000	0.00383	0.900	0.949	0.0088	0.898	0.948	0.00916
2	0.000	0.00888	0.514	0.599	0.386	0.902	0.949	0.00542
3	-0.020	0.0237	0.304	0.318	0.735	0.999	1.00	0.0988
4	-0.020	0.0261	0.193	0.228	0.746	0.970	0.999	0.0820
5	-0.600	1.06	0.000	0.000	0.900	0.000	0.000	0.900
6	-0.600	1.06	0.000	0.000	0.900	0.000	0.000	0.900

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

Table 3: Summary for Experiment Condition 3: Symmetric Probabilities, 10% Contamination, No Overdispersion

Tanh Results

	True	Coeff.	Coeff.	H-W Coverage		Hessian Coverage		OPG Coverage	
Coefficient	Coeff.	Mean	RMSE	90%	95%	90%	95%	90%	95%
β_{10}	-1	-1.00	0.0418	0.883	0.943	0.890	0.945	0.908	0.950
β_{11}	1	1.00	0.00351	0.887	0.946	0.900	0.949	0.919	0.954
β_{20}	-1	-1.00	0.00419	0.896	0.941	0.906	0.948	0.919	0.958
β_{21}	1	1.00	0.00356	0.873	0.935	0.889	0.939	0.909	0.952
β_{30}	-1	-1.00	0.00426	0.889	0.939	0.902	0.951	0.923	0.969
β_{31}	1	1.00	0.00365	0.871	0.934	0.879	0.941	0.908	0.952

Nonrobust Maximum Likelihood Results

	True	Coeff.	Coeff.	Coverage: $\sigma^2 = 1$		Coverage using $\hat{\sigma}^2$	
Coefficient	Coeff.	Mean	RMSE	90%	95%	90%	95%
β_{10}	-1	-1.05	0.0493	0.00	0.00	0.996	1.00
β_{11}	1	0.98	0.0202	0.001	0.001	1.00	1.00
β_{20}	-1	-1.00	0.0041	0.91	0.953	1.00	1.00
β_{21}	1	1.00	0.00338	0.91	0.951	1.00	1.00
β_{30}	-1	-0.953	0.0474	0.00	0.00	0.997	1.00
β_{31}	1	1.02	0.0176	0.00	0.001	1.00	1.00

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

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

Tanh Results

	True	Coeff.	Coeff.	H-W Coverage		Hessian Coverage		OPG Coverage	
Coefficient	Coeff.	Mean	RMSE	90%	95%	90%	95%	90%	95%
β_{10}	-3.5	-3.50	0.0111	0.877	0.920	0.873	0.926	0.887	0.937
β_{11}	1	1.00	0.00684	0.878	0.938	0.886	0.947	0.911	0.954
β_{20}	-3	-3.00	0.00849	0.877	0.941	0.883	0.949	0.908	0.958
β_{21}	1	1.00	0.00523	0.881	0.938	0.905	0.947	0.929	0.961
β_{30}	-1	-1.00	0.00362	0.883	0.942	0.901	0.951	0.911	0.962
β_{31}	1	1.00	0.0028	0.882	0.941	0.887	0.949	0.915	0.958

Nonrobust Maximum Likelihood Results

	True	Coeff.	Coeff.	Coverage: $\sigma^2 = 1$		Coverage using $\hat{\sigma}^2$	
Coefficient	Coeff.	Mean	RMSE	90%	95%	90%	95%
β_{10}	-3.5	-2.09	1.41	0.00	0.00	0.00	0.00
β_{11}	1	-0.31	1.31	0.00	0.00	0.00	0.00
β_{20}	-3	-1.73	1.27	0.00	0.00	0.00	0.00
β_{21}	1	-0.183	1.18	0.00	0.00	0.00	0.00
β_{30}	-1	-0.407	0.593	0.00	0.00	0.00	0.00
β_{31}	1	0.404	0.596	0.00	0.00	0.00	0.00

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

Table 5: Overdispersed Multinomial Regression Model for 2000 Election Vote Counts, Florida Counties, Robust Estimates (Initial Model)

	Candidate							
Regressor	Buchanan		Nader		Gore		Bush	
Constant	-0.589	(0.151)	1.1	(0.206)	3.42	(0.179)	4.02	(0.165)
Principal Component	-0.0481	(0.0263)	-0.0157	(0.0168)	0.0215	(0.011)	-0.0096	(0.00851)
1996 Vote Proportion Variables:								
Perot	4.32	(1.37)	—		—		—	
Clinton + Nader	—		0.00973	(0.441)	—		—	
Clinton	—		—		2.86	(0.377)	—	
Dole	—		—		—		1.88	(0.349)
Voter Registration Proportion Variables:								
Δ Republican (2000$-$1996)	13.7	(1.27)	—		—		2.759	(0.804)
Green (2000)	—		1210	(122)	—		—	
Δ Democrat (2000$-$1996)	—		—		0.992	(1.01)	—	

Note: Entries are tanh estimates of coefficient parameters of the overdispersed multinomial regression model for $J = 5$ choices (sandwich standard errors in parentheses). For each of the four choices above, there are four regressors. Each of the choices has a constant and a principle component regressor, but the past vote and the registration variables vary over the choices (see equation (1)). $n = 67$ counties. LQD dispersion estimate: $\hat{\sigma}_{\text{LQD}} = 5.61$. tanh dispersion estimate: $\hat{\sigma}_{\text{tanh}} = 4.53$.

Table 6: Outlier Florida Counties in the 2000 Election, Five Candidate Categories (Initial Model)

	Candidate					
County	Buchanan	Nader	Gore	Bush	Other	Total Votes
Alachua	0.52	−10.59	1.74	3.98	5.18	85,729
Broward	−1.20	−0.18	−7.01	7.53	−1.78	575,143
Duval	−0.91	−4.11	−5.99	6.85	1.52	264,636
Escambia	0.92	−2.14	4.06	−3.57	−0.08	116,648
Hillsborough	−0.62	−1.36	−4.04	4.29	1.09	360,295
Leon	−0.24	−6.89	6.41	−4.07	0.40	103,124
Marion	1.72	1.25	3.83	−4.76	3.49	102,956
Martin	−0.25	0.38	4.39	−4.28	−1.05	62,013
Miami-Dade	−4.64	−6.30	−28.43	31.28	−4.82	625,449
Orange	−0.45	−4.49	0.75	0.65	−0.75	280,125
Palm Beach	20.76	−1.41	−1.29	−0.15	−0.55	433,186
Pasco	1.26	4.59	−6.06	4.74	0.11	142,731
Pinellas	1.72	1.45	−10.26	9.55	2.03	398,472

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 Regression Model for 2000 Election Vote Counts, Florida Counties, Robust Estimates (Full Model)

	Candidate							
Regressor	Buchanan		Nader		Gore		Bush	
Constant	−0.312	(0.212)	1.05	(0.19)	3.37	(0.159)	4.13	(0.137)
Proportion Cuban	−5.73	(2.45)	0.284	(0.402)	2.03	(0.34)	2.75	(0.29)
Principal Component	−0.0254	(0.0257)	−0.00686	(0.0175)	−0.0246	(0.0111)	0.00383	(0.00988)
1996 Vote Proportion Variables:								
Perot	3.38	(1.74)	—		—		—	
Clinton + Nader	—		0.323	(0.42)	—		—	
Clinton	—		—		3.13	(0.323)	—	
Dole	—		—		—		1.71	(0.275)
Voter Registration Proportion Variables:								
Δ Republican (2000−1996)	14.4	(1.46)	—		—		2.04	(0.744)
Green (2000)	—		1270	(110)	—		—	
Δ Democrat (2000−1996)	—		—		1.80	(0.910)	—	

Note: Entries are tanh estimates of coefficient parameters of the overdispersed multinomial regression model for $J = 5$ choices (sandwich standard errors in parentheses). For each of the four choices above, there are five regressors. Each of the choices has a constant, a principle component and proportion Cuban-American regressor, but the past vote and the registration variables vary over the choices (see equation (2)). $n = 67$ counties. LQD dispersion estimate: $\hat{\sigma}_{\mathrm{LQD}} = 5.06$. tanh dispersion estimate: $\hat{\sigma}_{\mathrm{tanh}} = 4.45$.

Table 8: Outlier Florida Counties in the 2000 Election, Five Candidate Categories (Full Model)

	Candidate					
County	Buchanan	Nader	Gore	Bush	Other	Total Votes
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 9: Overdispersed Binomial Regression Model for 2000 Election Results, Miami-Dade Census Tracts, Robust Estimates

Regressor	Gore	
Constant	−1.87	(0.0798)
Proportion of 1996 Presidential Vote for Clinton	3.37	(0.180)
Proportion Registered Democrat in 2000	1.54	(0.227)
Proportion Cuban	−0.20	(0.0858)
Proportion Hispanic	−0.777	(0.0922)

Note: Entries are tanh estimates of coefficient parameters of the overdispersed multinomial regression model using census-tract-level data from the 2000 election (sandwich standard errors in parentheses). $n = 346$ census tracts. LQD dispersion estimate: $\hat{\sigma}_{\text{LQD}} = 2.55$. tanh dispersion estimate: $\hat{\sigma}_{\text{tanh}} = 2.39$.