

A Fast, Easy, and Efficient Estimator for Multiparty Electoral Data

James Honaker

*Department of Political Science,
University of California, Los Angeles,
Los Angeles, CA 90095-1472
e-mail: tercer@ucla.edu*

Jonathan N. Katz

*Division of Humanities and Social Science,
California Institute of Technology, Pasadena, CA 91125
e-mail: jkatz@caltech.edu*

Gary King

*Department of Government,
Harvard University, Cambridge, MA 02138
e-mail: king@harvard.edu*

Katz and King have previously developed a model for predicting or explaining aggregate electoral results in multiparty democracies. Their model is, in principle, analogous to what least-squares regression provides American political researchers in that two-party system. Katz and King applied their model to three-party elections in England and revealed a variety of new features of incumbency advantage and sources of party support. Although the mathematics of their statistical model covers any number of political parties, it is computationally demanding, and hence slow and numerically imprecise, with more than three parties. In this paper we produce an approximate method that works in practice with many parties without making too many theoretical compromises. Our approach is to treat the problem as one of missing data. This allows us to use a modification of the fast EMis algorithm of King, Honaker, Joseph, and Scheve and to provide easy-to-use software, while retaining the attractive features of the Katz and King model, such as the t distribution and explicit models for uncontested seats.

1 Introduction

We offer a computationally feasible algorithm, and easy-to-use software, that approximates Katz and King's (KK) (1999) full information maximum likelihood (FIML) model of district-level multiparty electoral data. Our model provides a tool analogous to what least-squares regression provides American political researchers in two-party American

Authors' note: An earlier version of the paper was presented at the annual meetings of the American Political Science Association, Washington, DC, 2000. For research support, we gratefully acknowledge the John M. Olin Foundation, the National Science Foundation (IIS-9874747), the National Institutes of Aging (P01 AG17625-01), and the World Health Organization.

Copyright 2002 by the Society for Political Methodology

elections. That is, scholars can use our method and software with aggregate multiparty electoral data (and related explanatory variables also coded or aggregated at the district level) to explain, predict, or infer about counterfactuals in real election results.

The main advantage of our approach is that it is considerably faster and scales up to many more parties without much difficulty, loss of speed, or numerical imprecision. In theory, the KK method is optimal (conditional on the model) but slow; in practice, our approach is faster, with better numerical precision and the same mean squared error. We thus believe that the result is a useful tool for comparative political research. In addition to treating vote share as a compositional variable, our approach also allows us to model vote share as t distributed and enables us to use election data where not all parties contest in every district. These can be important features of political science data.

We first introduce our notation and the basic KK model (Section 2) and an overview of our proposed approach (Section 3, with technical details given in the Appendix). We then present Monte Carlo comparisons of KK with our method in three-party data and replicate KK's empirical results with our methods (Section 4). We then conduct Monte Carlo experiments on data with more parties than would be computationally feasible under the KK model (Section 5). Finally, we replicate a real empirical article and show how the substantive results change when our improved method is applied (Section 6).

2 Notation and the Full Information Approach

Suppose that we wish to predict or explain the vote share of certain parties in certain districts. Let V_{ij} denote the fraction of the vote in district i ($i = 1, \dots, N$) for political party j ($j = 1, \dots, J$). Thus, for prediction or explanation, the relevant outcome is a *set* of J dependent variables for each district i . Of course, these J variables are closely related, which can be seen most easily in the two-party special case, where $V_{i1} = 1 - V_{i2}$. In general, multiparty electoral data are an example of *compositional data*, where the set of variables falls on the simplex, which means that each vote proportion falls between 0 and 1,

$$V_{ij} \in [0, 1] \quad \text{for all } i \text{ and } j \quad (1)$$

and the set of votes in a district sums to 1,

$$\sum_{j=1}^J V_{ij} = 1 \quad \text{for all } i. \quad (2)$$

Following Aitchison (1986), KK avoid the complications induced by these constraints by modeling the $J - 1$ log ratios of the vote variables $Y_{ij} = \ln(V_{ij}/V_{iJ})$, for $j = 1, \dots, J - 1$. The advantage of this approach is that the set of Y_{ij} variables is individually and collectively unconstrained, making modeling much easier. After modeling, estimates are mapped back onto the simplex and the results are recovered in their original scale of interest.

KK depart from Aitchison's approach of modeling the log ratios via a multivariate Normal in two important ways. First, they use a multivariate t distribution to model the log ratios. They show that this model, which becomes the additive logistic t on the scale of the V 's, fits the data far better than the Normal in electoral data. Second, they add a component of the model to cope with partially contested or uncontested district elections. We follow KK in setting as the goal of the analysis predicting or explaining the *effective vote*, the values of V_{ij} we would observe if all parties were contesting all J districts (Gelman and King 1994). To use information from district elections that are not fully contested, we assume that a party that chooses not to contest an election would have received fewer votes than any of the parties that did run if it had contested the election. This assumption is justified

in KK and is plausible for a wide range of circumstances. Moreover, the machinery used to modify the model could easily be adjusted to suit qualitative knowledge in different electoral circumstances.

The multivariate t distribution adds a few complications, but the main difficulty in estimating the KK model is implementing the assumptions required for the partially contested district elections. This is the feature of the model that is most uniquely related to political science data, and ignoring it by listwise deleting these districts is not a reasonable option. (As we explain below, these compositional data are not “missing completely at random.”)

There are two computational issues involving partially contested district elections. First, the KK model computes the likelihood for these observations by treating them as censored and using the assumption above. For example, suppose that parties 2 and 3 (of $J = 3$) contest the election in district i . Then the effective votes V_{i1} , V_{i2} , and V_{i3} are unobserved. However, our assumptions imply that $V_{i1} < \min(V_{i2}, V_{i3})$ and hence $Y_{i1} < \min(0, Y_{i2})$ or, equivalently, $Y_1 < 0$ and $Y_1 < Y_2$. This makes the likelihood function for this observation $L_i^{23} = \Pr(Y_{i1} < 0, Y_{i2} > Y_{i1} | \psi)$ wherever the set of parties contesting is $P_i = \{2, 3\}$.¹ The problem is that this likelihood requires an analytically intractable integration.² KK computed this integration numerically, which, while slow, is not overwhelmingly time-consuming when $J = 3$. Unfortunately, the curse of dimensionality makes this function exponentially slower, or considerably more imprecise, as J increases. The imprecision in this numerical integration is one of the chief reasons our method outperforms KK, either in mean squared error for quick integrations or in terms of time for accurate numerical integrations.

A second computational problem with the KK approach is exponentially increasing possible combinations of parties that could contest elections in each district. For the three-party case, the likelihood has eight logically possible pieces, based on the set of parties that could contest: $\{1, 2, 3\}$, $\{2, 3\}$, $\{1, 3\}$, $\{1, 2\}$, $\{1\}$, $\{2\}$, $\{3\}$, and (in theory anyway) $\{\}$. In general, there are 2^J distinct parts to the likelihood, which for large J makes hard-coding the likelihood function time-consuming. If $J = 10$, the likelihood has 1024 pieces; with $J = 20$, the likelihood has over a million parts, although of course the largest number of parts that would ever enter the model in any case is n .

These problems make direct extensions of KK’s analytical approach impractical. KK tried direct MCMC approaches, which were about 20 times slower for three parties, and they were much more difficult to automate. While the MCMC approach seemed to scale better than KK, the base speed was so slow that the method would likely deter political scientists from using it. This is an approach we do not follow up on in this paper, although with continuing advances and research in MCMC, there is some hope that future researchers may make this a feasible option. Another approach might be to simulate higher-dimensional integrals, but this does not seem promising beyond five or six parties, nor does it address the combinatorial problem.

3 Overview of Our Alternative Approach

We now provide an intuitive overview, saving the technical details of our approach for the Appendix. The central feature of our approach is to treat the problem of predicting or explaining the effective vote in partially contested districts as a missing data problem. This

¹The vector ψ is the set of means, covariates, and parameters of the t distribution.

²That is, $L_i^{23} = \int_{-\infty}^0 \int_{Y_{i1}}^{\infty} T(Y_{i1}, Y_{i2} | \psi) dY_{i2} dY_{i1} = \int_{-\infty}^0 T(Y_{i1} | \mu_{i1}, \sigma_{i1}, \nu) [1 - F_T(Y_{i1} | \mu_{i2|1}, \sigma_{i2|1}, \nu + 1)] dY_{i1}$, where F_T is the cumulative distribution function of the (univariate) t , and $\mu_{i2|1}$ and $\sigma_{i2|1}$ are the conditional moments defined in KK.

enables us to build on the literature of multiple imputation rather than the harder to analyze application-specific approaches, as in KK (see King et al. 2001).

In districts without all parties contesting, Y_{ij} is missing for all j , since the effective vote would differ for all parties, not just the one not contesting. If these cells in our data matrix were ignorably missing (that is, missing at random or missing completely at random), such as is typically assumed for survey responses, we might impute them with the usual multivariate Normal multiple-imputation models [such as Ameila (see Honaker et al. 2000)]. We would then follow the usual strategy of multiple imputation: create M data sets with the same values for the fully contested districts but different imputed values for the partially contested districts. The variance in each imputed value across the data sets reflects our uncertainty in the imputation. Analysts would then use whatever method they would apply if all the data were observed to each of the M data sets and use the usual procedure for combining the results of the separate imputations.

Unfortunately, two problems make this strategy incorrect without some modification. First, KK find that a Normal model does not fit multiparty electoral data. We fix this problem by creating a multivariate t imputation model, using the EMis algorithm. The Appendix describes this technique. This multivariate t imputation model would be applied at the level of the Y 's, and so the compositional constraint that the V 's form a simplex is automatically satisfied.

However, this setup needs to be further modified to include constraints regarding the partially contested districts. We have studied this issue and have found that to do it as part of the imputation procedure would be very time-consuming, not easy to automate, and thus difficult to use. We therefore adopt a somewhat nonstandard procedure of sampling subject to constraints given the original t imputations. That is, to generate the imputations, we use the EMis algorithm to draw from the posterior distribution of the multivariate t model, check whether the imputation fits the constraint, and discard it and draw another if the constraint is violated. The procedure is repeated until M imputations are drawn. This procedure leads one to reinterpret the parameters of the imputation model to be of the untruncated t density, corresponding to the truncated t density that interests us. Since, in this application, the parameters of the imputation model are not of substantive interest, this step has few important consequences other than speeding computation. The cost of this sampling correction is that it will be a poorer approximation to the KK model when the constraints have a strong effect. Our software therefore reports this information as a diagnostic as explained in Section A.3 in the Appendix. The sampling correction also has a benefit: since it is a two-step approach, and thus any modeling assumptions affect only the uncontested and partially contested districts rather than all the observations, this procedure is likely to be more robust to model misspecification than the KK FIML method. Nevertheless, whatever compromises this procedure requires seem worth consideration given the enormous computational savings that result.

The results of the imputation algorithm with the simulation-based correction for partially contested districts make available a set of M imputed data sets. These can be treated as if the effective vote were fully observed and analyzed accordingly (and then combined as usual in multiple imputation). No special attention need be paid to which districts are fully observed.

To be logically consistent, the analysis model will need to be a t -based multiple regression analysis on the multiply imputed data. Unfortunately, t regression analysis is an iterative procedure and so can also be time-consuming. However, the model is equivalent to iterative weighted least squares, and if the weights are known, only the first iteration is necessary. In our case, the weights are functions of the data and the degrees of freedom parameter, and an approximate estimate of that parameter is an output from our multivariate t imputation

procedure. We can treat the weights as known conditional on imputation j and, as a result, are able to make the analysis noniterative and indeed no slower than an ordinary weighted least-squares analysis. Section A.4 in the Appendix explains this procedure.

Scholars wishing to use our procedure will thus find it fairly straightforward. Researchers will use the Amelia imputation software and input the observed voting data for all of the parties, along with any other covariates to be used in the analysis model or other variables that might help predict in the imputation model (these covariates can also have missing elements). The output from Amelia will be (multiply) imputed data constituting the effective vote for each party, along with their covariates, and the missing values of the covariates (if any) are also imputed. Amelia will also output appropriate weights. Once the imputation procedure is complete, any ordinary regression software can be used with the imputed data and weights to perform all subsequent analyses. The only complication is that a set of M (usually five or so) analyses on the M sets of data needs to be performed separately and the results averaged, as in multiple imputation. Software exists that makes this step transparent. For example, with the Clarify package (King et al. 2000), this step is completely automatic. So for users already accustomed to Clarify, they need only to add a preprocessing step with Amelia before their usual Clarify runs.

4 Monte Carlo Comparison of KK and Our Method

We now compare our model to the KK model under each of two data generation processes via Monte Carlo simulations. We generate a three-party system (so that the KK model can be run easily) where party 2 does not contest some districts, and parties 1 and 3 always contest. The six relevant quantities of vote data include the observed V_j and effective V_j^* vote for parties $j = 1, 2, 3$. If party 2 contests, $V_j = V_j^*$ is the vote party j receives (and we observe) in a district. If party 2 does not contest, V_j does not necessarily equal the unobserved V_j^* . We also transform these into log vote ratios: $Y_{13} = \ln(V_1/V_3)$, $Y_{23} = \ln(V_2/V_3)$, $Y_{13}^* = \ln(V_1^*/V_3^*)$, and $Y_{23}^* = \ln(V_2^*/V_3^*)$.

We are interested in modeling Y_{13}^* and Y_{23}^* , which are observed when party 2 contests and treated as missing otherwise. We also define Y_{13}^2 as the log vote ratio for party 1 relative to party 3 that is observed when party 2 does not run. To avoid implausible assumptions, KK do not use any information in this variable. In contrast, in our approach we use the best linear approximation to the relationship between this variable and the other variables in the model to help impute partially contested districts. As such, for our Monte Carlos we first generate the effective vote, along with Y_{13}^2 , as

$$\mathbf{D} = \{Y_{13}^*, Y_{23}^*, Y_{13}^2, X\} \stackrel{\text{iid}}{\sim} t(\mu, \Psi, \nu), \quad (3)$$

where X are covariates.³ Then we impose uncontestedness as structural missingness according to one of two rules that we now describe.

³For each simulation, we set the parameters near those for the U.K. elections case: $\mu = \{0.5, 0.4, -0.7, 0.6, 0.6, 1\}$, $\nu = 5$, and

$$\Sigma = \begin{pmatrix} 2.5600 & 2.1092 & 0.4417 & -1.2457 & 2.3808 & 2.4080 \\ 2.1092 & 2.5600 & 0.3194 & -0.8435 & 1.9863 & 2.7037 \\ 0.4417 & 0.3194 & 0.1600 & -0.2947 & 0.4991 & 0.3418 \\ -1.2457 & -0.8435 & -0.2947 & 2.5600 & -1.2285 & -0.8909 \\ 2.3808 & 1.9863 & 0.4991 & -1.2285 & 2.5600 & 2.3037 \\ 2.4080 & 2.7037 & 0.3418 & -0.8909 & 2.3037 & 4.0000 \end{pmatrix}.$$

The assumption of multiple imputation is that the missingness mechanism is “missing at random” (MAR), which means that elements of the data set that are missing can be predicted from the data included in the imputation model. If which elements are missing is dependent also upon the missing values themselves, the data missingness mechanism is said to be “nonignorable” (NI). Under the KK assumptions, the missing effective vote in uncontested districts depends on the missing information and so, sans covariates, is NI. The key assumption is that a party that does not run would have received fewer votes if it had run than the parties that did field candidates, although this assumption could easily be modified for use in a country where different processes were at work. In addition, if sufficiently informative covariates are included, this NI process might be made MAR. To study these issues, we run Monte Carlos under these two data generation processes. In both we draw 100 districts.

MAR: Here we set the covariates and coefficients and then calculate the probability that the third party receives the fewest votes. The n districts with the highest probability of the third party being the smallest vote getter we then make partially contested by making the third party not run. We run different simulations, with values of n varying from 0 to 50. This can be thought of as the third party having the same information as we the researcher and making the best guesses about where it will fail and withdrawing from those based on their guesses, where n is a function of their resources.

NI: Here we draw the actual effective votes, and the n districts with the lowest effective votes for the third party we then make partially contested with the third party not running. We again vary the value of n from 0 to 50. This can be thought of as the third party knowing exactly where it will fail and withdrawing appropriately without error or risk. Note that the probability that the third party will receive the fewest votes is not completely predictable on the basis of the covariates.

We present the results in several stages, beginning with the mean square error (MSE) for each model under the two data generation processes.⁴ Figure 1 gives the MSE (averaged over the four coefficients in the model) vertically and the degree of contestedness horizontally. The left graph is the MAR process and the right graph is from data that are NI. The main point of this figure is the comparison between the KK model (the dashed line) and our method (the solid line): in both data generation processes and for all levels of contestedness, our method has lower MSE than KK. The improvement is nearly zero without uncontestedness and noticeably larger as uncontestedness grows.

So that we can ascertain which features of our method are having the largest effects on the improvement over KK, we also provide results from estimating our method without the rejection sampler (dashed–dotted line) and without including the observed vote Y_{13}^2 in the imputation stage (dotted line). In the MAR graph, neither of these omissions hurts our method much, and so the main improvement is clearly the algorithm that does the imputation.

In the NI graphs, the rejection sampler appears to have a large effect, since removing it (i.e., as shown by the dashed–dotted line) increases the MSE. It is still below that of KK, but not by much until very high levels of contestedness. However, removing Y_{13}^2 (the dotted line) has very little effect and so we infer that this feature—which was the only information used in our model but not in the KK model—does not account for the vast majority of the improvements over KK.

⁴We find very similar patterns when dividing MSE into bias and variance, and so we save space by not presenting them separately. We provide more detailed information about the distribution of the coefficients in Fig. 2, which we discuss below.

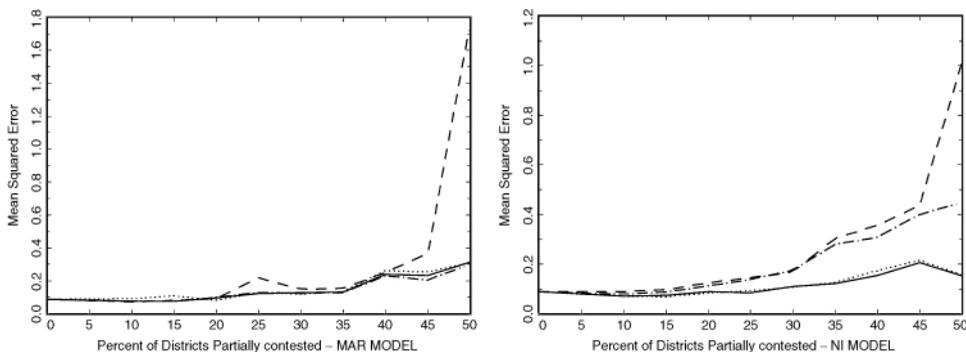


Fig. 1 Mean square error comparisons. The MSE is plotted for four methods of estimation under the MAR (left graph) and NI (right graph) data generation processes. The dashed line is the FIML model of KK. The solid line is the imputation model of this paper, while the dotted line is the imputation model without Y_{13}^2 , and the dashed-dotted line is our method without the rejection sampler.

To explore the comparative error of the two models, we also look at the distributions of the individual coefficients. For the constant term (in the left graphs) and the first coefficient (in the right graphs), and for MAR (at the top) and NI (at the bottom), Fig. 2 gives box plots for the distribution of our Monte Carlo simulations for each level of uncontestedness, with the KK model box plots in boldface. In each graph, the truth is labeled with a horizontal dotted line.

In nearly all runs, the KK model has more variation (longer lines) and is more biased (the median is on average farther from the dotted line) than our method. More interestingly, note the outliers in the KK model that are not in our method. These can be seen most clearly in all but the bottom-left graph. Indeed, only in the NI constant term (bottom left) do we find a clear pattern as contestedness increases. In both methods, there is more bias with high levels of contestedness, but considerably less with ours. In other runs (not shown), we have found for these parameters that this is a small sample result and that the bias in both methods vanish as n increases. We suspect that the remaining bias and much of the variance in the KK model is due to numerical instability, inability to converge on ν , and imprecision in numerical integration.⁵ In contrast, the small remaining bias in our method is likely due to running the rejection sampler as a separate stage, which is the cost of the compromise we suffer to have faster convergence and the ability to automate the program so that it would be easy to use.

4.1 Estimating Degrees of Freedom

To examine the ability of our algorithm to estimate ν correctly we ran a Monte Carlo experiment over a large range of ν values. For each value of ν we drew 1000 data sets, in each of which 20% of 100 districts were not fully contested according to the NI model. We then estimated $\hat{\nu}$ in these data sets, with ν itself varying from 4 to 2600. There are two possible sources of error. First, $\hat{\nu}$ may deviate from ν simply because the data set has a

⁵To explore the effects of imprecision of numerical integration, we replace the methods used by KK which we used in Fig. 1 with midpoint quadrature. For low numbers of terms of quadrature the FIML method does much more poorly than our method, while for very many terms to quadrature it exceeds the performance of our method. For details and figures see the Web Appendix to this paper, available at the *Political Analysis* Web site.

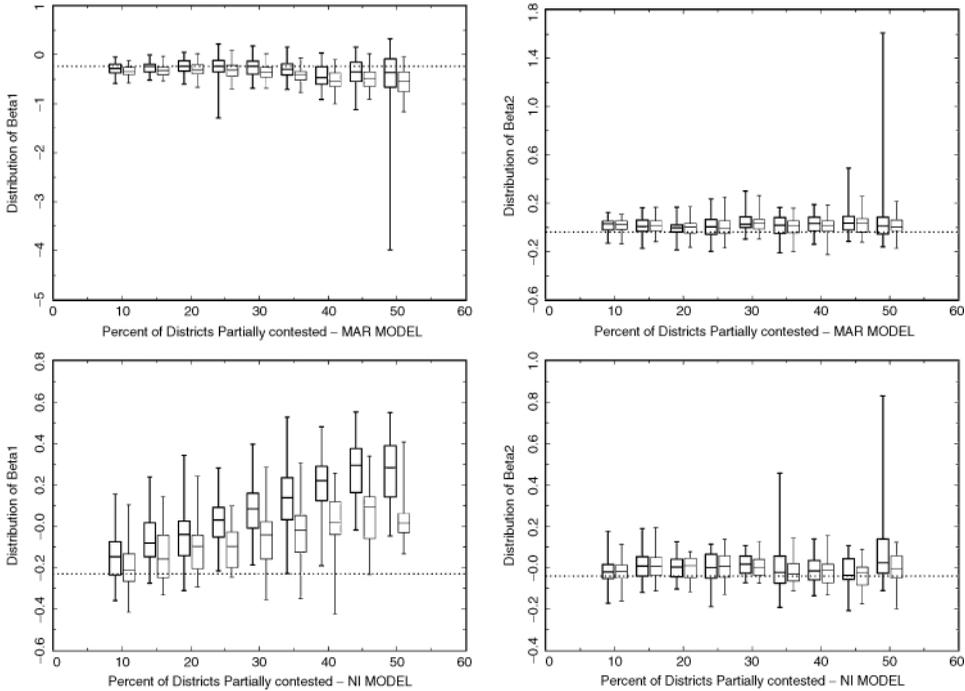


Fig. 2 Distribution of coefficients across Monte Carlo simulations. For the constant term (labeled beta1, in the left graphs) and the coefficient on the first explanatory variable (labeled beta2, in the right graphs), each box plot gives the distribution of the estimates for a given level of untestedness. The truth is labeled with a horizontal dotted line. Box plots for the KK model are in boldface and ours are in normal type.

small sample of only 100 districts. Second, error may be introduced because not all districts are fully contested (and by the mechanism that creates this missingness) and because the estimation routine is numerically imprecise or theoretically approximate. The left graph in Fig. 3 shows the results of this Monte Carlo for low values of ν , that is, where the data are not Normally distributed. The box plots show the 80% confidence intervals of $\hat{\nu}$ from our algorithm. The horizontal line at the top of the figure shows $\hat{\nu} = 30$ and thus the density above this line represents finding the data to be Normally distributed when in fact it is not. Similarly, in the right graph in Fig. 3 we show results for large values (10–2600) of ν . The x axis represents the actual value of ν as before, except it is now on a log scale, and the y value is the percentage of the simulations where $\hat{\nu} > 30$. Thus we see that for large values of ν the data are properly estimated to be Normally distributed, and the chance of misidentifying the data to be nonnormal decreases with increasing ν .

4.2 Replicating KK's Empirical Results

We ran our algorithm on the series of English elections KK used to estimate the effect of incumbency on vote share in each party. The results are presented in Fig. 4, which includes a direct replication of the same figure in KK. The arrows in the foreground (i.e., below) represent the effects from our replication of the KK model, while those in the background (above) represent the effects we calculate with our method. The vertical distance of each arrow above the line indicates the advantage of running an incumbent, compared to a

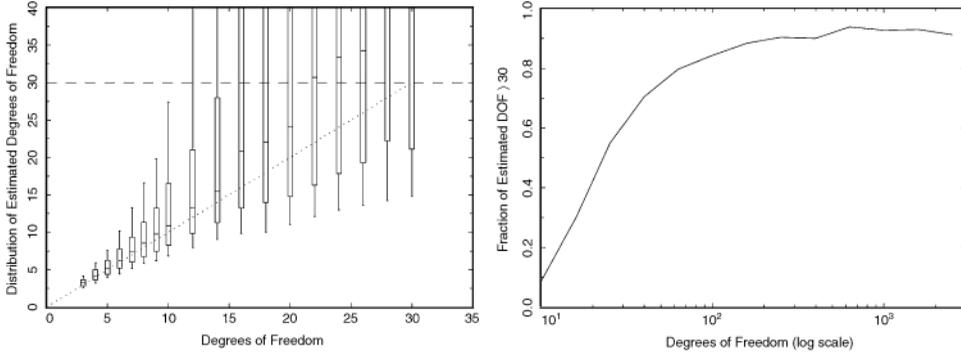


Fig. 3 Plot of estimated degrees of freedom against actual degrees of freedom. Data are simulated for 20% of 100 districts nonignorably partially contested. For low values, $\nu < 30$, we see, in the left graph, the correspondence between ν and $\hat{\nu}$. For large values we see, in the right graph, the fraction of the 1000 simulations where $\hat{\nu} > 30$ and thus the data would be considered Normally distributed. This fraction reaches 90% by roughly $\nu = 100$.

nonincumbent, to the respective party. The direction of the arrow shows from which of the other parties support is being drawn. The direction and magnitude of our estimates seem to match well the estimates of KK, although some variability should be expected, as our model is slightly different and also imputes the missing covariates in the original data set. Also of note is that we get similar results despite the fact that for our estimates we do not use the empirical Bayes procedures that KK used to reduce variance.

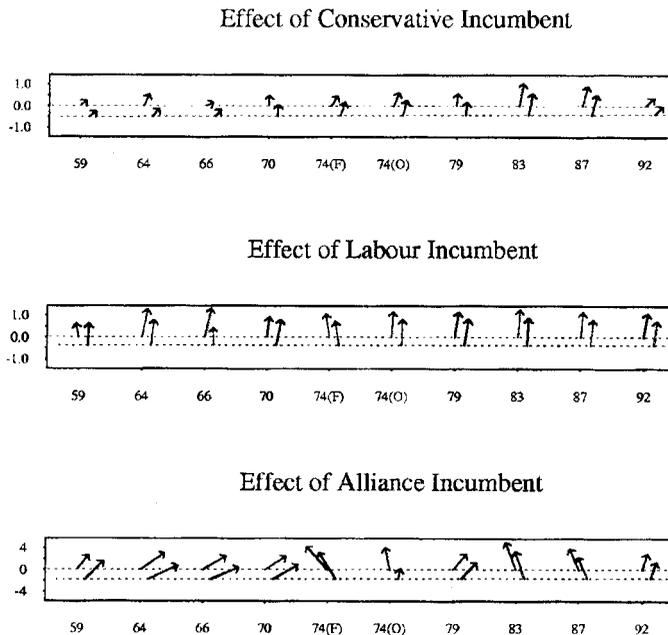


Fig. 4 The foreground arrows represent the effects of incumbency in the original model, while the background arrows are the effects calculated with our imputation model. The vertical distance of each arrow represents the advantage to that party of running an incumbent, while the direction shows from which party the support is being drawn.

5 Monte Carlo Evaluation with Many Parties

We also provide an example of estimates from a simulated party system with more than three parties. In this simulation, there are five parties and four covariates, X_1, \dots, X_4 , a setup for which KK's model is computationally infeasible. The first three parties contest every district. The fourth and fifth parties contest districts conditional on covariates X_1 and X_2 , respectively. The model of interest is the effect of X_4 on the effective vote shares of the five parties. Covariate X_3 is a variable useful in prediction and thus added to the imputation model but not the analysis model. The covariates are assumed to be completely observed, although scattered MAR missingness would pose no problem to the imputation model.

The log ratios of the vote shares are treated as different variables for every pattern of contestation. Thus, in addition to the covariates, the variables in the data set to be imputed are the log ratios of the effective votes, the log ratios of the contesting parties when party 4 does not contest, the log ratios when party 5 does not contest, and the log ratios when neither party 4 nor party 5 contests.⁶ By necessity only one of the latter four sets of variables can be observed in any particular district, and the other three sets of log ratios must be missing.

We draw 500 districts from a set of sufficient statistics defined as "truth." The true coefficient on X_4 was determined at this point; then after determining (conditional on X_1 and X_2) which districts would be only partially contested, we recomputed this coefficient employing listwise deletion on all of the partially contested districts (that is, running the model only on the districts for which the actual vote shares of all parties were known). We then imputed the effective votes for the partially contested districts with our algorithm and again computed this coefficient.

The small vertical bar in each panel in Fig. 5 gives the true coefficient each method seeks to estimate. The solid line in each panel is a kernel density plot (a smooth version of a histogram) of the distribution of a coefficient estimated based on the fully observed effective vote. This density is on average equal to the bar representing the truth which shows the near unbiasedness of the basic model. The distribution after listwise deletion (the dotted line) is heavily biased and, in the first panel, is even of a different sign (i.e., to the left of the long vertical line drawn at 0). In contrast, the density from our method, estimated on the basis of censored data due to uncontestedness (represented by the dashed line), is approximately unbiased but has a slightly higher variance than the estimate based on the fully observed effective vote data. Overall, our method recovers the truth very well in this high-dimensional case, and far better than a method based on deleting partially contested districts.

⁶The unique rows in the matrix of missingness patterns, \mathbf{R}_2 , are as follows.

$$\mathbf{R}_2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

The first four columns in \mathbf{R}_2 represent the log vote ratios; the next two groups of three columns each correspond to the remaining three log vote ratios when party 4 and then party 5, respectively, do not contest. Columns 11 and 12 are log vote ratios when neither party 4 nor party 5 contests, and the final four columns correspond to the fully observed covariates. The first row in \mathbf{R}_2 represents the pattern of missingness in the districts where all parties contest. The first four variables are the log vote ratios of the $J - 1$ parties, and the last four variables are the covariates. In row 2, party 4 decides not to contest some district. Thus the effective vote in the first four rows is unobserved. The three variables which are observed are the log vote share ratios of the $J_{-4} - 1$ parties which do contest. Similarly the next row represents the missingness of information in districts where party 5 does not contest, and the final row is where both party 4 and party 5 do not contest.

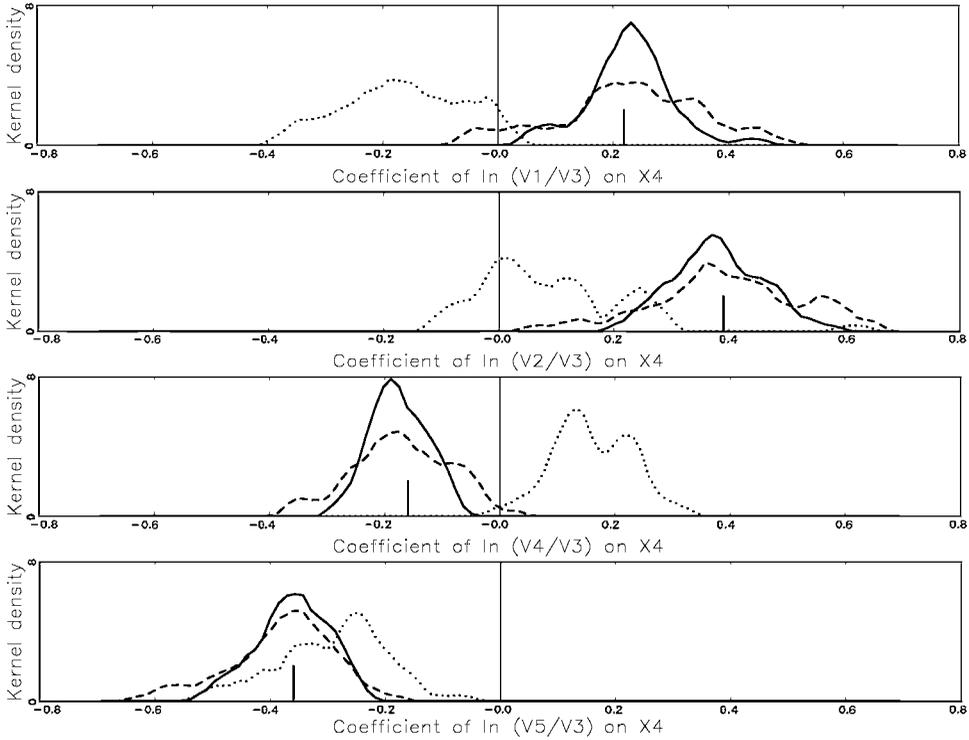


Fig. 5 Distribution over Monte Carlo simulations. Kernel density plots of a coefficient in 50 simulations of a five-party race applied to the effective votes (solid line), listwise deletion on the partially contested districts (dotted line), and our method (dashed). The true coefficient is represented by a small vertical bar. Note that our method is approximately unbiased, with only slightly higher variance than the method applied to the effective vote.

6 The 1993 Parliamentary Election in Poland

We now replicate and reanalyze Gibson and Cielecka's (1995) work on the 1993 parliamentary election in Poland. We ran our model on 12 parties in the 1993 parliamentary election and an "other" category composed of several smaller parties. We also ran a set of OLS models treating each party's vote share as a separate dependent variable, just as Gibson and Cielecka. We were able to replicate their results exactly.

We present differences between the results from the two methods first by comparing the marginal effects of all the variables in the two models and then by examining the change in the primary substantive finding. Figure 6 plots estimates from our method (horizontally) by OLS (vertically) for all coefficients (translated to first differences for a unit change from our method so as to be directly comparable to a coefficient in an OLS model) from all equations. The dashed 45° line marks the place where we would find equality between the two methods. When one method finds a first difference of 0, the other tends to as well, but when a larger effect is apparently detected, the OLS coefficient often veers far from our more accurate method. There is some evidence that the farther from 0 the OLS coefficient, the bigger the bias (this can be seen from the pattern of heteroskedasticity around the dashed line taking the shape of a bowtie tilted at an angle). Dots in the upper left and lower right are sign reversals between the two methods.

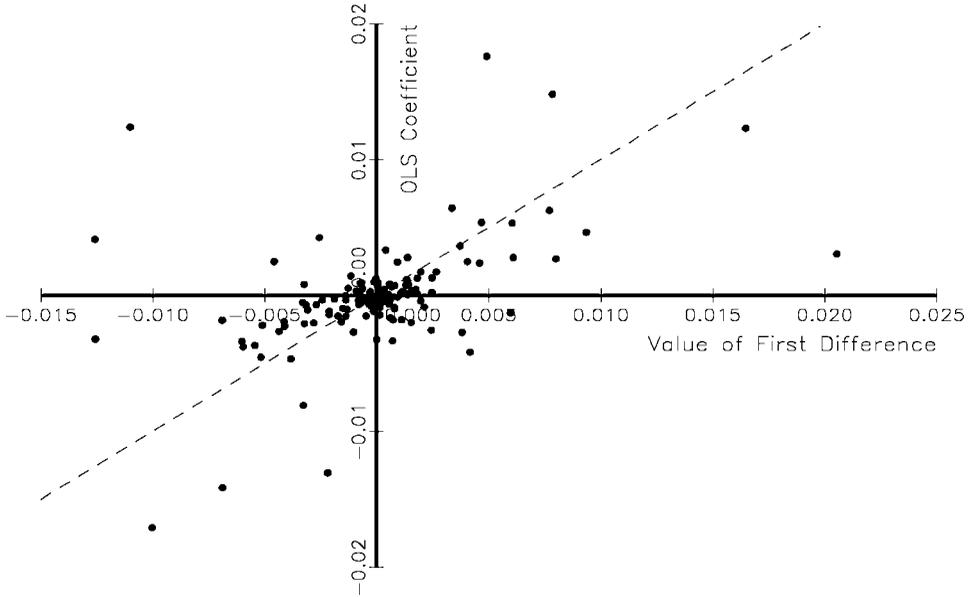


Fig. 6 Plot of first differences computed by OLS and our method. Each point is one regression coefficient, and the distance from the 45° line indicates how far apart the estimates from the two methods are.

We now turn to the central substantive point of the Gibson and Cielecka (1995) article. For reform-minded economists, Poland had up until 1993 been the leading example of a postcommunist country showing real economic growth brought about by their recommended dramatic “shock therapy” transition and privatization. Economic reform in Poland was held up as a model for other postcommunist states. The Polish electorate, however, since it put the communists back in power, seemed less happy with these changes than did the economists. One response of the economists was that “shock therapy” should have been implemented earlier and quicker; if it had, the reasoning goes, growth would have been quick enough and large enough that the reform party might have stayed in power. Gibson and Cielecka use their OLS analyses to test this hypothesis. That is, they examine the counterfactual: How would the reformers have fared had growth been more profound (a cumulative of 15% instead of the 10% actually seen since the end of the recession brought on by transition to a market economy) and unemployment not risen by 2.3%? They find that the main reform party, the Democratic Union (UD), would have gained a couple of points, but not nearly enough to close the 10 point gap between the UD and the chief postcommunist party, the Democratic Left Alliance (SLD). We reanalyze this result with our improved model (the KK approach being computationally infeasible with this many parties).

We run the same “first difference” with our model by increasing the growth in each voivodship (a district in Poland) by 5%, and reducing unemployment by 2.3%. The differences are shown in the last two columns in Table 1. Our method agrees with Gibson and Cielecka’s main finding that the reform party would not have gained control had the economy been much stronger, but very interestingly we find that this improved economic condition does shuffle votes profoundly among the postcommunist parties, with the PSL almost replacing the SLD as the largest vote share recipient. According to our results, and in contrary

Table 1 The effects of a better economy: Comparing results from OLS and our method

<i>Party</i>	<i>Acronym</i>	<i>Bloc</i>	<i>1993 vote share</i>	<i>Change in vote share</i>	<i>OLS predicted change</i>
Democratic Left Alliance	SLD	Postcommunist	20.41	-1.16	-1.41
Polish Peasant Party	PSL	Postcommunist	15.40	+3.25	-0.58
Democratic Union	UD	Proreform	10.59	-0.92	+0.86
Nonparty Bloc to Support Reform	BBWR	Proreform	5.41	-0.05	+0.78
Union of Labor	UP	Antireform	7.28	-0.79	-0.20
Confederation for an Independent Poland	KPN	Antireform	5.77	-0.03	+0.82
Fatherland Catholic	KKW	Center-right	6.37	+0.11	+0.51
Center Alliance	PC	Center-right	4.42	-0.21	-0.10
Liberal Democratic Congress	KLD	Center-right	3.99	-0.65	0.00
Polish Peasant Party—Party Alliance	PL		2.37	-0.30	-0.53
Party X	PX		2.74	-0.30	+0.01
Self-Defense	SL		2.78	+0.22	-0.49
Other			12.47	+0.83	+0.67
Total			100.00	0.00	0.34

to OLS, the question the election answered was which postcommunist party would win, not how the reform party would fare. That the state of the economy was of such importance to the battle between the two largest postcommunist parties was not picked up by the OLS model. (Of course Gibson and Cielecka cannot be held accountable for the differences between their approach and our improved method since our methods were not available at the time of their work).

6.1 *The t Compared to the Normal Distribution*

It is useful to remember that the Normal distribution is virtually nested within the t distribution, and so the choice between the two distributions need not be a matter of speculation or taste. The Normal distribution is the limiting special case of the t as ν approaches infinity. If $\hat{\nu} > 30$, then for all practical purposes, the data can be considered Normally distributed. For $\nu < 30$ the difference between the $t(\nu)$ and the Normal distribution [that is, $t(\infty)$] is considered nonnegligible. For example, $\hat{\nu}$ varies from 4 to 8 in the series of English elections and is 6.7 in the Polish data.

KK find that the t distribution fits the data set of English elections better than the Normal distribution. They find that the 95 and 50% confidence regions cover about 95 and 49% of the English elections in their data set, whereas for the Normal these regions cover 91 and 64% respectively. Thus the confidence region for the Normal is too large in the region of the peak and too small in the region of the tails, for which the t corrects.

In our example of the Polish elections, we find that the t distribution significantly alters the predicted effects. Comparing the results presented in Table 2 we see that when we use the Normal distribution (in both the imputation and the analysis models), several of the predicted effects on the expected vote share from an improved economy (again, increasing the growth in each voivodship by 5% and reducing unemployment by 2.3%) are changed

Table 2 The effects of a better economy: Comparing results from the t and Normal distributions

<i>Party</i>	<i>Acronym</i>	<i>Bloc</i>	<i>1993 vote share</i>	<i>Change in t dist.</i>	<i>Change in Normal dist.</i>
Democratic Left Alliance	SLD	Postcommunist	20.41	-1.16	-0.79
Polish Peasant Party	PSL	Postcommunist	15.40	+3.25	+0.32
Democratic Union	UD	Proreform	10.59	-0.92	+0.48
Nonparty Bloc to Support Reform	BBWR	Proreform	5.41	-0.05	+0.04
Union of Labor	UP	Antireform	7.28	-0.79	+0.21
Confederation for an Independent Poland	KPN	Antireform	5.77	-0.03	-0.27
Fatherland Catholic	KKW	Center-right	6.37	+0.11	+0.22
Center Alliance	PC	Center-right	4.42	-0.21	-0.38
Liberal Democratic Congress	KLD	Center-right	3.99	-0.65	-0.27
Polish Peasant Party—Party Alliance	PL		2.37	-0.30	-0.79
Party X	PX		2.74	-0.30	-0.07
Self-Defense	SL		2.78	+0.22	-0.05
Other			12.47	+0.83	+1.32
Total			100.00	0.00	0.03

considerably, including the key result that this dramatically aids the postcommunist PSL party.

7 Concluding Remark

We offer a computationally feasible algorithm for analyzing multiparty vote share data. The advantage of our approach is that it is considerably faster, while scaling up in practice to many more parties without much difficulty, loss of speed, or numerical imprecision. The approach of treating the estimation problem as a “missing data” problem also allows for greater flexibility in distributional assumptions and increased ease of imposing qualitatively understood restrictions on the effective vote, while also providing imputations of missing values among any of the vote or explanatory variables in the original data set. We believe that the result is a useful tool for comparative political research. In addition to treating vote share as a compositional variable, our approach also allows us to model vote share as t distributed and enables us to use election data where not all parties contest in every district. These can be important features of political science data.

Appendix

We briefly summarize some useful properties of the multivariate t density (Section A.1) and use these to summarize changes in the EMis algorithm we made to accommodate t -distributed data (Section A.2). Then we explain how to incorporate constraints for partially contested or uncontested seats (Section A.3) and summarize the use of the t regression analysis model as a model of weighted least squares where the weights are products of the imputation model (Section A.4). A more complete version of this Appendix as well as several other topics is available in the Web Appendix, available at the *Political Analysis* Web site.

A.1 Some Useful Properties of the t Distribution

We offer here a brief summary of properties of the multivariate t distribution that we find useful. If Y_i is distributed

$$Y_i \stackrel{\text{ind}}{\sim} N(\mu, \Psi/u_i) \quad (\text{A1})$$

$$u \stackrel{\text{iid}}{\sim} \chi_v^2/v \quad (\text{A2})$$

where $v > 0$, then Y is distributed as

$$Y \stackrel{\text{iid}}{\sim} t(\mu, \Psi, v). \quad (\text{A3})$$

The complete-data likelihood, for known weights, is then separable:

$$L(\mu, \Psi, v | Y, u) = L_N(\mu, \Psi | Y, u) + L_G(v | u) \quad (\text{A4})$$

where

$$L_G(v | u) = -n \ln \left(\Gamma \left(\frac{v}{2} \right) \right) + \frac{nv}{2} \ln \left(\frac{v}{2} \right) + \frac{v}{2} \sum_{i=1}^n (\ln(u_i) - u_i). \quad (\text{A5})$$

A.2 An EMis Algorithm for t -Distributed Data

We followed the framework of the EMis algorithm⁷ to impute the effective vote in constituencies where not all the parties ran and to deal with missingness we had in the covariates. The EM algorithm itself is often implemented under the assumption that the data are distributed normally, but this distributional assumption can be changed. The EM algorithm retains its simplicity if the E and, particularly, the M steps are noniterative themselves and do not involve hard to maximize likelihoods. This can be done easily with the t distribution by the use of the decomposition in Eq. (A1). We take the vector of weights u to be an additional variable (completely unobserved) to be imputed in the data set, and the degrees of freedom v as an additional element to θ . The t -distributed EM algorithm then resembles the normally distributed EM algorithm and can be driven with the same shortcuts, such as the sweep operator (Schafer 1997), except that the sums and sums of squares and cross-products computed for the M step need to be appropriately weighted by u . Furthermore, by exploiting the separability in (A4), the ECME algorithm (Liu 1994; Liu and Rubin 1994) allows faster, more accurate convergence in v by using two conditional likelihoods as follows.

The E step in ECME is the same as the E step in EM. The elements of Y_{mis} are filled in with their expected values from current estimates of μ and Ψ as in the EM algorithm.⁸ The

⁷See the Web Appendix, available at the *Political Analysis* Web site, for details on the EM methods and the EMis algorithm.

⁸Two families of EM algorithms are possible. In one, the completed data are stored (Beale and Little 1975); in the other, sufficient statistics (sums, sums of squares, and sums of cross products) are stored (Dempster et al. 1977). Done properly they are equivalent. [A mixture of the two is also possible, storing sufficient statistics for nearly completed observations and raw data otherwise (Little and Rubin 1987).] We opt for the first of these methods in the exposition in this paper and in our code because it seems conceptually simpler and more intuitive, and was a faster implementation in GAUSS, as it can be written to draw on GAUSS's strength in large matrix algebra computations and avoid GAUSS's weakness in looping.

vector of weights u^{t+1} is similarly created from the expectation

$$E(u_i^{t+1}) = \frac{p_i + v^{(t)}}{\delta_{i,\text{obs}}^{(t+1)} + v^{(t)}} \quad (\text{A6})$$

where p_i is the number of variables and δ , known as the Mahalanobis distance, is given by

$$\delta_{i,\text{obs}}^{(t+1)} = (Y_{i,\text{obs}} - \mu_{i,\text{obs}})' \Psi_{i,\text{obs}}^{-1} (Y_{i,\text{obs}} - \mu_{i,\text{obs}}). \quad (\text{A7})$$

Thus observations which can be considered as outliers have large Mahalanobis distances and are down-weighted.

After the E step are two conditional maximization steps (CM). First, we maximize the Q-function (the constrained expected log-likelihood) over $\theta_1 = (\mu, \Psi)$ given v . Then we maximize the L-function (the constrained actual log-likelihood) over $\theta_2 = v$ given $\theta_1 = (\mu, \Psi)$. To do this a one-dimensional search is implemented over Eq. (A5). This function is globally concave and has an analytical derivative, making it simple to maximize with a search such as Newton's method.

Under mild conditions the ECME algorithm has the convergent properties of GEM algorithms, although it is not itself a special case (Liu and Rubin 1994). It would be GEM if we maximized the Q-function over $\theta_2 = v$ given $\theta_1 = (\mu, \Psi)$ instead, but in most problems this approach leads to much more rapid convergence over v (Liu 1994). Indeed, with the t distribution, since the likelihood is separable by Eq. (A4), if we maximized the Q-function rather than the L-function, we would again have exactly the EM algorithm.

The maximum of the posterior provided by the ECME algorithm substitutes for the value that would be provided by EM in the EMis algorithm. For the importance resampling in the applications that follow we used a covering distribution resembling the ‘‘witch’s hat’’ distribution, with a t -distributed peak trailing to a constant-valued ‘‘brim’’ on the joint μ and Σ parameters, and a χ^2 distribution on $(v - 2)$ with mean $(\hat{v} - 2)$. By monitoring the distribution of the importance ratio and studying simulated data, we were confident that the true distribution had been properly covered. This was also confirmed strongly, albeit indirectly, in the analyses presented in this paper.

A.3 Imposing Uncontestedness Constraints

Sometimes imputations from missing data models are not appropriate to the user’s analysis or fail known bounds or identities of the data. For some special cases, such as with ordinal and nominal variables, it is possible to transform the normal model posterior directly to address the constraint of discreteness in a logically consistent way. We have analogous, although somewhat more complicated, constraints to implement.

The KK model imposes the constraint that ‘‘the noncontesting parties would have received fewer votes than the parties that did nominate candidates’’ and thus the effective vote in some district of any party that did not contest in that district must be lower than the effective vote of all other parties that did run candidates in that district. We impose this constraint in the imputation model by rejection sampling/resampling from the t model. For a given imputed data set, $j \in 1, \dots, M$, with sufficient statistics θ_j , each observation is checked as to whether it meets the model constraint.⁹ In each round, any observation, y_i , which fails

⁹To do this, we partition the imputed effective vote of party j in partially contested district i by $V_{ij} \in R_i^+$ if j originally ran and $V_{ij} \in R_i^-$ if the party did not run a candidate. The Boolean is then $\max(R_i^-) < \min(R_i^+)$.

the constraint is redrawn from $P(\theta_j | y_i, u_i)$. The number of failing observations, which is a useful diagnostic and reported by our software, is necessarily nonincreasing in each round. This is iterated until all observations pass. This approach can be tailored, with different check functions, to a broad range of analyst constraints that might fit in any particular application. As discussed in Section 3, the distribution from which the final imputations are drawn will then be the truncation (to the limits of the constraints) of the unbounded distribution that maximizes the likelihood of the observed data, whereas the FIML model derives the truncated distribution that maximizes the likelihood of the observed data.

A.4 The t Regression Analysis Model

Part of the appeal of the multiple imputation framework is that it separates the model of missingness from the model of analysis. Once the data sets are imputed, the user can apply whatever model he or she would have used if the data set had arrived fully observed. In the present application, the researcher can analyze the data as if all parties contested elections in every district, which would normally require the application of a multivariate t regression model.

As an easier alternative, a reasonable approximation would be to use t regressions conditional on the weights output from the imputation stage. This means that the user needs to run only one (noniterative) weighted least-squares analysis, for each imputation, and to average the results as in multiple imputation. Thus, Amelia will provide (say) five sets of imputed effective vote data, along with any covariates provided (with their missing data, if any, also imputed) and a weight. The user will then run a set of weighted least-squares regressions, using any statistical package. The dependent variable is the log ratios of the effective votes, the weight is as provided by Amelia, and the results are averaged.

References

- Aitchison, J. 1986. *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.
- Beale, E. M. L., and R. J. A. Little. 1975. "Missing Data in Multivariate Analysis." *Journal of the Royal Statistical Society Series B* 37:129–145.
- Dempster, A. P., N. M. Laird, and D. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion)." *Journal of the Royal Statistical Society, Series B* 39:1–38.
- Gelman, Andrew, and Gary King. 1994. "A Unified Method of Evaluating Electoral Systems and Redistricting Plans." *American Journal of Political Science* 38(2):514–554.
- Gibson, John, and Anna Cielecka. 1995. "Economic Influences on the Political Support for Market Reform in Post-communist Transitions: Some Evidence from the 1993 Polish Parliamentary Elections." *Europe-Asia Studies* 47(5):765–785.
- Honaker, James, Anne Joseph, Gary King, and Kenneth Scheve. 2000. *Amelia: A Program for Missing Data (Gauss Version)*. Cambridge, MA: Harvard University. <http://gking.harvard.edu>.
- Katz, Jonathan, and Gary King. 1999. "A Statistical Model for Multiparty Electoral Data." *American Political Science Review* 93(1):15–32.
- King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1):49–69.
- King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analysis: Improving Interpretation and Presentation." *American Journal of Political Science* 44:341–355.
- Little, Roderick J. A., and Donald Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.
- Liu, Chuanhai. 1994. *Statistical Analysis Using the Multivariate t Distribution*. Dissertation. Cambridge, MA: Harvard University.
- Liu, Chuanhai, and Donald Rubin. 1994. "A Simple Extension to EM and ECM with Faster Monotone Convergence." *Biometrika* 81:633–648.
- Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.